Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA

ACTA DE EXAMEN DE GRADO

No. 00134
Matrícula: 2133802652

DESCRIPCION DE DATOS EN EL SIMPLEX VIA VARIABLES DIRECCIONALES

MARCO ANTONIO SANCHEZ PEREZ
ALUMNO

REVISÓ

LIC. JULIO CESAR DE LARA ISASSI
DIRECTOR DE SISTEMAS ESCOLARES

En México, D.F., se presentaron a las 11:00 horas del día 12 del mes de febrero del año 2016 en la Unidad Iztapalapa de la Universidad Autónoma Metropolitana, los suscritos miembros del jurado:

DR. GABRIEL ARCANGEL RODRIGUEZ YAM
DR. GABRIEL NUÑEZ ANTONIO
DR. GABRIEL ESCARELA PEREZ

Bajo la Presidencia del primero y con carácter de Secretario el último, se reunieron para proceder al Examen de Grado cuya denominación aparece al margen, para la obtención del grado de:

MAESTRO EN CIENCIAS (MATEMÁTICAS APLICADAS E INDUSTRIALES)

DE: MARCO ANTONIO SANCHEZ PEREZ

y de acuerdo con el artículo 78 fracción III del Reglamento de Estudios Superiores de la Universidad Autónoma Metropolitana, los miembros del jurado resolvieron:

Aprobar

Acto continuo, el presidente del jurado comunicó al interesado el resultado de la evaluación y, en caso aprobatorio, le fue tomada la protesta.

DIRECTOR DE LA DIVISIÓN DE CBI

DR. JOSE GILBERTO CORDOBA HERRERA

PRESIDENTE

DR. GABRIEL ARCANGEL RODRIGUEZ YAM

VOCAL

DR. GABRIEL NUÑEZ ANTONIO

SECRETARIO

DR. GABRIEL ESCARELA PEREZ

Maestría en Ciencias Matemáticas

Aplicadas e Industriales

Descripción de datos en el simplex vía variables direccionales

Tesis para obtener el titulo de:

Maestro en Ciencias

Presenta:

Marco Antonio Sánchez Pérez

Asesor:

Dr. Gabriel Núñez Antonio

Febrero 12, 2016.

A Dios

Por todo lo que me ha dado.

A mis padres

A mi mamá, por sus oraciones, por preocuparse y amarme con tanta ternura y devoción. Me siento orgulloso que tú seas mi madre; te amo Isabel Pérez Ramírez.
A mi padre Roberto, por el valor y el coraje que has tenido para levantarme ante cualquier adversidad, por las enseñanzas que me has dado y por darme ánimos para seguir adelante. Muchas gracias, papá.
A mi hermana y segunda madre Belem, por todos sus cuidados y consejos que me ha dado.

A mis hermanos

Gloria, Guadalupe, Roberto y Denisse por ser parte de mi vida y representar la unidad familiar.

A Melina

Por todo el apoyo incondicional que me ha dado.

Hay hombres que luchan un día y son buenos. Hay otros que luchan un año y son mejores. Hay quienes luchan muchos años, y son muy buenos. Pero hay los que luchan toda la vida, esos son los imprescindibles.

Bertolt Brecht.

AGRADECIMIENTOS:

- Antes que nada quiero expresar mi más sincero agradecimiento a mi director de tesis, el Dr. Gabriel Núñez Antonio, por haberme dado la oportunidad de trabajar con él, por haber tenido la paciencia necesaria para ayudarme, por ser demasiado accesible en todo momento y por sus valiosos consejos personales.

- A mi segunda casa la Universidad Autónoma Metropolitana, en particular al Departamento de Matemáticas, por todo el apoyo brindado.

- A CONACYT por la beca que me proporcionaron, porque gracias a ello, fue posible la realización de mis estudios de maestría.

- A mis sinodales el Dr. Gabriel Arcángel Rodríguez Yam, el Dr. Gabriel Escarela Pérez, y el Dr. Gabriel Núñez Antonio. Por sus valiosos observaciones a este trabajo.

- A mi amigo Paco y a mi sobrina Karen por su valiosa ayuda técnica.

- A mis familiares y amigos por estar pendientes de mi formación académica y sobre todo por creer en mi.

- A mis profesores y compañeros de la UAM-I por los ánimos que me dieron para seguir con este trabajo.

Índice

Resumen

Los *datos direccionales* tienen que ver con observaciones de vectores unitarios en el espacio q-dimensional, este tipo de datos se pueden representar de diversas maneras, una de ellas es a través de puntos sobre la esfera unitaria. Otra manera de representar datos direccionales es a través de ángulos. Por otro lado, los *datos composicionales* son vectores cuyas componentes son no-negativas y cuya suma se restringe a un valor constante k, esta restricción hace que el espacio muestral asociado a los datos composicionales sea el *simplex q-dimensional*. Para trabajar este tipo de variables una propuesta es mapear variables composicionales sobre la hiper-esfera unitaria $q-dimensional$, ver Mardia y Jupp (2000), y usar distribuciones asociadas a variables direccionales. En este proyecto de investigación se propone emplear métodos y procedimientos definidos para *datos direccionales* en la descripción de *datos composicionales*. Esta propuesta implica, entre otras cosas, la implementación de procedimientos de inferencia Bayesianos para datos direccionales basados en la *distribución Normal proyectada q-dimensional*.

Palabras clave: Datos Direccionales, Datos Composicionales, Distribución Normal Proyectada, Estadística Bayesiana.

Introducción

En la modelación de fenómenos reales el investigador se puede encontrar con *variables composicionales*. Por ejemplo, en petrología, el anásis estadístico de la composición total geoquímica de rocas es fundamental. Comúnmente tales composiciones son expresadas como porcentajes de peso de óxidos mayores o como porcentajes de peso de algunos minerales básicos. En economía, es importante el análisis de la composición del 100 % de algún portafolio de inversión. En procesos electorales, es importante estimar el porcentaje de votos de cada fuerza política, etc. En este contexto, debido a la restricción de que los componentes deben sumar una constante c y de no negatividad, el *simplex unitario d-dimensional* resulta ser el espacio muestral comúnmente asociado a este tipo de datos. Los datos direccionales tienen que ver con observaciones que son vectores unitarios en el espacio k-dimensional. Los datos direccionales en el plano 2-dimensional se denominan *datos circulares* y, las direcciones en el plano 3-dimensional se denominan *datos esféricos*. Así, los espacios muestrales más comunes son el círculo unitario o la esfera unitaria. En la etapa inicial del planteamiento y modelación de cualquier fenómeno real, como un problema estadístico, es crucial el reconocimiento y la definición de un espacio muestral adecuado para el análisis de los datos con los que se este trabajando. Históricamente, la estructura algebraica de $\mathbb{R}^k$ resultó familiar y muy intuitiva para la modelación estadística de datos, propiciando el desarrollo de una gama extensa de métodos apropiados y con una interpretación natural. Por otro lado, hasta que se reconoció la topología propia de la esfera unitaria, $\mathbb{S}^k$, Fisher y Watson en sus trabajos de los años 50's del siglo pasado (ver, por ejemplo, Fisher (1953)) iniciaron el desarrollo para el análisis estadístico de *datos direccionales*. Por su parte, solo a partir de los trabajos seminales del profesor J. Aitchison en los años 80's del siglo XX se dispone de una propuesta metodológica muy general para el tratamiento y análisis de *datos composicionales*.

Los datos direccionales se pueden representar de diversas maneras, una de ellas es mediante puntos sobre la esfera unitaria $\mathbb{S}^k = \{\boldsymbol{u} \in \mathbb{R}^k : \boldsymbol{u}'\boldsymbol{u} = 1\}$. Otra manera de representar datos direccionales es a través de ángulos. Así, en el caso general, para cualquier valor de q se pueden emplear coordenadas hiper-esféricas para describir datos direccionales por medio de $(q-1)$ ángulos.

En la literatura, se han propuesto varios modelos para describir el comportamiento probabilístico de datos de tipo direccional. Sin pérdida de generalidad, estos modelos se pueden agrupar en tres grandes categorías: *modelos generados por proyecciones*, dentro de los cuales la distribución más representativa es la distribución Normal proyectada; *modelos wrapped o "envueltos"*, que incluyen a la Normal envuelta, la Cauchy envuelta y la Poisson envuelta, por citar algunos, y *modelos tipo von Mises-Fisher* cuya distribución principal es la distribución von Mises-Fisher. La distribución de von Mises-Fisher para datos circulares es conocida como distribución von Mises y es una de las distribuciones más relevantes en el análisis inferencial para datos circulares. En el caso de datos esféricos la distribución von Mises-Fisher es conocida como distribución Fisher.

De manera reciente, se ha vuelto a tener un desarrollo importante en las propuestas metodológicas para el análisis de datos direccionales. Ver por ejemplo, Wang y Gelfand (2013) y Nuñez-Antonio *et al.* (2015) y las referencias allí incluidas. Los procedimientos Bayesianos han contribuido a la posibilidad de llevar a cabo inferencias bayesianas sobre todos los parámetros involucrados en modelos cada vez más complejos.

Por otro lado, el espacio muestral natural asociado a los datos composicionales es el simplex *positivo unitario q-dimensional* definido como

$$S^q = \{\boldsymbol{x} = (x_1, \ldots, x_q) |\ x_1 \geq 0, \ldots, x_q \geq 0\ ;\ \sum_{i=1}^{q} x_i = 1\}.$$

En este trabajo nos referiremos al simplex *positivo unitario q-dimensional* simplemente como el *simplex q-dimensional*.

Las restricciones de nonegatividad y de suma constante introducen un reto importante en la modelación de variables aleatorias en el simplex. El enfoque propuesto por Aitchison (1982) basado en *log-cocientes* para el análisis de datos composicionales, ha sido la fuente de varias discusiones en las últimas décadas. Lo anterior se debe, en parte, a la enorme aplicación que tienen este tipo de datos en varios campos del conocimiento. La aproximación de Aitchison ha permitido realizar análisis estadísticos frecuentistas una vez que se ha aplicado una transformación a los datos originales. Lo anterior, con la idea de llevar las variables composicionales a espacios más manejables como $\mathbb{R}^k$ y posteriormente regresar al simplex a través de algunas *isometrías*. Esto

da una clara ventaja en el análisis, debido a la gran cantidad de procedimientos disponibles para variables definidas en $\mathbb{R}^k$.

Si bien la propuesta por Aitchison (1982) para el tratamiento de datos composicionales es muy reciente y ha ofrecido un panorama general para el análisis de variables en el simplex, aún quedan aspectos que resolver desde el punto de vista metodológico y que aún hoy en día son tema de discusión. Ver por ejemplo, Pawlowsky-Glahn y Egozcue (2001) y las referencias allí incluidas. Adicional a estos temas, en la práctica se pueden encontrar datos composicionales donde alguna o algunas de las componentes son cero. Por ejemplo, en el análisis geológico de rocas se puede tener interés en los componentes de óxidos mayores. Sin embargo, no en todas las rocas se puede observar la presencia de todos los óxidos considerados, lo cual lleva a tener datos con componentes cero. En el análisis de gastos de las familias, si se definen categorías de gasto de alimentación, salud, escolaridad, diversión, viajes al extranjero, autos, equipo de cómputo; claramente no todas las familias presentan gasto en todos los rubros (ya sea por imposibilidad o por otra causa). Las situaciones anteriores son ejemplos de que en la práctica el investigador puede tener conjuntos de datos composicionales con la presencia de ceros. En esta situación el enfoque propuesto por Aitchison (1982) no es aplicable ya que los log-cocientes no están definidos para estos datos. El tratamiento de cero hoy en día sigue siendo tema de discusión (ver por ejemplo, Scely y Welsh (2011)). Uno de los enfoques propuestos para el tratamiento de ceros es transformar las variables composicionales en el simplex S^q sobre la esferia unitaria $\mathbb{S}^{(q-1)}$, ver por ejemplo Mardia y Jupp (2000), Wang *et al.* (2007), Scely y Welsh (2011), y posteriormente emplear distribuciones y procedimientos asociados a variables direccionales.

En este proyecto de investigación se propone emplear métodos y procedimientos definidos para *datos direccionales* para describir *datos composicionales*. Esta propuesta implica, entre otras cosas, la implementación de procedimientos de inferencia Bayesianos para datos direccionales basados en la distribución Normal proyectada *q-dimensional*.

La estructura de esta tesis es la siguiente. En el Capítulo 1 se ofrece una introducción al análisis de datos direccionales y al análisis de datos composicionales en general, haciendo énfasis en los principales problemas que presentan en la modelación de los datos composicionales. Adicionalmente, se presenta una breve introducción al enfoque Bayesiano de la estadística y finalmente se exponen algunas ideas sobre métodos numéricos y de simulación empleados en la inferencia Bayesiana, los cuales se emplean en las propuestas metodológicas de este trabajo. En el Capítulo 2 se define y analiza ampliamente la distribución Normal proyectada q-variada, $NP_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, con $\boldsymbol{\Sigma} = \boldsymbol{I}$. Se muestra una forma de derivar esta distribución y se ana-

lizan de manera exhaustiva los casos $q = 2$ (caso circular) y $q = 3$ (caso esférico). Por último se desarrolla y presenta una propuesta de análisis Bayesiano para la distribución Normal proyectada en cualquier dimensión. En el Capítulo 3 se presenta nuestra propuesta para analizar datos composicionales vía variables direccionales. Particularmente, se ofrece la construcción de una matriz direccional análoga a la matriz de variación composicional definida en Aitchison (1986). La matriz anterior, resuelve los problemas descriptivos de datos composicionales cuando alguna componente del dato composicional es cero. Por otra parte, también se considera modelar datos composicionales a través de realizar inferencias en el modelo Normal proyectado definido para datos direccionales. En el Capítulo 4 se desarrollan algunos ejemplos que ilustran los procedimientos propuestos tanto a nivel descriptivo como a través de modelos. En el Capítulo 5 se presentan las conclusiones y limitaciones de este trabajo y se dan algunas líneas futuras de investigación que pueden contribuir al análisis de datos composicionales a través de metodologías y modelos propuestos inicialmente para variables direccionales. Finalmente, en los Apéndices A.1 y A.2 se presentan demostraciones derivadas en este trabajo y los programas desarrollados para la implementación de los procedimientos propuestos en este trabajo.

Capítulo 1

Preliminares

En muchas situaciones el investigador se puede encontrar con datos asociados a espacios muestrales diferentes a $\mathbb{R}^k$. En la Sección 1.1 se introduce el concepto de datos direccionales. Éstos son datos cuyo espacio muestral es la circunferencia de la esfera unitaria $\mathbb{S}^q$. En la Sección 1.2, se estudia el concepto de datos composicionales. Ambos tipos de datos requieren de análisis específicos propios de su diferente naturaleza topológica, respecto a $\mathbb{R}^k$. Y en la Sección 1.3 se incluye una breve introducción a la estadística bayesiana el cual es el enfoque empleado en esta tesis.

1.1. Datos direccionales

Los datos direccionales tienen que ver con observaciones de vectores unitarios en el espacio q-dimensional. Cuando $q = 2$ los datos direccionales se denominan datos circulares. Cuando $q = 3$ se denominan datos esféricos, y cuando $q > 3$ en general se denominan datos direccionales.

Los datos direccionales se usan en varias áreas de la ciencia, tal como en meteorología (análisis de dirección del viento), biología (dirección de navegación de animales acuáticos), psicología (estudio de mapa mentales), geología (estimación de rotaciones relativas de las placas tectónicas), astronomía (distancia entre cuerpos celestes), entre otras. Los datos direccionales se pueden representar de diversas maneras, una de ellas es a través de puntos sobre la esfera unitaria $\mathbb{S}^q := \{\boldsymbol{u} \in \mathbb{R}^q : \boldsymbol{u'u}=1\}$. Se debe notar que como $u \in \mathbb{S}$ es un vector unitario q-dimensional, éste también puede ser definido utilizando $q-1$ ángulos. Así, en el caso general, para cualquier valor de q se pueden utilizar coordenadas hiper-esféricas definidas por

$$\boldsymbol{u} = \begin{bmatrix} \cos(\theta_1) \\ \operatorname{sen}(\theta_1)\cos(\theta_2) \\ \operatorname{sen}(\theta_1)\operatorname{sen}(\theta_2)\cos(\theta_3) \\ \vdots \\ \operatorname{sen}(\theta_1)\cdots\cos(\theta_{q-1}) \\ \operatorname{sen}(\theta_1)\cdots\operatorname{sen}(\theta_{q-1}) \end{bmatrix} \tag{1.1}$$

donde $\theta_i \in [0,\pi]$ para $i = 1,\ldots,q-2$ y $\theta_{q-1} \in [0,2\pi]$. En particular, cuando $q = 2$, las coordenadas hiper-esféricas se denominan coordenadas polares definidas por $\boldsymbol{u} = (\cos(\theta), \operatorname{sen}(\theta))'$, con $\theta \in [0,2\pi]$, y cuando $q = 3$ las coordenadas hiper-ésfericas se llaman coordenadas esféricas $\boldsymbol{u} = (\cos(\theta), \operatorname{sen}(\theta)\cos(\phi), \operatorname{sen}(\theta)\operatorname{sen}(\phi))'$, con $\theta \in [0,\pi]$ y $\phi \in [0,2\pi]$. Se debe notar que para el caso de coordenadas esféricas $\theta = \theta_1$ y $\phi = \theta_2$ en la ecuación (1.1).

Dado $\boldsymbol{u} = (u_1,\cdots,u_q) \in \mathbb{R}^q$ se pueden obtener los ángulos $\boldsymbol{\theta} = (\theta_1,\cdots,\theta_{q-1})$ que lo definen bajo la siguiente transformación

$$\boldsymbol{\theta} = \begin{bmatrix} tan^{-1}\left(\sqrt{\sum_{i=2}^{q} u_i^2}/x_1\right) \\ tan^{-1}\left(\sqrt{\sum_{i=3}^{q} u_i^2}/u_2\right) \\ tan^{-1}\left(\sqrt{\sum_{i=4}^{q} u_i^2}/u_3\right) \\ \vdots \\ tan^{-1}\left(\sqrt{\sum_{i=q-1}^{q} u_i^2}/u_{q-2}\right) \\ tan^{-1}\left(u_q/u_{q-1}\right) \end{bmatrix} \tag{1.2}$$

donde $\tan^{-1}(\cdot)$ toma valores en $(-\frac{\pi}{2}, \frac{\pi}{2})$.

1.1.1. Medidas Descriptivas

Sean $\boldsymbol{u}_1,\cdots,\boldsymbol{u}_n$ puntos sobre la esfera unitaria $\mathbb{S}^q$. Entonces estos puntos pueden ser resumidos a través de su media muestral $\bar{\boldsymbol{u}} \in \mathbb{R}^q$, dada por

$$\overline{\boldsymbol{u}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{u}_i.$$

La *dirección media* de un conjunto de datos direccionales se da en términos de sus ángulos de la siguiente manera. Sean $\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_n$ los ángulos asociados a los vectores aleatorios unitarios $\boldsymbol{u}_1,\cdots,\boldsymbol{u}_n$. La *dirección media* $\overline{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_n$ se define como la dirección de la resultante $\boldsymbol{u}_1 + \cdots + \boldsymbol{u}_n$ de los

vectores $\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n$. La cual es también la *dirección del centro de masa*, $\overline{\boldsymbol{u}}$, de $\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n$.

Por ejemplo, en el caso de datos circulares, las coordenadas cartesianas de $\boldsymbol{u}_i$ son $(\cos(\theta_i), \operatorname{sen}(\theta_i))$ para $i = 1, \cdots, n$. Entonces las coordenadas cartesianas del centro de masa de $\boldsymbol{u}$ resultan ser $(\bar{u}_1, \bar{u}_2)$, donde

$$\bar{u}_1 = \frac{1}{n}\sum_{i=1}^{n}\cos(\theta_i) \qquad \bar{u}_2 = \frac{1}{n}\sum_{i=1}^{n}\operatorname{sen}(\theta_i) \tag{1.3}$$

Así, $\bar{\theta}$ es la solución a las ecuaciones

$$\bar{u}_1 = \bar{R}\cos(\bar{\theta}) \qquad \bar{u}_2 = \bar{R}\operatorname{sen}(\bar{\theta}), \tag{1.4}$$

donde

$$\bar{R} = (\bar{u_1}^2 + \bar{u_2}^2)^{\frac{1}{2}},$$

es la *longitud media resultante*, cabe mencionar que $0 \leq \bar{R} \leq 1$. Si las direcciones $\theta_1, \cdots, \theta_n$ están fuertemente agrupadas entonces $\bar{R}$ puede ser casi 1. Por otra parte, si $\theta_1, \cdots, \theta_n$ están dispersas $\bar{R}$ puede ser casi 0.
Si $\bar{R} = 0$ entonces $\bar{\theta}$ no esta definida. Cuando $\bar{R} > 0$, $\bar{\theta}$ está dada explícitamente por

$$\bar{\theta} = \begin{cases} \tan^{-1}\left(\bar{X}_2/\bar{X}_1\right) & \text{si } \bar{X}_1 \geq 0 \\ \tan^{-1}\left(\bar{X}_2/\bar{X}_1\right) + \pi & \text{si } \bar{X}_1 < 0. \end{cases}$$

Para el caso de datos esféricos (ver Mardia y Jupp (2000)), las coordenadas cartesianas de $\boldsymbol{u}_i$ son $(\cos(\theta), \operatorname{sen}(\theta)\cos(\phi), \operatorname{sen}(\theta)\operatorname{sen}(\phi)$ para $i = 1, \cdots, n$, entonces las coordenadas cartesianas del centro de masa de $\boldsymbol{U}$ son $\{\bar{X}_1, \bar{X}_2, \bar{X}_3\}$, donde

$$\bar{X}_1 = \frac{1}{n}\sum_{i=1}^{n}\cos(\theta_i), \quad \bar{X}_2 = \frac{1}{n}\sum_{i=1}^{n}\operatorname{sen}(\theta_i)\cos(\phi_i), \quad \bar{X}_3 = \frac{1}{n}\sum_{i=1}^{n}\operatorname{sen}(\theta_i)\operatorname{sen}(\phi_i). \tag{1.5}$$

Por tanto, $\bar{\theta}$ y $\bar{\phi}$ son las soluciones de las ecuaciones

$$\bar{X}_1 = \bar{R}\cos(\bar{\theta}) \qquad \bar{X}_2 = \bar{R}\operatorname{sen}(\bar{\theta})\cos(\bar{\phi}) \qquad \bar{X}_3 = \bar{R}\operatorname{sen}(\bar{\theta})\operatorname{sen}(\bar{\phi}). \tag{1.6}$$

Análogamente

$$\bar{R} = (\bar{X_1}^2 + \bar{X_2}^2 + \bar{X_3}^2)^{\frac{1}{2}}$$

Cuando $\bar{R} > 0$, $\bar{\theta}$ y $\bar{\phi}$ están dadas explícitamente por

$$\overline{\theta} = \begin{cases} \tan^{-1}\left((\bar{X}_2 + \bar{X}_3)^{\frac{1}{2}}/\bar{X}_1\right) & \text{si } \bar{X}_1 \geq 0 \\ \tan^{-1}\left((\bar{X}_2 + \bar{X}_3)^{\frac{1}{2}}/\bar{X}_1\right) + \pi & \text{si } \bar{X}_1 < 0 \end{cases}$$

$$\bar{\phi} = \begin{cases} \tan^{-1}\left(\bar{X}_3/\bar{X}_2\right) & \text{si } \bar{X}_2 \geq 0 \\ \tan^{-1}\left(\bar{X}_3/\bar{X}_2\right) + \pi & \text{si } \bar{X}_2 < 0. \end{cases}$$

Para propósitos descriptivos e inferenciales, la longitud media resultante es más importante que cualquier medida de dispersión. Sin embargo, a veces es útil contar con medidas de dispersión para datos direccionales. La más simple de éstas es la *varianza muestra direccional*, la cual se define como

$$V = 1 - \bar{R}$$

con $0 \leq V \leq 1$. Otra medida de gran ayuda es la *desviación estándar direccional* que está dada por

$$v = \{-2log(1 - V)\}^{\frac{1}{2}}.$$

1.1.2. Métodos Gráficos

Gráficamente los datos circulares se representar como puntos en la circunferencia de un circulo unitario, a este tipo de gráficos se les llama *histograma circulares*. En la Figura 1.1 se muestra los datos de 76 tortugas después de depositar sus huevos. Los datos fueron tomados de la Tabla 1.5 de Mardia y Jupp (2000)

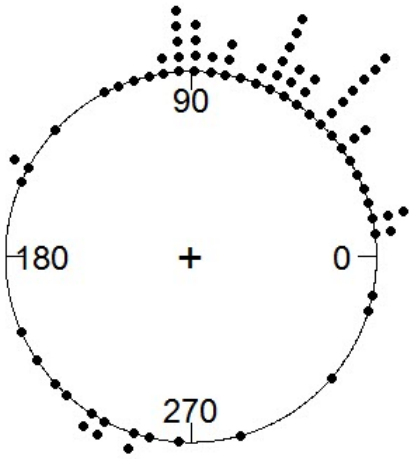

Figura 1.1: Histograma circular asociado a la orientación de 76 tortugas después de depositar sus huevos en la playa.

El *diagrama de rosas* es una variante muy útil del histograma circular, en el cual se remplazan las barras por sectores que parecen pétalos de rosas. El radio de cada sector debe ser proporcional a la frecuencia del grupo correspondiente. Ver Figura 1.2

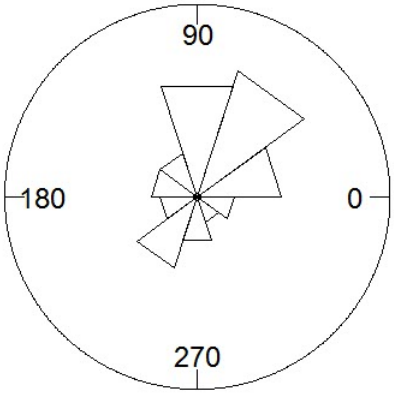

Figura 1.2: Diagrama de rosas asociado a la orientación de 76 tortugas de depositar sus huevos en la playa.

1.2. Datos Composicionales

Los datos composicionales son datos que describen cuantitativamente las partes de un todo y aportan solo información relativa entre sus componentes. Los datos composicionales aparecen en forma de vectores de dos o mas componentes todas ellas no negativas, además estas componentes se expresan como porporciones, porcentajes o partes por millón de algún todo y cuya suma es un valor constante k (igual a 1, 100 o 10^6, respectivamente).

Los datos composicionales surgen en diversas ciencias tales como, la geología (análisis de los minerales presentes en las rocas), economía (portafolios de inversión) y química (distribución de contaminantes en agua, aire y suelo).

Un dato composicional es un vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$ constituido por d partes o componentes no negativas cuyo espacio muestral es el símplex S^d definido como:

$$S^d := \{\boldsymbol{x} = (x_1, x_2, \ldots, x_d) |\ x_1 \geq 0, \ldots, x_d \geq 0\ ;\ \sum_{i=1}^{d} x_i = k\}.$$

Este espacio muestral tiene una estructura de espacio vectorial inducida por las operaciones de *perturbación* y *potenciación* que se definen líneas abajo. Tanto la operación perturbación, como la operación potenciación, hacen uso del operador *clausura* $C(\cdot)$, el cual se define para todo vector $\boldsymbol{z} \in \mathbb{R}_+^d$ como

$$C(\boldsymbol{z}) = \left(\frac{kz_1}{\sum_{i=1}^{d} z_i}, ..., \frac{kz_d}{\sum_{i=1}^{d} z_i} \right) \tag{1.7}$$

Es decir, el operador clausura lleva a todo vector no negativo D-dimensional al espacio símplex S^d. Con el operador clausura se pueden definir las operaciones de perturbación y potenciación como sigue. Sean $\boldsymbol{x}, \boldsymbol{y} \in S^d$, cuyas componentes se denotan por x_i, y_i, respectivamente y $\alpha \in \mathbb{R}$. La *perturbación* $\oplus$ de $\boldsymbol{x}$ con $\boldsymbol{y}$ se define como

$$\boldsymbol{x} \oplus \boldsymbol{y} = C(x_1 y_1, x_2 y_2, \ldots, x_d y_d)'.$$

La perturbación es una operación fundamental encargada de describir el cambio composicional en el símplex, la perturbación es el equivalente a la traslación o suma en espacios reales. La *potenciación* $\otimes$ de $\boldsymbol{x}$ con el escalar α se define como $\alpha \otimes \boldsymbol{x} = C(x_1^\alpha, x_2^\alpha, \ldots, x_d^\alpha)'$. La potenciación es el equivalente al producto por escalar en espacios reales. Las operaciones $\oplus$ y $\otimes$ definidas en S^d cumplen los requisitos de las operaciones de un espacio vectorial. Al

definir un producto escalar, una norma y distancia en el espacio muestral del símplex, éste tendrá estructura de espacio métrico. Para definir estas operaciones se necesita definir las transformaciones *clr* y *ilr*.

En los datos composicionales solo las razones o cocientes entre las componentes aportan información, es decir, los cocientes $x_i/x_j (i, j = 1, 2, \ldots, d; i \neq j)$. Analizar las magnitudes absolutas de las partes $x_1, x_2, \ldots, x_d$ carece de sentido. Este principio se denomina *invariancia por cambios de escala.*

Considere un dato composicional cuyas dos componentes $x = (x_1, x_2)$ en una muestra tiene el valor de 5 % y 10 % respectivamente, y en otra muestra tienen el valor de 50 % y 55 %. En ambas muestras la distancia euclidiana es la misma, hay un incremento de 5 unidades, pero el incremento relativo en la primera muestra es del 50 % mientras que en la segunda muestra el incremento relativo es de solo el 10 %.

La metodología que propone Aitchison (1986) para establecer la geometría del símplex se basa en transformaciones log-cocientes. Un primer intento de representar composiciones en forma de log-cocientes es la *transformación logística-aditiva* definida en Aitchison (1986) abreviada por *alr*. Si $\boldsymbol{x} \in S^d$ entonces la transformación *alr* se define como sigue

$$alr(\boldsymbol{x}) := log\left(\frac{x_1}{x_d}, \frac{x_2}{x_d}, \ldots, \frac{x_{d-1}}{x_d}\right). \tag{1.8}$$

Se observa que el cociente elimina las constantes de clausura o unidades que puedan multiplicar a las partes y el logaritmo se hace cargo del principio de *escala relativa* el cual dice que, cada una de las partes de una composición tiene escala relativa. Esta transformación adolece de la falta de simetría al elegir una de las partes, en este caso la última, como denominador común, violando el principio de *invariancia por permutación de las partes* el cual establece que, las conclusiones de un análisis composicional no deben depender de la ordenación de las partes.

Para superar la asimetría de la transformación *alr* se define la representación *log-cociente centrada clr* Aitchison (1986), como sigue

$$clr(\boldsymbol{x}) := log\left(\frac{x_1}{g(\boldsymbol{x})}, \frac{x_2}{g(\boldsymbol{x})}, \ldots, \frac{x_d}{g(\boldsymbol{x})}\right), \tag{1.9}$$

donde

$$g(\boldsymbol{x}) = \left(\prod_{i=1}^{d} x_i\right)^{\frac{1}{d}}.$$

Nótese que *clr* transforma $\boldsymbol{x} \in S^d$ a $clr\boldsymbol{x} \in \mathbb{R}^d$. Conociendo $\boldsymbol{z}$=clr($\boldsymbol{x}$) se

puede obtener $\boldsymbol{x}$ mediante la transformación inversa $clr^{-1}(\cdot)$ definida por

$$\boldsymbol{x} = clr^{-1}(\boldsymbol{z}) := C(exp(\boldsymbol{z})).$$

La transformación *clr* define una estructura métrica en S^d. Así, el producto escalar, norma y distancia de Aitchison en S^d están dados por

$$< \boldsymbol{x}, \boldsymbol{y} >_A = < clr(\boldsymbol{x}), clr(\boldsymbol{y}) >_e,$$
$$||\boldsymbol{x}||_A = ||clr(\boldsymbol{x})||_e,$$
$$d_A(\boldsymbol{x}, \boldsymbol{y}) = d_e(clr(\boldsymbol{x}), clr(\boldsymbol{y})),$$

donde el sub-sufijo e representa el producto escalar, la norma y la distancia en el espacio euclideano. Estas definiciones constituyen la métrica de Aitchison sobre el símplex.

El producto escalar, la norma y la distancia de Aitchison obedecen a los principios de análisis composicional (invarianza por escala, coherencia subcomposicional, escala relativa e invarianza por perturbación) se convierten así en instrumentos de análisis libres de incoherencias. Pero además dan una estructura euclídea al símplex. Esto sugiere utilizar los instrumentos habituales en esos espacios: bases ortonormales, representación por coordenadas (ortonormales), proyecciones ortogonales, etc. Para dar este paso es conveniente disponer de algún método para construir bases ortonormales y las correspondientes coordenadas.

Una base ortonormal en S^d es un conjunto de composiciones $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{d-1}$ tal que $< \boldsymbol{e}_i, \boldsymbol{e}_j >_A = 0$ cuando $i \neq j$ y $||e_i||_A = 1$. Fijada la base, las coordenadas de una composición son

$$\boldsymbol{w} = ilr(\boldsymbol{x}) := (< \boldsymbol{x}, \boldsymbol{e}_1 >_A, < \boldsymbol{x}, \boldsymbol{e}_2 >_A, \ldots, < \boldsymbol{x}, \boldsymbol{e}_{D-1} >_A) \tag{1.10}$$

de las cuales se pueden recuperar la composición $\boldsymbol{x}$ mediante

$$\boldsymbol{x} = ilr^{-1}(\boldsymbol{w}) = \oplus_{i=1}^{d-1}(w_i \otimes e_i)$$

La construcción de coordenadas ortonormales se llama transformación log-cociente isométrica, abreviada *ilr*. Al igual que la transformación *clr*, esta transformación tiene las siguientes propiedades

$$< \boldsymbol{x}, \boldsymbol{y} >_A = < ilr(\boldsymbol{x}), ilr(\boldsymbol{y}) >_e,$$
$$||\boldsymbol{x}||_A = ||ilr(\boldsymbol{x})||_e,$$
$$d_A(\boldsymbol{x}, \boldsymbol{y}) = d_e(ilr(\boldsymbol{x}), ilr(\boldsymbol{y})).$$

Cabe resaltar que tanto la transformación *clr*, como la transformación ilr son transformaciones isométricas.

1.2.1. Medidas Descriptivas

La estadística sintetiza la información en una muestra de datos en cualquier espacio muestral. Así, el vector de promedios y la matriz de varianzas-covarianzas son los estadísticos más frecuentes en escenarios multivariantes. A continuación se introducen medidas descriptivas y métodos gráficos usados en el análisis de datos composicionales.

El *centro* o *media composicional* de un conjunto de datos $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in S^d$ con n observaciones es la composición $\bar{\boldsymbol{x}}$ que se define como

$$\bar{\boldsymbol{x}} = C\left[exp\left(\frac{1}{n}\sum_{i=1}^{n} ln(x_{ij})\right)\right] = C\left[g_1, g_2, \ldots, g_d\right] \tag{1.11}$$

donde $j = 1, 2, \ldots, d$ y

$$g_j = \left(\prod_{i=1}^{n} x_{ij}\right)^{\frac{1}{n}}.$$

Por otro lado, como una medida global de dispersión se puede usar la **varianza total** definida como

$$totVar[X] = \frac{1}{d}\sum_{i=1}^{d-1}\sum_{j=i+1}^{d} Var[log(X_i/X_j)]$$

La estimación del centro, de la varianza total y sus componentes, puede hacerse en coordenadas ilr, y las propiedades de estos estimadores corresponderán a las de los estimadores de medias y varianzas reales (Pawlowsky-Glahn y Egozcue (2001), Pawlowsky-Glahn y Egozcue (2002)). El análisis de la variabilidad de una muestra se realiza utilizando la matriz de variación composicional de todos los log-cocientes simples, llamada en Aitchison (1986) *matriz de variación composicional*, como se define a continuación.

Definición 1.1 *Para una composición* $\boldsymbol{X}$ *de d-partes y n observaciones* $(x_1, \ldots, x_n)$ *la matriz de variación composicional esta dado por*

$$\begin{array}{c|cccccc} & 1 & 2 & 3 & \cdots & d-1 & d \\ 1 & \cdot & \tau_{12} & \tau_{13} & \cdots & \tau_{1(d-1)} & \tau_{1d} \\ 2 & \xi_{12} & \cdot & \tau_{23} & \cdots & \tau_{2(d-1)} & \tau_{2d} \\ 3 & \xi_{13} & \xi_{23} & \cdot & \cdots & \tau_{3(d-1)} & \tau_{3d} \\ \vdots & & & & & & \\ d-1 & \xi_{1(d-1)} & \xi_{2(d-1)} & \xi_{3(d-1)} & \cdots & \cdot & \tau_{(d-1)d} \\ d & \xi_{1d} & \xi_{2d} & \xi_{3d} & \cdots & \xi_{(d-1)d} & \cdot \end{array}$$

donde $\xi_{ij} = E\{log(x_i/x_j)\}$ *y* $\tau_{ij} = var\{log(x_i/x_j)\}$.

Esta matriz de variación se puede estimar mediante $\hat{\xi}_{ij}$ y $\hat{\tau}_{ij}$ que son los estimadores de la media de los log cocientes ξ_{ij} y la varianza de los log cocientes τ_{ij} respectivamente, dados por: $n\hat{\xi}_{ij} = \sum_{r=1}^{n} log(x_{ri}/x_{rj})$ y $(n-1)\hat{\tau}_{ij} = \sum_{r=1}^{n}\{log(x_{ri}/x_{rj})\}^2 - n(\hat{\xi}_{ij})^2$. La matriz de variación composicional se puede interpretar como sigue (Aitchison (1986)). Si $\hat{\xi}_{x_i x_j}$ es positivo indica que el porcentaje de x_i en la composición tiende a tener mayor peso, que el de x_j. Sin embargo, si $\hat{\xi}_{x_i x_j} < \sqrt{\hat{\tau}_{x_i x_j}}$, entonces para un número sustancial de replicas $log(\frac{x_i}{x_j})$ es negativo con un porcentaje de x_j excediendo al de la componente x_i. Por otro lado, un valor de $\hat{\tau}_{x_i x_j}$ pequeño, muestra que hay poca variabilidad relativa entre las componentes x_i y x_j. Además, si se tiene que $\hat{\xi}_{x_i x_j} > \sqrt{\hat{\tau}_{x_i x_j}}$ esto indica que la proporción de la componente x_i es apreciablemente más grande que la proporción de la componente x_j.

1.2.2. Métodos Gráficos

Los datos composicionales, para ciertas dimensiones, pueden ser representados por diagramas ternarios (diagramas de dispersión de tres componentes) y también como secuencias de diagramas de barras. A continuación se explicara en que consiste cada uno de estos métodos gráficos.

Diagrama ternarios

Cuando los datos composicionales constan de 3 partes suelen representarse mediante *diagramas ternarios*. Estos consisten en representar los datos composicionales sobre un triángulo equilátero, en el cual se puede graficar el porcentaje que tiene cada uno de los componentes del vector, entre más alejado este el dato de un vértice del triángulo significa que ese dato tiene un valor muy pequeño de esa componente y viceversa si el dato está muy cercano al vértice significa que ese dato tiene un valor alto de tal componente, ver Figura 1.3.

Diagrama de barras

Un *diagrama de barras* es una representación de todas las partes de la composición. En un diagrama de barras, uno representa la cantidad de cada parte de un *individuo* como una barra dentro de un conjunto. Es decir, las barras se apilan con la altura correspondiente de cada una de las partes de la composición del individuo hasta añadir el total de la composición ver Figura 1.4.

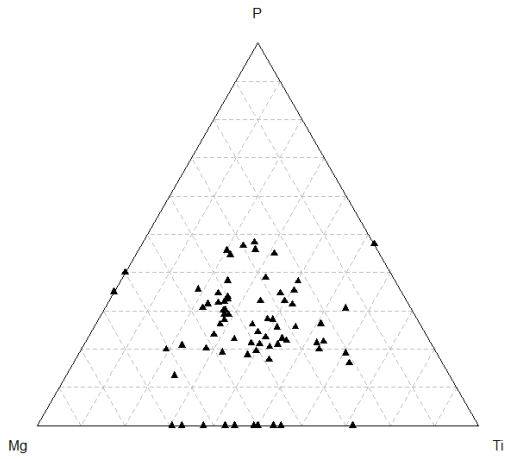

Figura 1.3: Diagrama ternario asociado a datos de cerámica.

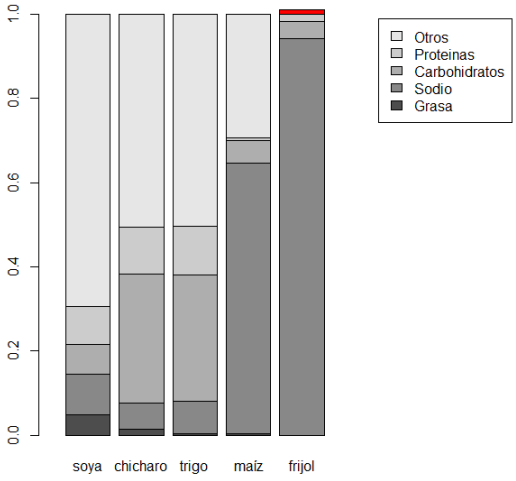

Figura 1.4: Diagrama de barras asociado a 5 datos composicionales cuyas componentes son $X = (Grasa, Sodio, Carbohidratos, Proteinas, Otros)$.

Para ver más métodos gráficos el lector se puede referir por ejemplo a Geral van den Boogaart y Tolosana-Delgado (2013).

1.2.3. Modelos de probabilidad para datos composicionales

En inferencia estadística paramétrica se especifican modelos de probabilidad para el estudio de una población. En el análisis estadístico de datos composicionales la distribución de probabilidad más usada es la distribución logística aditiva, también llamada normal en el simplex. Otras distribuciones que se usan frecuentemente son la distribución de Dirichlet y sus variantes, tales como las distribuciones beta-multivariadas. Para más sobre estas distribuciones véanse (Aitchison (1986), Geral van den Boogaart y Tolosana-Delgado (2013), Pawlowsky-Glahn *et al.* (2015)).

En esta sección se hablará sobre la distribución normal en el simplex, pero antes se introducen algunos conceptos relevantes.

Cuando se trata con datos composicionales no puede prescindirse de la geometría de su espacio muestral S^D y, en particular, de la distancia de Aitchison en S^D. Siguiendo el planteamiento en Pawlowsky-Glahn y Egozcue (2001), se define la esperanza o centro de $\mathbf{X}$ y la varianza total en S^D.

Definición 1.2 *Sea* $\boldsymbol{X}$ *una composición aleatoria. La esperanza en* S^D *o centro de* $\boldsymbol{X}$ *es*

$$E[\boldsymbol{X}] = cen[\boldsymbol{X}] = \underset{z \in S^D}{\arg\min}\{Var(\boldsymbol{X}, z)\}$$

y el valor mínimo alcanzado es la varianza total

$$totvar[\boldsymbol{X}] = \min_{z \in S^D}\{Var(\boldsymbol{X}, z)\},$$

donde $Var(\boldsymbol{X}, z) = E[d_A^2(\boldsymbol{X}, z)]$.

En Pawlowsky-Glahn y Egozcue (2001) se dan dos resultados esenciales para $E[\boldsymbol{X}]$ y $totvar[\boldsymbol{X}]$.

Proposición 1.1 *Si* $h : \xi \to \mathbb{R}^m$ *es una isometría y* $h(\boldsymbol{X}) = Y \in \mathbb{R}^m$ *entonces*

- $E[\boldsymbol{X}] = h^{-1}(E[h(\boldsymbol{X})])$.
- $totvar[\boldsymbol{X}] = \sum_{i=1}^{m} Var[Y_i]$.

Es decir

$$E[\boldsymbol{X}] = ilr^{-1}(E[ilr(X)]) = clr^{-1}(E[clr(X)]),$$

$$totVar[X] = \sum_{i=1}^{D} Var[clr_i(X)] = \sum_{j=1}^{D-1} Var[ilr_j(X)],$$

ya que, tanto la transformación *clr* y *ilr* son transformaciones isométricas, es decir, transformaciones que preservan distancias.

1.2.4. La distribución normal en el simplex

La *distribución logística normal aditiva* fue introducida en (Aitchison (1986)). La idea básica fue representar la composición aleatoria en términos de la transformación log-cociente aditiva (*alr*) y asumir que la composición bajo esta transformación sigue una distribución normal multivariada. La misma distribución es obtenida, si la composición bajo la transformación *ilr* sigue una distribución normal multivariada, con la desventaja que las coordenadas *ilr* corresponden a una base ortonormal en el simplex. Por esta razón, el nombre de la distribución logística normal aditiva fue simplificado a distribución logística normal y más tarde cambiado a distribución normal en el simplex (Mateu-Figueras *et al.* (2013)).

Definición 1.3 *Dado una composición aleatoria* $\boldsymbol{X} \in S^D$*, se dice que* $\boldsymbol{X}$ *sigue una distribución normal en el simplex* S^D *si el vector aleatorio de coordenadas ortonormales ilr(*$\boldsymbol{X}$*) sigue una distribución normal multivariante en* $\mathbb{R}^{D-1}$*, es decir,* $ilr\boldsymbol{X} \sim N_{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Por lo tanto, si $\mathbf{X}$ sigue una distribución normal en el simplex, su función de densidad de probabilidad esta dada por

$$f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D-1}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(ilr(\boldsymbol{x}) - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(ilr(\boldsymbol{x}) - \boldsymbol{\mu})'\}.$$

Con $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$, su vector de media y matriz de varianza, respectivamente. El parámetro $\boldsymbol{\mu}$ es el vector de medias de las coordenadas ilr($\mathbf{X}$). El centro de una composición $\mathbf{X}$ bajo este modelo esta dado por la Proposición (1.1). Es decir,

$$cen[\boldsymbol{X}] = ilr^{-1}(E[ilr(\boldsymbol{X})]) = ilr^{-1}(\boldsymbol{\mu}),$$

1.2.5. Principales Problemas

En el análisis de datos composicionales existen algunos problemas importantes que no han sido resueltos satisfactoriamente y que siguen siendo temas de análisis. A continuación se discuten alguno de ellos.

Chayes [1960] identifica como una de las principales dificultades, la restricción de la suma constante. Esta restricción impide la aplicación de los procedimientos estadísticos habituales. Se puede notar que, el cambio en una de las partes provoca el cambio en por lo menos una de las partes restantes.

Karl Pearson [1987] ya manifestaba la imposibilidad de interpretar correctamente las covarianzas y los coeficientes de correlación para datos composicionales. La matriz de correlación no puede analizarse para los datos

composicionales por que presentan necesariamente correlaciones negativas no nulas, esto debido a la restricción de suma constante. Pearson calificó estas correlaciones como correlaciones espurias. Por otra parte, si se analiza la matriz de covarianzas entre las partes de una composición, $\Sigma = \{cov(x_i, x_j) : i, j = 1, 2, \ldots, D\}$, se obtiene que

$$cov(x_i, x_1) + cov(x_i, x_2) + \cdots + cov(x_i, x_D) = 0 \qquad i = 1, 2, \ldots, D.$$

Por otra parte $cov(x_i, x_i) = var(x_i) > 0$, esto provoca que necesariamente debe haber una covarianza $cov(x_i, x_j)(i \neq j)$ de signo negativo.

Otro problema que hay en el análisis de datos composicionales es el principio de *coherencia subcomposicional*. Para entender este principio se establece la siguiente definición.

Definición 1.4 *Si S es un subconjunto cualquiera de las partes $1, 2, \ldots, D$ de un dato composicional $\boldsymbol{x} \in S^D$ y $\boldsymbol{x}_S$ simboliza el subvector formado por las correspondientes partes de $\boldsymbol{x}$, entonces $\boldsymbol{x_s} = C(\boldsymbol{x}_S)$ recibe el nombre de subcomposición de las S partes de $\boldsymbol{x}$, donde $C(\cdot)$ es el operador clausura.*

El principio de *coherencia subcomposicional* establece lo siguiente: Cuando se examina un subconjunto de las partes de una composición, una subcomposición, se requiere que los resultados del análisis no sean contradictorios con los obtenidos por la composición original.

Aitchison (1986) muestra este problema mediante un sencillo ejemplo. Se consideran dos científicos A y B que analizan muestras de tierra. Para cada una de las muestras, el científico A calcula un dato composcional de 4 partes (animal, vegetal, mineral, agua). El científico B elimina el agua de las muestras y calcula datos composicionales de 3 partes (animal, vegetal, mineral). Se puede notar que los datos del científico B son subcomposiciones de los datos del científico A. Los datos obtenidos son:

$(x_1; x_2; x_3; x_4)$	$(s_1; s_2; s_3)$
(0.1;0.2;0.1;0.6)	(0.25;0.50;0.25)
(0.2;0.1;0.1;0.6)	(0.50;0.25;0.25)
(0.3;0.3;0.2;0.2)	(0.375;0.375;0.25)

Al calcular la correlación de los datos composicionales entre las partes animal y vegetal se obtiene que para el científico A $corr(x_1, x_2)$=0.5 mientras que para el científico B es de $corr(s_1, s_2) = -1$. Así, se puede observar un problema de coherencia subcomposicional.

Otro problema muy importante es el análisis de datos composicionales con componentes nulas, ya que el análisis de los datos composicionales se hace mediante las trasformaciones log-cocientes. A continuación se presenta una breve introducción al tratamiento existente en la literatura sobre la presencia de ceros en variables composicionales.

Ceros en datos composicionales

Usualmente los datos composicionales contienen valores ceros, que pueden o no corresponder a la verdadera ausencia de una componente. Las formas más comunes de los ceros son los que resultan de tomar muestras muy pequeñas e instrumentos insuficientemente precisos. En esta sección se expuso que la teoría de Atchinson para el análisis de datos composicionales se basa en transformaciones de log-cocientes. Sin embargo, ¿que pasaría si uno o más componentes del dato composicional fueran ceros?. El lector puede constatar que las transformaciones propuestas anteriormente ya no se pueden utilizar, ya que $\ln(0)$ no esta definido y peor aún, la división entre cero no esta definida. A continuación se dará una breve explicación de la clasificación de los ceros en los datos composicionales (ceros esenciales, ceros por conteo y ceros redondeados) según su naturaleza, así mismo, se explicarán los diferentes tratamientos que se han propuesto para lidiar con esta dificultad en el análisis de datos composicionales.

Ceros esenciales o estructurales

Los ceros esenciales son aquellos ceros que son verdaderos en los datos composicionales. Por ejemplo, en la planeación de gastos de las familias, el dinero se reparte en productos de la canasta básica, transporte, diversiones, ropa, tabaco y alcohol y otros gastos. Hay familias que no fuman ni beben alcohol, por lo tanto, para este tipo de familias la componente de tabaco y alcohol del dato composicional será cero. Éste es un ejemplo de un cero esencial, en Aitchison y Kay (2003), Bacom-Shone (2003), y Bacom-Shone (2008) el lector puede ver las metodologías que se han propuesto para resolver los problemas con este tipo de ceros.

Ceros por redondeo

Este tipo de ceros aparece sobre todo en aquellos estudios en los que las variables son continuas (por ejemplo, porcentajes por pesos, tiempos, gastos, o longitudes).

Los ceros por redondeo corresponden con valores que no han podido observarse por limitaciones en los instrumentos de medida o en el procedimiento

de recolección y tratamiento de los datos, o incluso por políticas que impiden que se registren cuantías pequeñas que no superan cierto umbral de detección.

Tratamiento de ceros por redondeo

Las técnicas para tratar este tipo de ceros se dividen en técnicas no paramétricas y paramétricas. El grupo de técnicas no paramétricas para ceros redondeados consiste esencialmente en una familia de estrategias de imputación, aquí imputación es equivalente a forzar al conjunto de datos incompletos, a un conjunto de datos completos, poniendo una cantidad a cada valor faltante o cero redondeado. Por otro lado, el grupo de técnicas paramétricas para ceros redondeados (por ejemplo el algoritmo EM) se basa en modelos paramétricos para datos multivariantes.

Métodos de reemplazo no paramétrico

Cuando los valores que faltan son en realidad datos censurados, es decir, los valores de algunos componentes se indican como *menor que* un valor umbral dado, una simple imputación puede ser considerada. Para el tratamiento de ceros por redondeo se tienen dos técnicas de reemplazo no paramétrico, una de ellas es el reemplazo aditivo propuesto en Aitchison (1986) y la otra es el reemplazo multiplicativo propuesta en Martín-Fernández *et al.* (2003)

Primero se explicará en que consiste el método de reemplazo aditivo y después el método de reemplazo multiplicativo.

Sea $\boldsymbol{x} \in S^D$ que contiene Z ceros redondeados, entonces $\boldsymbol{x}$ puede ser representado por una nueva composición $r \in S^D$ sin ceros de acuerdo con la siguiente regla de sustitución.

$$r_j = \begin{cases} \frac{\delta(Z+1)(D-Z)}{D^2}, & \text{si } x_j = 0, \\ x_j - \frac{\delta(Z+1)Z}{D^2}, & \text{si } x_j > 0, \end{cases} \tag{1.12}$$

donde δ es un valor pequeño, inferior a un umbral determinado. Observemos que se puede generalizar 1.12 para umbrales diferentes, es decir, un umbral distinto para cada x_j. La estrategia de reemplazo multiplicativo consiste en lo siguiente. Sea $\boldsymbol{x} \in S^D$ que contiene Z ceros redondeados, se propone reemplazar $\boldsymbol{x}$ con una composición $r \in S^D$ sin ceros usando la expresión

$$r_j = \begin{cases} \delta_j, & \text{si } x_j = 0, \\ \left(1 - \frac{\sum_{k|x_k=0} \delta_k}{c}\right) x_j, & \text{si } x_j > 0, \end{cases} \tag{1.13}$$

donde δ_j es el valor imputado por parte de x_j y c es la constante de la restricción de la suma. Este reemplazo es natural en el sentido de que, si los valores imputados δ_j en una composición $\boldsymbol{x}$ son iguales a los verdaderos valores censurados, entonces r recupera la composición verdadera, sin embargo, el valor imputado no depende de la cantidad de ceros ni de las componentes del dato composicional.

Martín-Fernández *et al.* (2003) recomiendan utilizar δ_j igual al 65 porciento del valor del umbral, esto debido a pruebas de sensibilidad sobre el parámetro.

Métodos de reemplazo paramétrico

El algoritmo EM (Expectation - Maximization), es una herramienta muy conocida para resolver problemas que involucran datos no observados en un espacio real. Sin embargo, en su forma estándar, no es capaz de resolver el problema de ceros redondeados. Por este motivo, en Palarea-Albaladejo *et al.* (2007) y Palarea-Albaladejo y Martín-Fernández (2008) se introduce una modificación al algoritmo EM que junto con la transformación log-cociente aditiva (1.8) genera estimaciones adecuadas para valores por debajo del límite de detección. Considérese un conjunto de n datos composicionales $\boldsymbol{X} = [x_{ij}]$ y D componentes. Es decir un conjunto de n renglones y D columnas que incluyan ceros redondeados. Al conjunto de datos $\boldsymbol{X}$ se le aplica la transformación *alr* para llevar los datos al espacio real sin restricción y así tener los datos $Y = [y_{ij}]$. Si $x_{ij} < \varepsilon_{ij}$ la transformación *alr* produce un valor y_{ij} que es un dato faltante en Y, donde ε_{ij} es el umbral para el cual valores menores que él se consideran ceros por redondeo. El procedimiento supone que la composición aleatoria $\boldsymbol{X}$ se distribuye de acuerdo a un modelo *logistico normal aditivo*, es decir, si al aplicar la transformación $\boldsymbol{Y}$=alr($\boldsymbol{X}$), $\boldsymbol{Y}$ sigue una distribución $N_{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, entonces se dice que $\boldsymbol{X}$ tiene una distribución logistico normal aditiva. Después en la iteración t-ésima el paso E modificado sustituye los valores y_{ij} en el conjunto de datos Y por:

$$y_{ij}^{(t)} = \begin{cases} y_{ij}, & \text{si } y_{ij} \geq \Psi_{ij}, \\ E[y_{ij}|y_{i,-j}, y_{ij} < \Phi_{ij}, \mu^t, \Sigma^t], & \text{si } y_{ij} < \Psi_{ij}, \end{cases} \tag{1.14}$$

donde $\Psi_{ij} = ln(\frac{\varepsilon_{ij}}{x_{iD}})$ y

$$E[y_{ij}|y_{i,-j}, y_{ij} < \Phi_{ij}, \mu^t, \Sigma^t] = y_{i,-j}^T \beta_j - \sigma_j \frac{\phi(\frac{\Psi_{ij} - y_{i,-j}^T \beta_j}{\sigma_j})}{\Phi(\frac{\Psi_{ij} - y_{i,-j}^T \beta_j}{\sigma_j})}, \tag{1.15}$$

donde $y_{i,-j}$ denota el conjunto de variables observadas para la fila i de la matriz de datos $\boldsymbol{Y}$, Φ y ϕ denotan las funciones de distribucion y densidad de una normal estándar respectivamente, σ_j^2 es la varianza condicional de la variable y_j, y β_j denota el vector de coeficientes de la regresión lineal de y_{ij} sobre $y_{i,-j}$.

En Martín-Fernández *et al.* (2001) comparan el método de reemplazo multiplicativo con el algoritmo EM modificado, donde se explica que los dos métodos dan resultados similares cuando hay una proporcion de ceros menor del diez porciento del total del conjunto de datos, y que si hay una mayor cantidad de este porcentaje se debe utilizar el algoritmo EM modificado.

Ceros por conteo

Los últimos avances en el tratamiento de la composición de ceros se han centrado sobre todo en los ceros de carácter esencial y en el caso de ceros por redondeo. Cuando tenemos ceros por conteo en un conjunto de datos tenemos un nuevo tipo de cero relacionado con el problema de muestreo. Las piezas no son observadas debido al tamaño limitado de la muestra. Un ejemplo donde podemos encontrar este tipo de ceros es al momento de tratar de estimar la cantidad de peces de cada especie en un lago. Se toma una muestra de N pescados de un lago donde se sabe que hay k especies de peces, pero debido a el pequeño tamaño de muestra que se puede tomar, al hacer el conteo no se registra una o mas especies de peces. Por lo tanto, nuestro objetivo consiste en estimar los valores de θ_j (promedio de la especie j) con el fin de obtener los verdaderos valores de $N\theta_j$ rompiendo las limitaciones de la muestra.

La metodología propuesta en Daunis-Estadella *et al.* (2008) para el tratamiento de ceros por conteo se basa en la estimación bayesiana y en el método de reemplazo multiplicativo.

$$x_j^* = \begin{cases} \frac{\alpha_j}{N+s}, & \text{si } x_j = 0, \\ x_j \left(1 - \frac{\sum_{x_k=0} \alpha_j}{N+s}\right), & \text{si } x_j > 0, \end{cases} \tag{1.16}$$

Bajo esta propuesta todos los porcentajes ceros son reemplazados por su valor esperado a posteori y los porcentajes distintos de cero se multiplican por un factor de acuerdo con el número de ceros que hay en el dato composicional.

1.3. Estadística bayesiana

La estadística bayesiana es una alternativa a la estadística clásica en los problemas de inferencia como son la estimación de parámetros, prueba de

hipótesis y regresión. Tanto el enfoque clásico de inferencia, como el enfoque bayesiano se desarrollan en presencia de observaciones $\boldsymbol{x}$ cuyos valores son inicialmente inciertos y se describen a través de una función de densidad de probabilidad $f(x|\theta)$. El parámetro θ sirve como una referencia para la elección de posibles distribuciones para las observaciones, que representan las características de interés que se desearía saber, con el fin de obtener una descripción completa del fenómeno que se este analizando.

Para hacer inferencia se necesita extraer una muestra aleatoria de una población que se distribuya de acuerdo con la densidad $f(x|\theta)$. El parámetro θ es la cantidad de interés que tiene un significado relevante en el problema en estudio. Además, es probable que el investigador tenga algún conocimiento sobre este parámetro. Es posible que este conocimiento sea incorporado formalmente en el análisis. En esta parte, el enfoque bayesiano y el enfoque clásico divergen. El enfoque bayesiano permite incorporar esta información para el análisis a través de una medida $p(\theta)$, incluso cuando esta información no es precisa. En la inferencia bayesiana la especificación de un modelo contiene dos ingredientes: la verosimilitud $f(x|\theta)$ y la distribución $p(\theta)$. La distribución $p(\theta)$ se puede especificar con la ayuda de parámetros constantes. Estas constantes se denominan hiperparámetros, ya que son los parámetros de la distribución de los parámetros. Inicialmente, los hiperparámetros se asumen conocidos. El segundo ingrediente $p(\theta)$ es llamado *distribución inicial*, esta especifica una medida sobre θ antes de observar los valores de $\boldsymbol{x}$. Una vez que el problema se da en esta forma, es natural que la inferencia debe basarse en la distribución de probabilidad de θ después de observar los valores de $\boldsymbol{x}$, que pasan a formar parte del conjunto de información disponible. Esta distribución $p(\theta|x)$ se llama *distribución final* y se puede obtener por medio del teorema de Bayes de la siguiente manera:

$$p(\theta|x) = \frac{f(x|\theta)p(\theta)}{f(x)} \tag{1.17}$$

donde

$$f(x) = \int f(x|\theta)p(\theta)d\theta$$

El término $f(x)$ es sólo una constante de normalización que no depende de θ, por lo cual la distribución final (1.17) se puede escribir como

$$p(\theta|x) = f(x|\theta)p(\theta) \tag{1.18}$$

Las inferencias se hacen sobre $p(\theta|x)$ en lugar de $f(x|\theta)$; es decir, sobre la distribución de probabilidad (del valor desconocido) del parámetro dados

los datos observados, en vez de la distribución de los datos dado el valor del parámetro. Para una revisión detallada sobre estadística bayesiana se puede consultar Bolstad (2007), Leonard y Hsu (1999), Berry (1996), Bernado (1994), Berger (1985) y Box y Tiao (1973).

A continuación se muestra el proceso de obtención de la distribución final, a través del teorema de Bayes.

Ejemplo 1.1 *Sea* $\boldsymbol{y} = \{y_1, \ldots, y_n\}$ *una muestra aleatoria de una población* $Normal(\mu, 1)$*, con* μ *desconocido. Se propone como distribución inicial una distribución* $Normal(\mu|\mu_0, \lambda_0)$*. El objetivo es encontrar la distribución final de* μ*, es decir,* $f(\mu|\mathbf{y})$ *.*

Dada la muestra aleatoria $\{y_1, \ldots, y_n\}$ del modelo $Normal(\mu, 1)$ entonces se puede derivar la función de verosimilitud, la cual resulta ser:

$$l(y_1, \ldots, y_n|\mu) = \frac{1}{(2\pi)^{\frac{n}{2}}} exp\{-\frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2\}.$$

Ocupando la siguiente igualdad

$$\sum_{i=1}^{n}(y_i - \mu)^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2 + n(\overline{y} - \mu)^2,$$

la función de verosimilitud se puede escribir como

$$l(y_1, \ldots, y_n|\mu) = \frac{1}{(2\pi)^{\frac{n}{2}}} exp\{-\frac{1}{2}\sum_{i=1}^{n}(y_i - \overline{y})^2 - \frac{n}{2}(\overline{y} - \mu)^2\}.$$

Así,

$$\begin{aligned}
&f(\mu|y) = f(y|\mu)f(\mu) = \textstyle\prod_{i=1}^{n} N(y_i|\mu, 1)N(\mu|\mu_0, \lambda_0) \\
&\propto exp\{-\tfrac{1}{2}\textstyle\sum_{i=1}^{n}(y_i - \overline{y})^2 + n(\overline{y} - \mu)^2\}exp\{-\tfrac{\lambda_0}{2}(\mu - \mu_0)^2\} \\
&= exp\{-\tfrac{1}{2}\textstyle\sum_{i=1}^{n}(y_i - \overline{y})^2\}exp\{-\tfrac{1}{2}n(\overline{y} - \mu)^2\}exp\{-\tfrac{\lambda_0}{2}(\mu - \mu_0)^2\} \\
&\propto exp\{-\tfrac{1}{2}n(\overline{y} - \mu)^2\}exp\{-\tfrac{\lambda_0}{2}(\mu - \mu_0)^2\},
\end{aligned}$$

trabajando con la expresión $n(\overline{y} - \mu)^2 + \frac{\lambda_0}{2}(\mu - \mu_0)^2$ se tiene

$$
\begin{aligned}
n(\overline{y}-\mu)^2+\tfrac{\lambda_0}{2}(\mu-\mu_0)^2 &= n(\overline{y}^2-2\overline{y}\mu+\mu^2)+\tfrac{\lambda_0}{2}(\mu^2-2\mu\mu_0+\mu_0^2) \\
&= (n+\lambda_0)\mu^2-2(n\overline{y}+\lambda_0\mu_0)\mu+n\overline{y}^2+\lambda_0\mu_0^2 \\
&= (n+\lambda_0)[\mu^2-2\tfrac{n\overline{y}+\lambda_0\mu_0}{n+\lambda_0}\mu+(\tfrac{n\overline{y}+\lambda_0\mu_0}{n+\lambda_0})^2]+n\overline{y}^2+\lambda_0\mu_0^2 \\
&- (\tfrac{n\overline{y}+\lambda_0\mu_0}{n+\lambda_0})^2 \\
&= [\mu-(\tfrac{n\overline{y}+\lambda_0\mu_0}{n+\lambda_0})]^2+n\overline{y}^2+\lambda_0\mu_0^2-(\tfrac{n\overline{y}+\lambda_0\mu_0}{n+\lambda_0})^2.
\end{aligned}
$$

Por lo anterior,

$$
\begin{aligned}
&f(\mu|\boldsymbol{y}) \propto exp\{-\tfrac{1}{2}n(\overline{y}-\mu)^2\}exp\{-\tfrac{\lambda_0}{2}(\mu-\mu_0)^2\} \\
&= exp\{[\mu-(\tfrac{n\overline{y}+\lambda_0\mu_0}{n+\lambda_0})]^2+n\overline{y}^2+\lambda_0\mu_0^2-(\tfrac{n\overline{y}+\lambda_0\mu_0}{n+\lambda_0})^2\} \\
&\propto exp\{-\tfrac{1}{2}(n+\lambda_0)[\mu-\tfrac{n\overline{y}+\lambda_0\mu_0}{n+\lambda_0}]^2\},
\end{aligned}
$$

lo cual implica que:

$$
f(\mu|\boldsymbol{y}) \sim Normal(\frac{n\overline{y}+\lambda_0\mu_0}{n+\lambda_0}, n+\lambda_0).
$$

Es decir, la distribución final de μ (dado los datos) es una distribución Normal con media y varianza dadas por

$$
\mu_n = \frac{n\overline{y}+\lambda_0\mu_0}{n+\lambda_0},
$$

$$
\lambda_n = n+\lambda_0.
$$

1.3.1. Métodos Numéricos y de Simulación

Dado un modelo observado $f(x|\theta)$ y una distribución inicial $p(\theta)$, con $\theta = (\theta_1, \theta_2, \ldots, \theta_k)'$, se puede obtener, vía el teorema de Bayes, la distribución final $\pi(\theta|\boldsymbol{y})$ y a partir de ésta realizar inferencia sobre θ. Por ejemplo, se puede tener interés en obtener las densidades marginales de algunas componentes de θ o bien, las densidades finales de alguna función de estos componentes, tales como cocientes o productos; o en general explorar y resumir

distribuciones finales. Salvo en ciertos casos donde los procedimientos analíticos son posibles, la aplicación de las técnicas Bayesianas para la solución de problemas involucra el uso de métodos numéricos y/o técnicas de simulación.

La clave para la implementación de la solución, en la mayoría de los problemas es, por un lado, la habilidad para obtener o evaluar cierto número de integraciones y, por otro lado, la posibilidad de obtener muestras de alguna distribución final. Por lo anterior, a continuación se discuten algunas técnicas de simulación y métodos numéricos.

Métodos MCMC

Los métodos de Monte Carlo vía Cadenas de Markov (MCMC) son algoritmos que permiten obtener una muestra de una distribución de probabilidad π sin necesidad de simular directamente de dicha distribución. Para ello estos métodos se basan en la construcción de unas cadenas de Markov cuya distribución de equilibrio es precisamente π.

Dado un punto inicial arbitrario $x^{(0)}$, se construye una cadena de Markov ergódica $X^{(t)}$ con $t \in \mathbb{N}$ cuya distribución estacionaria es la distribución de interés π. Esto garantiza que, para un valor $l \in \mathbb{N}$ suficientemente grande, se cumple $X^{(l)}, X^{(l+1)}, X^{(l+2)}$ tengan una distribución f. Se debe notar que $X^{(l)}$ y $X^{(l+1)}$ no son independientes; sin embargo, para un cierto valor $k \in \mathbb{N}$, se puede considerar que $X^{(l)}$ y $X^{(l+k)}$ son aproximadamente independientes. Por lo tanto si se simula dicha cadena y se define una nueva cadena como $Y^{(t)} = X^{(l+kt)}$, entonces se obtiene una muestra aproximadamente independiente de la distribución π. El valor l determina el tiempo necesario para que la cadena converja a la distribución estacionaria, mientras, que el valor k indica cada cuántas simulaciones se deben considerar para tomar una nueva observación, por tanto, la cantidad final de simulaciones va ha estar determinado por estos dos valores. Dos de los algoritmos más comunes de los métodos MCMC son el muestreo de Gibbs y Metropolis los cuales se describen a continuación.

Muestreo de Gibbs

El muestreo de Gibbs es un método MCMC donde el *kernel de transición* está formado por todas las distribuciones condicionales completas. Supongamos que la distribución de interés es $\pi(\theta)$ donde $\theta = (\theta_1, \ldots, \theta_d)'$, donde cada una de las componentes θ_i puede ser un escalar, un vector o una matriz. Considerese también que la distribución condicional completa $\pi_i(\theta_i) = \pi_i(\theta_i|\theta{-}i),\ i = 1, \ldots, d$ es conocida y se puede simular. El problema a resolver es obtener de π simulaciones directas cuando las generaciones

son costosas, complicadas o simplemente no están disponible. Sin embargo, cuando las generaciones de π_i son posibles, el muestreo de Gibbs provee una alternativa basada en generaciones sucesivas de las distribuciones condicionales completas. El algoritmo se puede describirse de la siguiente manera:

1. Inicializar el contador de las iteraciones de la cadena $j = 1$ e inicializar valores del conjunto $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_d^{(0)})'$;

2. Obtener un nuevo valor $\theta^{(j)} = (\theta_1^{(j)}, \ldots, \theta_d^{(j)})'$ a partir de $\theta^{(j-1)}$ generando valores mediante

$$\begin{aligned}
\theta_1^{(j)} &\sim \pi(\theta_1|\theta_2^{(j-1)}, \ldots, \theta_d^{(j-1)}) \\
\theta_2^{(j)} &\sim \pi(\theta_2|\theta_1^{(j)}, \theta_3^{(j-1)}, \ldots, \theta_d^{(j-1)}) \\
&\vdots \\
\theta_d^{(j)} &\sim \pi(\theta_d|\theta_1^{(j)}, \ldots, \theta_{d-1}^{(j)})
\end{aligned}$$

3. Cambiar el contador j a $j + 1$ y volver al paso 2 hasta lograr la convergencia.

Cuando se alcanza la convergencia, el valor resultante θ seguirá la distribución de probabilidad π. Conforme el número de iteraciones incrementa, la cadena se aproxima a su condición de equilibrio. Por lo tanto, tomar n valores con k saltos de esta cadena después de un período de calentamiento l proporcionará una muestra de observaciones de π.

Metropolis-Hasting

Este algoritmo construye una cadena de Markov con espacio de estados $\boldsymbol{X}$ y distribución de equilibrio $\pi(x)$, definiendo la probabilidad de transición de $x^t = x$ a x^{t+1} de la siguiente manera.

Si $q(x^*, x)$ es una función de probabilidad de transición, de manera que, si $x^t = x$, x^* obtenida de $q(x^*, x)$ se considera como un valor candidato para $x^{t+1} = x^*$. Con cierta probabilidad $\alpha(x^*, x)$ se acepta $x^{t+1} = x^*$. Si se rechaza el valor candidato de $q(x^*, x)$, entonces $x^{t+1} = x$.

Así el algoritmo de M-H se puede describir como sigue:

(1) Generar $x^* \sim q(x^*, x)$;

(2) Generar $u \sim U(0, 1)$
si $u \leq \alpha(x^*, x)$ entonces $x^{t+1} = x^*$,
si $u > \alpha(x^*, x)$ entonces $x^{t+1} = x$.

La construcción anterior define una cadena de Markov con probabilidades de transición dadas por

$$p(x^*, x) = \begin{cases} q(x^*, x)\alpha(x^*, x) & \text{si } x^* \neq x \\ 1 - \sum_{x^{**}} q(x^{**}, x)\alpha(x^{**}, x) & \text{si } x^* = x \end{cases} \tag{1.19}$$

Si se define

$$\alpha(x^*, x) = \begin{cases} min\left[\frac{\pi(x^*)q(x,x^*)}{\pi(x)q(x^*,x)}, 1\right] & \text{si } \pi(x)q(x^*, x) > 0 \\ 1 & \text{si } \pi(x)q(x^*, x) = 0 \end{cases} \tag{1.20}$$

se puede verificar que $\pi(x)p(x, x^*) = \pi(x^*)p(x^*, x)$, lo cual junto con la condición de que $q(x^*, x)$ sea ergódica y aperiodica, es una condición suficiente para que $\pi(x)$ sea la distribución de equilibrio de la cadena construida.

Es importante notar que $\alpha(\cdot, \cdot)$ solo depende de $\pi(x)$ a través del cociente $\frac{\pi(x^*)}{\pi(x)}$.

Lo anterior es vital en contextos donde $\pi(x)$ es una distribución final, lo que significa que la constante de normalización no es necesaria para la implementación de este algoritmo. Claramente, diferentes selecciones específicas de $q(x^*, x)$ dan como resultado diferentes algoritmos. En particular si $q(x^*, x) = q(x, x^*)$ se tiene

$$\alpha(x^*, x) = min\left[\frac{\pi(x^*)}{\pi(x)}, 1\right]$$

el cual es llamado el algoritmo de Metropolis. Para mayores detalles de este y otros algoritmos MCMC el lector puede referirse, por ejemplo, a Gamerman y Lopes (2006).

Para finalizar esta sección, a continuación se presenta el método de Newton-Raphson. Aunque este método no es un procedimiento de simulación, se puede aplicar para encontrar raíces de la función $u - F(x)$,

donde $u \sim U(0, 1)$ y F es la función de distribución de la cual se quiere simular. Lo anterior es una técnica que genera observaciones $x \sim F(x)$, de manera numérica.

Newton-Raphson

El método de Newton-Raphson es uno de los más utilizados para localizar raíces ya que en general es muy eficiente y siempre converge para una función polinomial. Se requiere que las funciones sean diferenciables, y por tanto, continuas, para poder aplicar este método. Se debe partir de un valor inicial

x_i cercano a la raíz. La fórmula de Newton-Raphson para el calculo de raíces se deduce a partir de la fórmula de la pendiente de una recta. Es decir,

$$m = \frac{f(x_2) - f(x_1)}{x_2 - x_1} = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} = \frac{0 - f(x_i)}{x_{i+1} - x_i}$$

entonces se tiene que

$$m(x_{i+1} - x_i) = -f(x_i)$$

$$x_{i+1} = x_i - \frac{f(x_i)}{m}$$

En este punto es cuando se ocupa la hipótesis de que la función sea diferenciable, ya que la derivada de una función en un punto x se define como la pendiente de la recta tangente en dicho punto, es decir, $m = f(x)$. Por tanto la fórmula anterior se puede escribir como

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \tag{1.21}$$

El método de Newton-Raphson es convergente cuadráticamente, es decir, el número de cifras decimales correctas se duplica aproximadamente en cada interacción. Cuando el método de Newton-Raphson converge, se obtienen resultados en relativamente pocas interacciones, ya que para raíces no repetidas este método converge con orden 2.

De esto puede afirmarse que de cada iteración duplica aproximadamente el número de dígitos correctos. Sin embargo el método de Newton-Raphson algunas veces no converge, sino que oscila. Esto ocurre si no hay raíz real, si la raíz es un punto de inflexión o si el valor inicial esta muy alejado de la raíz buscada.

Capítulo 2

El Modelo Normal Proyectado

Una forma simple de generar distribuciones de probabilidad sobre la esfera unitaria q-dimensional es proyectando radialmente distribuciones de probabilidad originalmente definidas sobre el espacio q-dimensional. Una distribución relevante, dentro de esta familia de modelos, es la *distribución Normal proyectada q- variada*. El modelo *Normal proyectado q-variado* se obtiene al proyectar radialmente una distribución de probabilidad normal multivariada de dimensión q. La distribución normal proyectada $NP_q(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ es muy versátil ya que puede modelar comportamientos simétricos, asimétricos, unimodales y/o multimodales. En este trabajo nos enfocaremos sobre el modelo normal proyectado q-variado con $\boldsymbol{\Lambda} = \mathbf{I}$. Para el caso $q = 2$, el lector puede referir al trabajo de Nuñez Antonio (2010). Adicionalmente, se verá la forma de enlazar este modelo con el análisis de datos composicionales. En las siguientes secciones se presenta una forma de derivar distribuciones $NP_q(\boldsymbol{\mu}, \boldsymbol{I})$ poniendo énfasis en el caso circular ($q = 2$) y en el caso de datos esféricos ($q = 3$) ver Nuñez-Antonio y Gutiérrez-Peña (2005).

2.1. Especificación del Modelo

Sea $\boldsymbol{X}$ un vector aleatorio q-dimensional tal que $Pr(\boldsymbol{X} = 0) = 0$. Entonces $\boldsymbol{U} = ||\boldsymbol{X}||^{-1}\boldsymbol{X}$ es un punto aleatorio sobre la esfera unitaria q-dimensional $\mathbb{S}^q$. Un ejemplo importante es el caso en el que $\boldsymbol{X}$ tiene una distribución normal q-variada con vector de medias $\boldsymbol{\mu}$ y matriz de precisión $\boldsymbol{\Lambda}$, denotada por $N_q(\boldsymbol{\mu}, \boldsymbol{\Lambda})$, en cuyo caso se dice que $\boldsymbol{U}$ tiene una distribución normal proyectada, denotado por $NP_q(\boldsymbol{\mu}, \boldsymbol{\Lambda})$. Este modelo trata las observaciones direccionales como proyecciones sobre la esfera unitaria de vectores parcialmente no observados de una normal multivariada. La versión más sim-

ple del modelo produce una distribución comparable a una distribución von Mises, que es la distribución de probabilidad usualmente asumida para datos circulares.

Si $\boldsymbol{U}$ es una dirección aleatoria en $\mathbb{R}^q$, entonces su *dirección media* es el vector unitario $\eta = E(\boldsymbol{U})/\rho$, donde $\rho = ||E(\boldsymbol{U})||$, $E(\cdot)$ representa la esperanza usual para vectores aleatorios, y $||\cdot||$ representa la norma Euclideana. El parámetro ρ, es llamado la *longitud de la media resultante* y sirve como una media de concentración para distribuciones direccionales, su valor esta acotado entre cero y uno, cuando su valor es cercano a cero los datos están muy dispersos, mientras que si su valor es cercano a uno los datos esta muy concentrados. Ya que $\boldsymbol{U}$ es un vector unitario q-dimensional, como se mencionó en el capítulo anterior, en este también se puede representar mediante *q-1* ángulos.

2.1.1. La Distribución Normal Proyectada: Caso Circular

En esta sección se presentan propiedades del modelo normal proyectado para el caso bivariado, así mismo, en la siguiente sección se analizan estas propiedades para ver cómo van cambiando conforme la dimensión aumenta.

Definición 2.1 *Sea $\boldsymbol{Y}$ un vector aleatorio con distribución de probabilidad normal bivariada con vector de medias $\boldsymbol{\mu}$ y matriz de precisión $\boldsymbol{I}$. Entonces, la función de densidad de probabilidad de $\boldsymbol{Y}$ está dada por*

$$N_2(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{I}) = \frac{|\boldsymbol{I}|^{1/2}}{(2\pi)} \exp\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{I}(\boldsymbol{y}-\boldsymbol{\mu})\}.$$

En seguida se obtiene la función de densidad conjunta de la transformación $y = r(\cos(\theta), \text{sen}(\theta))$. Sin embargo, antes se enuncia el teorema de cambio de variable ya que es de mucha utilidad en esta y en las demás secciones correspondientes a este capítulo.

Teorema 2.1 *Sea X una variable aleatoria con función de densidad $f_x(\cdot)$. Sea $A = \{x : f_X(x) > 0\}$. Se asume que*

(i) $y = g(x)$ define una transformación a trozos uno a uno (inyectiva) de A a D, es decir, A se puede descomponer en conjuntos finitos disjuntos $A_1, A_2, \ldots, A_m$ tal que $y = g(x)$ define una trasformación uno a uno para cada A_i en D.

(ii) La derivada de $x = g^{-1}(y)$ con respecto a y es continua y distinta de cero $\forall\ y\ \in D$ donde $g^{-1}(y)$ es la función inversa de $g(x)$

Entonces $Y = g(X)$ *es una variable aleatoria continua con función de densidad dada por*

$$f_Y(y) = \sum_{i=1}^{m} \left| \frac{d}{dy} g_i^{-1}(y) \right| f_X(g_i^{-1}) I_D(y)$$

La demostración de este teorema se puede revisar, por ejemplo, en Mood *et al.* (1974)

Proposición 2.1 *Sea* $\boldsymbol{Y}$ *un vector aleatorio con distribución* $N_2(\boldsymbol{\mu}, \boldsymbol{I})$*, si se define la transformación*

$$\boldsymbol{y} = r[\cos(\theta), sen(\theta)]' = r\boldsymbol{v}'$$

donde $\theta \in [0, 2\pi]$ *y* $r \in \mathbb{R}^+$*. Entonces, la función de densidad conjunta de la transformación* (r, θ)*, está dada por*

$$f(r, \theta | \boldsymbol{\mu}, \boldsymbol{I}) = r \ K_1 \ \exp\left(-\frac{1}{2} \ r^2 \ + \ b \ r\right),$$

donde $b = \boldsymbol{v}'\boldsymbol{\mu}$*, y*

$$K_1 = \frac{|\boldsymbol{I}|^{1/2}}{2\pi} \exp(-\frac{1}{2} \ \|\boldsymbol{\mu}\|^2).$$

Demostración:

Empleando el teorema de cambio de variable, y dado que el determinante del jacobiano de la transformación $|J(r, \theta)|$ *es r, se tiene*

$$\begin{aligned}
f(r, \theta | \boldsymbol{\mu}, \boldsymbol{I}) &= f(rv' | \boldsymbol{\mu}, \boldsymbol{I}) |J(r, \theta)| \\
&= (2\pi)^{-1} |\boldsymbol{I}|^{1/2} \ \exp\{-\tfrac{1}{2}[(r\boldsymbol{v} - \boldsymbol{\mu})'(r\boldsymbol{v} - \boldsymbol{\mu})]\} \ r \\
&= r(2\pi)^{-1} \ \exp\{-\tfrac{1}{2}\left[(\boldsymbol{v'v})r^2 \ - \ 2(\boldsymbol{v'\mu})r \ + \ \boldsymbol{\mu'\mu}\right]\} \\
&= r(2\pi)^{-1} \exp\{-\tfrac{1}{2}\|\boldsymbol{\mu}\|^{\boldsymbol{2}}\} \exp\{-\tfrac{1}{2}\left[r^2 \ - \ 2\boldsymbol{v'\mu} \ r\right]\} \\
&= r(2\pi)^{-1} \exp\{-\tfrac{1}{2}\|\boldsymbol{\mu}\|^{\boldsymbol{2}}\} \exp\{-\tfrac{1}{2}\left[r^2 \ - \ 2b \ r\right]\} \\
&= K_1 \ r \exp\{-\tfrac{1}{2}\left[r^2 \ - \ 2b \ r\right]\}
\end{aligned}$$

□

Dada la función de densidad conjunta $f(r, \theta | \boldsymbol{\mu}, \boldsymbol{I})$ se puede dar la función de densidad de la Normal Proyectada, es decir, del ángulo aleatorio θ corresponde a la densidad marginal de θ de $f(r, \theta, \boldsymbol{\mu}, \boldsymbol{I})$. Esta distribución se deriva en la siguiente proposición.

Proposición 2.2 *Bajo las mismas condiciones de la Proposición (2.1), la función de densidad del ángulo aleatorio Θ, es decir, la densidad de probabilidad de la correspondiente normal proyectada, está dada por*

$$\begin{aligned} NP(\theta|\boldsymbol{\mu},\boldsymbol{I}) &= \int_0^\infty f(r,\theta|\boldsymbol{\mu},\boldsymbol{I})dr \\ &= K_1\ [1\ + \tfrac{b}{\phi(b)}\Phi(b)]\ 1_{[0,2\pi]}(\theta) \end{aligned}$$

donde $\Phi(\cdot)$ y $\phi(\cdot)$ denotan las funciones de distribución y de densidad de una normal estándar, respectivamente.

DEMOSTRACIÓN:

Se tiene que

$$\begin{aligned} NP(\theta|\boldsymbol{\mu},\boldsymbol{I}) &= \int_0^\infty f(r,\theta|\boldsymbol{\mu},\boldsymbol{I})dr \\ &= K_1 \int_0^\infty\ r\exp\{-\tfrac{1}{2}\,[r^2\ -\ 2b\ r]\} \end{aligned}$$

Integrando por partes se tiene

$$u = \exp\{\ b\ r\} \qquad dv = r\ \exp\{-\tfrac{1}{2}r^2\}dr$$

$$du = (b\exp\{\ br\})dr \qquad v = -\exp\{-\tfrac{1}{2}r^2\}$$

entonces

$$\begin{aligned} K_1 \int_0^\infty\ r\exp\{-\tfrac{1}{2}\,[r\ -\ 2b\ r]\} &= K_1[1 + b\int_0^\infty \exp\{-\tfrac{1}{2}\,[r\ -\ 2b\ r]\}] \\ &= K_1[1 + b\int_0^\infty \exp\{-\tfrac{1}{2}\,[r\ -b]^2 + \tfrac{b^2}{2}\}] \\ &= K_1[1 + b\exp\{\tfrac{b^2}{2}\}\int_0^\infty \exp\{-\tfrac{1}{2}\,[r\ -b]^2\}] \\ &= K_1[1 + b\exp\{\tfrac{b^2}{2}\}\int_{-b}^\infty \exp\{-\tfrac{1}{2}\,[s]^2\}] \\ &= K_1[1 + b\phi(b)^{-1}\int_{-\infty}^b \exp\{-\tfrac{1}{2}\,[s]^2\}] \\ &= K_1[1 + \tfrac{b}{\phi(b)}\Phi(b)] \end{aligned}$$

□

A partir de $f(\theta|\boldsymbol{\mu},\boldsymbol{I}) = NP(\theta|\boldsymbol{\mu},\boldsymbol{I})$ y $f(r,\theta|\boldsymbol{\mu},\boldsymbol{I})$ se puede obtener la densidad condicional de R dado Θ, la cual es relevante para llevar acabo los procedimientos de inferencia propuestos en este trabajo.

Proposición 2.3 *Bajo las mismas condiciones de la Proposición (2.1), la función de densidad condicional de R dado Θ está dada por*

$$f(r|\theta,\boldsymbol{\mu},\boldsymbol{I}) = \frac{r\ \exp\left(-\frac{1}{2}\ [r^2\ -\ 2\ b\ r]\right)}{1\ +\frac{b}{\phi(b)}\Phi(b)}\ 1_{(0,\infty)}(r).$$

Demostración:
El resultado se sigue de la igualdad

$$f(r|\theta,\boldsymbol{\mu},\boldsymbol{I}) = \frac{f(r,\theta|\boldsymbol{\mu},\boldsymbol{I})}{NP(\theta|\boldsymbol{\mu},\boldsymbol{I})}$$

□

La función de densidad condicional acumulada de R dado Θ al igual que la función de densidad condicional de R dado Θ se empleará posteriormente para simular las variables latentes $R_i,\ i = 1,\ldots,n$ mediante el método de Newton-Raphson. Por lo anterior, a continuación se deriva esta distribución..

Proposición 2.4 *Bajo las mismas condiciones de la Proposición (2.1), la función de densidad condicional acumulada de R dado Θ está dada por*

$$\begin{aligned} F(r|\theta,\boldsymbol{\mu},\boldsymbol{I}) &= \frac{\int_0^r w\ \exp\left(-\frac{1}{2}\ [w^2\ -\ 2\ b\ w]\right)dw}{1\ +\frac{b}{\phi(b)}\Phi(b)} \\ &= \frac{1-exp\{-\frac{1}{2}[r^2-2br]\}+b[\phi(b)]^{-1}[\Phi(r-b)-\Phi(-b)]}{1\ +\frac{b}{\phi(b)}\Phi(b)}\,1_{(0,\infty)}(r). \end{aligned}$$

Demostración:

Se tiene que

$$F(r|\theta,\boldsymbol{\mu},\boldsymbol{I}) = \frac{\int_0^r w\ \exp\left(-\frac{1}{2}\ [w^2\ -\ 2\ b\ w]\right)dw}{1\ +\frac{b}{\phi(b)}\Phi(b)}$$

Se puede notar que lo que se tiene que resolver es la integral, ya que, el denominador es constante con respecto a el diferencial, si se toma

$$u = \exp\{\ b\ w\} \qquad dv = w\ \exp\{-\tfrac{1}{2}w^2\}dx$$

$$du = (b\exp\{\ bw\})dw \qquad v = -\exp\{-\tfrac{1}{2}w^2\}$$

entonces

$$\begin{aligned}\int_0^r w\exp\{-\tfrac{1}{2}[w - 2b\ w]\}dw &= 1 - exp\{-\tfrac{1}{2}[r^2 - 2br]\} \\ &+ b\int_0^r \exp\{-\tfrac{1}{2}[w^2 - 2b\ w]\}dw \\ &= 1 - exp\{-\tfrac{1}{2}[r^2 - 2br]\} \\ &+ b\sqrt{2\pi}\exp\{\tfrac{b^2}{2}\}\tfrac{1}{\sqrt{2\pi}}\int_0^r \exp\{-\tfrac{1}{2}[w - b]^2 dw\}\end{aligned}$$

Resolviendo la integral

$$\begin{aligned}\tfrac{1}{\sqrt{2\pi}}\int_0^r \exp\{-\tfrac{1}{2}[w - b]^2 dw\} &= \tfrac{1}{\sqrt{2\pi}}\int_{-b}^{r-b}\exp\{-\tfrac{1}{2}(u)^2\}du \\ &= \Phi(r-b) - \Phi(-b)\end{aligned}$$

Por tanto se tiene

$$\begin{aligned}F(r|\theta,\boldsymbol{\mu},\boldsymbol{I}) &= \frac{\int_0^r w\ \exp(-\tfrac{1}{2}\ [w^2 - 2\ b\ w])dw}{1 + \frac{b}{\phi(b)}\Phi(b)} \\ &= \frac{1 - exp\{-\frac{1}{2}[r^2 - 2br]\} + b[\phi(b)]^{-1}[\Phi(r-b) - \Phi(-b)]}{1 + \frac{b}{\phi(b)}\Phi(b)}.\end{aligned}$$

□

La dirección media para el modelo Normal Proyectado en el caso circular viene dada por $\frac{\boldsymbol{\mu}}{||\boldsymbol{\mu}||}$ mientras que la longitud media resultante esta dada por $\rho = \sqrt{\pi\gamma/2}e^{-\gamma}[I_0(\gamma) + I_1(\gamma)]$ donde $\gamma = ||\mu||^2/4$ y $I_q(\cdot)$ es la función de Bessel modificada de primer tipo y orden q, ver Presnell y Rumcheva (2008).

2.1.2. La Distribución Normal Proyectada: Caso Esférico

Al igual que en el caso circular ($q = 2$) en esta sección se derivan las propiedades para el caso esférico. Este caso $q = 3$ junto con el caso $q = 2$ permitirá comprender de mejor manera las propiedades del modelo normal proyectado en el caso q-dimensional.

Definición 2.2 *Sea $\boldsymbol{Y}$ un vector aleatorio con distribución de probabilidad normal 3-dimensional con vector de medias $\boldsymbol{\mu}$ y matriz de precisión $\boldsymbol{I}$. Entonces, la función de densidad de probabilidad de $\boldsymbol{Y}$ está dada por*

$$N_3(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{I}) = \frac{|\boldsymbol{I}|^{1/2}}{(2\pi)^{3/2}}\exp\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{I}(\boldsymbol{y}-\boldsymbol{\mu})\}.$$

Recordar que para el caso esférico ($q = 3$) se requieren dos ángulos (θ, ϕ) para definir un vector en $\mathbb{R}^3$. Al igual que en la sección anterior, a continuación se derivan las densidades correspondientes al modelo esférico (3-dimensional).

Proposición 2.5 *Sea* $\boldsymbol{Y}$ *un vector aleatorio con distribución* $N_3(\boldsymbol{\mu}, \boldsymbol{I})$*, si se define la transformación*

$$\boldsymbol{y} = r[\cos\theta, \cos(\phi)sen(\theta), sen(\phi)\, sen(\theta)]' = r\boldsymbol{v}'$$

donde $\theta \in [0, \pi]$, $\phi \in [0, 2\pi]$ *y* $r \in \mathbb{R}^+$*. Entonces, la función de densidad conjunta de la transformación* (r, θ, ϕ)*, está dada por*

$$f(r, \theta, \phi|\boldsymbol{\mu}, \boldsymbol{I}) = r^2\ K_2\ \exp\,(-\frac{1}{2}\ r^2\ +\ br),$$

donde $b = \boldsymbol{v}'\boldsymbol{\mu}$*, y*

$$K_2 = \frac{1}{(2\pi)^{\frac{3}{2}}} \exp(-\frac{1}{2}\ \|\boldsymbol{\mu}\|^2) sen(\theta).$$

Demostración:

Empleando el teorema de cambio de variable, y dado que el determinante del jacobiano de la transformación $|J(r, \theta, \phi)|$ *es* $r^2\, sen(\theta)$*, se tiene*

$$\begin{aligned}
f(r, \theta, \phi|\boldsymbol{\mu}, \boldsymbol{I}) &= (2\pi)^{-\frac{3}{2}}|\boldsymbol{I}|^{1/2}\ \exp\{-\tfrac{1}{2}[(r\boldsymbol{v} - \boldsymbol{\mu})'(r\boldsymbol{v} - \boldsymbol{\mu})]\}\ |J(r, \theta, \phi)| \\
&= (2\pi)^{-\frac{3}{2}}|\boldsymbol{I}|^{1/2}\ \exp\{-\tfrac{1}{2}[(r\boldsymbol{v} - \boldsymbol{\mu})'(r\boldsymbol{v} - \boldsymbol{\mu})]\}\ r^2 sen(\theta) \\
&= r^2(2\pi)^{-\frac{3}{2}} sen(\theta)\ \exp\{-\tfrac{1}{2}\,[(\boldsymbol{v'v})r^2\ -\ 2(\boldsymbol{v'\mu})r\ +\ \boldsymbol{\mu'\mu}]\} \\
&= r^2(2\pi)^{-\frac{3}{2}} sen(\theta) \exp\{-\tfrac{1}{2}\|\boldsymbol{\mu}\|^{\boldsymbol{2}}\} \exp\{-\tfrac{1}{2}\,[r^2\ -\ 2\boldsymbol{v'\mu}\ r]\} \\
&= r^2(2\pi)^{-\frac{3}{2}} sen(\theta) \exp\{-\tfrac{1}{2}\|\boldsymbol{\mu}\|^{\boldsymbol{2}}\} \exp\{-\tfrac{1}{2}\,[r^2\ -\ 2b\ r]\}
\end{aligned}$$

□

Se puede notar que los cambios de la función de densidad conjunta para el caso esféricos con respecto al caso circular son:

- El exponente de r aumento en uno, es decir, el exponente de r depende de la dimensión.
- A diferencia de K_1, en K_2 aparece un termino que depende del ángulo, sen(θ), además de que el exponente que acompaña al termino 2π se incrementa, el incremento del exponente en la variable r y la aparición de $sen\theta$ viene dado del jacobiano de la transformación.

Proposición 2.6 *Bajo las mismas hipótesis de la Proposición (2.5), la función de densidad del vector aleatorio* (Θ, Φ)*, es decir, la densidad de probabilidad de la correspondiente Normal proyectada, está dada por*

$$\begin{aligned} NP(\theta, \phi|\boldsymbol{\mu}, \boldsymbol{I}) &= \int_0^\infty f(r, \theta, \phi|\boldsymbol{\mu}, \boldsymbol{I})dr \\ &= K_2\ [b\ + (b^2+1)\Phi(b)[\phi(b)]^{-1}]\ 1_{[0,\pi]}(\theta)1_{[0,2\pi]}(\phi) \end{aligned}$$

donde $\Phi(\cdot)$ *y* $\phi(\cdot)$ *denotan las funciones de distribución y de densidad de una Normal estándar, respectivamente, y* K_2 *es como en la preposición (2.5).*

Demostración:

$$\begin{aligned} NP(\theta, \phi|\boldsymbol{\mu}, \boldsymbol{I}) &= \int_0^\infty f(r, \theta, \phi|\boldsymbol{\mu}, \boldsymbol{I})dr \\ &= K_2 \int_0^\infty r^2 \exp\{-\frac{1}{2}r^2\}\ \exp\{b\ r\}\ dr. \end{aligned}$$

Trabajando sólo con la integral y haciendo integración por partes

$$u = r\exp\{\ b\ r\} \qquad dv = r\ \exp\{-\frac{1}{2}r^2\}dr$$

$$du = (br\exp\{\ br\} + \exp\{\ br\})dr \qquad v = -\exp\{-\frac{1}{2}r^2\}$$

entonces, tenemos que

$$\begin{aligned} \int_0^\infty r^2 exp\left\{-\tfrac{1}{2}r^2\right\} exp\left\{br\right\} dr &= b\int_0^\infty r\, exp\left\{-\tfrac{1}{2}r^2\right\} exp\left\{br\right\} dr \\ &+ \int_0^\infty exp\left\{-\tfrac{1}{2}r^2\right\} exp\left\{br\right\} dr \end{aligned}$$

El resultado de la primera integral ya lo conocemos gracias a la Proposición (2.2) Resolviendo la segunda integral tenemos

$$\begin{aligned} \int_0^\infty exp\left\{-\tfrac{1}{2}r^2\right\} exp\left\{br\right\} dr &= \sqrt{2\ \pi} exp\left\{\tfrac{b^2}{2}\right\} \frac{1}{\sqrt{2\ \pi}} \int_0^\infty exp\left\{-\tfrac{1}{2}(r-b)^2\right\} dr \\ &= [\phi(b)]^{-1}\Phi(b) \end{aligned}$$

Por tanto tenemos que

$$\begin{aligned} NP(\theta|\boldsymbol{\mu}, \boldsymbol{I}) &= K_2\ \int_0^\infty r^2 \exp\{-\tfrac{1}{2}r^2\}\ \exp\{b\ r\}\ dr \\ &= K_2\ [b\ + (b^2+1)\Phi(b)[\phi(b)]^{-1}]\ 1_{(0,2\pi]}(\theta)1_{(0,\pi]}(\phi) \end{aligned}$$

□

Existe una relación al calcular la densidad de probabilidad de la normal proyectada para el caso esférico con el del caso circular y esta es que al resolver la integral de la Proposición (2.6), aparece la integral que se tuvo que calcular para obtener la densidad de probabilidad de la normal proyectada para el caso circular.

Proposición 2.7 *Bajo las mismas hipótesis de la Proposición (2.5), la función de densidad condicional de R dado (Θ, Φ) está dada por*

$$f(r|\theta,\phi,\boldsymbol{\mu},\boldsymbol{I}) = \frac{r^2 \ \exp\left(-\frac{1}{2}\ [r^2 \ - \ 2\ b\ r]\right)}{[b \ + (b^2+1)\Phi(b)[\phi(b)]^{-1}]}\ 1_{(0,\infty)}(r).$$

Demostración:
El resultado se obtiene de la igualdad

$$\begin{aligned} f(r,|\boldsymbol{\mu},\boldsymbol{I}) &= \frac{f(r,\theta,\phi|\boldsymbol{\mu},\boldsymbol{I})}{NP(\theta,\phi|\boldsymbol{\mu},\boldsymbol{I})} \\ &= \frac{r^2 K_2 \exp\{-\frac{1}{2}\,[r^2 \ - \ 2br]\}}{K_2\ [b \ + (b^2+1)\Phi(b)[\phi(b)]^{-1}]} \\ &= \frac{r^2 \exp\{-\frac{1}{2}\,[r^2 \ - \ 2br]\}}{[b \ + (b^2+1)\Phi(b)[\phi(b)]^{-1}]} \end{aligned}$$

□

Los cambios que se pueden notar para el caso circular y esférico en la función de densidad condicional de R dado (Θ, Φ) es el incremento del exponente r, y que el denominador en ambos casos es distinto, esto debido al cálculo de diferentes integrales.

Proposición 2.8 *Bajo las mismas hipótesis de la Proposición (2.5), La función de densidad condicional acumulada de R dado (Θ, Φ) está dada por*

$$\begin{aligned} F(r|\theta,\phi,\boldsymbol{\mu},\boldsymbol{I}) &= \frac{\int_0^r w^2 \ \exp\left(-\frac{1}{2}\ [w^2 \ - \ 2\ b\ w]\right)dw}{b \ + (b^2+1)\Phi(b)[\phi(b)]^{-1}} \\ &= \frac{b-(b+r)\exp\left(-\frac{1}{2}[r^2-2br]\right)}{[b \ + (b^2+1)\Phi(b)[\phi(b)]^{-1}]}\,1_{(0,\infty)}(r) \\ &+ \frac{(b^2+1)[\phi(b)]^{-1}[\Phi(r-b)-\Phi(-b)]}{[b \ + (b^2+1)\Phi(b)[\phi(b)]^{-1}]}\,1_{(0,\infty)}(r). \end{aligned}$$

donde $\Phi(\cdot)$ y $\phi(\cdot)$ denotan las funciones de distribución y de densidad de una Normal estándar, respectivamente.

DEMOSTRACIÓN:
Se sabe que

$$F(r|\theta,\phi,\boldsymbol{\mu},\boldsymbol{I}) \;=\; \frac{\int_0^r w^2\ \exp\left(-\frac{1}{2}\ [w^2\ -\ 2\ b\ w]\right)dw}{b\ +(b^2+1)\Phi(b)[\phi(b)]^{-1}}$$

Por demostrar

$$\begin{aligned}\int_0^r w^2\ \exp\left(-\frac{1}{2}\ [w^2\ -\ 2\ b\ w]\right)dw \;&=\; b-(b+r)\exp\left(-\frac{1}{2}[r^2-2br]\right)\\ &+\;\; (b^2+1)[\phi(b)]^{-1}[\Phi(r-b)-\Phi(-b)]\end{aligned}$$

Mediante el método integración por partes se resolverá la integral. Sea

$$u = w\exp\{\ b\ w\} \qquad\qquad dv = w\ \exp\{-\tfrac{1}{2}w^2\}dw$$

$$du = (b\exp\{\ bw\}+\exp\{\ bw\})dw \qquad v = -\exp\{-\tfrac{1}{2}w^2\}$$

entonces

$$\begin{aligned}\int_0^r\ w^2\exp\{-\frac{1}{2}\,[w\ -\ 2\boldsymbol{b}\ w]\}dx \;&=\; -r\exp\{-\frac{1}{2}[r^2-2br]\}\\ &+\;\; b\textstyle\int_0^r w\exp\{-\frac{1}{2}\,[w^2\ -\ 2\boldsymbol{b}\ w]\}dw\\ &+\;\; \displaystyle\int_0^r \exp\{-\frac{1}{2}\left[w^2\ -\ 2\boldsymbol{b}\ w\right]\}dw.\end{aligned}$$

Pero, $\displaystyle\int_0^r w\exp\{-\frac{1}{2}\left[w^2\ -\ 2\boldsymbol{b}\ w\right]\}dw$ como $\displaystyle\int_0^r \exp\{-\frac{1}{2}\left[w^2\ -\ 2\boldsymbol{b}\ xw\right]\}dw$ se calcularon en la demostración de la preposición (2.4)

Por tanto se tiene que

$$\begin{aligned}\int_0^r w^2\ \exp\left(-\frac{1}{2}\ [w^2\ -\ 2\ b\ w]\right)dw \;&=\; b-(b+r)\exp\left(-\frac{1}{2}[r^2-2br]\right)\\ &+\;\; (b^2+1)[\phi(b)]^{-1}[\Phi(r-b)-\Phi(-b)]\end{aligned}$$

□

En el cálculo de la función de densidad condicional acumulada de R dado (Θ, Φ) para el caso esférico se puede observar que, esta depende de la integral que se calcula para $f(r|\theta, \phi, \mu, \boldsymbol{I})$ en el caso circular. Esto es muy importante ya que a partir de estas observaciones el cálculo de la función de densidad del ángulo aleatorio $\boldsymbol{\Theta}$, la función de densidad condicional de R dado $\boldsymbol{\Theta}$ y la función de densidad condicional acumulada de R dado $\boldsymbol{\Theta}$ se pueden obtener de forma recursiva, teniendo como caso base los resultados para el caso circular y el caso esférico.

La dirección media para el modelo Normal Proyectado en el caso esférico viene dada por $\frac{\boldsymbol{\mu}}{||\boldsymbol{\mu}||}$ mientras que la longitud media resultante esta dada por $\rho = \frac{2}{\gamma}\phi(\gamma) + \left(1 - \frac{1}{\gamma^2}\right)[2\Phi(\gamma) - 1]$ donde $\gamma = ||\mu||$ y $\phi(\cdot), \Phi(\cdot)$ representan las funciones de densidad y distribución de una normal estándar respectivamente ver Presnell y Rumcheva (2008).

2.1.3. La Distribución Normal Proyectada: Caso q-dimensional

Después de derivar las distribuciones asociadas a la normal proyectada para el caso circular ($q = 2$) y el caso esférico ($q = 3$), a continuación se presentan los resultados correspondientes al caso q-variado.

Definición 2.3 *Sea $\boldsymbol{Y}$ un vector aleatorio q-variado con distribución de probabilidad normal q-variada con vector de medias $\boldsymbol{\mu}$ y matriz de precisión $\boldsymbol{I}$. Entonces, la función de densidad de probabilidad de $\boldsymbol{Y}$ está dada por*

$$N_q(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{I}) = \frac{|\boldsymbol{I}|^{1/2}}{(2\pi)^{q/2}} \exp\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})'\boldsymbol{I}(\boldsymbol{y} - \boldsymbol{\mu})\}.$$

.

Proposición 2.9 *Sea $\boldsymbol{Y}$ un vector aleatorio con distribución $N_q(\boldsymbol{\mu}, \boldsymbol{I})$, si se define la transformación*

$$\boldsymbol{y} = r \begin{bmatrix} \cos(\theta_1) \\ sen(\theta_1)\cos(\theta_2) \\ sen(\theta_1)\, sen(\theta_2)\cos(\theta_3) \\ \vdots\ sen(\theta_1)\cdots\cos(\theta_{q-1}) \\ sen(\theta_1)\cdots sen(\theta_{q-1}) \end{bmatrix} = r\boldsymbol{v}$$

donde $\theta_{q-1} \in [0, 2\pi]$, $\theta_k \in [0, \pi]$ $\forall\ k = 1, 2, \ldots, q-2$. y $r \in \mathbb{R}^+$. Entonces, la función de densidad conjunta de la transformación $(r, \boldsymbol{\theta})$,con $\boldsymbol{\theta} =$

$(\theta_1, \theta_2, \ldots, \theta_{q-1})$ *está dada por*

$$f(r, \boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{I}) = \frac{|\boldsymbol{I}|^{1/2}}{(2\pi)^{\frac{q}{2}}} \exp(-\frac{1}{2} \|\boldsymbol{\mu}\|^2) \exp(-\frac{1}{2} r^2 + b\, r)|J(r, \boldsymbol{\theta})|,$$

donde $b = \boldsymbol{v}'\boldsymbol{\mu}$

Se debe notar que bajo la transformación propuesta en la Proposición (2.9) el determinante del jacobiano resulta ser

$$|J(r, \theta)| = r^{q-1} sen^{q-2}(\theta_1) sen^{q-3}(\theta_2) \cdots sen^{q-(q-2)}(\theta_{q-3}) sen^{q-(q-1)}(\theta_{q-2}).$$

Es decir,

$$|J(r, \theta)| = r^{q-1} \prod_{i=1}^{q-2} sen^{(q-(i-1))}\theta_i.$$

Esta expresión se puede re-escribir de la siguiente manera
$|J(r, \theta)| = |J(r)||J(\theta)|$ donde

$$|J(r)| = r^{q-1} \quad \text{y} \quad |J(\theta)| = \prod_{i=1}^{q-2} sen^{(q-(i-1))}\theta_i$$

de esta manera la función de densidad conjunta queda de la siguiente forma

$$f(r, \boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{I}) = r^{q-1} K_q \exp(-\frac{1}{2} r^2 + b\, r),$$

donde $K_q = \frac{|\boldsymbol{I}|^{1/2}}{(2\pi)^{\frac{q}{2}}} \exp(-\frac{1}{2} \|\boldsymbol{\mu}\|^2)|J(\theta)|$

Proposición 2.10 *Bajo las mismas premisas de la Proposición (2.9), la función de densidad del ángulo aleatorio* $\boldsymbol{\Theta}$*, es decir, la densidad de probabilidad de la correspondiente Normal proyectada q-variada, está dada por*

$$NP(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{I}) = \int_0^\infty f(r, \boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{I}) dr$$

$$= K_q[(q-2) \int_0^\infty r^{q-3} exp\left\{-\frac{1}{2}r^2\right\} exp\{br\}\, dr$$

$$+b \int_0^\infty r^{q-2} exp\left\{-\frac{1}{2}r^2\right\} exp\{br\}\, dr] 1_{[0,\pi]}(\theta) 1_{[0,\pi]}(\theta_1) \cdots 1_{[0,2\pi]}(\theta_{q-1})$$

Demostración*: Ver Apéndice A.1*

Como se mencionó en la sección anterior el cálculo de la densidad de la Normal proyectada para cualquier valor de q se puede realizar de manera recursiva, solo se necesitan conocer los desarrollos para, $q = 2$ y $q = 3$. Con los resultados de estos dos casos se puede obtener la función de densidad del ángulo aleatorio $\boldsymbol{\Theta}$ para cualquier q, por ejemplo, si se desea saber el valor para el caso $q = 5$ se necesita el valor de la integral para el caso $q = 4$ y $q = 3$, pero el caso $q = 4$ se obtiene con base al caso $q = 2$ y $q = 3$.

Proposición 2.11 *Bajo las mismas premisas de la Proposición (2.9), la función de densidad condicional de R dado $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_{q-1})$ está dada por*

$$f(r|\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{I}) = \frac{r^{q-1} \ \exp\left(-\frac{1}{2}\ [r^2 \ - \ 2\ b\ r]\right)}{\displaystyle\int_0^\infty r^{q-1} exp\left\{-\frac{1}{2}r^2\right\} exp\left\{br\right\} dr}\ 1_{(0,\infty)}(r).$$

Demostración:

El resultado se obtiene de la igualdad

$$f(r, |\boldsymbol{\mu}, \boldsymbol{I}) \ = \ \frac{f(r, \boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{I})}{NP(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{I})}$$

□

Como se mencionó para poder emplear el método de Newton-Raphson se necesita tanto $F(r|\cdot)$ como $F'(r|\cdot)$ por eso es necesario el calculo de $f(r|\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{I})$ y el de $F(r|\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{I})$ ya que a partir de estas dos funciones, y por medio del teorema de transformación integral de probabilidad, es posible simular las variables latentes $R's$, y con esto estar en conclusiones de hacer inferencia sobre el parámetro $\boldsymbol{\mu}$.

Proposición 2.12 *Bajo las mismas premisas de la Proposición (2.9) la función de densidad condicional acumulada de R dado $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_{q-1})$ está dada por*

$$
\begin{aligned}
F(r|\theta,\boldsymbol{\mu},\boldsymbol{I}) &= \frac{\int_0^r w^{q-1}\ \exp\left(-\frac{1}{2}\ [w^2\ -\ 2\ b\ w]\right)dw}{\int_0^\infty r^{q-1} exp\left\{-\frac{1}{2}r^2\right\} exp\left\{-\frac{1}{2}r^2\right\} dr} \\
&= \frac{-r^{q-2}\ \exp\left(-\frac{1}{2}\ [r^2\ -\ 2\ b\ r]\right)}{\int_0^\infty r^{q-1} exp\left\{-\frac{1}{2}r^2\right\} exp\left\{-\frac{1}{2}r^2\right\} dr} \\
&+ \frac{b\int_0^r w^{q-2}\ \exp\left(-\frac{1}{2}\ [w^2\ -\ 2\ b\ w]\right)dw}{\int_0^\infty r^{q-1} exp\left\{-\frac{1}{2}r^2\right\} exp\left\{-\frac{1}{2}r^2\right\} dr} \\
&+ \frac{(q-2)\int_0^r w^{q-3}\ \exp\left(-\frac{1}{2}\ [w^2\ -\ 2\ b\ w]\right)dw}{\int_0^\infty r^{q-1} exp\left\{-\frac{1}{2}r^2\right\} exp\left\{-\frac{1}{2}r^2\right\} dr}
\end{aligned}
$$

DEMOSTRACIÓN*: Ver Apéndice A.1*

En la siguiente sección se discute cómo aplicar el método de Newton-Raphson dentro de un muestreo de Gibbs para llevar acabo inferencia sobre el parámetro de interés $\boldsymbol{\mu}$

2.2. Análisis bayesiano del Modelo Normal Proyectado

El enfoque que se utilizará se basa en la introducción de variables latentes adecuadas para definir una distribución conjunta de μ. Esta distribución conjunta será construida de una manera tal que se asegure que se puede simular de todas las densidades condicionales requeridas para un muestreo de Gibbs.

Sea

$$\boldsymbol{X} \sim N_q(\cdot|\boldsymbol{\mu},\boldsymbol{I}) \tag{2.1}$$

como ya se vio en la sección anterior a partir de (2.1) se puede obtener vía coordenas hiper-esféricas, la densidad conjunta de $(R,\boldsymbol{\Theta})$ con $R = ||\boldsymbol{X}||$ de donde se deriva que $\boldsymbol{\Theta} \sim NP(\cdot|\boldsymbol{\mu},\boldsymbol{I})$.

El problema entonces se puede expresar como sigue: dada una muestra $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$ de $NP_q(\cdot|\boldsymbol{\mu}, \boldsymbol{I})$. ¿Cómo se puede hacer inferencias sobre $\boldsymbol{\mu}$?. Si se pudieran observar $(R_1, \boldsymbol{\theta}_1), \ldots, (R_n, \boldsymbol{\theta}_n)$, entonces se podría hacer inferencias sobre μ, pero el problema es que solo los ángulos $(\theta_1, \ldots, \theta_n)$ son observados.

La estructura de este modelo sugiere que se debe tratar a las normas no observadas $R_i = ||X_i||$ $i = 1, \ldots, n$ como variables latentes. Así, los datos completos del modelo seguirían un modelo normal multivariado. Este fue el enfoque seguido por Presnell *et al.* (1998) ,Nuñez-Antonio y Gutiérrez-Peña (2005), los cuales solo analizaron el caso $q = 2$. En el modelo normal multivariado si $T_n = (X_1, \ldots, X_n)$ es una muestra del modelo $N_p(X|\mu, I)$ y se considera el modelo $N_p(\mu|\mu_0, \lambda_0 I)$ como la distribución inicial (donde μ_0 es un vector p-dimensional con j-th elemento $\mu_{0j} \in \mathbb{R}$ y $\lambda_0 > 0$ entonces por el Ejemplo (1.1) la distribución final de μ esta dada por

$$f(\mu|T_n) = N_p(\mu|\mu_F, \Lambda_F)$$

donde μ_F es un vector p-dimensional con j-ésimo elemento definido por $(\lambda_0\mu_{0j} + n\overline{x}_{j.})/(\lambda_0 + n)$; aquí $\overline{x}_{j.}$ es la media aritmética del j-ésimo componente de los vectores $x_1, \ldots, x_n$. La matriz Λ_F de dimensión pxp es diagonal con elementos dados por $\lambda_0 + n$ en la j-ésima entrada con $j = 1, \ldots, p$. Como se vio en la sección anterior, a partir de la Ecuación (2.1) y considerando a R como una variable latente definida sobre el intervalo $(0, \infty)$, $X = g(R, \Theta)$ se puede definir la distribución conjunta de la transformación dada por la Proposición (2.9). Para implementar el muestreo de Gibbs se necesitan todas las distribuciones condicionales que fueron descritas en la sección anterior. Se debe notar que la densidad condicional completa de $(\boldsymbol{\mu}|r, \theta)$ está dada por

$$f(\boldsymbol{\mu}|\boldsymbol{r}, \boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_n}) = N_p(\boldsymbol{\mu}_F, \boldsymbol{\Lambda}_F) \tag{2.2}$$

donde $\boldsymbol{r} = (r_1, \ldots, r_n)$ ϵ $\mathbb{R}^n_+$. Se debe notar que los R_i $i = 1, \ldots, n$ son condicionalmente independientes dado los Θ_j.

Por otra parte se pueden simular R_i mediante del método de Newton Raphson ocupando la *transformación integral probabilidad* bajo la cual se desprende que si x es una variable continua con función de distribución F entonces $u = F(x) \sim U[0, 1]$. Este resultado se demuestra como

$$F_u(y) = Pr(u \leq y) = Pr(F(x) \leq y) = Pr(x \leq F^{-1}(y)) = F(F^{-1}(y)) = y,$$

si $0 < y < 1$.

Recordando que el método de Newton-Raphson sirve para encontrar raíces en una ecuación, este se puede emplear de la siguiente manera para simular variables latentes $R's$.

- Se simula $u \sim U(0,1)$.

- Se resuelve $g(r) = F(r) - u = 0$ con $g'(r) = f(r)$.

Por Newton-Raphson se tiene que

$$r_{i+1} = r_i - \frac{g(r_i)}{g'(r_i)} \tag{2.3}$$

Aquí el valor inicial r_0 se toma como la moda $f(r|\boldsymbol{\theta}, \boldsymbol{\mu})$

Finalmente, se puede usar la condicional completa (2.2) y el procedimiento descrito anteriormente en un muestreo de Gibbs para obtener una muestra de la distribución final conjunta, de $\boldsymbol{\mu}$ y $R_i,\ i = 1, \ldots, n$.

A continuación se presentan un par de ejemplos

2.2.1. Ejemplos numéricos

Ejemplo 2.1 *En este ejemplo se simuló una muestra de tamaño 100 de una normal proyectada 3-dimensional con* $\boldsymbol{\mu} = (3.5, 5, 2.7)$. *En este caso la verdadera media direccional está dada por* $(\theta, \phi) = (58.37°, 28.37°)$ *y* $||\boldsymbol{\mu}|| = 6.67$. *Para llevar a cabo el análisis Bayesiano descrito en las secciones anteriores, se considero una distribución inicial no informativa para* $\boldsymbol{\mu}$ *de* $\mu_0 = \mathbf{0}$ *y* $\lambda_0 = 0{,}0001$. *Los resultados de la distribución marginal para* θ, ϕ *y* $||\boldsymbol{\mu}||$ *son presentados en la Figura 2.1.*

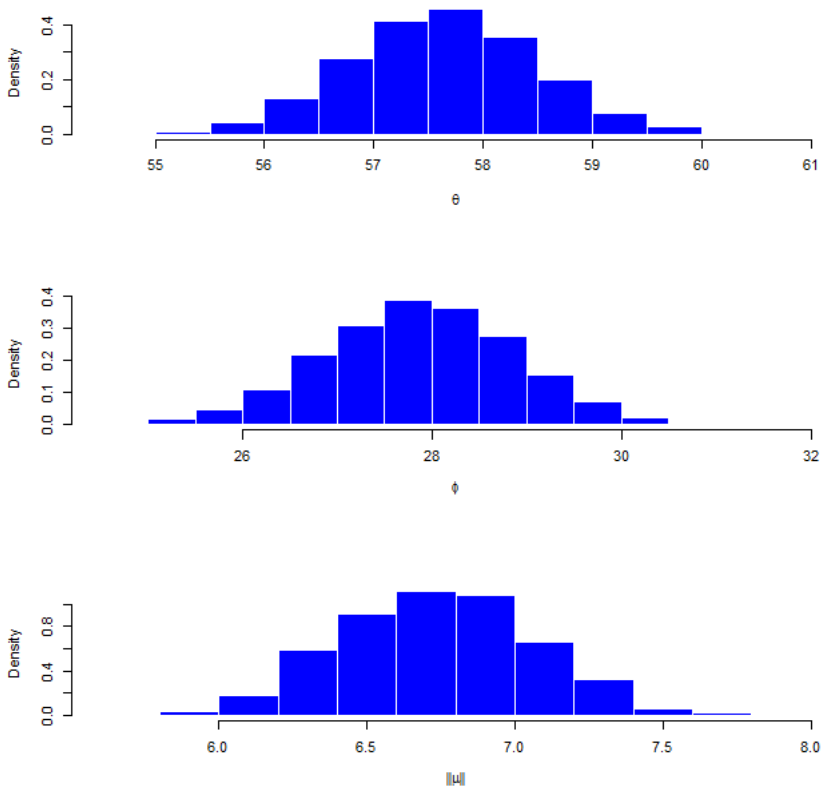

Figura 2.1: Distribuciones finales de θ, ϕ y $||\boldsymbol{\mu}||$.

Ejemplo 2.2 *En este segundo ejemplo, se simuló una muestra de tamaño 75 de una normal proyectada de dimensión 4 con* $\boldsymbol{\mu} = (-1.5, 3, -2.7, 1.9)$. *En este caso, la verdadera media direccional está dada por* $(\theta, \phi, \omega) = (71.41°, 47.74°, 144.87°)$ *y* $||\boldsymbol{\mu}|| = 4.71$. *Al igual que el valor de* $\mu_0 = \mathbf{0}$ *y* $\lambda_0 = 0{,}0001$. *Los resultados de las distribuciones marginales,en el ejemplo anterior se tomó para* θ, ϕ, ω *y* $||\boldsymbol{\mu}||$ *se presentan en la Figura 2.2.*

De los resultados de la Figura 2.1 y 2.2 se puede ver que bajo la metodología propuesta se pueden realizar inferencias adecuadas para los parámetros de un modelo Normal Proyectado $NP_q(\boldsymbol{\mu}, \boldsymbol{I})$ *de dimensión q. Hay que señalar que el algoritmo de simulación tiene un desempeño mucho más eficiente que el propuesto en Nuñez-Antonio y Gutiérrez-Peña (2005) al emplear un método de búsqueda de raíces como el Newton-Raphson para simular las variables latentes* $R_i,\ i = 1, \ldots, n$ *en lugar de un algoritmo Metropolis-Hastings usado por Nuñez-Antonio y Gutiérrez-Peña (2005).*

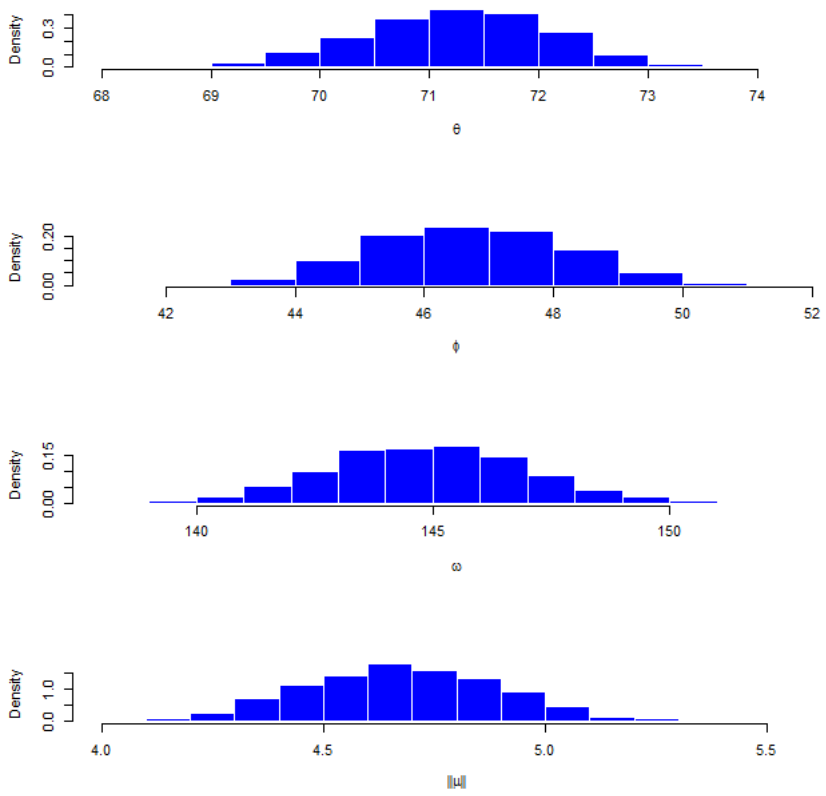

Figura 2.2: Distribuciones finales de θ, ϕ, ω y $||\mu||$

Capítulo 3

Análisis de datos composicionales vía variables direccionales

Como se mencionó en el Capítulo 1, en el análisis de los datos composicionales es necesario tener en cuenta su naturaleza composicional así como la presencia de posibles ceros. La propuesta de Aitchison (1986) para analizar datos composicionales se basa en una transformación del logaritmo de cocientes (*logratio*) y es una propuesta muy utilizada en varias disciplinas. Aunque la propuesta topológica de Aitchison (1986) para trabajar con el espacio muestral asociado a variables composicionales (el simplex) resulto en un enfoque general para el tratamiento de variables composicionales, bajo este enfoque aún quedaron temas por resolver. Entre los temas por resolver se encuentra la presencia de componentes con valor cero que se presentan en problemas aplicados. Lo anterior, dado que la transformación logratio no está definida cuando alguna componente de la variable composicional es cero. Aitchison (1986) ofreció ciertas alternativas para tratar con este problema, pero que señalo que difícilmente se podría resolver de manera satisfactoria. Hoy en día, el tratamiento de ceros es un tema actual y para el cual se siguen haciendo propuestas para su tratamiento. Ver por ejemplo,Neocleous *et al.* (2011), Wang *et al.* (2007) y las referencias allí incluidas.

En este trabajo de investigación se propone una forma de describir datos composicionales a través de variables direccionales. Específicamente, después de aplicar alguna transformación que vaya del simplex unitario a la esfera unitaria, la propuesta consiste, por un lado, en modelar los datos transformados por medio de técnicas estadísticas desarrolladas para variables angulares definidas en la esfera unitaria. Aunque la teoría topológica, desde el punto de vista estadístico, para tratar con el variables definidas en el simplex aún se encuentra en desarrollo (ver, por ejemplo,Pawlowsky-Glahn y Egozcue (2001),

Pawlowsky-Glahn y Egozcue (2002)) en este trabajo también se revisan los conceptos asociados a la propuesta de modelos estadísticos direccionales para describir variables composicionales.

3.1. Antecedentes

La mayoría de las propuestas existentes en la literatura para el tratamiento de ceros se basa en remplazar los ceros por algún pequeño valor δ o alguna variante de este procedimiento. Sin embargo, como se menciona en Fry *et al.* (2000), en muchas áreas y situaciones reales se deben considerar procedimientos de remplazo específicos que tomen en cuenta los datos del área de aplicación correspondiente . Por su parte, en Neocleous *et al.* (2011) se analizan los casos de remplazar las componentes cero de los datos composicionales por 0.0001 y el empleo de una transformación que denominan *log-log complementaria*. Esta transformación involucra tomar logaritmos de los datos transformados una vez aplicada la transformación logratio. Sin embargo, esta transformación solo es adecuada cuando todos los logratio son negativos. Por otro lado, Wang *et al.* (2007) propone una aproximación hiperesférica para tratar con la restricción de la no-negatividad de las componentes de variables composicionales y la presencia de componentes cero en la práctica. Analizan su propuesta en el contexto específico de modelos econométricos. Sin embargo, como ellos mismos señalan su procedimiento falla si alguna de las componentes toma el valor de 1 y el resto toma el valor de 0. Lo anterior, nos lleva a que no solo de debe tomar en cuenta la transformación propuesta, sino también la forma de analizar los datos transformados, ya sea bajo la estructura de un modelo paramétrico o bajo un enfoque no paramétrico.

Como se definio anteriormente, mientras un dato composicional $\boldsymbol{x}$ en el símplex unitario S^D de D-partes $\boldsymbol{x} = (x_1, x_2, \ldots, x_D)$ se caracteriza por tener componentes no negativos y por la restricción de suma constante de sus partes, es decir, si

$$S^D = \{\boldsymbol{x} = (x_1, x_2, \ldots, x_d, x_D) |\ x_1 \geq 0, \ldots, x_D \geq 0\ ;\ \sum_{i=1}^{D} x_i = 1\}, \quad (3.1)$$

un vector direccional $\boldsymbol{u} = (u_1, u_2, \ldots, u_D)$ direccional de dimensión D esta sujeto a la restricción

$$u_1^2 + u_2^2 + \cdots + u_D^2 = 1. \quad (3.2)$$

Es reconocido (ver, por ejemplo, Mardia y Jupp (2000)) que la transformación

$$\boldsymbol{u} = (\sqrt{x_1}, \sqrt{x_2}, \cdots, \sqrt{x_D})', \quad (3.3)$$

donde $\boldsymbol{x} = (x_1, \ldots, x_D) \in S^D$, mapea el dato composicional $\boldsymbol{x}$ sobre la superficie de la hiper-esfera unitaria $(D-1)$- dimensional. Lo anterior, abre la posibilidad de emplear la teoría desarrollada para variables direccionales para la modelación de vectores composicionales. Así, un enfoque alternativo a la transformación logratio de Aitchison (1986) y a las transformaciones como la log-log complementaria es emplear una transformación esférica (como la raíz cuadrada) sobre el vector $\boldsymbol{x}$ y entonces aplicar técnicas de análisis de variables direccionales a los datos transformados.

En las siguientes secciones se expone la propuesta de analizar datos composicionales a través de técnicas y modelos definidos para datos direccionales. Particularmente, se analizan dos transformaciones que van del símplex unitario a la esfera unitaria. Adicionalmente, se presenta la manera de analizar características de variables composicionales a través de modelos para datos direccionales.

3.2. Transformaciones hiperesféricas

3.2.1. Transformación raíz cuadrada

Como se mencionó anteriormente, la raíz cuadrada de un vector composicionale $\boldsymbol{x} \in S^D$ produce una variable en la esfera unitaria $\mathbb{S}^D$. Ver ecuación (3.3). Es decir, primero los vectores composicionales $\boldsymbol{x} = (x_1, \ldots, x_D)$ son transformados tomando

$$u_j = \sqrt{x_j}, \quad j = 1, \ldots, D. \tag{3.4}$$

Como se mencionó en el Capítulo 1, dado que el vector $\boldsymbol{u}$ resulta ser un vector en $\mathbb{S}^D$, entonces se puede expresar de manera equivalente a través de $D-1$ ángulos por medio de la transformación

$$\begin{aligned}
u_1 &= sen\theta_2 sen\theta_3 sen\theta_4 \cdots sen\theta_D \\
u_2 &= cos\theta_2 sen\theta_3 sen\theta_4 \cdots sen\theta_D \\
u_3 &= cos\theta_3 sen\theta_4 \cdots sen\theta_D \\
&\vdots \\
u_{D-2} &= cos\theta_{D-2} sen\theta_{D-1} sen\theta_D \\
u_{D-1} &= cos\theta_{D-1} sen\theta_D \\
u_D &= cos\theta_D
\end{aligned} \tag{3.5}$$

donde $0 < \theta_j \leq \frac{\pi}{2},\ j = 2, 3, \ldots, D$. Se debe notar que bajo esta transformación se tiene una reducción en la dimensionalidad del espacio correspondiente. Es decir, las D-componentes correlacionadas se transforman en $D-1$ ángulos independientes θ_j. Bajo esta transformación las componentes θ_j del vector de ángulos de la ecuación (3.5) se pueden obtener de manera recursiva, por ejemplo, utilizando la siguiente transformación

$$\begin{aligned} \theta_D &= arcos(u_D) \\ \theta_j &= arcos\left(\frac{u_j}{\prod_{i=j+1}^{D} sen\theta_{j+1}}\right) \end{aligned} \tag{3.6}$$

con $j = 2, 3, \ldots, D-1$. Dado que sólo se considera la raíz cuadrada positiva otra transformación que se puede emplear para definir los ángulos asociados a la variable direccional $\boldsymbol{u}$ es

$$\theta_j = tan^{-1}\left(\frac{\sqrt{\sum_{i=j+1}^{D} u_i^2}}{u_j}\right) \tag{3.7}$$

con $j = 1, 2, \ldots, D-1$.

Cabe mencionar que la transformación raíz cuadrada no preserva el ángulo *natural* del dato composicional, esto se puede apreciar en el siguiente ejemplo y en la Figura 3.1.

Ejemplo 3.1 *Sea* $\boldsymbol{x} = (0.3, 0.7) \in S^2$ *aplicando la transformación raíz cuadrada a* $\boldsymbol{x}$ *se tiene* $\boldsymbol{w} = \sqrt{\boldsymbol{x}} = (\sqrt{0.3}, \sqrt{0.7}) \in \mathbb{S}^2$. *Aplicando la trasformación (3.8), tanto a* $\boldsymbol{x}$ *como a* $\boldsymbol{w}$, *se obtiene que el ángulo correspondiente a* $\boldsymbol{x}$ *es* $\theta_x = 66.80°$, *mientras que, el ángulo formado por las componentes de* $\boldsymbol{w}$ *es* $\theta_w = 56.79°$.

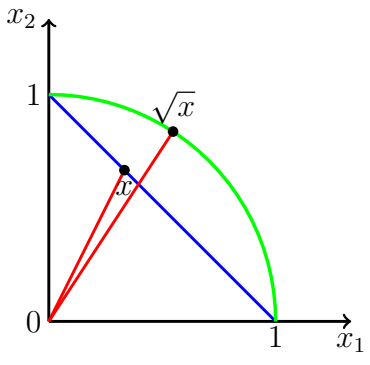

Figura 3.1: Transformación de un dato composicional a dato direccional mediante la transformación raíz cuadrada

Aunque esta transformación no conserva el ángulo natural asociado al dato composicional, esta es invariante permutacionalmente (ver Egozcue (2009)) e invariante ante cambios de escala. Sin embargo, ésta viola el principio de coherencia subcomposicional (Scely y Welsh (2011))

3.2.2. Transformación proyectada

Sea $\boldsymbol{X}$ un vector composicional, $\boldsymbol{X} \in S^D$ y $\boldsymbol{U} = \boldsymbol{X}/R$ donde $R = ||\boldsymbol{X}||$, entonces $\boldsymbol{U}$ pertenece a la esfera unitaria $\mathbb{S}^D$. Como ya se ha mencionado, desde que que $||\boldsymbol{U}|| = 1$ se sigue que $\mathbf{U}$ puede ser representado por $D-1$ ángulos mediante una transformación esférica (ver 1.1). Como los componentes x_j son no negativos, se pueden obtener los ángulos asociados a las componentes u_j del vector direccional $\mathbf{U}$ por medio de las transformaciones (3.6) o (3.7).

Se debe notar que con esta transformación si se preserva el ángulo *natural* ángulo asociado al dato composicional $\boldsymbol{X}$, tal y como se muestra en la Figura 3.2.

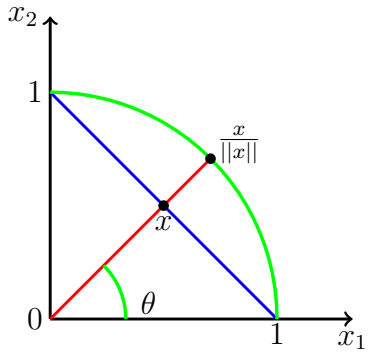

Figura 3.2: Transformación de un dato composicional a dato direccional mediante la transformación proyectada

Por otro lado, el vector de ángulos $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{D-1})$ se puede obtener directamente a partir del dato composicional $\boldsymbol{X}$ sin pasar por la transformación proyectada $\boldsymbol{U} = \boldsymbol{X}/R$. Lo anterior, empleando directamente la transformación (3.7) a las componentes del vector $\boldsymbol{X}$. Es decir, tomando

$$\theta_j = tan^{-1}\left(\frac{\sqrt{\sum_{i=j+1}^{D} x_i^2}}{x_j}\right) \tag{3.8}$$

con $j = 1, 2, \ldots, D-1$. Dado que $\boldsymbol{X} = R\boldsymbol{U}$, entonces $x_j = Ru_j$ por lo que se tiene

$$\theta_j = tan^{-1}\left(\frac{\sqrt{\sum_{i=j+1}^{D} x_i^2}}{x_j}\right) = tan^{-1}\left(\frac{\sqrt{\sum_{i=j+1}^{D}(Ru_i)^2}}{Ru_j}\right)$$

$$= tan^{-1}\left(\frac{R\sqrt{\sum_{i=j+1}^{D} u_i^2}}{Ru_j}\right) = tan^{-1}\left(\frac{\sqrt{\sum_{i=j+1}^{D} u_i^2}}{u_j}\right)$$

3.3. El Enfoque propuesto

El análisis de datos composicionales propuesto por Aitchison (1986) se basa principalmente en el análisis de cocientes sobre las componentes del vector composicional $\boldsymbol{x}$, particularmente, se utilizan transformaciones log-cocientes como la transformación *logcociente aditiva* $(log(x_1/x_D), \ldots, log(x_{D-1}/x_D))$. Sin embargo, este enfoque requiere que $x_j > 0 \; \forall \; \jmath = 1, \ldots, D$. En la práctica, a menudo los datos composicionales contienen 0's y este tipo de transformaciones no se puede aplicar sin adoptar alguna estrategia como colapsar categorías, modelar los 0's separadamente o remplazándolos por pequeños valores positivos. En este trabajo se propone aplicar al correspondiente vector composicional $\boldsymbol{x}$, algunas de las transformaciones revisadas en las secciones anteriores y una vez que se obtenga el vector de ángulos asociados $\boldsymbol{\Theta} = t(\boldsymbol{x})$ emplear procedimientos de análisis para datos direccionales. Este enfoque tiene la ventaja de tratar los 0's de manera natural. Por lo que en términos aplicados tiene mayor alcance que las transformaciones basadas en cocientes.

3.3.1. Propuesta de análisis descriptivo

Una manera resumida de describir el comportamiento de un conjunto de datos composicionales es a través de su matriz de variación, ver Capítulo 1, Sección 1.1. En este trabajo se propone describir los datos composicionales a través de una matriz asociada a la matriz de variación la cual llamaremos *matriz de variación direccional para datos composicionales*, cuya interpretación es análoga a la matriz de variación composicional. Antes de presentar la matriz de variación direccional para datos composicionales se explica la idea de la cual se parte para construir dicha matriz en el caso de un vector composiiconal $\boldsymbol{x} = (x, y)$ en S^2. En este caso, después de aplicar alguna transformación de coordenadas polares, se tendrá asociado solo un ángulo θ.

Sea $\boldsymbol{\Theta} = (\theta_1, \ldots, \theta_n)$ los n ángulos asociados con la matriz$\boldsymbol{X}_{n\times 2} = (x, y)$ de n datos composicionales 2-dimensionales. Si se toma el cociente $\frac{y_i}{x_i} = k,\ i = 1, \ldots, n$, y se aplica la transformación $x_i = cos\theta,\ y_i = sen\theta$, se tienen cuatro resultados posibles

- Si $\theta_i = tan^{-1}(\frac{y_i}{x_i}) \in (0, \frac{\pi}{4}]$ es porque $y_i \leq x_i$ y $\log(\frac{x_i}{y_i}) \geq 0$. En este caso, se dice que el porcentaje de la componente X es más grande que el porcentaje de la componente Y, en el dato composicional i.
- Si $\theta_i = tan^{-1}(\frac{y_i}{x_i}) \in (\frac{\pi}{4}, \frac{\pi}{2}]$ es porque $y_i > x_i$ y $\log(\frac{x_i}{y_i}) < 0$. En este caso, se dice que el porcentaje de la componente X es más chico que el porcentaje de la componente Y, en el dato composicional i.
- Si $\theta_i = tan^{-1}(\frac{y_i}{x_i}) = 0$ es porque $y_i = 0$ y aunque $\log(\frac{x_i}{y_i})$ no esta definido, se puede concluir que el porcentaje de la componente X representa el $100\,\%$ del dato composicional.
- Si $\theta_i = tan^{-1}(\frac{y_i}{x_i}) = \frac{\pi}{2}$ es porque $x_i = 0$ y aunqnue $\log(\frac{x_i}{y_i})$ no esta definido, se puede concluir que el porcentaje de la componente Y representa el $100\,\%$ del dato composicional.

Se debe notar que bajo esta definición se recupera la interpretación de la matriz de variación. Además, en la matriz de variación se tiene problemas para describir los datos composicionales dado que el logaritmo de cero y el cociente entre cero no están definidos. Por otro lado, mediante la matriz de variación direccional para datos composicionales se solucionan estos problemas, ya que el ángulo formado por el cociente si esta definido (0 ó $\frac{\pi}{2}$) sin importar que alguna de las componentes sea cero.

A continuación se define la matriz de variación direccional para datos composicionales D-dimensionales.

Definición 3.1 *Para una composición $\boldsymbol{X}$ de D-partes $(x_1, \ldots, x_D)$ y n observaciones, si tomamos la transformación $x_i = cos\theta, x_j = sen\theta,\ i < j, j = 1, 2, \ldots, D,\ i = j+1, \ldots, D-1$, la matriz de variación direccional para datos composicionales está dada por*

	1	2	3	$\cdots$	$D-1$	D	
1	$\cdot$	η_{12}	η_{13}	$\cdots$	$\eta_{1(D-1)}$	η_{1D}	$Varianzas$
2	ζ_{12}	$\cdot$	η_{23}	$\cdots$	$\eta_{2(D-1)}$	η_{2D}	
3	ζ_{13}	ζ_{23}	$\cdot$	$\cdots$	$\eta_{3(D-1)}$	η_{3D}	
$\vdots$							
$D-1$	$\zeta_{1(D-1)}$	$\zeta_{2(D-1)}$	$\zeta_{3(D-1)}$	$\cdots$	$\cdot$	$\eta_{(D-1)D}$	
D	ζ_{1D}	ζ_{2D}	ζ_{3D}	$\cdots$	$\zeta_{(D-1)D}$	$\cdot$	
$Medias$							

donde $\zeta_{ij} = E\{tan^{-1}(\frac{x_j}{x_i})\}$ *y* $\eta_{ij} = var\{tan^{-1}(\frac{x_j}{x_i}))\}$.

Esta matriz de variación se puede estimar mediante $\hat{\zeta}_{ij}$ y $\hat{\eta}_{ij}$ que son los estimadores de la dirección media ζ_{ij} y la varianza direccional η_{ij} respectivamente, dados por:

$$\hat{\zeta}_{ij} = tan^{-1}\left(\frac{\bar{X}_j}{\bar{X}_i}\right) \qquad \hat{\eta}_{ij} = 1 - (\bar{X}_i^2 + \bar{X}_j^2)^{\frac{1}{2}}$$

donde

$$\bar{X}_i = \frac{1}{n}\sum_{k=1}^{n} cos\theta_k \qquad \bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} sen\theta_k.$$

Ver Sección 1.1.1, Capítulo 1. Se debe notar que θ_k es el ángulo entre las componentes x_i y x_j de la k-ésima observación.

3.3.2. Descripción de variables composicionales a través del Modelo Normal proyectado

Uno de los principales problemas que se considera en este trabajo es modelar variables composicionales $\boldsymbol{x}$ en el simplex unitario S^q a través de modelos para datos direccionales. La restricción y el hecho de que realizaciones de $\boldsymbol{x}$ puedan incluir al 0 complica substancialmente el análisis de datos composicionales.

Varias distribuciones como la multinomial, la Dirichlet y Dirichlet generalizadas se han sugerido para modelar datos composicionales desde que ellas incorporan la restricción $\sum_{j=1}^{q} x_j = 1$. Sin embargo, como señala Aitchison (1986), para incorporar esta restricción , estos modelos imponen estructuras de correlación restrictivas. En su lugar Aitchison (1986) sugiere llevar (transformar) la variable composicional a $\mathbb{R}^{p-1}$, usando por ejemplo la transformación logcociente aditiva (alr) y tratar este vector transformado como un vector con distribución normal multivariada. Este enfoque resuelve el problema de tener restricciones en la estructura de covarianza, pero, como se señalo anteriormente, requiere que $x_j > 0 \; \forall \; \jmath = 1, \ldots, D$.

En la literatura se han propuesto transformaciones hiperfesféricas, como la transformación raíz cuadrada $\boldsymbol{u} = \sqrt{\boldsymbol{x}}$ (ver Sección anterior) de datos composicionales hacia la superficie de la esfera unitaria $(q-1)$-dimensional y así usar distribuciones de datos direccionales para modelar los datos composicionales. Por ejemplo, Wang *et al.* (2007) propone un enfoque *ad hoc* en el contexto de modelos econométrico para modelar $\boldsymbol{u}$ sin emplear algún modelo paramétrico y sin hacer referencia a la metodlogía para datos composiconales.

Por su parte Stephens (1982) modela el vector $\boldsymbol{u}$ usando una distribución von Mises-Fisher($\boldsymbol{\mu}, \kappa$) q-dimensional. Sin embargo, Mardia (1976) señala que al igual que la distribución multinomial, se tiene una estructura de correlación restrictiva. De manera reciente, Scely y Welsh (2011) emplean la distribución Kent para modelar la variable $\boldsymbol{u}$ como la variable de respuesta en un contexto de regresión. Al igual que Stephens (1982) ofrecen estimaciones (aproximaciones) solo en el caso κ grande. Como ellos mismos señalan su enfoque es apropiado cuando la mayoría de los datos se distribuyen lejos de las fronteras del ortante positivo y se concentran principalmente dentro del ortante. Cuando alguna de las componentes de $\boldsymbol{u}$ se distribuyen cercanas a 0, aparecen problemas de frontera y en esos casos se necesitan enfoques alternativos desde que el modelo Kent ajustado, así como el von Mises-Fisher puede tomar valores fuera del ortante positivo.

El modelo propuesto

En este trabajo de investigación nosotros aplicamos la transformación raíz cuadrada y la transformación proyectada, ver Secciones (3.2.1) y (3.2.2), respectivamente, a un vector composicional $\boldsymbol{x} \in S^D$ y posteriormente modelamos el vector de ángulos resultante $\boldsymbol{\Theta} = (\theta_1, \ldots, \theta_{D-1})$ como una variable con distribución Normal Proyectada $(D-1)$-dimensional, $NP_{D-1}(\boldsymbol{\mu}, \boldsymbol{I})$. Ver Capítulo 2.

Es decir si $\boldsymbol{X}$ es una variable en el simplex unitario D-dimensional, S^D, entonces,

$$f(\boldsymbol{\theta}(\boldsymbol{u})|\boldsymbol{\mu}) = NP_{D-1}(\boldsymbol{\theta}(\boldsymbol{u})|\boldsymbol{\mu}, \boldsymbol{I}).$$

Donde

$$\theta_j = tan^{-1}\left(\frac{\sqrt{\sum_{i=j+1}^{D} u_i^2}}{u_j}\right),$$

en el caso de la transformación raíz cuadrada,

$$\boldsymbol{u} = (\sqrt{x_1}, \sqrt{x_2}, \cdots, \sqrt{x_D})'.$$

Y en el caso de la transformación proyectada,

$$\theta_j = tan^{-1}\left(\frac{\sqrt{\sum_{i=j+1}^{D} x_i^2}}{x_j}\right).$$

con $j = 1, 2, \ldots, D-1$.

En el Capítulo 1, se mostró que

$$E(\boldsymbol{X}) = cen[\boldsymbol{X}] = ilr^{-1}(E[ilr(\boldsymbol{X})]) = ilr^{-1}(\boldsymbol{\mu_s}),$$

donde el parámetro $\boldsymbol{\mu_s}$ es el vector medio asociado a la distribución normal en el simplex (logística aditiva). Sin embargo, como ya se ha mencionado cuando existen componentes 0 del vector composicional $\boldsymbol{X}$, $cen[\boldsymbol{X}]$ no está definida.

En este trabajo se propone estimar el valor esperado de variables composicionales llevando a cabo inferencias bayesianas sobre el parámetro $\boldsymbol{\mu}$ del modelo Normal Proyectado, $NP_{D-1}(\boldsymbol{\theta(u)}|\boldsymbol{\mu}, \boldsymbol{I})$. Se debe hacer notar que bajo este enfoque la presencia de valores 0 en las componentes de $\boldsymbol{X}$ son tratados de manera natural a través de la transformación hiperesférica. La propuesta para estimar $E[\boldsymbol{X}]$, donde $\boldsymbol{X} \in S^D$, por medio de las transformaciones proyectada y raíz cuadrada es la siguiente:

- Dado $\boldsymbol{X} \in S^D$, hacer $\boldsymbol{U} = \boldsymbol{X}/||\boldsymbol{X}||$ y considerar $\boldsymbol{U(\theta)} \sim NP(\boldsymbol{\mu}, \boldsymbol{I})$. Mediante la introducción de variables latentes y emplenado los procedimientos MCMC descritos en el Capítulo 2, llevar a cabo inferencias (estimación puntual y construccción de intervalos de probabilidad) sobre el parámetro $\boldsymbol{\mu}$. Finalmente, obtener una aproximación a $E[\boldsymbol{X}]$ aplicando la operación clausura, definida en el simplex, al vector $\boldsymbol{\mu}$. Es decir, $\hat{E}[\boldsymbol{X}] = C(\boldsymbol{\mu})$.

- Dado $\boldsymbol{X} \in S^D$, hacer $\boldsymbol{U} = \sqrt{\boldsymbol{X}}$ y suponer que $\boldsymbol{U} \sim NP(\boldsymbol{\mu}, \boldsymbol{I})$. Una vez más, mediante la introducción de variables latentes y emplenado los procedimientos MCMC descritos en el Capìtulo 2, llevar a cabo inferencias (estimación puntual y construccción de intervalos de probabilidad) sobre el parámetro $\boldsymbol{\mu}$. En este caso al tomar un estimador $\boldsymbol{\mu^*}$ derivado de la distribución final $f(\boldsymbol{\mu}|\boldsymbol{\Theta})$, como un estimador del parámetro $\boldsymbol{\mu}$, aproximar $E[\boldsymbol{X}]$ a través de $\left(\frac{\mu^*}{||\mu^*||}\right)^2$, es decir, $\hat{E}[\boldsymbol{X}] = \left(\frac{\mu^*}{||\mu^*||}\right)^2$.

Como una medida la variabilidad de las dos aproximaciones, se puede considera la longitud media resultante (definida en el Capítulo 1) para ambas estimaciones con la varianza total definida en el simplex.

Capítulo 4

Ejemplos

Cuando se tiene un conjunto de datos composicionales desde un inicio es conveniente realizar una exploración de los mismos mediante algunas estadísticas descriptivas, por medio de las cuales se puede tener una idea de la importancia de aquellas variables con mayor variabilidad dentro de la composición. Como una primera aproximación se pueden comparar los centros muestrales para determinar la identificación de los componentes que pueden tener más peso en la composición. La matriz de variación composicional muestra el porcentaje de variación total expresada como pares de log-cocientes de los componentes, así como sus respectivas medias. Por otro lado, es importante analizar la media composicional por medio de la media direccional y la matriz de variación direccional para datos composicionales. Adicionalmente, se comparará la media composicional mediante la media direccional de la siguiente forma:

- Para el caso de la transformación proyectada se calculará la media direccional $\bar{\boldsymbol{\theta}} = (\bar{\theta}_1, \bar{\theta}_2, \ldots, \bar{\theta}_{D-1})$, posteriormente se aplicará la operación clausura $C(\cdot)$ a la transformación (1.1). Es decir,

$$\bar{\boldsymbol{u}} = C\left(cos\bar{\theta}_1, sen\bar{\theta}_1 cos\bar{\theta}_2, \ldots, sen\bar{\theta}_1 \cdots cos\bar{\theta}_{D-1}, sen\bar{\theta}_1 \cdots sen\bar{\theta}_{D-1}\right).$$

- Para el caso de la transformación raíz cuadrada se calculará la media direccional $\bar{\boldsymbol{\theta}}$, posteriormente cada componente de la transformación (1.1) se elevarán al cuadrado, las componentes del vector que resulta serían

$$\bar{\boldsymbol{w}} = ((cos\bar{\theta}_1)^2, (sen\bar{\theta}_1 cos\bar{\theta}_2)^2, \ldots, (sen\bar{\theta}_1 \cdots cos\bar{\theta}_{D-1})^2, (sen\bar{\theta}_1 \cdots sen\bar{\theta}_{D-1})^2).$$

Ejemplo 4.1 *Para este ejemplo se consideró un conjunto de 25 composiciones minerales de roca de tipo hongite. Los datos fueron tomados de Aitchison (1986) apéndice D (pag. 355) y cada composición* $\boldsymbol{X} = (x_1, x_2, x_3, x_4, x_5)$ *consiste en el porcentaje en peso de cinco minerales, albita* (x_1)*, blandite* (x_2)*, cornite* (x_3)*, daubite* (x_4) *y endite* (x_5)*. Como un análisis descriptivo de estos datos se obtendrán la media composicional vía la media direccional y la matriz de variación direccional para datos composicionales.*

Tabla 4.1: Comparación de la media composicional para los datos de hongite.

Medias	Albita	Blandite	Cornite	Daubite	Endite
Geométrica	0.4880	0.2157	0.1025	0.1059	0.0878
Proyectada	0.4397	0.2199	0.1583	0.0981	0.0839
Raíz	0.4627	0.2211	0.1297	0.1014	0.0851

Como se puede apreciar en la Tabla 4.1 los valores de la media composicional y los valores de la media composicional bajo las transformaciones proyectada y raíz cuadrada, son distintos, lo anterior, se debe a que la nube de puntos tiene una forma aproximadamente cóncava, esto se puede ver en las Figuras 4.1 y 4.2. Si los datos están concentrados, la media composicional y la media composicional bajo las transformaciones propuestas en este trabajo, resultarán muy similares. Esto se puede apreciar de mejor manera en el Ejemplo 4.2.

Tabla 4.2: Matriz de variación composicional

Medias/Varianzas	x_1	x_2	x_3	x_4	x_5
x_1	·	0.264	1.559	0.076	0.143
x_2	0.816	·	3.052	0.538	0.673
x_3	1.560	0.744	·	1.150	0.940
x_4	1.528	0.712	-0.032	·	0.179
x_5	1.715	0.899	0.155	0.187	·

La matriz de variación composicional para los datos de hongite esta dada en la Tabla 4.2. De la Tabla 4.2 se puede observar que la variación relativa más grande se da entre la componenete blandite (x_2) y la cornite (x_3) con $\hat{\tau}_{x_2x_3} = 3.052$. Por otro lado, el valor positivo $\hat{\xi}_{x_2x_3} = 0.744$ sugiere que x_2 tiene

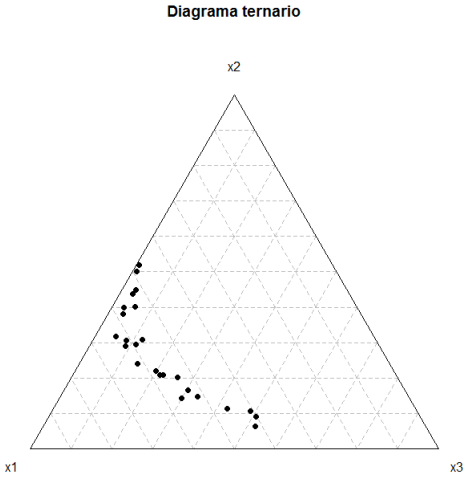

Figura 4.1: Diagrama ternario de las componentes x_1, x_2 y x_3. Ejemplo 4.1.

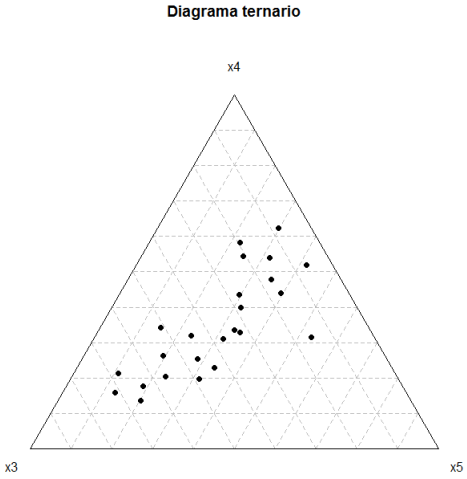

Figura 4.2: Diagrama ternario de las componentes x_3, x_4 y x_5. Ejemplo 4.1.

un mayor peso promedio en la composición que x_3. Sin embargo, dado que $\hat{\xi}_{x_2x_3} < \sqrt{\hat{\tau}_{x_2x_3}}$, se espera que para un número sustancial de réplicas $log(\frac{x_2}{x_3})$ sea negativo. Lo que indica que cierto porcentaje de x_3 excederá al de la componente x_2. Por otro lado, se tiene que $\hat{\tau}_{x_1x_4} = 0.076$, esto muestra que hay poca variabilidad relativa entre las componentes x_1 y x_4. Sin embargo, el hecho de que $\hat{\xi}_{x_1x_4} = 1.528$, y que $\hat{\xi}_{x_1x_4} > \sqrt{\hat{\tau}_{x_1x_4}}$ indica que la proporción de componente x_1 es consistentemente más grande que la proporción de la componente x_4. También se puede notar que la variabilidad relativa que hay entre la componente x_3 y x_4 es alta $\hat{\tau}_{x_3x_4}$=1.150, además $\hat{\xi}_{x_3x_4}$= -0.032 estos dos valores indican que en promedio las dos componentes tienen el mismo peso en la composición. Finalmente, para los datos de hongite el valor de la varianza total resulta ser 1.71481.

A continuación se presenta un análisis similar de los datos de hongite usando procedimientos para datos direccionales. La matriz de variación direccional para datos composicionales (ver Capítulo 3, Sección 3.3.1) se presenta en la Tabla 4.3. En esta tabla se puede observar, entre otras cosas, que la variación relativa más grande se da entre las componentes x_2 y la x_3 con un valor de $\hat{\eta}_{x_2x_3} = 0.119$. Además, la dirección media entre estas dos componentes es $\hat{\zeta}_{x_2x_3} = 25.439°$. Por lo tanto, se puede decir que en este conjunto de datos la componente x_2 tiende a tener mayor peso en la composición que

Tabla 4.3: Matriz de variación direccional para datos composicionales del Ejemplo 4.1.

	x_1	x_2	x_3	x_4	x_5
x_1	·	0.017	0.040	0.002	0.002
x_2	25.439	·	0.119	0.036	0.035
x_3	18.555	34.378	·	0.081	0.067
x_4	12.640	28.804	44.873	·	0.018
x_5	10.834	25.901	41.348	40.091	·

la componente x_3. Aunque inicialmente el valor de $\hat{\eta}_{x_2x_3} = 0.119$ se podría pensar grande, hay que recordar que $\hat{\eta}$ solo puede tomar valores en $[0, 1]$ (ver Sección 1.1.1). Así, en este caso podemos decir que para estos datos el porcentaje x_2 dentro de la composición es mayor que el de la componente x_3.

Por otro lado, se tiene que $\hat{\eta}_{x_1x_4} = 0.002$, lo cual muestra que la variabilidad relativa entre las componentes x_1 y x_4 es muy baja. Adicionalmente, el hecho de que $\hat{\eta}_{x_1x_4} = 12.640°$, indica que la proporción de componente x_1 es consistentemente más grande que la proporción de la componente x_4. Finalmente, se debe notar que auqnue la variabilidad relativa que hay entre la componente x_3 y x_4 no es *muy grande*, $\hat{\eta}_{x_3x_4}$=0.081, dado que $\hat{\eta}_{x_3x_4}$= 44.873° toma un valor cercano a 45° estos sugiere que en promedio las dos componentes representan el mismo peso en la composición. Se debe notar que todas estas interpretaciones están en concordancia con las que se pueden derivar de la matriz de variación composicional.

Ejemplo 4.2 *En este ejemplo se realiza el análisis de un conjunto de 73 datos de diferentes tipos de cerámica provenientes del sitio arqueológico de los Teteles de Ocotitla (ver, Argote-Espino y López-García (2014)). Las componentes de los datos composicionales de cerámica 10-dimensionales son* $\boldsymbol{X} = (Oxigeno\ (O), Sodio\ (Na), Magnesio\ (Mg),\ Aluminio\ (Al), Silicio\ (Si), Fosforo\ (P), Potasio\ (K), Calcio\ (Ca), Titanio\ (Ti), Hierro\ (Fe))$.

En la Tabla 4.4 se puede observar que las componente *Na*, *Mg*, *P* y *Ti* presentan ceros composicionales. En este caso al calcular la media composicional se tienen problemas en las componentes que presenten ceros, ya que para esas componentes la media composicional sería cero, y las demás componentes incrementarían su valor, y por tanto el análisis para la media composicional sería erróneo. Adicionalmente, para el cálculo de la matriz de variación composicional se tendrían problemas al aplicar los log-cocientes entre las partes que tengan ceros composicionales.

Tabla 4.4: Cantidad de ceros en cada componente.

O	Na	Mg	Al	Si	P	K	Ca	Ti	Fe
0	2	1	0	0	12	0	0	2	0

En Argote-Espino y López-García (2014) se aplica un algoritmo conocido como IRMI, el cual es utilizado para dar un valor a los ceros composicionales de la muestra. Por su parte, el paquete R-*compositions*, imputa ciertos valores a los ceros composicionales. En este ejemplo se presenta el análisis con dichos valores imputados para dar una estimación de la media composicional y la matriz de variación, lo anterior, con el fin de compararlos con el análisis vía variables direccionales bajo el cual se conservan los datos originales que incluyen los ceros, sin necesidad de remplazarlos $artificialmente$.

Tabla 4.5: Medias estimadas para los datos del Ejemplo 4.2.

Medias	O	Na	Mg	Al	Si
Compositions	0.258	0.535	0.015	0.006	0.097
IRMI	0.259	0.536	0.015	0.006	0.097
Proyectada	0.259	0.535	0.015	0.006	0.097
Raíz	0.258	0.535	0.014	0.006	0.097

Medias	P	K	Ca	Ti	Fe
Compositions	0.006	0.012	0.024	0.006	0.041
IRMI	0.004	0.012	0.024	0.006	0.041
Proyectada	0.004	0.013	0.025	0.006	0.041
Raíz	0.003	0.013	0.024	0.006	0.041

Del análisis de la Tabla 4.5, se pueden derivar las siguientes observaciones:

- Las medias composicionales estimadas para todas las componentes son muy similares. Lo anterior, se debe quizá a que los datos no tienen forma cóncava.
- Las componentes que tienen mayor peso en la muestra composicional son O y Si, en estas dos componentes se tiene en promedio el 78 % del peso de la composición

- Las medias de las componentes *Na* y *K* son parecidas, siendo la media de la componente *Na* mayor que la media de la componente *K*, lo mismo ocurre al comparar la media de las componentes *Mg* y *Ti*.
- La media de la componente *P* es la que tiene menor peso, a excepción de la estimación usando la función *compositions*.

Tabla 4.6: Matriz de variación composicional para el vector composicional del Ejemplo 4.2.

	O	Na	Mg	Al	Si
O	·	0.085	0.110	0.006	0.007
Na	3.578	·	0.200	0.076	0.085
Mg	4.465	0.887	·	0.127	0.108
Al	1.710	-1.868	-2.755	·	0.008
Si	0.729	-2.849	-3.736	-0.982	·
P	4.418	0.840	-0.048	2.707	3.689
K	3.761	0.184	-0.704	2.051	3.033
Ca	3.099	-0.479	-1.366	1.389	2.371
Ti	4.488	0.910	0.022	2.777	3.759
Fe	2.576	-1.002	-1.889	0.866	1.848

	P	K	Ca	Ti	Fe
O	0.704	0.088	0.043	0.260	0.040
Na	0.694	0.173	0.146	0.340	0.104
Mg	0.816	0.144	0.181	0.417	0.182
Al	0.714	0.097	0.035	0.243	0.033
Si	0.721	0.084	0.040	0.247	0.038
P	·	0.876	0.812	0.917	0.635
K	-0.656	·	0.133	0.320	0.128
Ca	-1.318	-0.662	·	0.219	0.059
Ti	0.070	0.726	1.388	·	0.226
Fe	-1.841	-1.185	-0.522	-1.911	·

En la Tabla 4.6 se analiza matriz de variación composicional, los valores imputados a los ceros composicionales son los que genera el paquete *compo-*

sitions, el lector puede revisar la estimación de la matriz de variación composicional para el algoritmo IRM en Argote-Espino y López-García (2014).

De la Tabla 4.6 se puede observar que las varianzas relativas de los log-cocientes más grandes son aquellas donde la componente P esta presente, esto se debe a los valores imputados donde hay un cero composicional, ya que cualquier cambio en alguna de las componentes afecta a por lo menos una componente del dato composicional. También se puede notar que entre las variabilidades relativas más grandes están la de las componentes Ti con Mg y Ti con K. Dado que para la mayoría de los log-cocientes se tiene que $\hat{\xi}_{x_i x_j} > \sqrt{\hat{\tau}_{x_i x_j}}$, se puede decir que, si $\hat{\xi}_{x_i x_j} > 0$ entonces la proporción de componente x_i es considerablemente más grande que la proporción de la componente x_j, y caso contrario si $\hat{\xi}_{x_i x_j} < 0$.

Se debe poner mucha atención para el caso $\hat{\xi}_{x_{Mg} x_P}$ =-0.048, ya que este caso $\hat{\xi}_{x_{Mg} x_P} < 0$, entonces se podría concluir que la componente P tiene mayor peso en promedio que Mg en la composición $\mathbf{X}$. Sin embargo, esto no necesariamente es correcto, ya que este signo negativo se debe en parte a los valores imputados en los ceros de la componente P. Ver análisis de la Tabla 4.7.

Uno de los motivos de describir variables composicionales a través de variables direccionales es la presencia de ceros en los datos. A continuación se analiza el mismo conjunto de datos de cerámica sin necesidad de imputar algún valor a los ceros composicionales.

En la Tabla 4.7 se puede ver que la variación relativa más grande, esta entre la componente Ti y la componente P teniendo un valor de 0.064. Además, que la dirección media entre estas dos componentes es $\hat{\zeta}_{PTi} = 55.313°$ esto indica que Ti tiende a tener mayor peso que P en la composición. Se puede notar que las interpretaciones anteriores, están en concordancia con los derivados a partir de la matriz de variación composicional (Tabla 4.6), con la excepción de que en este análisis no se asigna (remplaza) ningún valor a los ceros observados.

De la Tabla 4.7, también se desprende que la variación relativa más grande es la formada por las componentes Mg y P cuyo valor es 0.053 y la dirección media entre ellas es $\hat{\zeta}_{MgP} = 32.739°$ lo cual indica que la componente Mg tiene mayor peso promedio en la composición que P. Este resultado claramente se contrapone a la conclusión mediante la matriz de variación composicional la cual considera valores remplazados de los ceros observados (Tabla 4.6).

Se ha visto hasta ahora que el método propuesto para el cálculo de la media composicional mediante variables direccionales es adecuado. Sin embargo, hay que señalar que esto ocurre en general cuando los datos composicionales tengan forma no cóncava.Adicionalmente, bajo este esquema se resuelve el

problema de presencia de ceros composicionales. Así mismo, se puede apreciar que el análisis descriptivo de los datos composicionales mediante la matriz de variación direccional para datos composicionales, es una buena opción para describir datos composicionales, cuando se tienen o no ceros composicionales en el conjunto de datos.

Tabla 4.7: Matriz de variación direccional para datos composicionales para el vector composicional del Ejemplo 4.2.

	O	Na	Mg	Al	Si
O	·	2.851e-05	5.894e-06	9.834e-05	5.070e-04
Na	1.561	·	2.683e-02	7.597e-04	1.175e-04
Mg	0.661	24.282	·	2.099e-04	2.540e-05
Al	10.279	81.507	86.342	·	3.989e-04
Si	25.803	86.781	88.628	69.405	·
P	0.442	15.740	32.739	2.439	0.093
K	1.391	41.654	63.580	7.672	2.873
Ca	2.634	59.596	75.539	14.198	5.434
Ti	0.647	23.817	43.129	3.573	1.346
Fe	4.435	70.101	81.314	23.039	9.111

	P	K	Ca	Ti	Fe
O	1.086e-05	2.649e-05	3.895-05	1.484e-05	0.0001
Na	1.509e-02	2.429e-02	1.005e-02	3.254e-02	0.006
Mg	5.328e-02	1.061e-02	4.667e-03	3.962e-02	0.001
Al	3.345e-04	9.038e-04	8.666e-04	4.297e-04	0.002
Si	5.098e-05	1.033e-04	1.375e-04	6.220e-05	0.0004
P	·	2.022e-02	4.902e-03	6.533e-02	0.002
K	72.078	·	1.137e-02	2.082e-02	0.005
Ca	80.559	62.026	·	4.561e-03	0.005
Ti	55.313	25.596	13.750	·	0.002
Fe	83.991	72.201	58.968	81.602	·

4.1. Simulaciones

Para poder llevar a cabo inferencias estadísticas a partir de un conjunto de datos en general es necesario proponer un modelo probabilístico, el cual pueda describir adecuadamente el comportamiento de los datos con los que se esta trabajando. En esta sección se presentan un par de ejemplos con datos simulados. Lo anterior, asumiendo, por un lado, un modelo normal en el simplex (logístico aditivo) y por otro lado un modelo Normal proyectado.

Ejemplo 4.3 *Para este ejemplo se simuló una muestra de tamaño 100 de datos composicionales* $\boldsymbol{X} = (x_1, x_2, x_3)$ *que se distribuyen normal en el simplex con* $\boldsymbol{\mu} = (0.2553065, -0.2631642)$ *y matriz de varianzas y covarianzas*

$$\boldsymbol{\Sigma} = \begin{pmatrix} 0.18009947 & 0.04932677 \\ 0.04932677 & 0.17457692 \end{pmatrix}$$

Los datos simulados se pueden observar en el diagrama ternario de la Figura 4.3

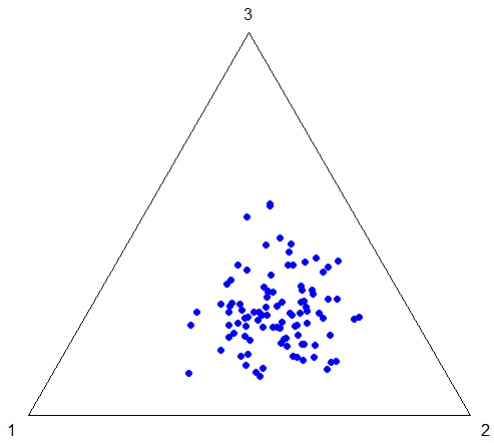

Figura 4.3: Diagrama ternario para los datos simulados del Ejemplo 4.3.

El valor esperado bajo este modelo esta dado por

$$\boldsymbol{\mu}_s = E[\boldsymbol{X}] = ilr^{-1}(\boldsymbol{\mu}) = (0.3027864, 0.4344520, 0.2627617)$$

y su valor estimado, usando el conjunto de los 100 datos a través de la media geométrica, esta dado por

$$\hat{\boldsymbol{\mu}}_s = (0.2970625, 0.4369151, 0.2660224).$$

Por otro lado, si se considera un modelo Normal proyectado, $NP_3(\boldsymbol{\mu^1}, \boldsymbol{I})$ y la transformación proyectada definida en la Sección 3.2.2, se genera una muestra de la distribución final de $\boldsymbol{\mu^1}$ y se toma la Clausura del vector de medianas de la muestra obtenida como estimador de $\boldsymbol{\mu}_s$ bajo este modelo es (2.753, 3.862, 2.498). Ver Sección 3.3.2. Es decir,

$$\hat{\boldsymbol{\mu}}_s^1 = C(\hat{\boldsymbol{\mu}}^{\mathbf{1}}) = (0.2997201, 0.4348037, 0.265462).$$

Para calcular un estimador bajo el modelo normal proyectado, $NP_3(\boldsymbol{\mu}^1, \boldsymbol{I})$, con la transformación raíz cuadrada, se toma como estimador de $\boldsymbol{\mu}_s$ el cuadrado del vector de medias normalizado, derivado de la muestra obtenida de la distribución final del parámetro $\boldsymbol{\mu^1}$. Ver, Sección 3.3.2. Así,

$$\hat{\boldsymbol{\mu}}_s^2 = (0.2941683, 0.4423807, 0.2634510).$$

En la Tabla 4.8 se dan intervalos bayesianos del 95 % para el verdadero valor de $E[\boldsymbol{X}]$ (el cual es $\boldsymbol{\mu}_s = (0.302, 0.434, 0.263)$) de los datos simulados bajo el modelo normal en el simplex. Como se podrá observar el valor verdadero de $\boldsymbol{\mu}_s$ cae en el intervalo de calculado bajo las dos transformaciones.

Tabla 4.8: Intervalos bayesianos del 95 % de credibilidad para $\boldsymbol{\mu}_s = E[\boldsymbol{X}]$.

Proyectada	Raíz
(0.2837986, 0.3196554)	(0.2805776, 0.3184318)
(0.4080646, 0.4448228)	(0.4119754, 0.4518156)
(0.2527507, 0.2886972)	(0.2501062, 0.2863924)

De los resultados anteriores, se puede observar que bajo el enfoque de datos direccionales se pueden realizar inferencias adecuadas para estimar la esperanza de un vector composicional.

Nota: Para dar una muestra de la distribución final del parámetro $\boldsymbol{\mu}^1$ del modelo $NP(\boldsymbol{\mu}^1, \boldsymbol{I})$ bajo la transformación proyectada se tomó *burn in* $l = 10000$ y *thinning* $k = 5$, se eligió dar el valor $\boldsymbol{\mu}_0 = \mathbf{0}$ y $\lambda = 0{,}0001$ como parámetros de la distribución inicial del parámetro $\boldsymbol{\mu}^1$. En al Figura 4.4 se puede apreciar el promedio ergódico, la función de autocorrelación y el histograma de la distribución final para cada una de las componentes del parámetro $\boldsymbol{\mu}^1$.

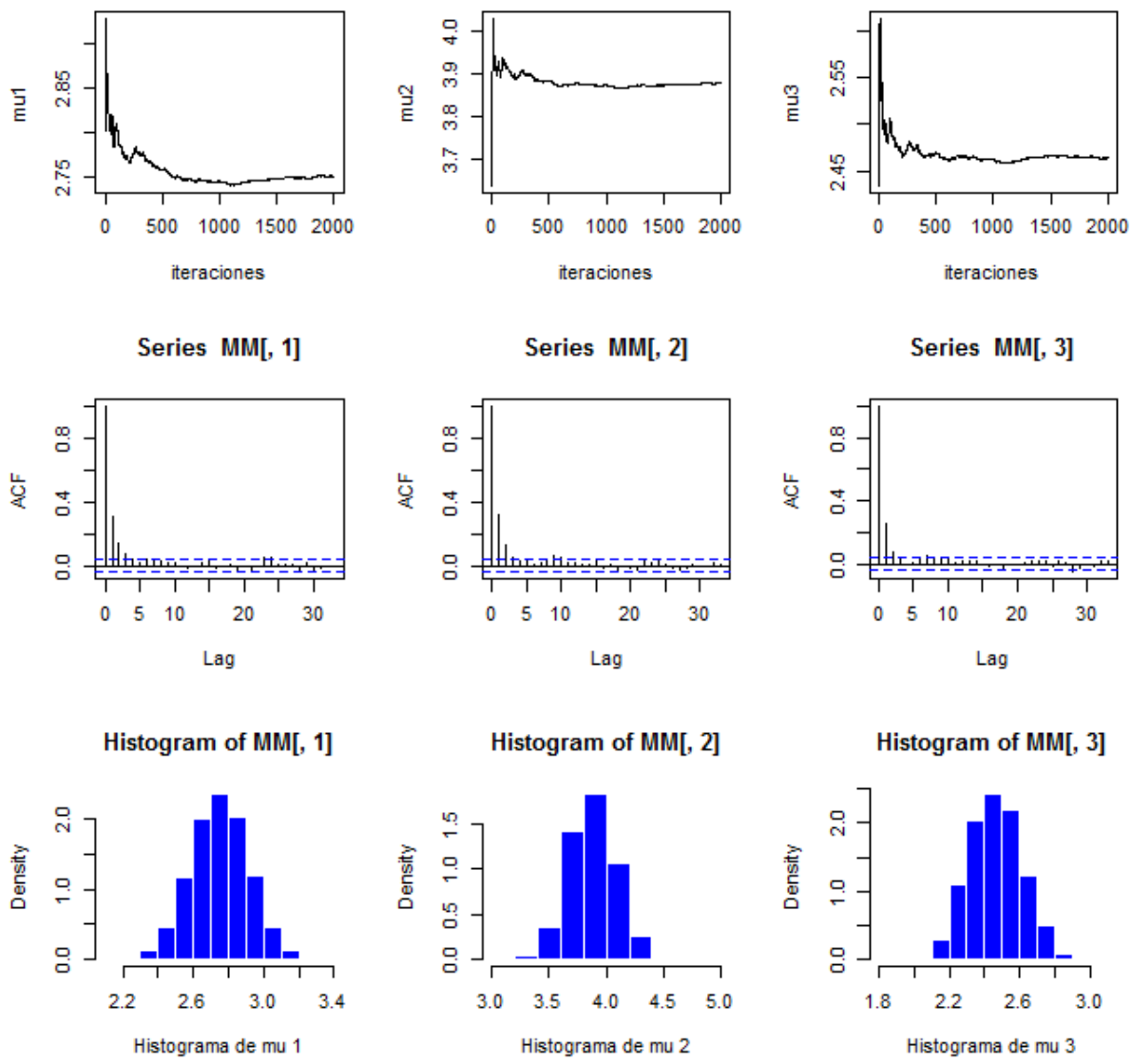

Figura 4.4: Promedios ergódicos,función de autocorrelación e histogramas de la distribución final del parámetro $\boldsymbol{\mu}$.

Ejemplo 4.4 *Para este ejemplo se simuló una muestra de 100 de datos composicionales* $\boldsymbol{Y} = (y_1, y_2, y_3)$ *bajo un modelo* $NP(\boldsymbol{\mu}, \boldsymbol{I})$ *con* $\boldsymbol{\mu} = (2.7, 3.4, 5.4)$. *Los datos se pueden visualizar en la Figura 4.5.*

Para este caso se tiene que $E[\boldsymbol{X}] = (0.2347826, 0.4695652, 0.2956522)$. Los estimadores correspondientes bajo un modelo normal en el simplex y considerando un modelo Normal proyectado (empleando las transformaciones proyectadas y raíaz cuadarada) resultaron, respectivamente:

$$\hat{\boldsymbol{\mu}}_s = (0.2338249, 0.4701506, 0.2960246).$$
$$\hat{\boldsymbol{\mu}}_s^1 = (0.2400169, 0.4667291, 0.293254).$$
$$\hat{\boldsymbol{\mu}}_s^2 = (0.2324327, 0.4749451, 0.2926223).$$

Figura 4.5: Diagrama ternario de los datos simulados del Ejemplo 4.4.

Tabla 4.9: Intervalos bayesianos del 95 % de credibilidad para $E[\boldsymbol{X}]$ bajo el modelo normal proyectado.

Proyectada	Raíz
(0.2250657, 0.2543093)	(0.2170373, 0.2473211)
(0.4519982, 0.4814332)	(0.4564992, 0.4928852)
(0.2786409, 0.3080116)	(0.2768831, 0.3090530)

Como se puede ver, en este ejemplo el empleo del enfoque direccional para describir datos composicionales también resulta adecuado. En general, los

ejemplos presentados muestran que las inferencias obtenidas bajo un modelo Normal proyectado resultan adecuadas para estimar la esperanza de variables composicionales.

Capítulo 5

Conclusiones y Perspectivas

El estudio estadístico de datos en espacios topológicos diferentes a $\mathbb{R}^k$ resulta muy interesante, pero representa todo un desafío. No solo por la complejidad natural de cada espacio, sino también, por el reto que implica implementar los procedimientos correspondientes para el análisis de fenómenos reales.

En el contexto de datos composicionales, hoy en día aún se tiene discusión sobre algunos temas básicos que van desde la propuesta de modelos probabilísticos para su descripción hasta la forma de llevar a cabo inferencias en los modelos propuestos. Aunque el enfoque propuesto por Aitchison (1982) sentó las bases para un análisis general, éste esta basado principalmente en transformaciones que involucran logaritmos de cocientes. Lo anterior, tiene como consecuencia que ante la presencia de ceros en las componentes del dato composicional se tenga que utilizar algún procedimiento *ad hoc* o alguno que no sea de carácter general. Por su parte, aunque aún no se tiene completamente desarrollada todo una teoría como la que existe para datos en $\mathbb{R}^k$, el análisis de datos direccionales ha tenido un gran avance en los últimos años. Incluso en esta área ya se cuenta con modelos no-paramétricos bayesianos para su descripción.

En este trabajo se revisaron los conceptos asociados a la modelación tanto de datos composicionales como de datos direccionales. En el caso de datos direccionales se construyó la distribución Normal proyectada $NP_q(\boldsymbol{\mu}, \boldsymbol{I})$ para el caso general q-dimensional y se derivaron todas las distribuciones condicionales completas que permitieron llevar a cabo inferencias para todos los parámetros en el modelo. Lo anterior, a través de técnicas de Monte Carlo vía Cadenas de Markov, como el muestreo de Gibbs. Se debe señalar que para poder llevar a cabo estos procedimientos en dimensiones altas se implementó un paso interno de Newton-Raphson para simular de algunas variables

latentes. Esta variante resulta en algoritmos más rápidos que los que presentan Nuñez-Antonio y Gutiérrez-Peña (2005) . Además, estos autores solo analizan el caso circular (q=2).

A nivel descriptivo se propuso la construcción de la matriz de variación direccional para describir datos composicionales, la cual evita el problema de la presencia de ceros. Por otro lado, aunque la transformación raíz cuadrada posee propiedades atractivas, en este trabajo también se exploró la transformación proyectada. Ambas transformaciones *mapean* datos cuyo espacio muestral es el simplex q-dimensional al ortante positivo de la hiper-esfera unitaria de dimensión q. Lo anterior, da la posibilidad de analizar los datos composicionales utilizando procedimientos definidos para variables direccionales. Particularmente, nosotros llevamos acabo este tipo de procedimientos empleando el modelo Normal proyectado $NP_q(\boldsymbol{\mu}, \boldsymbol{I})$. Se debe señalar que no cualquier conjunto de datos composicionales se puede describir con este enfoque. Lo anterior, dado que el modelo $NP_q(\boldsymbol{\mu}, \boldsymbol{I})$ solo es capaz de describir comportamientos *no-cóncavos* de conjuntos de datos, lo cual se debe a la estructura de covarianza de la matriz indentidad. Sin embargo, a pesar de que el modelo $NP_q(\boldsymbol{\mu}, \boldsymbol{I})$ parece restrictivo, este resulta muy simple y es comparable a modelos más complejos como el von Mises-Fisher y la distribución Kent. Se debe señalar que el poder usar el modelo $NP_q(\boldsymbol{\mu}, \boldsymbol{I})$ para describir ciertos comportamientos de datos composicionales abre la posibilidad de usar modelos de *mezclas de Normales proyectadas* (ver Nuñez-Antonio *et al.* (2015)) para modelar cualquier comportamiento de datos composicionales. Así, el uso del modelo $NP_q(\boldsymbol{\mu}, \boldsymbol{I})$ para describir datos composicionales puede resultar atractivo.

A pesar de las limitaciones y/o desventajas que se puedan tener, los modelos y las metodología desarrolladas en este trabajo de investigación representan una opción a considerar para la descripción de datos composicionales. Finalmente, dado que los modelos discutidos aquí se basan en una distribución normal multivariada bajo proyección, estos se pueden extender de manera más natural al estudio de datos composicionales en contextos más generales como los de regresión.

Bibliografía

Aitchison Jhon. y Kay Jhon W . Possible solutions of some essential zero problems in compositonal data analysis. *Coda Work 03*, 2003.

Aitchison John . The statistical analysis of compostional data (with discussion). *Journal of hte Royal Statistical Society*, 44:139–177, 1982.

Aitchison John . *The statistical analysis of compositional data*. Chapman and Hall, 1986.

Argote-Espino Denisse L. y López-García Pedro A . Análisis de datos composicionales para el estudio de materiales arqueológicos, 2014.

Bacom-Shone Jhon . Modelling structural zeros in compositional data. *Coda Work 03*, 2003.

Bacom-Shone Jhon . Discrete and continuos compostional. *Coda Work 08*, 2008.

Berger J.O . *Statistical Decision Theory and Byesian Analysis*. New York: Springer Verlag, 1985.

Bernado A.F.M , J.M y Smith. *Bayesian Theory*. Chichester: Wiley, 1994.

Berry D.A . *Statistics: A Bayesian Perspective*. Belmont: Duxbury Press, 1996.

Bolstad W.M . *Introduction to Bayesian Statistics*. Hoboken,N.J: Wiley-Interscience, 2007.

Box G.E.P. y Tiao G.C . *Bayesian Inference in Statistical Analysis*. Reading, Mass: Addison-Wesley, 1973.

Daunis-Estadella Josep , Martín-Fernández Josep Antoni , y Palarea-Albaladejo Javier . Bayesian tools for count zeros in compositional data sets. 2008.

Egozcue Juan José . On the harker variation diagrams. *Mathematica Geosciences*, pages 829–834, 2009.

Fisher R . Dispersion on a sphere. *Proceedings of the Royal Society of London Series A*, 217:295–305, 1953. doi: 10.1098/rspa.1953.0064.

Fry J. , Fry T. , y McLaren K . Compositional data analysis and zeros in micro data. *Applied Economics 32*, pages 953–959, 2000.

Gamerman Dani. y Lopes Hedibert. F . *Markov Chain Monte Carlo Stochastic Simulation for Bayesian Inference*. Chapman y Hall/CRC, 2006.

Geral van den Boogaart K. y Tolosana-Delgado Raimon . *Analyzing Compositional Data with R*. Springer, 2013.

Leonard T. y Hsu J . *Bayesian Methods*. New York: Cambridge University Press, 1999.

Mardia K.V . Discussion on 'the ordering of multivariate data'. *Journal of hte Royal Statistical Society*, pages 346–347, 1976.

Mardia K.V. y Jupp P.E . *Directional Statistics*. Chischester: Wiley, 2000.

Martín-Fernández Josep Antoni , Palarea-Albaladejo Javier , y Barceló-Vidal C . Técnicas composicionales para concentraciones geoquímicas por deabajo del límite de detección. *Boletín Geolólogico y Minero*, 2001.

Martín-Fernández Josep Antoni , Barceló-Vidal C. , y Pawlowsky-Glahn Vera . Dealing whit zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, 2003.

Mateu-Figueras Gloria. , Pawlowsky-Glahn Vera. , y Egozcue Juan Jose . The normal distribution in some constrained sample space. *Statistis and Operations Research Transactions*, 37(1):29–26, 2013.

Mood Alexander , Graybill Franklin , y Boes Duanes . *Introduction to the theory of statistics*. McGraw-Hill, 1974.

Neocleous Tereza. , Aitken Colin. , y Zadora Grzegorz . Transformations for compositional data with zeros with and application to foresic evidence evaluation. *Chenometrics and Intelligent Laboratory Systems*, pages 77–85, 2011.

Nuñez Antonio Gabriel . *Análisis Bayesiano de Modelos Lineales para Datos Direccionales considerando la Distribución Normal bajo Proyección*. PhD thesis, Universidad Autónoma Metropolitana, 2010.

Nuñez-Antonio Gabriel. , Concepción-Ausín Ma. , y Wiper Michael Peter . Bayesian nonparametric models of circular variables based on dirichlet process mixtures of normal distributions. *Journal of Agricultural Biological and Environmental Statistics*, 20(1):47–64, 2015.

Nuñez-Antonio Gabriel y Gutiérrez-Peña Eduardo . A bayesian analysis of directional data using the projected normal distribution. *Journal of Applied Statistics*, pages 995–1001, 2005.

Palarea-Albaladejo Javier y Martín-Fernández Josep Antoni . A modified em alr-algorithm for replacing rounded zeros in compositional data sets. *Computer and Gosciences*, 2008.

Palarea-Albaladejo Javier. , Martín-Fernández Josep Antoni. , y Gómez-García Juan . A parametric approach for dealing with compositional rounded zeros. *International Association for Mathematical Geology*, 2007.

Pawlowsky-Glahn V. y Egozcue J.J . Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*, pages 384–398, 2001.

Pawlowsky-Glahn V. y Egozcue Juan Jose . Blu estimators and compositional data. *Mathematical Geology*, 34(3):259–274, 2002.

Pawlowsky-Glahn Vera. , Egozcue Juan José. , y Tolosana-Delgado Raimon . *Modeling and Analysis of Compositonal Data*. Wiley, 2015.

Presnell Brett y Rumcheva Pavlina . The mean resultant length of the spherically projected normal distribution. *Statistics and Probability Letters*, 78:557–563, 2008.

Presnell Brett. , Morrison Scott P. , y Littell Ramon C . Projected multivariate linear models for directional data. *Journal of the American Statistical Association*, 93(443):1068–1077, 1998.

Scely J.L. y Welsh A.H . Regression for compositional data by using distribution defined on the hypersphere. *Journal of the Royal Statistical Society*, pages 351–375, 2011.

Stephens M.A . Use of the von mises distribution to analyses continuous proportions. *Biometrika*, pages 197–203, 1982.

Wang F. y Gelfand A.E . Directional data analysis under the general projected normal distribution. *Stat Methodol.*, pages 113–127, 2013.

Wang Huiwen , Qiang Lui , y Mok Henry. M.K . A hyperspherical transformation forecasting model for compositional data. *European Journal of Operational Rearch*, pages 459–468, 2007.

Apéndice

A.1

En este Apéndice se presentan las demostraciones del Capítulo 2.

Proposición 5.1 *Bajo las mismas premisas de la Proposición (2.9), la función de densidad del ángulo aleatorio* $\boldsymbol{\Theta}$*, es decir, la densidad de probabilidad de la correspondiente Normal proyectada q-variada, está dada por*

$$NP(\boldsymbol{\theta}|\boldsymbol{\mu},\boldsymbol{I}) = \int_0^\infty f(r,\boldsymbol{\theta}|\boldsymbol{\mu},\boldsymbol{I})dr$$

$$= K_q[(q-2)\int_0^\infty r^{q-3} exp\left\{-\frac{1}{2}r^2\right\} exp\left\{br\right\} dr$$

$$+b\int_0^\infty r^{q-2} exp\left\{-\frac{1}{2}r^2\right\} exp\left\{br\right\} dr] 1_{[0,2\pi]}(\theta) 1_{[0,\pi]}(\theta_1)\cdots 1_{[0,\pi]}(\theta_{q-1})$$

Demostración:

$$\begin{aligned} NP(\theta,\phi|\boldsymbol{\mu},\boldsymbol{I}) &= \int_0^\infty f(r,\theta,\phi|\boldsymbol{\mu},\boldsymbol{I})dr \\ &= K_q \int_0^\infty r^{q-1}\exp\{-\frac{1}{2}r^2\}\ \exp\{b\ r\}\ dr. \end{aligned}$$

Trabajando sólo con la integral y haciendo integración por partes

$$u = r^{q-2}\exp\{\ b\ r\} \qquad dv = r\ \exp\{-\frac{1}{2}r^2\}dr$$

$$du = (br^{q-2}\exp\{\ br\} + (q-2)r^{q-3}\exp\{\ br\})dr \qquad v = -\exp\{-\frac{1}{2}r^2\}$$

entonces, tenemos que

$$\begin{aligned}\int_0^\infty r^{q-1} exp\left\{-\tfrac{1}{2}r^2\right\} exp\{br\}\, dr &= b\int_0^\infty r^{q-2} exp\left\{-\tfrac{1}{2}r^2\right\} exp\{br\}\, dr \\ &+ (q-2)\int_0^\infty r^{q-3} exp\left\{-\tfrac{1}{2}r^2\right\} exp\{br\}\, dr\end{aligned}$$

□

Proposición 5.2 *Bajo las mismas premisas de la Proposición (2.9) la función de densidad condicional acumulada de R dado $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_{q-1})$ está dada por*

$$\begin{aligned}F(r|\theta, \boldsymbol{\mu}, \boldsymbol{I}) &= \frac{\int_0^r w^{q-1}\ \exp(-\frac{1}{2}\ [w^2\ -\ 2\ b\ w])dw}{\int_0^\infty r^{q-1} exp\left\{-\frac{1}{2}r^2\right\} exp\left\{-\frac{1}{2}r^2\right\} dr} \\ &= \frac{-r^{q-2}\ \exp(-\frac{1}{2}\ [r^2\ -\ 2\ b\ r])}{\int_0^\infty r^{q-1} exp\left\{-\frac{1}{2}r^2\right\} exp\left\{-\frac{1}{2}r^2\right\} dr} \\ &+ \frac{b\int_0^r w^{q-2}\ \exp(-\frac{1}{2}\ [w^2\ -\ 2\ b\ w])dw}{\int_0^\infty r^{q-1} exp\left\{-\frac{1}{2}r^2\right\} exp\left\{-\frac{1}{2}r^2\right\} dr} \\ &+ \frac{(q-2)\int_0^r w^{q-3}\ \exp(-\frac{1}{2}\ [w^2\ -\ 2\ b\ w])dw}{\int_0^\infty r^{q-1} exp\left\{-\frac{1}{2}r^2\right\} exp\left\{-\frac{1}{2}r^2\right\} dr}\end{aligned}$$

Demostración:
Se sabe que

$$F(r|\theta, \boldsymbol{\mu}, \boldsymbol{I}) = \frac{\int_0^r w^{q-1}\ \exp(-\frac{1}{2}\ [w^2\ -\ 2\ b\ w])dw}{\int_0^\infty r^{q-1} exp\left\{-\frac{1}{2}r^2\right\} exp\left\{-\frac{1}{2}r^2\right\} dr}$$

Por demostrar

$$\begin{aligned}\int_0^r w^{q-1}\ \exp\,(-\frac{1}{2}\ [w^2\ -\ 2\ b\ w])dw\ &=\ -r^{q-2}\ \exp\,(-\frac{1}{2}\ [r^2\ -\ 2\ b\ r])\\ &+\ b\int_0^r w^{q-2}\ \exp\,(-\frac{1}{2}\ [w^2\ -\ 2\ b\ w])dw\\ &+\ (q-2)\int_0^r w^{q-3}\ \exp\,(-\frac{1}{2}\ [w^2\ -\ 2\ b\ w])dw\end{aligned}$$

Haciendo integración por partes y tomando

$$u = w^{q-2}\exp\{\ b\ w\} \qquad dv = w\ \exp\{-\frac{1}{2}w^2\}dw$$

$$du = (bw^{q-2}\exp\{\ bw\} + (q-2)w^{q-3}\exp\{\ bw\})dw \qquad v = -\exp\{-\frac{1}{2}w^2\}$$

Se obtiene el resultado.

A.2

Este Apéndice contiene los programas necesarios para implementar la metodología propuesta en este trabajo de investigación. Además, estos programas permiten llevar a cabo y reproducir todos los resultados y ejemplos obtenidos. Los programas NR3.R, NR4.R y MatrizCD.R corren en R en su version 3.1.0. Los programas requieren de los R-paquetes MASS, CircStat y compositions.

Contenido del Apéndice A.2

NR3.R

```
# Programa NR3.R
# En este programa se implementan
# el método de Newton-Rap, el muestreo de Gibbs,
# para tomar muestras de las distribuciones finales
# relativas a la distribución NP3(mu,I)


#-------------------------------------------------------------


# Constante de Normalización
kn<-function(bb)
        {
        drop(bb + ((bb^2 + 1)*pnorm(bb)/dnorm(bb)))
        }


Dbd<-function(t1,t2,mu1,mu2,mu3)
         {
         trans<-c(cos(t1),cos(t2)*sin(t1),sin(t2)*sin(t1))
         mu<-c(mu1,mu2,mu3)
         drop(crossprod(trans,mu))
        }

# Función para calcular la moda de frct

moda0<-function(bb)
        {
        #moda<-(bb+sqrt((bb^2)+4*3))/2
        moda<-(bb+sqrt((bb^2)+4))/2
        moda
        }

# Función de densidad condicional de r dado theta
frct<-function(r,bb)
         {
        ( (r^2)*exp( -0.5*( (r^2) - 2*bb*r)  )  )/kn(bb)
```

```
        }


# Función de densidad condicional acumulada de r dado theta.
Frct<-function(r,bb)
        {
        Frdt<-( ( bb-(bb+r)*exp(bb*r-((r^2)/2)) ) +
        (bb^2+1)*(1/dnorm(bb))*(pnorm(r-bb)-pnorm(-bb)) ) /kn(bb)
        Frdt
        }


# Métodо de Newton-Rap
NR<-function(g,f,tt1,tt2,mu1,mu2,mu3,tam.muestra,no.ite=2)
        {
  N<-tam.muestra
  ite<-no.ite
  bb<<-Dbd(tt1,tt2,mu1,mu2,mu3)
  r0<-moda0(bb)
  r.n<-r0
  u<-runif(N)
  for (i in 1:ite)
        {
        r.n<-r.n-((g(r.n,bb)-u)/(f(r.n,bb)))
        }
  rdt<-r.n
        }


###---------------- Muestro de Gibss ------------------------------
#Paqutes necesarios

library(MASS)
library(CircStats)
library(compositions)
x1<- rnorm(100,3.5,1)
x2 <- rnorm(100,5,1)
x3 <- rnorm(100,2.7,1)
# comando para volver un vector no-negativo en un vector
#composicional
z<-acomp(cbind(x1,x2,x3))
#z<-sqrt(z1)
#ángulos asociados a los datos composicionales
t1<-(atan2(sqrt((z[,2])^2+(z[,3]^2)),z[,1]))
t2<-(atan2(z[,3],z[,2]))
theta<-cbind(t1,t2)
n<-length(theta)/2
circ.plot(t1,stack=TRUE, bins=150, shrink=1.5,main='theta')
circ.plot(t2,stack=TRUE, bins=150, shrink=1.5,main='phi')
#comando para el diagrama ternario
plot(z,pch = 16,col = "blue")
#write(c(t1,t2), file="datos.r",100)
#the<-scan(file="datos.r")
#theta<- matrix(the,nrow=100,ncol=2)
#theta[,1]
#theta[,2]
#n<-length(theta)
#---------------termina de generar datos
datosx<-cbind(cos(theta[,1]),cos(theta[,2])*sin(theta[,1]),
sin(theta[,2])*sin(theta[,1]))

# Especificación de los valores para la distribución a priori
```

```
mu0<-c(0.0,0.0,0.0)
lambda0<-0.0001

# Valores iniciales de la variable latente

r<-rep(1,n)

# Número de iteraciones

  tm<-2000
# Saltos
  t.lag<-7
  kk<-tm*t.lag
  print(paste(" Total iterations =", kk, "..."))

# Matriz para obtener la muestra final

  MM<-matrix(0,tm,3)

# Periodo de calentamiento
burn<-10000

for(k in 1:(burn+kk))
        {

        x<-r*datosx
        # Sampling of vector mu.
        lambdaF<-(n+lambda0)
        mu1.e<-( n*mean(x[,1]) + lambda0*mu0[1] )/lambdaF
        mu2.e<-( n*mean(x[,2]) + lambda0*mu0[2] )/lambdaF
        mu3.e<-( n*mean(x[,3]) + lambda0*mu0[3] )/lambdaF
        desv<-sqrt( 1/lambdaF )

        mu.e<-c(rnorm(1,mu1.e,desv),rnorm(1,mu2.e,desv),
        rnorm(1,mu3.e,desv))

        # Sampling of vector r

        for(j in 1:n)
                {
                t.e1<-theta[j,1]
                t.e2<-theta[j,2]
                b<-Dbd(t.e1,t.e2,mu.e[1],mu.e[2],mu.e[3])
                # Usando Newton Rapson.
                r[j]<-NR(Frct,frct,t.e1,t.e2,mu.e[1],mu.e[2],
                mu.e[3],1)
}

#Values of each iteration
if(k>burn){
flag1<-(k/500)-trunc(k/500)
if(flag1==0){print(k-burn)}
flag2<-((k-burn)/t.lag)-trunc((k-burn)/t.lag)
if(flag2==0)
{
ii<-(k-burn)/t.lag
MM[ii,]<-mu.e
}
}

#####---------- Gibss completed ----------
```

```
}

###------------------------------ Diagnostico------------------------

med.erg<-cbind(cumsum(MM[,1]),cumsum(MM[,2]),cumsum(MM[,3])
  )/(1:tm)

par(mfrow=c(3,3))
plot(med.erg[,1],type="l",xlab="iteraciones",ylab="mu1")
plot(med.erg[,2],type="l",xlab="iteraciones",ylab="mu2")
plot(med.erg[,3],type="l",xlab="iteraciones",ylab="mu3")
#library(examples)
acf(MM[,1])
acf(MM[,2])
acf(MM[,3])
hist(MM[,1],freq=F,col="blue",border="white",xlab =
'Histograma de mu 1')
hist(MM[,2],freq=F,col="blue",border="white",xlab =
'Histograma de mu 2')
hist(MM[,3],freq=F,col="blue",border="white",xlab =
'Histograma de mu 3')
```

NR4.R

```
# Programa NR4.R
# En este programa se implementan
# El método de Newton-Rap, el muestreo de Gibbs,
# para tomar muestras de las densidades  posteriori
# relativas a la distribución NP4(mu,I)

#-----------------------------------------------------------------
# Constante de Normalización

kn<-function(bb)
        {
        drop(2+bb^2 + ((bb^3 + 3*bb)*pnorm(bb)/dnorm(bb)))
        }


Dbd<-function(t1,t2,t3,mu1,mu2,mu3,mu4)
         {
         trans<-c(cos(t1),cos(t2)*sin(t1),cos(t3)*sin(t2)*sin(t1)
         ,sin(t3)*sin(t2)*sin(t1))
         mu<-c(mu1,mu2,mu3,mu4)
         drop(crossprod(trans,mu))
         }


# Función para calcular la moda de frct
moda0<-function(bb)
        {
        moda<-(bb+sqrt((bb^2)+4*4))/2
        moda
        }

# Función de densidad condicional de r dado theta
frct<-function(r,bb)
         {
        ( (r^3)*exp( -0.5*( (r^2) - 2*bb*r)  )  )/kn(bb)
        }

# Función de densidad condicional acumulada de r dado theta.
Frct<-function(r,bb)
        {
        Frdt<-( ( (2+bb^2)-(2+bb^2+bb*r+r^2)*exp(bb*r-((r^2)/2)) ) + (bb^3+3*bb)*(1/dnorm(bb)
        Frdt
        }


# Método de Newton-Rap
NR<-function(g,f,tt1,tt2,tt3,mu1,mu2,mu3,mu4,tam.muestra,no.ite=2)
        {
  N<-tam.muestra
  ite<-no.ite
  bb<<-Dbd(tt1,tt2,tt3,mu1,mu2,mu3,mu4)
  r0<-moda0(bb)
  r.n<-r0
  u<-runif(N)
  for (i in 1:ite)
        {
        r.n<-r.n-((g(r.n,bb)-u)/(f(r.n,bb)))
        }
  rdt<-r.n
```

```
        }

###---------------- Muestro de Gibss ------------------------------
#Paqutes necesarios
library(MASS)
library(CircStats)
library(compositions)


#--------------Generar datos

v <- rnorm(50,2.9,1)
w <- rnorm(50,7,1)
x <- rnorm(50,2.8,1)
y <- rnorm(50,3.5,1)
# comando para volver un vector no-negativo en un vector
# composicional
z<-acomp(cbind(v,w,x,y))


t1<-(atan2(sqrt(z[,2]^2+z[,3]^2+z[,4]^2),z[,1]))%%(2*pi)
t2<-(atan2(sqrt(z[,3]^2+z[,4]^2),z[,2]))%%(pi)
t3<-(atan2(z[,4],z[,3]))%%(pi)
par(mfrow=c(1,3))
circ.plot(t1,stack=TRUE, bins=175, shrink=1.3)
circ.plot(t2,stack=TRUE, bins=175, shrink=1.3)
circ.plot(t3,stack=TRUE, bins=175, shrink=1.3)

theta<-cbind(t1,t2,t3)
n<-length(theta)/3

#write(c(t1,t2,t3), file="datos4.r",100)
#the<-scan(file="datos4.r")
#theta<- matrix(the,nrow=100,ncol=3)

#---------------termina de generar datos
datosx<-cbind(cos(theta[,1]),cos(theta[,2])*sin(theta[,1]),
cos(theta[,3])*sin(theta[,2])*sin(theta[,1]),sin(theta[,3]
)*sin(theta[,2])*sin(theta[,1]))

# Especificación de los valores para la distribución a priori

mu0<-c(0.0,0.0,0.0,0.0)
lambda0<-0.0001

# Valores iniciales de la variable latente

r<-rep(1,n)

# Número de iteraciones

  tm<-2000

# Saltos
  t.lag<-7
  kk<-tm*t.lag
  print(paste(" Total iterations =", kk, "..."))

# Matriz para obtener la muestra final

  MM<-matrix(0,tm,4)
```

```
# Periodo de calentamiento
burn<-10000

for(k in 1:(burn+kk))
        {

        x<-r*datosx
        # Sampling of vector mu.
        lambdaF<-(n+lambda0)
        mu1.e<-( n*mean(x[,1]) + lambda0*mu0[1] )/lambdaF
        mu2.e<-( n*mean(x[,2]) + lambda0*mu0[2] )/lambdaF
        mu3.e<-( n*mean(x[,3]) + lambda0*mu0[3] )/lambdaF
        mu4.e<-( n*mean(x[,4]) + lambda0*mu0[4] )/lambdaF
        desv<-sqrt( 1/lambdaF )

        mu.e<-c(rnorm(1,mu1.e,desv),rnorm(1,mu2.e,desv),
        rnorm(1,mu3.e,desv),rnorm(1,mu4.e,desv))

        # Sampling of vector r

        for(j in 1:n)
                {
                t.e1<-theta[j,1]
                t.e2<-theta[j,2]
                t.e3<-theta[j,3]
                b<-Dbd(t.e1,t.e2,t.e3,mu.e[1],mu.e[2],mu.e[3],
                mu.e[4])
                # Usando Newton Rapson.
                #r[j]<-ifelse( k<=burn,((2+b^2+(b^3+3*b
                )*(pnorm(b)*(dnorm(b)^(-1))))/(b+(b^2+1)*(
                pnorm(b)*(dnorm(b))^(-1)))),
                NR(Frct,frct,t.e1,t.e2,t.e3,mu.e[1],
                mu.e[2],mu.e[3],mu.e[4],5)   )
                r[j]<-NR(Frct,frct,t.e1,t.e2,t.e3,mu.e[1],
                mu.e[2],mu.e[3],mu.e[4],1)

}


#Values of each iteration
if(k>burn){
flag1<-(k/500)-trunc(k/500)
if(flag1==0){print(k-burn)}
flag2<-((k-burn)/t.lag)-trunc((k-burn)/t.lag)
if(flag2==0)
{
ii<-(k-burn)/t.lag
MM[ii,]<-mu.e
}
}

#####---------- Gibss completed ------------------------------
}



###------------------------------ Diagnostico-------------------

med.erg<-cbind(cumsum(MM[,1]),cumsum(MM[,2]),cumsum(MM[,3]),
cumsum(MM[,4]) )/(1:tm)
```

```
#win.graph()
par(mfrow=c(3,4))
plot(med.erg[,1],type="l",xlab="iteraciones",ylab="mu1")
plot(med.erg[,2],type="l",xlab="iteraciones",ylab="mu2")
plot(med.erg[,3],type="l",xlab="iteraciones",ylab="mu3")
plot(med.erg[,4],type="l",xlab="iteraciones",ylab="mu4")
#library(examples)
acf(MM[,1])
acf(MM[,2])
acf(MM[,3])
acf(MM[,4])
hist(MM[,1],freq=F)
hist(MM[,2],freq=F)
hist(MM[,3],freq=F)
hist(MM[,4],freq=F)
```

MatrizCD.R

```
# Programa para obtener la matriz de variación composicional
# y la matriz de variación direccional para datos
# composicionales.

library(MASS)
library(CircStats)
library(compositions)

#------función para sacar la varianza del log-cociente(x_i,x_j)
var.com <- function(a1,a2){
n<-length(a1)
varcom <- (sum((log(a1/a2))^2) - n*(med.com(a1,a2))^2)/(n-1)
varcom
}

#-------función para sacar la media direccional del ángulo_ij
med.dir <- function(a1,a2){
theta<-(atan2(a2,a1))%%(pi)
n<-length(theta)
s1 <-sum(cos(theta))/n
s2 <-sum(sin(theta))/n
R<-sqrt((s1)^2+(s2)^2)
c.a <- cbind(s1*R,s2*R)
m.at1<-(atan2(c.a[,2],c.a[,1]))
#V <- 1-R
#v<-sqrt(-2*log(R))
m.at1<-(atan2(c.a[,2],c.a[,1]))
m.at1*180/pi
}



#---------función para sacar la longitud media
# resultante R_ij y la varianza V=1-R
lon.dir <- function(a1,a2){
theta<-(atan2(a2,a1))%%(pi)
n<-length(theta)
s1 <-sum(cos(theta))/n
s2 <-sum(sin(theta))/n
R<-sqrt((s1)^2+(s2)^2)
c.a <- cbind(s1*R,s2*R)
m.at1<-(atan2(c.a[,2],c.a[,1]))
V <- 1-R
V
}

#--función para sacar la media del log-cociente(x_i,x_j)
med.com <- function(a1,a2){
n<-length(a1)
y<-sum(log(a1/a2))/n
return (y)
}

# compositional variation array
mc<-function(z){
com.var.ar<-matrix(NA,dim(z)[2],dim(z)[2])
for(i in 1:dim(z)[2]-1){
k<- i+1
for(j in k:dim(z)[2]){
```

```
com.var.ar[i,j]<-var.com(z[,i],z[,j])
com.var.ar[j,i]<-med.com(z[,i],z[,j])
}
}
com.var.ar
}


md<-function(z){
dir.var.ar<-matrix(NA,dim(z)[2],dim(z)[2])
for(i in 1:dim(z)[2]-1){
k<- i+1
for(j in k:dim(z)[2]){
dir.var.ar[i,j]<-lon.dir(z[,i],z[,j])
dir.var.ar[j,i]<-med.dir(z[,i],z[,j])
}
}
dir.var.ar
}

#-------Programa principal-----------

# leer datos
z<-acomp(read.csv("nuevosdatos.csv",header = TRUE, sep = ","))


m.c<-mc(z)
#matriz de variación composicional
m.c
m.d<-md(z)
# matriz de variación direccional para datos composicionales
m.d
```