

Aspectos geométricos de las poblaciones y los individuos estadísticos

Antonio Miñarro Alonso

ADVERTIMENT. La consulta d'aquesta tesi queda condicionada a l'acceptació de les següents condicions d'ús: La difusió d'aquesta tesi per mitjà del servei TDX (www.tdx.cat) ha estat autoritzada pels titulars dels drets de propietat intel·lectual únicament per a usos privats emmarcats en activitats d'investigació i docència. No s'autoritza la seva reproducció amb finalitats de lucre ni la seva difusió i posada a disposició des d'un lloc aliè al servei TDX. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX (framing). Aquesta reserva de drets afecta tant al resum de presentació de la tesi com als seus continguts. En la utilització o cita de parts de la tesi és obligat indicar el nom de la persona autora.

ADVERTENCIA. La consulta de esta tesis queda condicionada a la aceptación de las siguientes condiciones de uso: La difusión de esta tesis por medio del servicio TDR (www.tdx.cat) ha sido autorizada por los titulares de los derechos de propiedad intelectual únicamente para usos privados enmarcados en actividades de investigación y docencia. No se autoriza su reproducción con finalidades de lucro ni su difusión y puesta a disposición desde un sitio ajeno al servicio TDR. No se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR (framing). Esta reserva de derechos afecta tanto al resumen de presentación de la tesis como a sus contenidos. En la utilización o cita de partes de la tesis es obligado indicar el nombre de la persona autora.

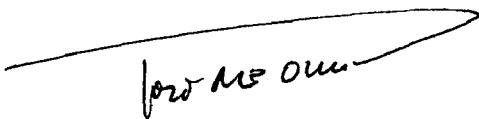
WARNING. On having consulted this thesis you're accepting the following use conditions: Spreading this thesis by the TDX (www.tdx.cat) service has been authorized by the titular of the intellectual property rights only for private uses placed in investigation and teaching activities. Reproduction with lucrative aims is not authorized neither its spreading and availability from a site foreign to the TDX service. Introducing its content in a window or frame foreign to the TDX service is not authorized (framing). This rights affect to the presentation summary of the thesis as well as to its contents. In the using or citation of parts of the thesis it's obliged to indicate the name of the author.

Aspectos geométricos de las poblaciones y los individuos estadísticos.

Memoria presentada para optar al
grado de Doctor en Biología
por la Universidad de Barcelona, por

Antonio Miñarro Alonso

Visto bueno
El director

A handwritten signature in black ink, reading "Josep Maria Oller Sala". The signature is written in a cursive style and is positioned below the text "Visto bueno El director".

Dr. Josep Maria Oller Sala
Profesor titular de Estadística.
Departamento de Estadística.
Universidad de Barcelona.

Barcelona, 21 de Enero de 1.991

*a mis padres
y a Lourdes*

Prólogo.

Cuando iniciamos la elaboración de la presente memoria, nuestro objetivo principal era el estudio probabilístico de una clase de funciones de densidad de probabilidad que combinara las propiedades de capacidad de adaptación a una gran variedad de datos y a un mismo tiempo que poseyera propiedades geométricas sencillas que facilitarían la aplicación sobre ellas de los métodos estadísticos basados en las técnicas de geometría diferencial. El primer contratiempo con el que tuvimos que enfrentarnos fue la aparición de problemas al aplicar sobre las mencionadas familias los métodos clásicos de estimación de parámetros, básicamente el método de la máxima verosimilitud. Los resultados obtenidos al aplicar de forma directa la estimación máximo verosímil eran insatisfactorios para nosotros. Dada la extraordinaria flexibilidad de las funciones empleadas, la misma filosofía de la máxima verosimilitud chocaba con el propósito de obtener estimaciones útiles a partir de una muestra de datos aleatorios, fuimos conscientes de la insuficiencia de la maximización de la verosimilitud de una muestra, sin ninguna otra consideración complementaria. Creíamos necesaria la búsqueda de alguna alternativa razonable al procedimiento utilizado hasta el momento. Tal alternativa podría consistir en una modificación no drástica, en el sentido de algunas tentativas ya efectuadas por algunos investigadores mediante la adición de alguna función penalizadora a la verosimilitud, o bien intentado complementar la idea de maximizar la verosimilitud con la de maximizar o minimizar alguna otra cantidad relacionada de forma que se eliminaran los inconvenientes presentados. Qué cantidad o cantidades deberían ser éstas es una cuestión a la que todavía no hemos encontrado una respuesta satisfactoria, sin embargo surgen nombres o ideas, como por ejemplo los de complejidad, estabilidad, incertidumbre, ... , que sin estar hasta el momento y para nosotros, definidos de una forma clara, estamos convencidos de que deben desempeñar un papel importante en un futuro no muy lejano.

También existía la posibilidad de desarrollar un método basado en consideraciones totalmente diferentes. La línea de trabajo que lleva nuestro equipo de investigación nos inducía a considerar una aproximación geométrica al problema de la estimación. Para tal fin nos propusimos realizar un estudio que significara una nueva visión geométrica de los modelos estadísticos, en particular de los individuos estadísticos y de las variables aleatorias. En gran medida intentamos la construcción de una geometría sobre el espacio muestral, que permitiera desarrollar bajo consideraciones meramente geométricas, las técnicas habituales de la inferencia estadística, como por ejemplo la estimación de parámetros, el contraste de hipótesis o las técnicas de clasificación de los individuos. Tal construcción permite ofrecer una visión global y unificadora de las diferentes técnicas. Fruto de los estudios anteriores ha sido el desarrollo de las distancias entre individuos y de sus posteriores aplicaciones discutidas en la presente memoria.

La distancia inmediata con su simplicidad conduce a la recuperación de gran parte de los resultados de la estadística clásica, la distancia estructural, más compleja en su formulación y manipulación, es quizá más sugerente de cara a futuras interpretaciones al suponer un mayor aprovechamiento de la información y de la estructura geométrica más íntima de los modelos estadísticos. Está todavía por dilucidar la importancia que las distancias entre individuos puedan tener en una posterior definición de los conceptos de complejidad o simplicidad.

Como una solución práctica y a un tiempo sencilla a los problemas de estimación surgidos al trabajar con familias flexibles, hemos propuesto el algoritmo stepwise de estimación no paramétrica, cuyos resultados dentro del campo de la estadística clásica pueden ser considerados satisfactorios al trabajar con las familias de funciones de densidad previamente mencionadas.

Toda la programación informática necesaria para realizar los diferentes cálculos numéricos ha sido realizada, salvo que se indique lo contrario, por el autor de la presente memoria en los lenguajes de programación FORTRAN 77 y PL/I, e implementada en el ordenador IBM 3090/200 del Centro de Informática de la Universidad de Barcelona.

Quiero expresar de forma sincera, mi profundo agradecimiento al Dr. Josep María Oller por su constante ayuda, dedicación y paciencia en la dirección de la presente memoria y en la de nuestro equipo de investigación.

También quiero agradecer la colaboración prestada en uno u otro momento por todos y cada uno de mis compañeros del Departamento de Estadística.

Finalmente y de forma muy especial la ayuda y el aliento prestados por Lourdes LLeonar que han colaborado a aumentar el sentido y el valor de todo el trabajo desarrollado.

Índice General

I. Introducción	1
1.1. Geometría en la estadística	2
1.2. Geometría diferencial en estadística.	6
1.3. Geometría diferencial en inferencia estadística.	9
II. Geometría de los Modelos Estadísticos.	13
2.1. Modelo estadístico.	14
2.2. Distancia entre modelos estadísticos.	18
2.3. Distancia entre funciones de densidad.	21
2.4. Métrica Informacional, Distancia de Rao.	25
2.5. Conexiones y curvatura.	37
III. Distancias entre individuos estadísticos.	42
3.1. Una representación del Espacio tangente de un modelo estadístico y su dual.	43
3.1.1. Representación del Espacio tangente.	43
3.1.2. Representación del Espacio tangente dual.	51
3.2. Distancia inmediata.	53
3.3. Algunos ejemplos de distancias.	60
3.3.1. Distribuciones uniparamétricas.	60
3.3.2. Distribuciones multiparamétricas.	66
3.4. Distancia estructural	74
IV. Aplicaciones de las Distancias entre individuos.	82
4.1. Estimación de Parámetros.	83
4.1.1. Procedimiento general.	83
4.1.2. Estimación utilizando la distancia inmediata.	84
4.1.3. Estimación utilizando la distancia estructural.	88

4.2. Contraste de Hipótesis.	92
4.2.1. Consideraciones generales.	92
4.2.2. Hipótesis nula simple.	93
4.3. Discriminación y Clasificación.	98
V. Sobre una clase de funciones de densidad de probabilidad	99
5.1. Construcción y consideraciones básicas.	100
5.2. Geometría informacional del modelo.	104
5.3. Algunos ejemplos de familias.	113
5.3.1. Distribuciones multinomiales.	113
5.3.2. Sistema de funciones de Hermite.	115
5.3.3. Sistema de funciones de Laguerre.	117
5.3.4. Sistema de funciones de Legendre.	118
5.4. Consideraciones finales.	120
VI. Estimación no paramétrica de la densidad.	122
6.1. Estimación no paramétrica de la densidad.	123
6.2. Estimación máximo verosímil.	125
6.2.1. Planteamiento.	125
6.2.2. Resultados generales.	126
6.2.3. Resultados con las familias del apartado 5.3.2.	128
6.3. Un algoritmo stepwise de estimación no paramétrica.	141
6.3.1. Desarrollo del algoritmo.	141
6.3.2. Resultados.	144
6.3.3. Comparación con otros métodos.	167
6.4. Estimación minimizando la esperanza del cuadrado de la distancia.	171
6.4.1. Consideraciones generales.	171
6.4.2. Una aproximación a la distancia riemanniana.	172
6.4.3. Algoritmo de minimización numérica.	175
6.4.4. Resultados.	177
VII. Aplicaciones a la Biología.	183
7.1. Estimación del tamaño en <i>Octopus vulgaris</i>	184
7.1.1. Introducción.	184
7.1.2. Resultados.	186

7.2. Clasificación de patrones electroforéticos.	189
7.2.1. Introducción.	189
7.2.2. Resultados.	193
7.2.3. Discusión.	194
VIII. Resumen de Resultados.	198
Resumen de Resultados.	199
Bibliografía	203

I. Introducción

A modo de introducción efectuamos un pequeño recorrido por la historia de la aplicación de técnicas y conceptos geométricos en la estadística, con especial énfasis en las técnicas de geometría diferencial aplicadas a la obtención de distancias entre distribuciones y al estudio de la inferencia estadística.

- 1.1. Geometría en la Estadística.
- 1.2. Geometría diferencial en Estadística.
- 1.3. Geometría diferencial en inferencia estadística.

1.1. Geometría en la estadística

La utilización de técnicas y conceptos geométricos en estadística no es un fenómeno reciente. La representación de datos multidimensionales como puntos de un espacio euclídeo n -dimensional ha facilitado una comprensión intuitiva de las relaciones y propiedades estadísticas entre los mismos. Gran cantidad de conceptos habituales en estadística tienen su origen inmediato en conceptos geométricos: rectas y curvas de regresión, modelos lineales, distancia o curvatura, son sólo algunos de ellos.

Uno de los trabajos pioneros en el desarrollo de una formalización geométrica de la estadística se lo debemos a K. Pearson (1857-1936). En 1890 fue invitado a dar un ciclo de cuatro conferencias sobre geometría en el Gresham College, con el título global de "Ambito y Conceptos de la Ciencia Moderna". A su visión del uso de la representación geométrica se deben principalmente sus descubrimientos en el campo de la correlación, donde destacan la teoría general de correlación para tres variables, los coeficientes de las ecuaciones de regresión múltiple, o el coeficiente de correlación conocido como "r" de Pearson, Pearson (1896).

Otro gran nombre entre los pioneros de la geometría en la estadística es el de R.A. Fisher (1890-1962), quien desde sus primeros trabajos Fisher (1912,1913,1915) tiene a la geometría presente en ellos. Destaquemos entre sus contribuciones el criterio para ajustar curvas de frecuencia, Fisher(1912), o la obtención de la distribución del coeficiente de correlación, Fisher(1915). Posteriormente en varios trabajos se interesa por el estudio de la inferencia estadística, destacando la introducción de la estimación máximo verosímil así como una profundización en la teoría de la estimación donde aparecen los conceptos de eficiencia, suficiencia, cantidad de información en una observación y pérdida de información al utilizar un estadístico. Define como

estadísticos suficientes aquellos que poseen toda la información relevante contenida en los datos Fisher (1925). También introduce el concepto de estadístico auxiliar con la intención de recuperar parte de la información perdida al utilizar un estadístico. Tales conceptos, tan solo apuntados por Fisher, como veremos posteriormente tendrán gran importancia en el desarrollo posterior de una teoría geométrica de la estadística.

Un cambio cualitativo importante se produjo en el momento en que la geometría no solamente proporcionó una herramienta importante, sino que los conceptos geométricos entraron a formar parte plenamente de la estadística. Este cambio puede estar representado por la utilización en estadística del concepto geométrico de distancia. Uno de los objetivos básicos de la estadística es la realización de comparaciones entre diferentes objetos en base a la información que se posee sobre los mismos. Una forma natural e intuitiva de efectuar dicha comparación es a través de la definición de una distancia. Es muy amplio el abanico de posibilidades que se nos abre para aprovecharla con vistas a efectuar análisis estadísticos, por ejemplo, clasificaciones entre los objetos, realizar análisis discriminantes, representaciones gráficas de los mismos en dimensiones reducidas o incluso utilizarla para definir a través de ella procesos de inferencia estadística como puede ser la estimación de parámetros, realización de contrastes de hipótesis, etc.

La distancia euclídea habitual entre dos puntos de un espacio n -dimensional ha sido utilizada desde principios de siglo, Pearson (1901), Hotelling (1933). Sin embargo presenta como inconvenientes importantes el hecho de cambiar de forma no monótona si efectuamos un cambio de escala en las medidas de las variables, o también que ignora el hecho de que las variables aleatorias sean o no estocásticamente independientes. Para intentar paliar estos defectos pronto surgieron algunas alternativas, destaquemos entre ellas el coeficiente de semejanza racial de Pearson, Pearson (1926), que tiene la propiedad de resultar invariante frente a cambios de

escala, pero no frente a transformaciones lineales en general, también ignora el hecho de que las variables aleatorias sean estocásticamente dependientes o no. Posteriormente se definió la distancia de Mahalanobis, Mahalanobis (1936), construida entre poblaciones estadísticas donde tenemos definidas n variables aleatorias que se distribuyen según una distribución normal conjunta, con la hipótesis adicional de que todas las poblaciones poseen la misma matriz de covarianzas. La distancia de Mahalanobis presenta la ventaja de resultar invariante frente a transformaciones lineales biyectivas, en particular frente a cambios de escala, asocia además el concepto de ortogonalidad con el de independencia estocástica. La distancia de Mahalanobis ha jugado un destacado papel en muchas técnicas estadísticas, destacando el Análisis Canónico de Poblaciones o el Análisis Discriminante. Bose (1936) calculó la distribución exacta de la distancia de Mahalanobis. Posteriormente se han propuesto extensiones a otras distribuciones diferentes de la normal, sin embargo se pierden parte de las propiedades de la misma.

Muchas otras distancias han sido propuestas, sin ánimo de hacer un recuento exhaustivo citemos entre otras a la distancia de Bhattacharyya (1946) entre poblaciones en las cuales observamos n variables aleatorias que siguen una distribución multinomial de n sucesos mutuamente excluyentes. Bhattacharyya asocia a cada población un vector del espacio euclídeo n -dimensional de forma que los cosenos directores al cuadrado coinciden con los parámetros que definen la distribución multinomial asociada a cada población. La distancia equivale al ángulo formado entre los vectores asociados a dos poblaciones, o también a la distancia inducida por la distancia euclídea sobre una hiperesfera de radio 2 entre los dos puntos que definen los dos vectores al cortar la superficie de la hiperesfera. Rao (1949) propuso una extensión a la distancia de Bhattacharyya para el caso de distribuciones continuas.

I - Introducción

Otra distancia a mencionar puede ser la distancia ji-cuadrado, utilizada en el Análisis Factorial de Correspondencias, Benzecri (1976). Y finalmente, por sólo citar algunas de las más destacadas o útiles en algunos campos concretos, tenemos Matusita (1964), Cavalli-Sforza (1969), Prevosti et al. (1975), Nei (1978), etc.

1.2. Geometría diferencial en estadística.

Una nueva visión se produjo con la introducción de las técnicas de geometría diferencial. La geometría diferencial, cuyos orígenes se remontan a los trabajos de Newton y Leibnitz, quienes establecieron los principios del Cálculo infinitesimal, experimentó un gran avance a partir de los trabajos de Riemann, Levi-Civita, Ricci y Christoffel, quienes entre otros, proporcionaron un enfoque general y dotaron a la geometría diferencial de las herramientas necesarias para su expansión. Los trabajos de Einstein sobre el Principio General de la Relatividad supusieron un gran desarrollo y proporcionaron una nueva visión sobre la utilización de geometrías no euclídeas en la realidad natural.

Rao (1945) fue el pionero en utilizar un enfoque geométrico diferencial en la construcción de una distancia entre modelos estadísticos. El método desarrollado por Rao consiste en la introducción de una métrica Riemanniana en la variedad paramétrica asociada a los parámetros de las funciones de densidad. Para ello utiliza como matriz de la métrica, la matriz de información de Fisher. Siguiendo las técnicas habituales de la geometría Riemanniana, es posible la obtención de la distancia geodésica entre dos puntos cualesquiera de la variedad, representación de diferentes modelos estadísticos. En su trabajo, Rao, notablemente influido por los descubrimientos de Mahalanobis y Bhattacharyya, estudia la construcción de distancias entre distribuciones normales univariantes y multinomiales. La distancia de Mahalanobis aparece como un caso particular de la distancia de Rao en el caso de distanciar poblaciones normales multivariantes con matriz de covarianzas común. La distancia de Rao resulta invariante frente a transformaciones no singulares tanto de las variables como de los parámetros.

I - Introducción

En la misma línea que Rao y por la misma época destacamos los trabajos de Jeffreys (1946), también en la línea de introducir una métrica Riemanniana a través de la matriz de información de Fisher, definiendo distribuciones a priori no informativas e invariantes.

Las dificultades matemáticas inherentes a la aplicación del método desarrollado por Rao fueron lo suficientemente importantes para detener el desarrollo de la obtención de nuevas distancias entre diferentes distribuciones durante bastante tiempo.

Un esfuerzo encaminado a la obtención de la distancia de Rao para un gran grupo de distribuciones de probabilidad, al menos las más conocidas, no se llevó a cabo hasta años después con los trabajos de Atkinson y Mitchell (1981), influidos por los trabajos de Efron y Amari, entre otros. Que redescubriendo la aplicación del enfoque geométrico diferencial en estadística, obtuvieron la distancia de Rao para algunas de las distribuciones uniparamétricas más conocidas, Poisson, binomial, exponencial, gamma, normal univariante con media conocida, normal univariante con varianza conocida; para la distribución normal univariante, la normal multivariante con matriz de covarianzas conocida, la normal multivariante con vector de medias conocido, y la multinomial, coincidiendo en este último caso con la distancia de Hellinger-Bhattacharyya. Posteriormente otros autores han ampliado la lista de distribuciones para las cuales se posee la distancia de Rao, Oller y Cuadras (1985) la obtienen para la multinomial negativa, Rios y Cuadras (1986) para los modelos lineales normales, Oller (1987) para la logística y valores extremos. En cuanto a la distribución normal multivariante todavía no ha sido posible obtener la expresión explícita de la distancia pero destaquemos los estudios realizados sobre la geometría del modelo normal multivariante debidos a Skovgaard (1981,1984), Oller (1982), Sato et al. (1979) en el modelo bivalente, Burbea y Oller (1988) estudiando los modelos lineales

I - Introducción

elípticos, Eriksen (1986) y Calvo (1988) obteniendo las ecuaciones de las geodésicas así como acotando la distancia de Rao para el modelo normal multivariante.

Otros estudios se han encaminado a obtener un método que proporcione una aproximación a la distancia de Rao útil para aquellos casos en que no se posea una expresión explícita de la distancia, Miñarro (1985), Miñarro y Oller (1990), Calvo y Oller (1990).

1.3. Geometría diferencial en inferencia estadística.

Trabajos posteriores de Rao se encaminaron al estudio de la estimación, clarificando algunos de los conceptos previamente introducidos por Fisher. En particular Rao (1961,1962) trata con los conceptos de eficiencia y eficiencia de segundo orden, conceptos relacionados con la recuperación de la información perdida al trabajar con estadísticos en lugar de con la muestra real utilizando derivadas superiores del logaritmo de la función de densidad. Fisher había apuntado tales conceptos, pero el desarrollo de Rao supuso una clarificación considerable. En Rao (1961,1962) podemos encontrar ejemplos relacionados con la pérdida de información, calculando explícitamente su valor en la distribución multinomial para diferentes métodos de estimación.

Varios autores han tratado a partir de entonces de construir una teoría geométrica de la estadística con aplicación a la inferencia. En 1959 en notas no publicadas, Amari señala que el espacio paramétrico correspondiente a la distribución normal univariante es un espacio paramétrico de curvatura constante y negativa. En la misma línea podemos destacar los siguientes trabajos, Amari (1968), Yoshizawa (1971), Takiyama (1974) o Sato et al. (1979).

Un nuevo concepto asociado al estudio de las variedades estadísticas fue introducido por el soviético Chentsov (1972), donde introduce una familia de conexiones afines en una variedad estadística. En los trabajos realizados hasta el momento tan solo se había utilizado la conexión de Levi-Civita. También probó que la información de Fisher y las conexiones afines son únicas en la variedad de distribuciones de probabilidad sobre un número finito de átomos. Con su teoría contribuyó a elucidar las estructuras geométricas de la familia exponencial. Sin

embargo en su trabajo no tuvo en cuenta el concepto de curvatura de una variedad, cosa que si hizo en un trabajo posterior de gran importancia Efron, Efron (1975).

A grandes rasgos el trabajo de Efron es una cuantificación de lo diferente que es un modelo estadístico de la familia exponencial, mediante la definición del concepto de curvatura de un modelo. Se señala que la curvatura juega un importante papel en la teoría asintótica de la inferencia estadística y está íntimamente relacionada con la teoría de la eficiencia de segundo orden de Rao y Fisher. Según la definición de Efron, la curvatura es idénticamente cero si la familia es exponencial, o mayor que cero en caso contrario. El objetivo también es mostrar que familias que poseen una pequeña curvatura gozan de las buenas propiedades estadísticas que son reconocidas para la familia exponencial, por ejemplo que el test localmente de potencia máxima es uniformemente de potencia máxima, que los estimadores máximo verosímiles son estadísticos suficientes o que se alcanza la cota de Cramer-Rao si hemos escogido la adecuada función para estimar. Efron consideró subfamilias uniparamétricas de familias exponenciales, a las que denominó familias exponenciales curvadas, notando además que cualquier familia con propiedades regulares puede ser localmente aproximada por una familia exponencial curva. Demostró para estas familias exponenciales curvadas que el estimador máximo verosímil minimiza la pérdida de información, interpretando esta pérdida como la curvatura de la subfamilia.

En el trabajo de Efron se usa el espacio paramétrico natural con el producto escalar definido por la matriz de información de Fisher en un punto p . Siendo necesaria la construcción de un nuevo espacio con otro producto escalar para estudiar la curvatura en un punto diferente. Para relacionar ambos espacios debemos utilizar una conexión afín. Dawid (1975) señala en comentarios al artículo de Efron que aunque Efron había usado la métrica Riemanniana definida por la matriz de información de Fisher, había utilizado, aunque no de modo explícito, una conexión afín no

Riemanniana, que pasó a recibir el nombre de conexión exponencial. Dawid también sugirió en su trabajo la posibilidad de introducir otra conexión afin que denominó conexión mezcla ("mixture connexion"). Los nombres de las conexiones provienen del hecho de que geodésicas con respecto a la conexión exponencial forman familias exponenciales, mientras que geodésicas con respecto a la conexión mezcla forman familias de mezclas. Otros trabajos generalizando estas ideas son Reads (1975) o Madsen (1979).

Amari (1980,1982) continúa en la línea del estudio de propiedades del modelo estadístico en escala no local mediante la introducción de conexiones afines entre diferentes espacios tangentes en puntos vecinos. Su trabajo se encamina a estudiar propiedades asintóticas de ordenes superiores. Introduce una familia de conexiones denominadas α -conexiones, donde para un valor de α igual a 1 obtenemos la conexión exponencial, y para α igual a -1 la conexión mezcla. Esta familia coincide con la introducida por Chentsov para la distribución multinomial. A partir de las α -conexiones tomando dos distribuciones de probabilidad pertenecientes a una familia de curvatura cero con respecto a la α -conexión, es posible definir una medida de divergencia denominada α -divergencia, la distancia de Hellinger, la información de Kullback-Leibler (1951) y otras muchas medidas de divergencia pertenecen a la clase de α -divergencias. La definición de las α -divergencias proporciona un nuevo método de estimación, según comenta Kass (1987) "El elemento de una familia exponencial curva que minimiza la α -divergencia desde un punto en el espacio paramétrico de la familia exponencial puede ser encontrado siguiendo la α -geodésica que contiene dicho punto y es perpendicular a la familia curva". Esto genera una nueva clase de estimadores de mínima α -divergencia, donde el estimador máximo verosímil sería el estimador de mínima -1-divergencia.

En otra dirección hemos de destacar los trabajos encaminados al estudio de técnicas inferenciales utilizando la distancia de Rao. A partir de la distancia es posible la construcción de contrastes de hipótesis, por ejemplo podemos contrastar la existencia de diferencias significativas entre dos poblaciones mediante un test sobre su distancia, considerando que no existen diferencias si podemos aceptar que su distancia es nula. Un primer planteamiento sobre este tema puede encontrarse en Oller (1982) y algunos de los trabajos más relevantes en este campo son Ríos y Cuadras (1986) con los modelos lineales normales, Burbea y Oller (1988) con los modelos lineales elípticos y Burbea y Oller (1989) obteniendo la distribución asintótica y planteando contrastes para ciertas familias de funciones de densidad.

II. Geometría de los Modelos Estadísticos.

En este capítulo presentamos una vez introducido el concepto de modelo estadístico, varias consideraciones sobre la introducción de distancias entre modelos basándonos en representaciones geométricas de los mismos. Introducimos mediante diversos enfoques la métrica informacional resumiendo sus propiedades así como también indicamos la importancia estadística de otros conceptos geométricos como la curvatura.

- 2.1. Modelo estadístico.
- 2.2. Distancia entre modelos estadísticos.
- 2.3. Distancia entre funciones de densidad.
- 2.4. Métrica informacional. Distancia de Rao.
- 2.5. Conexiones y curvatura.

2.1. Modelo estadístico.

Gran parte de los estudios estadísticos se llevan a cabo sobre observaciones efectuadas al realizar un experimento. Para manejar de forma efectiva tales observaciones, se hace necesaria una formalización matemática del proceso de observación. El concepto de modelo estadístico se ha convertido en la más útil herramienta para llevar a cabo tal propósito.

El primer paso consiste en la fijación de un espacio medible, mediante la construcción de una σ -álgebra de conjuntos sobre el espacio muestral asociado al experimento. Notaremos a dicho espacio medible por (χ, \mathcal{A}) . El álgebra de conjuntos es construida como representación de un álgebra de sucesos, donde los elementos de esta última son los denominados sucesos observables, condicionados por la capacidad de observación del experimentador y definidos como aquellos enunciados referentes a un experimento, de los cuales podemos afirmar si se han cumplido o no al efectuarlo. La representación se logra haciendo corresponder a cada suceso el conjunto de resultados posibles de la experiencia que verifican el enunciado que caracteriza al suceso mencionado.

Una vez construido el espacio medible efectuaremos la introducción de una medida sobre él. En el contexto de la probabilidad, la medida que se introduce busca cuantificar el grado de expectabilidad de cada uno de los sucesos del álgebra. Dicha medida, a la que denominaremos probabilidad, P , genera junto con el espacio medible una nueva estructura que denominaremos **Espacio de Probabilidad**, y que notaremos por (χ, \mathcal{A}, P) .

Sobre un mismo espacio medible es posible definir diferentes medidas de probabilidad, es decir, manteniendo inalterables los sucesos pero modificando el grado de expectabilidad de los mismos. Esto nos lleva a la existencia de una gran cantidad

de espacios de probabilidad que pueden reflejar un experimento. El investigador puede tener razones para restringir a un cierto subconjunto de los posibles espacios de probabilidad, aquellos que permiten reflejar los resultados de su experiencia. Definiremos de un modo general a un **Modelo estadístico** sobre un espacio medible (X, A) , como una familia de espacios de probabilidad, familia que generalmente es construida de forma que permita describir nuestro experimento.

$$M = \{(X, A, P) : P \in P_M\}$$

Con frecuencia surge la necesidad de efectuar comparaciones entre los diferentes espacios de probabilidad con vistas a efectuar diversos tipos de tratamientos estadísticos: procesos de inferencia, estudio de afinidades entre poblaciones, etc. Tal y como hemos señalado en el capítulo precedente, un buen número de investigadores se han inclinado por las técnicas geométricas para afrontar tales problemas, la filosofía de nuestro presente trabajo es seguir en esa misma línea, profundizando en las técnicas y formalizaciones geométricas relacionadas con la teoría de la probabilidad. Será por tanto importante poseer una representación geométrica adecuada de los modelos estadísticos. Dicha representación nos permitirá introducir, de forma natural, conceptos geométricos que nos serán de gran ayuda para implementar o interpretar técnicas propiamente estadísticas.

Entre los conceptos geométricos que consideramos más importantes se encuentra el de distancia, en los apartados siguientes vamos a estudiar la construcción de distancias entre espacios de probabilidad.

Previamente a considerar cualquier definición de distancia entre espacios de probabilidad vamos a citar, basándonos en consideraciones expuestas en Mahalanobis (1936), Rao (1945, 1949), Oller (1982) y Oller y Cuadras (1985,b), Oller (1989), algunas de las propiedades que creemos aconsejables que debe cumplir una distancia.

1. Debe estar relacionada con el concepto de información, puesto que la distancia entre varios objetos se basa en la información que poseemos sobre las analogías y diferencias entre ellos.

2. La distancia debe aumentar al aumentar la información accesible sobre los objetos, por ejemplo al aumentar el número de variables estudiadas en un experimento.

3. Debe poseer propiedades de invariancia frente a determinadas transformaciones de los espacios de probabilidad, en concreto frente a aquellas que mantengan invariante la cantidad de información. Véase al respecto Fisher (1925) y Kullback y Leibler (1951). En general, tal y como señala Oller (1989), es conveniente que la distancia entre espacios de probabilidad transformados mediante una función medible T sea menor que la distancia entre los espacios de probabilidad sin transformar, excepto si T es un estadístico suficiente, en cuyo caso la distancia debe ser invariante, y T puede ser llamada entonces, una transformación admisible.

4. Debe permitir una relación entre los conceptos de ortogonalidad e independencia estocástica. En particular si $E_1 = \{\chi_1, A_1, P_1\}$ y $E_2 = \{\chi_2, A_2, P_2\}$ son dos espacios de probabilidad, y construimos el espacio conjunto suponiendo independencia $E_1 \times E_2 = \{\chi_1 \times \chi_2, A_1 \otimes A_2, P_1 \times P_2\}$, la distancia en este último sería conveniente que fuera la suma $d^2(E_1 \times E_2, F_1 \times F_2) = d_1^2(E_1, F_1) + d_2^2(E_2, F_2)$.

5. Debe estar relacionada con algunos de los estadísticos clásicos conocidos.

Tal y como hemos señalado anteriormente, la diferencia entre los espacios de probabilidad pertenecientes a un modelo radica únicamente en la medida de

II - Geometría de los Modelos Estadísticos

probabilidad que se asigna a los sucesos del álgebra, manteniéndose inalterable el espacio de medida. Por tanto, si existe la posibilidad de introducir una distancia entre las diferentes medidas de probabilidad, por extensión, podemos considerar como distancia entre dos espacios de probabilidad aquella que existe entre sus respectivas medidas probabilísticas asociadas.

2.2. Distancia entre modelos estadísticos.

Una primera posibilidad para establecer una distancia entre medidas de probabilidad, consiste en trabajar directamente sobre las medidas.

Sea (χ, A) , un espacio medible sobre el que están definidas dos medidas probabilísticas, P_1 y P_2 . Definamos la distancia entre ambas medidas como:

$$d(P_1, P_2) = \sup_{a \in A} |P_1(a) - P_2(a)|$$

Una fórmula análoga a la expuesta anteriormente ha sido utilizada en trabajos de mecánica cuántica para definir una distancia entre posibles estados de un sistema físico. La equivalencia entre la formulación de la teoría de la probabilidad a partir de la teoría de la medida debida a Kolmogorov, con la formulación de von Neumann de la teoría cuántica, donde los estados de un sistema son equiparados a medidas de probabilidad, puede encontrarse por ejemplo en Mackey (1963) y Jauch (1968). La definición de la distancia entre estados puede verse en Misra (1974) y Hadjisavvas (1981).

Efectivamente, el resultado se puede considerar una distancia establecida directamente entre medidas. Es posible sin embargo, desarrollar otros métodos que permiten utilizar propiedades geométricas adicionales, en particular el concepto de ortogonalidad.

Una forma diferente de establecer distancias entre medidas de probabilidad la desarrollamos mediante una representación funcional del modelo estadístico. Trasladar el problema de distanciar medidas a distanciar funciones nos proporciona una mayor

comodidad de manejo matemático, y como veremos, nos permitirá una representación geométrica con buenas propiedades.

Hemos de destacar que no siempre existe una representación funcional de todo el modelo estadístico. Será factible dicha representación si existe una medida de referencia que domine a todas las medidas del modelo. Es decir si existe una medida μ tal que para cualquier otra medida P , $P(a) = 0$ para todo $a \in A$ para el que $\mu(a) = 0$. Sin embargo, aun cuando no exista dicha medida de referencia, siempre es posible encontrar una medida de referencia que dadas dos medidas, las domine a ambas, por ejemplo dadas dos medidas P_1 y P_2 , siempre podemos utilizar como medida que domina a ambas: $\frac{1}{2}(P_1 + P_2)$.

Si existe una medida global de referencia para todo el modelo, generalmente será elegida de acuerdo a la naturaleza del espacio muestral y del modelo estadístico. Por ejemplo podemos utilizar la medida de Lebesgue si el espacio muestral es \mathbf{R}^n , o bien una medida discreta sobre los conjuntos del espacio muestral si éste es discreto.

En ambos supuestos, tanto si existe una medida global, como si hemos de tomar una medida diferente para cada par de espacios de probabilidad, la representación funcional se basa en el procedimiento que describimos en el siguiente párrafo.

Sea μ una medida positiva acotada sobre la σ -álgebra A en χ , tal que domine a todas las medidas probabilísticas del modelo $P \ll \mu$ $P \in P_M$. Por el teorema de Radon-Nikodym, Halmos (1950), podemos representar las medidas del modelo por una familia de funciones medibles $D^\lambda \subset F$ a través de la aplicación:

$$\Phi_\lambda : P_M \rightarrow D^\lambda, \quad \Phi_\lambda(P) = \lambda \left(\frac{dP}{d\mu} \right)$$

donde λ es una función real C^∞ monótona estricta y F el conjunto de todas las funciones medibles sobre (χ, A) .

Denominaremos a D^λ la λ -representación del modelo estadístico. Podemos destacar los casos siguientes:

- $\lambda(x) = x$, donde D^λ representará una familia de funciones de densidad $p(x) \in \mathcal{L}^1(\mu)$, $0 \leq p(x) < \infty$, y dado que $D^\lambda \subset \mathcal{L}^1(\mu)$ llamaremos a $D^\lambda \equiv D^1$ la 1-representación del modelo estadístico.

- $\lambda(x) = 2\sqrt{x}$ en cuyo caso $D^\lambda \subset \mathcal{L}^2(\mu)$ y podemos llamar a $D^\lambda \equiv D^2$ la 2-representación del modelo estadístico.

- $\lambda(x) = \log x$, obteniendo la que denominaremos l -representación del modelo estadístico.

Cada espacio de probabilidad del modelo es asociado mediante Φ_λ a una función, introduciendo una distancia entre las funciones, por extensión la podemos considerar como la distancia entre las respectivas medidas probabilísticas asociadas.

2.3. Distancia entre funciones de densidad.

Vamos a centrarnos en el presente apartado en el estudio de una distancia entre funciones de densidad. Como hemos visto en el apartado anterior, las funciones de densidad pertenecen al espacio $\mathcal{L}^1(\mu)$. Un procedimiento para distanciar dos funciones consistirá en aplicar dichas funciones sobre $\mathcal{L}^2(\mu)$. Sea $\varphi: \mathbb{R}^+ \rightarrow \mathbb{R}$, tal que nos permite definir la aplicación de $\mathcal{L}^1(\mu)$ en $\mathcal{L}^2(\mu)$, del modo siguiente:

$$\begin{aligned} \psi: \mathcal{L}^1(\mu) &\rightarrow \mathcal{L}^2(\mu) \\ p &\rightarrow \psi(p) = \varphi \circ p \end{aligned} \quad [2.2]$$

Si no existe una medida de referencia global, μ varía para cada par de funciones que consideremos, los espacios $\mathcal{L}^2(\mu)$ también variarían, sin embargo cada par de funciones pueden ser aplicadas a un mismo espacio y distanciadas allí. Si existe una medida de referencia global, todas las funciones de densidad del modelo se pueden aplicar sobre el mismo espacio $\mathcal{L}^2(\mu)$ y podemos considerar nuestra acción como una aplicación global del espacio de funciones de densidad en el espacio de Hilbert $\mathcal{L}^2(\mu)$.

En $\mathcal{L}^2(\mu)$ se encuentran definidos los conceptos de producto escalar y ortogonalidad, y la distancia entre dos funciones p y q la obtendremos como la distancia euclídea habitual en \mathcal{L}^2

$$d^2(p, q) = d_{\mathcal{L}^2}^2(\psi(p), \psi(q)) = \int_{\mathcal{X}} (\varphi \circ p - \varphi \circ q)^2 d\mu$$

En este momento vamos a añadir una condición adicional a las ya mencionadas como deseables para las distancias. Una distancia debería ser invariante frente a cambios de la medida de referencia utilizada para construir las funciones de densidad. Con este nuevo supuesto, vamos a definir cual debe ser la función φ , para que la distancia sea invariante.

Proposición 2.1. Para cualquier espacio medible (χ, A) y para cualquier par de funciones de densidad p, q , la distancia natural de \mathcal{L}^2 entre funciones de densidad transformadas por [2.2] resulta invariante frente a cambios de la medida de referencia, si y solo si:

$$\varphi(x) = \alpha\sqrt{x} + \beta \quad x \geq 0$$

Es decir trabajando con la que hemos denominado 2-representación del modelo estadístico.

Demostración. Efectuemos un cambio en la medida de referencia y pasemos a considerar una nueva medida ν , tal que $\mu \ll \nu$. La nueva medida induce un cambio en las funciones de densidad

$$p = \frac{dP}{d\mu} \rightarrow \bar{p} = \frac{dP}{d\nu}$$

$$\bar{p} = \frac{dP}{d\nu} = \frac{dP}{d\mu} \frac{d\mu}{d\nu} = ph$$

donde $h \in \mathcal{L}^1(\nu)$, entonces, la invariancia en la distancia se logra si se cumple la siguiente igualdad:

$$\int_{\chi} \{\varphi(p(x)) - \varphi(q(x))\}^2 d\mu(x) = \int_{\chi} \{\varphi(p(x)h(x)) - \varphi(q(x)h(x))\}^2 d\nu(x) \quad [2.3]$$

Trivialmente podemos demostrar la condición de suficiencia.

$$\begin{aligned} \int_{\chi} (\alpha\sqrt{p(x)h(x)} + \beta - \alpha\sqrt{q(x)h(x)} - \beta)^2 d\nu(x) &= \int_{\chi} \alpha^2 (\sqrt{p(x)} - \sqrt{q(x)})^2 h(x) d\nu(x) = \\ \int_{\chi} \alpha^2 (\sqrt{p(x)} - \sqrt{q(x)})^2 d\mu(x) &= \int_{\chi} (\varphi(p(x)) - \varphi(q(x)))^2 d\mu(x) \end{aligned}$$

Vamos seguidamente a comprobar la condición necesaria.

Sea un espacio medible (χ, A) determinado tal que podamos definir dos conjuntos compactos M y N , pertenecientes al álgebra A tales que:

$$M \cap N = \emptyset \quad \mu(M) = m \quad \mu(N) = n$$

Dada la arbitrariedad de las funciones de densidad y de la función h definamos:

$$p(x) = \begin{cases} m^{-1} & \text{para } x \in M \\ 0 & \text{para } x \notin M \end{cases}$$

$$q(x) = \begin{cases} n^{-1} & \text{para } x \in N \\ 0 & \text{para } x \notin N \end{cases}$$

$$h(x) = \lambda \in \mathbf{R}^+ \quad \text{para todo } x \in \chi$$

donde p , q y h tienen el mismo significado que anteriormente.

De [2.3] se sigue

$$\begin{aligned} & \int_M (\varphi(m^{-1}) - \varphi(0))^2 d\mu + \int_N (\varphi(0) - \varphi(n^{-1}))^2 d\mu = \\ & = \int_M (\varphi(\lambda m^{-1}) - \varphi(0))^2 dv + \int_N (\varphi(0) - \varphi(\lambda n^{-1}))^2 dv \end{aligned}$$

Impongamos ahora la condición $m = n$

$$\lambda(\varphi(m^{-1}) - \varphi(0))^2 = (\varphi(\lambda m^{-1}) - \varphi(0))^2$$

$$\sqrt{\lambda} \varphi(m^{-1}) + \varphi(0)(1 - \sqrt{\lambda}) = \varphi(\lambda m^{-1})$$

Tomemos $m = 1$ y llamemos $\alpha = \varphi(1) - \varphi(0)$ y $\beta = \varphi(0)$.

$$\alpha\sqrt{\lambda} + \beta = \varphi(\lambda)$$

y por lo tanto, al demostrar la necesidad de la transformación $\varphi(x) = \alpha\sqrt{x} + \beta$ para ciertos espacios medibles, y también su suficiencia, queda demostrado el enunciado de la proposición ■.

La distancia resulta ser finalmente:

$$d^2(p, q) = \alpha^2 \int_X (\sqrt{p} - \sqrt{q})^2 d\mu = \alpha^2 (2 - \int_X \sqrt{pq} d\mu)$$

Conocida para una determinada constante de proporcionalidad, como **Distancia de Hellinger**.

Debido a las restricciones a que están sujetas las funciones de densidad $p, q \in \mathcal{L}^1(\mu)$:

$$p(x), q(x) \in \mathbf{R}^+$$

$$\int_X p(x) d\mu(x) = \int_X q(x) d\mu(x) = 1$$

el conjunto de densidades no es aplicado sobre todo el espacio $\mathcal{L}^2(\mu)$, sino sobre una porción de una superficie esférica de radio α en \mathcal{L}^2 . Para calcular la distancia entre dos densidades, podemos utilizar, no la distancia euclídea que corresponde a la métrica global de $\mathcal{L}^2(\mu)$, sino la métrica inducida sobre la superficie esférica. La curva que conecta ambas densidades y hace mínima su distancia corresponde a un arco de un círculo que pase por ellas. La distancia sobre este círculo en la esfera es igual a:

$$d^2(p, q) = \alpha \cos^{-1} \left(\int_X \sqrt{pq} d\mu \right)$$

Conocida como **Distancia de Hellinger-Bhattacharyya**, ver entre otros, Bhattacharyya (1943), Rao (1949), Dawid (1977) y Burbea (1986).

2.4. Métrica Informacional. Distancia de Rao.

Sea μ una medida aditiva σ -finita sobre la σ -álgebra de los subconjuntos de χ . Designemos por $\mathcal{L}_+^1(\mu)$ al espacio de funciones μ -medibles p sobre χ , tales que $p(x) \in \mathbf{R}^+$ para casi todo $x \in \chi$ respecto μ y $\|p\|_\mu = \int_\chi |p| d\mu < \infty$. Finalmente sea D el subconjunto convexo de $\mathcal{L}_+^1(\mu)$ formado por todas las funciones $p \in \mathcal{L}_+^1(\mu)$ tales que $\|p\|_\mu = 1$. D es interpretado como el conjunto de funciones de densidad correspondientes a la medida μ .

En bastantes ocasiones, las funciones de densidad que representan un modelo estadístico, pertenecen a una determinada familia paramétrica $F_{(f, \Theta)}$, que podemos representar frecuentemente a través de una función auxiliar f por:

$$F_{(f, \Theta)} = \{p \in D : p = f(\cdot, \theta) \theta \in \Theta\}$$

donde Θ es una variedad real C^∞ , n -dimensional y $f: \chi \times \Theta \rightarrow \mathbf{R}$ con las condiciones de ser $f(x, \theta) \geq 0$ en casi todas partes y $\int_\chi f(x, \theta) d\mu(x) = 1$, ambas para todo $\theta \in \Theta$.

Por ser Θ una n -variedad C^∞ existe una familia maximal de n -cartas locales (φ_i, U_i) cada una de ellas C^∞ respecto a todas las demás, tal que la unión de los U_i es Θ , donde $\varphi_i: U_i \rightarrow \mathbf{R}^n$. Podemos inducir en $F_{(f, \Theta)}$ una estructura de variedad C^∞ dimensional mediante la definición de la aplicación $\psi_f: \Theta \rightarrow D$ de forma que $\psi_f(\theta) = p = f(\cdot, \theta)$, resultando que $\psi_f(\Theta) = F_{(f, \Theta)}$, y construyendo una familia de cartas locales (ξ_i, W_i) donde $W_i = \psi_f(U_i)$ y $\xi_i = \varphi_i \circ \psi_f^{-1}|_{W_i}$. Como condición a verificar, $\psi_f|_{U_i}$ debe ser inyectiva, en particular se cumplirá si ψ_f es globalmente inyectiva, en caso contrario una misma función de densidad podría estar representada por diferentes coordenadas, careciendo por tanto de la propiedad de identificabilidad.

En muchas ocasiones, la variedad Θ será un abierto conexo de \mathbb{R}^n , y en tal caso considerando el n -subatlas (I, Θ) , existirá una n -carta local que a su vez será un n -subatlas sobre $F_{(f, \Theta)}$, $(\psi^{-1}, F_{(f, \Theta)})$ donde $\psi^{-1}: F_{(f, \Theta)} \rightarrow \Theta \subset \mathbb{R}^n$ $\psi^{-1}(p) = \bar{\theta}$ con $p = f(\cdot, \bar{\theta})$ y cada función de densidad vendrá ahora identificada por unas coordenadas en Θ que coincidirán con sus parámetros.

A partir de estas consideraciones podremos, y generalmente nos interesará, reducir el estudio de $F_{(f, \Theta)}$ al estudio de la variedad Θ .

La definición de la distancia entre dos puntos de una variedad Riemanniana conexa, como el ínfimo de las longitudes de las curvas quebradas C^∞ que unen ambos puntos requiere, como condición general para ser efectivamente una métrica, que la variedad sea Hausdorff. Al ser Θ , en el caso general de que sea un abierto conexo de \mathbb{R}^n , una variedad trivialmente Hausdorff y ψ^{-1} un homeomorfismo entre $F_{(f, \Theta)}$ y Θ , $F_{(f, \Theta)}$ resulta ser también una variedad Hausdorff.

Las ideas expresadas previamente pueden generalizarse de forma inmediata al considerar otras λ -representaciones adecuadas del modelo estadístico. Podemos definir el modelo a través de las variedades que notaremos por $(\varphi_{D^\lambda}, D^\lambda)$, donde φ_{D^λ} es un atlas maximal C^∞ sobre D^λ .

De ahora en adelante nos limitaremos al estudio de variedades que representen un modelo estadístico, a las que supondremos una serie de condiciones de regularidad:

A1. $\lambda(p(\mathbf{x}|\theta))$ será derivable tantas veces como sea necesario respecto a cada θ_i .

A2. Las $\frac{\partial \lambda(p(\mathbf{x}|\theta))}{\partial \theta_i}$ $i = 1, \dots, n$ pertenecen a un adecuado espacio $\mathcal{L}^\alpha(p(\cdot|\theta)d\mu)$.

A3. Las funciones $\frac{\partial \lambda(p(\mathbf{x}|\theta))}{\partial \theta_1}, \dots, \frac{\partial \lambda(p(\mathbf{x}|\theta))}{\partial \theta_n}$ son linealmente independientes.

A4. Las derivadas parciales $\partial/\partial\theta_i$ y la integración con respecto a la medida de referencia $p(\cdot|\theta)d\mu$ pueden ser intercambiadas.

Nótese que como consecuencia inmediata de A4, caso de considerar una variedad de funciones de densidad:

$$E\left(\frac{\partial \log p(\mathbf{x}|\theta)}{\partial \theta_i}\right) = 0 \quad i = 1, \dots, n$$

A la hora de distanciar dos densidades, tenemos la posibilidad de hacerlo no en la hipersuperficie de \mathcal{L}^2 inducida por todas las funciones de densidad, la que podríamos considerar en un sentido amplio, variedad infinito dimensional de todas las funciones de densidad relativas a μ , sino en la variedad inducida por cada familia paramétrica $F_{(\mathcal{U}, \Theta)}$, que a su vez están incluidas en la anterior.

Sean $p(\cdot|\theta)$ y $p(\cdot|\theta) + dp(\cdot|\theta)$ dos puntos contiguos de una variedad $F_{(\mathcal{U}, \Theta)}$ representados en el espacio paramétrico Θ por $(\theta_1, \theta_2, \dots, \theta_n)$ y $(\theta_1 + d\theta_1, \theta_2 + d\theta_2, \dots, \theta_n + d\theta_n)$ respectivamente. La distancia entre ambos puntos será, siguiendo el desarrollo elaborado en el apartado precedente:

$$ds^2 = \alpha^2 \int_{\mathcal{X}} (\sqrt{p(\mathbf{x}|\theta) + dp(\mathbf{x}|\theta)} - \sqrt{p(\mathbf{x}|\theta)})^2 d\mu(x)$$

Definamos $f(\theta) = \int_{\mathcal{X}} (\sqrt{p(\mathbf{x}|\theta)} - \sqrt{p(\mathbf{x}|\mathbf{a})})^2 d\mu(x)$ y efectuemos un desarrollo de Taylor de segundo orden en el punto $\theta = \mathbf{a}$. Es fácil comprobar que $f(\mathbf{a}) = 0$ y que $\frac{\partial f}{\partial \theta_i}(\mathbf{a}) = 0 \quad i = 1, \dots, n$ además:

$$\frac{\partial^2 f}{\partial \theta_i \partial \theta_j} = \frac{1}{2} \int_{\mathcal{X}} \frac{1}{p(\mathbf{x}|\theta)} \partial_i p(\mathbf{x}|\theta) \partial_j p(\mathbf{x}|\theta) d\mu(x)$$

por lo tanto y tomando la aproximación de segundo orden, la distancia entre $(\theta_1, \theta_2, \dots, \theta_n)$ y $(\theta_1 + d\theta_1, \theta_2 + d\theta_2, \dots, \theta_n + d\theta_n)$, la podemos expresar por

$$ds^2 = \frac{\alpha^2}{4} \sum_{i,j=1}^n \left(\int_{\mathcal{X}} \frac{1}{p(\mathbf{x}|\theta)} \partial_i p(\mathbf{x}|\theta) \partial_j p(\mathbf{x}|\theta) d\mu \right) d\theta_i d\theta_j$$

y denominando

$$g_{ij}(\theta) = \int_{\mathcal{X}} \frac{1}{p(\mathbf{x}|\theta)} \partial_i p(\mathbf{x}|\theta) \partial_j p(\mathbf{x}|\theta) d\mu(x) = \int_{\mathcal{X}} p(\mathbf{x}|\theta) \frac{\partial \log p(\mathbf{x}|\theta)}{\partial \theta_i} \frac{\partial \log p(\mathbf{x}|\theta)}{\partial \theta_j} d\mu(x) \quad [2.4]$$

podemos escribir

$$ds^2 = \frac{\alpha^2}{4} \sum_{i,j=1}^n g_{ij}(\theta) d\theta_i d\theta_j \quad [2.5]$$

La matriz $G = (g_{ij}(\theta))$ son las componentes de un tensor de segundo orden, covariante, simétrico y definido positivo que coincide con las componentes de la **matriz de información de Fisher**. Por tanto [2.5] es la métrica de un espacio de Riemann con tensor métrico $g_{ij}(\theta)$.

Tomando $\alpha = 2$ en [2.5], recuperamos la métrica propuesta por Rao (1945, 1949), conocida como métrica informacional.

$$ds^2 = \sum_{i,j=1}^n g_{ij}(\theta) d\theta_i d\theta_j \quad [2.6]$$

Bajo las condiciones de regularidad generales, la matriz de información de Fisher se puede escribir como:

$$g_{ij}(\theta) = - \int_{\mathcal{X}} p \partial_i \partial_j \log p d\mu = - E_{\theta}(\partial_i \partial_j \log p)$$

donde $\partial_i \partial_j \log p = \frac{\partial^2 \log p}{\partial \theta_i \partial \theta_j}$.

En Atkinson y Mitchell (1981), Burbea (1986), Burbea y Rao (1984), Oller y Cuadras (1985,b), entre otros, podemos encontrar ejemplos de la métrica informacional aplicada a familias concretas de distribuciones. Dicha métrica conduce a la denominada **Distancia de Rao** entre funciones de densidad.

Sea $\theta = \theta(t), t_1 \leq t \leq t_2$, una curva en Θ uniendo los puntos $\theta^1, \theta^2 \in \Theta$, con $\theta^j = \theta(t_j)$ ($j = 1, 2$). La distancia entre ambos puntos sobre esta curva es:

$$\rho(\theta^1, \theta^2) = \left| \int_{t_1}^{t_2} \left[\sum_{j=1}^n g_{ij}(\theta) \dot{\theta}_i \dot{\theta}_j \right]^{\frac{1}{2}} dt \right| \quad [2.7]$$

donde el punto indica derivación respecto al parámetro de la curva. Interesa especialmente, caso de existir, la curva que uniendo ambos puntos sea de menor longitud, a la que denominamos geodésica o curva geodésica informacional. Dicha curva se obtiene como solución de las ecuaciones de Euler-Lagrange:

$$\sum_{i=1}^n g_{ik}(\theta) \ddot{\theta}_i + \sum_{i,j=1}^n [ij,k] \dot{\theta}_i \dot{\theta}_j = 0 \quad k = 1, \dots, n$$

o bien:

$$\ddot{\theta}_i + \sum_{j,k=1}^n \Gamma_{jk}^i \dot{\theta}_j \dot{\theta}_k = 0 \quad i = 1, \dots, n$$

con las condiciones de contorno:

$$\theta_i(t_j) = \theta_i^j \quad i = 1, \dots, n; \quad j = 1, 2$$

Las cantidades $[ij,k]$ son conocidas como símbolos de Christoffel de primera especie y vienen dadas por:

$$[ij,k] = \frac{1}{2} \left[\frac{\partial g_{ik}}{\partial \theta_j} + \frac{\partial g_{jk}}{\partial \theta_i} - \frac{\partial g_{ij}}{\partial \theta_k} \right] \quad i, j, k = 1, \dots, n$$

y las cantidades Γ_{jk}^l son los denominados símbolos de Christoffel de segunda especie que vienen definidos por:

$$\Gamma_{jk}^l = \sum_{i=1}^n g^{il} [jk, i] \quad i, j, k = 1, \dots, n$$

La distancia ρ es la llamada distancia geodésica informacional o distancia de Rao, cuya expresión analítica para algunas distribuciones conocidas, se puede encontrar en las referencias mencionadas anteriormente.

En Atkinson y Mitchell (1981) podemos encontrar enfoques alternativos para obtener las ecuaciones de las geodésicas, por ejemplo mediante las ecuaciones de Hamilton-Jacobi, o reduciendo la métrica a la de alguna geometría conocida, por ejemplo la euclídea, hiperbólica o esférica.

Se comprueba fácilmente que dicha métrica posee las propiedades de invarianza frente a transformaciones admisibles de los parámetros, de las variables $x \in \chi$ y que la distancia se incrementa al agregar nuevas variables estocásticamente independientes.

En Burbea (1986) se generaliza el concepto de métrica informacional definiendo el tensor métrico como:

$$g_{ij}^{(f)}(\theta) = E_{\theta}[(f \circ p)(\partial_i \log p(\cdot|\theta))(\partial_j \log p(\cdot|\theta))]$$

y el elemento de línea por:

$$ds_f^2(\theta) = E_{\theta}[(f \circ p)(d \log p)^2]$$

donde f es una función continua y positiva sobre \mathbf{R}^+ y $\partial_i \log p(\cdot|\theta) \equiv \frac{\partial \log p(\cdot|\theta)}{\partial \theta_i}$. Dicha métrica es denominada métrica f -informacional. La elección de $f(x) = x^{\alpha-1}$ da lugar a

la métrica informacional de orden α , que coincide con la métrica informacional previa tomando $\alpha = 1$.

Entre las propiedades con que cuenta la métrica informacional, destaquemos las siguientes extraídas de Burbea (1986) y para más detalles Rao (1973)

1. Sean F_1 y F_2 las matrices de información asociadas a dos variables aleatorias independientes, X_1 y X_2 . Entonces $F = F_1 + F_2$ es la matriz de información debida a $X = X_1 + X_2$.

2. Sea F_T la matriz de información debida a una función T de X . Entonces $F - F_T$ es semidefinida positiva, y es la matriz cero si y sólo si T es un estadístico suficiente, Kullback y Leibler (1951).

3. Sea $p(\cdot|\theta) \in F(\mathcal{X}, \Theta)$ con la correspondiente matriz de información F Supongamos que $f = (f_1, f_2, \dots, f_m)$ es un vector de m estadísticos y definamos $g(\theta) = (g_1(\theta), \dots, g_m(\theta))$ por $g_i(\theta) = E(f_i)$ $i = 1, \dots, m$. Consideremos las matrices $m \times m$ y $m \times n$ $V = [V_{ij}]$ y $U = [U_{ij}]$ definidas por

$$V_{ij} = E_{\theta}[(f_i - g_i)(f_j - g_j)] \quad i, j = 1, \dots, m$$

$$U_{ij} = E_{\theta}[f_i \partial_j \log p] \quad i = 1, \dots, m; j = 1, \dots, n$$

Entonces:

(i) La matriz $m \times m$ $V - UF^{-1}U'$ es semidefinida positiva para todo $\theta \in \Theta$. La matriz es cero en algún $\theta \in \Theta$ si y solo si $f = (f_1, \dots, f_m)$ es de la forma $f_i = \sum_{k=1}^n \lambda_{ik} \partial_k \log p + E_{\theta}(f_i)$ $i = 1, \dots, m$.

(ii) Supongamos que $\partial_j \int_{\mathcal{X}} f_i(x) p(x|\theta) d\mu(x) = \int_{\mathcal{X}} f_i \partial_j p(x|\theta) d\mu(x)$ $i = 1, \dots, m; j = 1, \dots, n$. Entonces U es la matriz jacobiana de $g = (g_1, \dots, g_m)$ con respecto a

$\theta = (\theta_1, \dots, \theta_n)$. En particular, cuando $m = n$ y $g(\theta) = 0$, entonces $V - F^{-1}$ es semidefinida positiva.

La última propiedad constituye el teorema de Cramer-Rao.

Existen otros enfoques que conducen a una métrica análoga, en el capítulo siguiente veremos como llegar a ella a través del producto escalar definido en el espacio tangente a la variedad en un punto. Oller y Cuadras (1985,a) también llegan a un resultado similar partiendo de considerar la distancia entre dos puntos próximos $p(x|\theta)$ y $p(x|\theta + d\theta)$ proporcional a la esperanza del incremento de información al cuadrado

$$\Delta s^2 \propto E_{\theta}(\Delta I^2)$$

donde definimos de forma general la información, como una cierta función real de las funciones de densidad, $I(x|\theta) = \Phi(p(x|\theta))$, donde Φ es una función real, con dominio en los reales positivos y diferenciable tantas veces como sea necesario. Suponiendo las condiciones de regularidad habituales y siendo k una constante positiva obtenemos:

$$\begin{aligned} ds^2 &= kE\left(\left(\sum_{i=1}^n \frac{\partial \Phi(p(x|\theta))}{\partial \theta_i} d\theta_i\right)^2\right) = \\ &= k \sum_{i,j=1}^n E\left((\Phi'(p(x|\theta)))^2 \partial_i p(x|\theta) \partial_j p(x|\theta)\right) d\theta_i d\theta_j \end{aligned}$$

donde:

$$g_{ij}(\theta) = E\left((\Phi'(p(x|\theta)))^2 \partial_i p(x|\theta) \partial_j p(x|\theta)\right) \quad i, j = 1, \dots, n \quad [2.8]$$

son las componentes de un tensor covariante simétrico y definido o semidefinido positivo.

Optemos por la función Φ que mantenga invariante el tensor [2.8] al efectuar un cambio en la medida de referencia.

Proposición 2.2. El tensor métrico definido en [2.8] es invariante frente a transformaciones de la medida de referencia para cualquier espacio medible (χ, A) si y solo si

$$\Phi(x) = \alpha \log x + \beta$$

o lo que es lo mismo, trabajando con la l -representación del modelo estadístico.

Demostración. Efectuemos el siguiente cambio en la medida de referencia $\mu \rightarrow \nu$ donde $\mu \ll \nu$. La nueva medida induce un cambio en las funciones de densidad

$$p = \frac{dP}{d\mu} \rightarrow \bar{p} = \frac{dP}{d\nu}$$

$$\bar{p} = \frac{dP}{d\nu} = \frac{dP}{d\mu} \frac{d\mu}{d\nu} = ph$$

donde $h \in \mathcal{L}^1(\nu)$.

La invariancia en el tensor métrico se logra si

$$E((\Phi'(p))^2 \partial_{ip}(\mathbf{x}|\theta) \partial_{jp}(\mathbf{x}|\theta)) = E(h^2(\Phi'(hp))^2 \partial_{ip}(\mathbf{x}|\theta) \partial_{jp}(\mathbf{x}|\theta)) \quad i, j = 1, \dots, n$$

Demostremos la condición de suficiencia, si $\Phi(x) = \alpha \log x + \beta$, entonces

$$\begin{aligned} E(h^2(x)(\Phi'(h(x)p(\mathbf{x}|\theta)))^2 \partial_{ip}(\mathbf{x}|\theta) \partial_{jp}(\mathbf{x}|\theta)) &= E(h^2(x) \left(\frac{\alpha}{h(x)p(\mathbf{x}|\theta)}\right)^2 \partial_{ip}(\mathbf{x}|\theta) \partial_{jp}(\mathbf{x}|\theta)) = \\ &= E\left(\left(\frac{\alpha}{p(\mathbf{x}|\theta)}\right)^2 \partial_{ip}(\mathbf{x}|\theta) \partial_{jp}(\mathbf{x}|\theta)\right) = E((\Phi'(p(\mathbf{x}|\theta)))^2 \partial_{ip}(\mathbf{x}|\theta) \partial_{jp}(\mathbf{x}|\theta)) \end{aligned}$$

Veamos ahora la condición necesaria. Consideremos $h(x) = \lambda \in \mathbf{R}^+$ para todo $x \in \mathcal{X}$, bajo este supuesto se debe verificar:

$$E((\Phi'(p(\mathbf{x}|\theta)))^2 \partial_i p(\mathbf{x}|\theta) \partial_j p(\mathbf{x}|\theta)) = \lambda^2 E((\Phi'(\lambda p(\mathbf{x}|\theta)))^2 \partial_i p(\mathbf{x}|\theta) \partial_j p(\mathbf{x}|\theta))$$

Dada la arbitrariedad del espacio de probabilidad, podemos trabajar con el caso particular en que tenemos dos conjuntos disjuntos $M, N \in \mathcal{A}$ tales que $\mu(M) = \mu(N) = 1/2$ y

$$p(\mathbf{x}|\theta) = \begin{cases} \theta_1 & x \in M \\ \theta_2 & x \in N \\ 0 & \text{en caso contrario} \end{cases}$$

evidentemente $\theta_2 = 1 - \theta_1$. La condición a verificar se transforma en:

$$(\Phi'(\theta_1))^2 \theta_1 + (\Phi'(\theta_2))^2 \theta_2 = \lambda^2 [(\Phi'(\lambda \theta_1))^2 \theta_1 + (\Phi'(\lambda \theta_2))^2 \theta_2]$$

donde tomando en particular $\theta_1 = \theta_2 = \frac{1}{2}$, resulta:

$$|\Phi'(1/2)| = \lambda |\Phi'(\frac{\lambda}{2})|$$

o lo que es lo mismo:

$$\Phi'(\frac{\lambda}{2}) = \frac{\alpha}{\lambda}$$

obteniendo:

$$\Phi(x) = \alpha \log x + \beta \quad \blacksquare$$

Obteniéndose por tanto:

$$g_{ij}(\mathbf{0}) = k\alpha^2 E(\partial_i p(\mathbf{x}|\theta) \partial_j p(\mathbf{x}|\theta)) \quad i, j = 1, \dots, n$$

donde tomando $k\alpha^2 = 1$ recuperamos la métrica informacional [2.6].

Destaquemos finalmente que Burbea y Rao (1982) a partir de un funcional Φ -entropía, utilizando la entropía de Shannon obtienen un resultado análogo.

Sea Φ una función cóncava, de clase C^2 , con dominio en \mathbb{R}^+ . Se define el funcional Φ -entropía como:

$$H_{\Phi}(p) = \int_{\mathcal{X}} \Phi(p(x)) d\mu(x) \quad [2.9]$$

Definimos la función entropía de orden α , Φ_{α} , por:

$$\Phi_{\alpha}(s) = \begin{cases} (\alpha - 1)^{-1}(s - s^{\alpha}) & \alpha \neq 1 \\ -s \log s & \alpha = 1 \end{cases}$$

para $\alpha \in \mathbb{R}^+$ y $\Phi_{\alpha}(0) = 0$. Llamemos ahora $H_{\alpha} = H_{\Phi}$. Burbea y Rao demuestran a partir del Hessiano de $H_{\alpha}(p)$

$$\Delta_p H_{\Phi}(p) = 4 \int_{\mathcal{X}} \Phi''(p(x)) |f(x)|^2 d\mu(x)$$

que existe una métrica diferencial asociada con H_{α}

$$ds_{\alpha}^2(\theta) = -\frac{1}{\alpha} \int_{\mathcal{X}} \Phi''(p(x|\theta)) |\partial p(x|\theta)|^2 d\mu(x)$$

expresable como:

$$ds_{\alpha}^2(\theta) = \sum_{i,j=1}^n g_{ij}^{\alpha} d\theta_i d\theta_j$$

definida positiva y por tanto métrica de un espacio de Riemann denominada métrica entrópica de orden α . Tomando $\alpha = 1$, H_1 es la entropía de Shannon y la métrica, la ya familiar métrica informacional.

2.5. Conexiones y curvatura.

Una vez construida una representación geométrica de los modelos estadísticos, es importante el estudio de las propiedades que pueda poseer el espacio paramétrico asociado a los modelos, con vista a utilizar dichas propiedades en estudios estadísticos.

Además de la distancia, la **curvatura** es una noción geométrica que como veremos más adelante tiene gran importancia. Para introducir el concepto de curvatura es necesario previamente tener determinada una **conexión lineal** que nos define el concepto de paralelismo en la variedad, es decir, el desplazamiento de un vector paralelo a sí mismo a lo largo de una curva.

Recordemos que, en una variedad, los coeficientes de una conexión afin pueden ser definidos de forma arbitraria verificando ciertas condiciones de regularidad, puesto que representan los coeficientes de la combinación lineal que, usando como base, una base del espacio tangente a la variedad en un punto T_θ , nos proporciona el incremento que sufre un vector del espacio tangente en cada dirección al representarlo en otro espacio tangente a la variedad, en un punto infinitesimalmente desplazado del anterior.

Si $\theta = (\theta_1, \dots, \theta_n)$ es un punto de la variedad, T_θ el espacio tangente en θ , $T_{\theta+d\theta}$ el espacio tangente en $\theta + d\theta$, llamando $e_i(\theta)$ $i = 1, \dots, n$ a los vectores de la base de T_θ , al establecer una correspondencia entre T_θ y $T_{\theta+d\theta}$, al vector e_i se le hace corresponder el vector $e_i(\theta) + \delta e_i \in T_{\theta+d\theta}$ donde haciendo uso del convenio de sumación de los índices repetidos:

$$\delta e_i = d\theta_j \Gamma_{ji}^k e_k(\theta)$$

o bien:

$$D_{e_j} e_i = \Gamma_{ij}^k e_k(0)$$

donde $D_{e_j} e_i$ es la tasa de cambio de e_i en la dirección e_j , y es llamada la derivada covariante de e_i en la dirección e_j . Se puede representar mediante el producto escalar en el espacio tangente:

$$\langle D_{e_j} e_i, e_m \rangle = \Gamma_{ji}^k \langle e_k, e_m \rangle = \Gamma_{jlm}$$

Amari (1968,1980) basándose en trabajos anteriores, introduce una familia uniparamétrica de conexiones afines definida por:

$$\overset{\alpha}{\Gamma}_{ijk} = E_0[(\partial_i \partial_j \log p)(\partial_k \log p)] + \frac{1-\alpha}{2} T_{ijk} \quad \alpha \in \mathbb{R}$$

donde T_{ijk} es un tensor covariante de tercer orden definido por:

$$T_{ijk} = E_0[(\partial_i \log p)(\partial_j \log p)(\partial_k \log p)]$$

La α -curvatura de Riemann-Christoffel vendrá expresada, haciendo uso del convenio de sumación de los índices repetidos que utilizaremos libremente de ahora en adelante, por el tensor:

$$\overset{\alpha}{R}_{ijkl} = \overset{\alpha}{R}_{jkl}^m g_{mi}$$

donde:

$$\overset{\alpha}{R}_{ijk}^l = \partial_j \overset{\alpha}{\Gamma}_{ik}^l - \partial_k \overset{\alpha}{\Gamma}_{ij}^l + \overset{\alpha}{\Gamma}_{ik}^m \overset{\alpha}{\Gamma}_{mj}^l - \overset{\alpha}{\Gamma}_{ij}^m \overset{\alpha}{\Gamma}_{mk}^l$$

indicando ∂_i derivación parcial respecto θ_i y siendo $\overset{\alpha}{\Gamma}_{ij}^k = \overset{\alpha}{\Gamma}_{ijm} g^{mk}$, donde g^{ij} son los componentes de la inversa de la matriz de información de Fisher.

Finalmente, la α -curvatura informacional en las direcciones de $x = (x_1, \dots, x_n)$ e $y = (y_1, \dots, y_n)$ de \mathbb{R}^n viene dada por:

$$\kappa = \frac{R_{ijkl}^{\alpha} x_i y_j x_k y_l}{(g_{ik} g_{jl} - g_{il} g_{jk}) x_i y_j x_k y_l}$$

Se denomina variedad plana a aquella variedad que posea curvatura cero, y si la curvatura es constante, denominamos a la variedad isótropa, es decir la curvatura es independiente de la dirección.

De entre las diferentes α -conexiones, destaquemos la 1-conexión, que como hemos comentado en el capítulo I se encontraba implícita en el trabajo de Efron (1975). La curvatura asociada con la 1-conexión o conexión de Efron, se hace nula para aquellas distribuciones de probabilidad pertenecientes a la familia exponencial. Efron sugirió esta curvatura como un índice de lo "exponencial" que llegaba a ser una determinada distribución. Este índice es importante debido a las buenas propiedades estadísticas que posee la familia exponencial, por ejemplo que los estimadores máximo verosímiles de los parámetros alcanzan la cota de Cramer-Rao o que son estadísticos suficientes, por citar algunas.

Dawid (1975) propuso la -1-conexión o conexión de Dawid, cuya curvatura asociada se anula para el espacio formado por una mezcla de distribuciones de probabilidad linealmente independientes.

Finalmente en una variedad Riemanniana existe una única conexión afín, denominada conexión Riemanniana, libre de torsión, y en la que el desplazamiento paralelo de un vector no cambia su longitud, es decir, es compatible con el tensor métrico g_{ij} en el sentido de que conserva el producto escalar. Dicha conexión, también denominada conexión de Levi-Civita o conexión informacional se obtiene a partir de

la familia de α -conexiones mediante la elección $\alpha = 0$. Una conexión que cumple tales condiciones se denomina métrica, en general, las α -conexiones no son métricas excepto para $\alpha = 0$, Amari (1980). Destaquemos finalmente que la 0-conexión coincide con los símbolos de Christoffel de primera especie obtenidos a partir del tensor métrico, y que las "líneas rectas" con esta conexión son las curvas de longitud mínima o geodésicas.

Oller (1988) en una comunicación personal, construye la representación:

$$p \rightarrow \frac{1}{\beta} p^\beta \in \mathcal{L}_{\frac{1}{\beta}}(\mu)$$

Tomando $\partial_i \frac{1}{\beta} p^\beta$ $i = 1, \dots, n$ como base del espacio tangente en un punto $p \in E_\beta$ y considerando el producto escalar en E_β

$$\langle f, g \rangle = \int_{\mathcal{X}} f g p^\gamma d\mu$$

obtiene que el producto escalar es invariante frente a cambios de la medida de referencia μ si $\gamma = 1 - 2\beta$, y por tanto, es independiente del punto p elegido si y solo si $\beta = \frac{1}{2}$ es decir, la inclusión de p se realiza tal y como hemos visto en un espacio \mathcal{L}^2 .

Posteriormente introduce una familia biparamétrica de conexiones definida por:

$$\Gamma_{jk}^{(\beta, \gamma)} = \int_{\mathcal{X}} (\partial_i \partial_j \frac{1}{\beta} p^\beta) \partial_k \frac{1}{\beta} p^\beta p^\gamma d\mu$$

necesitándose $\gamma = 1 - 2\beta$ para obtener invariancia frente a cambios de la medida de referencia. Resulta finalmente la familia uniparamétrica

$$\Gamma_{jk}^\beta = (\beta - 1)E(\partial_i \log p \partial_j \log p \partial_k \log p) + E(\frac{1}{p} \partial_i \partial_j p \partial_k \log p)$$

y teniendo presente que

II - Geometría de los Modelos Estadísticos

$$E(\partial_i \partial_j \log p \partial_k \log p) = - E(\partial_i \log p \partial_j \log p \partial_k \log p) + E\left(\frac{1}{p} \partial_i \partial_j p \partial_k \log p\right)$$

la familia de β -conexiones coincide con la familia de α -conexiones de Amari teniendo en cuenta que $\alpha = 1 - 2\beta$, por tanto la conexión de Levi-Civita que corresponde a $\alpha = 0$ se obtiene también con $\beta = \frac{1}{2}$.

III. Distancias entre individuos estadísticos.

En este capítulo desarrollamos distancias entre valores muestrales, a los que denominaremos individuos estadísticos, pertenecientes a una misma población estadística, identificando a los individuos con formas lineales pertenecientes al espacio tangente a la variedad paramétrica en el punto correspondiente a la población. Presentamos dos construcciones alternativas basadas en diferentes distancias entre las formas lineales asociadas a los valores muestrales, y finalmente obtenemos expresiones explícitas de la distancia en algunos ejemplos correspondientes a distribuciones concretas.

- 3.1. Una representación del Espacio tangente de un modelo estadístico y su dual.
- 3.2. Distancia inmediata.
- 3.3. Algunos ejemplos de distancias.
- 3.4. Distancia estructural.

3.1. Una representación del Espacio tangente de un modelo estadístico y su dual.

3.1.1. Representación del Espacio tangente.

Recordemos que si M es una n -variedad C^∞ y señalamos por $C^\infty(q)$ el conjunto de funciones reales C^∞ en algún entorno de un punto $q \in M$, un **vector tangente** en q , Hicks (1965), es una función real X sobre $C^\infty(q)$ que verifica:

$$1) X(af+bg) = a(Xf) + b(Xg)$$

$$2) X(fg) = (Xf)g(q) + f(q)(Xg)$$

donde f y g son de $C^\infty(q)$ y a y b son de \mathbb{R} .

El **espacio tangente** a M en un punto q de M , representado por M_q es definido como el conjunto de todos los vectores tangentes en q .

El espacio tangente tiene una estructura de espacio vectorial sobre el cuerpo de los números reales con las siguientes operaciones:

$$(X + Y)f = Xf + Yf$$

$$(bX)f = b(Xf)$$

donde X e Y son de M_q , f de $C^\infty(q)$ y b de \mathbb{R} .

Dada una n -carta local (ξ, U) sobre M con $q \in U$ y el sistema de coordenadas inducido por ella $x_i = u_i \circ \xi$, $i = 1, \dots, n$ donde $u_i: \mathbb{R}^n \rightarrow \mathbb{R}$ $u_i(a_1, \dots, a_n) = a_i$, la familia de vectores tangentes en q representada por $(\frac{\partial}{\partial x_i})_q$ definida por:

$$\left(\frac{\partial}{\partial x_i}\right)_q f = \frac{\partial(f \circ \xi^{-1})}{\partial u_i}(\xi(q)) \equiv D_i(f \circ \xi^{-1})(\xi(q)) \quad i = 1, \dots, n$$

donde la diferenciación a la derecha es la usual en \mathbb{R}^n , forma una base de M_q .

Llamaremos a la base anterior, la base natural asociada al sistema de coordenadas x_i $i = 1, \dots, n$. De forma abreviada, cuando quede claro el sistema de coordenadas, representaremos cada vector de la base por $(\partial_i)_q$. Cualquier vector tangente X_q de M_q podrá ser expresado como una combinación lineal de los vectores $(\partial_i)_q$

$$X_q = \sum_{i=1}^n x_i (\partial_i)_q$$

donde x_i serán las coordenadas de X_q respecto a la base natural.

Consideremos ahora un modelo estadístico, con la misma notación básica que en el capítulo 2, y su λ -representación correspondiente $(\varphi_{D^\lambda}, D^\lambda)$. Sea (ξ, U) una carta local y sea q un punto de la variedad $q \in U \subset D^\lambda$. Notaremos por T_θ^λ al espacio tangente a D^λ en el punto q de coordenadas θ $q = \xi^{-1}(\theta)$. A la base natural de T_θ^λ , respecto del sistema de coordenadas definido por (ξ, U) , la notaremos por $(\partial_i)_q$ $i = 1, \dots, n$.

A continuación vamos a introducir una representación adecuada del espacio tangente a D^λ en q , de la cual presentamos una justificación detallada, que nos va a permitir introducir de forma natural un producto escalar definiendo una métrica Riemanniana en la variedad D^λ .

Admitamos que para todo $q \in D^\lambda$ existe una carta local (ξ, U) y una función $\Phi: \mathbf{R} \rightarrow \mathbf{R}^+$ tal que para todo $\beta \in \xi(U)$ se cumple que $\lambda(p(\cdot|\beta)) \in \mathcal{L}^2(\Phi(p(\cdot|\theta)) d\mu)$. Nótese que esta condición es fácil de verificar para cualquiera de las λ -representaciones habituales comentadas en el capítulo anterior. Consideremos también la aplicación $\xi^{-1}: W \subset \mathbf{R}^n \rightarrow \mathcal{L}^2(\Phi(p(\cdot|\theta)) d\mu)$ tal que $\xi^{-1}(\beta) = \lambda(p(\cdot|\beta))$, donde W es un abierto de \mathbf{R}^n que contiene a θ . La notación ξ^{-1} para la aplicación supone un relativo abuso del lenguaje, sin embargo la mantenemos con objeto de no recargar excesivamente la notación empleada.

Definamos ahora la función Ψ como $\Psi = f \circ \xi^{-1}$, Ψ es una función diferenciable sobre un abierto de \mathbf{R}^n que contiene el punto $\xi(q) = \theta$.

Impongamos a partir de ahora la siguiente condición de regularidad, útil para la demostración de la proposición que enunciaremos a continuación. Existe una función $g \in \mathcal{L}^2(\Phi(p(\cdot|\theta)) d\mu)$ tal que para todo $\theta \in \Theta$

$$\left| \frac{\partial \lambda(p(\mathbf{x}|\theta))}{\partial \theta_i} \right| \leq g(\mathbf{x}) \quad i = 1, \dots, n$$

para casi todo punto \mathbf{x} respecto μ .

Proposición 3.1. La transformación lineal $T: \mathbf{R}^n \rightarrow \mathcal{L}^2(\Phi(p(\cdot|\theta)) d\mu)$ definida del modo siguiente:

$$T(h) = \sum_{i=1}^n \frac{\partial \lambda(p(\cdot|\theta))}{\partial \theta_i} h_i$$

es la diferencial de ξ^{-1} en $\xi(q)$, es decir $D\xi^{-1}(\xi(q))(h) = T(h)$.

Demostración. Sea

$$\begin{aligned} \frac{\|\lambda(p(\cdot|\theta + h)) - \lambda(p(\cdot|\theta)) - T(h)\|^2}{\|h\|^2} &= \int_{\mathbf{x}} \left[\frac{\lambda(p(\mathbf{x}|\theta + h)) - \lambda(p(\mathbf{x}|\theta)) - T(h)}{\|h\|} \right]^2 \Phi(p(\mathbf{x}|\theta)) d\mu(\mathbf{x}) = \\ &= \int_{\mathbf{x}} \left[\frac{\lambda(p(\mathbf{x}|\theta + h)) - \lambda(p(\mathbf{x}|\theta)) - \sum_{i=1}^n \frac{\partial \lambda(p(\mathbf{x}|\theta))}{\partial \theta_i} h_i}{\|h\|} \right]^2 \Phi(p(\mathbf{x}|\theta)) d\mu(\mathbf{x}) \end{aligned}$$

Tomemos el límite cuando $h \rightarrow 0$:

$$\lim_{h \rightarrow 0} \int_{\mathbf{x}} \left[\frac{\lambda(p(\mathbf{x}|\theta + h)) - \lambda(p(\mathbf{x}|\theta)) - \sum_{i=1}^n \frac{\partial \lambda(p(\mathbf{x}|\theta))}{\partial \theta_i} h_i}{\|h\|} \right]^2 \Phi(p(\mathbf{x}|\theta)) d\mu(\mathbf{x})$$

Basándonos en el Teorema de la Convergencia Dominada de Lebesgue, Rudin (1966), al haber impuesto la condición de regularidad anterior, podemos conmutar la integral con el límite, resultando:

$$\int_{\mathbf{x}} \lim_{h \rightarrow 0} \left[\frac{\lambda(p(\mathbf{x}|\theta + h)) - \lambda(p(\mathbf{x}|\theta)) - \sum_{i=1}^n \frac{\partial \lambda(p(\mathbf{x}|\theta))}{\partial \theta_i} h_i}{\|h\|} \right]^2 \Phi(p(\mathbf{x}|\theta)) d\mu(\mathbf{x})$$

El límite interior de la integral es fácil ver que se anula en casi todas partes respecto μ , implicando por tanto:

$$\lim_{h \rightarrow 0} \frac{\|\lambda(p(\cdot|\theta + h)) - \lambda(p(\cdot|\theta)) - T(h)\|^2}{\|h\|^2} = 0$$

y la proposición queda demostrada ■

Llamemos E_{θ}^{λ} al subespacio generado por las n funciones $\frac{\partial \lambda(p(\cdot|\theta))}{\partial \theta_i}$ $i = 1, \dots, n$ y definamos sobre él una aplicación lineal H_f de la forma siguiente:

$$H_f \left(\frac{\partial \lambda(p(\cdot|\theta))}{\partial \theta_i} \right) = D(f \circ \xi^{-1})(\theta)(e_i) \quad i = 1, \dots, n$$

donde $e_i \quad i = 1, \dots, n$ es la base canónica de \mathbb{R}^n .

Por el teorema de Hahn-Banach, Rudin(1966), al ser E_0^λ un subespacio del espacio vectorial $\mathcal{L}^2(\Phi(p(\cdot|\theta)) d\mu)$, H_f puede ser extendido como aplicación lineal sobre todo el espacio vectorial, y bajo estas condiciones H_f puede ser interpretada como el diferencial de Fréchet de una extensión diferenciable de f , \tilde{f} , en un abierto de $\mathcal{L}^2(\Phi(p(\cdot|\theta))d\mu)$, restringido al subespacio E_0^λ .

Sea ahora X_q un vector de T_0^λ

$$X_q = x_1(\partial_1)_q + \dots + x_n(\partial_n)_q$$

Por las definiciones anteriores tendremos:

$$X_q f = \sum_{i=1}^n x_i (\partial_i)_q f = \sum_{i=1}^n x_i H_f \left(\frac{\partial \lambda(p(\cdot|\theta))}{\partial \theta_i} \right) = H_f \left(\sum_{i=1}^n x_i \frac{\partial \lambda(p(\cdot|\theta))}{\partial \theta_i} \right)$$

En estas condiciones cada vector puede ser representado por la combinación:

$$X_q \leftrightarrow \Xi(X_q) = x_1 \frac{\partial \lambda(p(\cdot|\theta))}{\partial \theta_1} + \dots + x_n \frac{\partial \lambda(p(\cdot|\theta))}{\partial \theta_n}$$

Podemos por tanto establecer una correspondencia natural entre el espacio T_0^λ y el espacio E_0^λ donde

$$E_0^\lambda = \left\langle \frac{\partial \lambda(p(\cdot|\theta))}{\partial \theta_1}, \dots, \frac{\partial \lambda(p(\cdot|\theta))}{\partial \theta_n} \right\rangle \quad [3.1]$$

donde $\langle \rangle$ indica espacio generado.

$$\begin{aligned} \Xi: T_0^\lambda &\rightarrow E_0^\lambda \\ X_q &\rightarrow \Xi(X_q) = x_1 \frac{\partial \lambda(p(\cdot|\theta))}{\partial \theta_1} + \dots + x_n \frac{\partial \lambda(p(\cdot|\theta))}{\partial \theta_n} \end{aligned}$$

El espacio E_0^λ es entonces un espacio vectorial de funciones medibles, variables aleatorias, donde podemos introducir de forma natural un producto escalar. Dados dos elementos $X, Y \in E_0^\lambda$ su producto escalar es:

$$\langle X, Y \rangle = \int_{\mathcal{X}} X(\mathbf{x})Y(\mathbf{x}) \Phi(p(\mathbf{x}|\mathbf{0})) d\mu(\mathbf{x})$$

donde $\langle \cdot, \cdot \rangle$ indica aquí el producto escalar.

Esto nos permite definir un producto escalar en T_0^λ a través de la representación por E_0^λ . Dados dos vectores $X_q, Y_q \in T_0^\lambda$

$$\begin{aligned} \langle X_q, Y_q \rangle &= \langle \Xi(X_q), \Xi(Y_q) \rangle = E(\Xi(X_q) \Xi(Y_q)) = \\ &= \int_{\mathcal{X}} \frac{\partial \lambda(p(\mathbf{x}|\mathbf{0}))}{\partial \theta_i} \frac{\partial \lambda(p(\mathbf{x}|\mathbf{0}))}{\partial \theta_j} \Phi(p(\mathbf{x}|\mathbf{0})) d\mu(\mathbf{x}) \quad i, j = 1, \dots, n \end{aligned} \quad [3.2]$$

Proposición 3.2. El producto escalar definido en [3.2] es invariante frente a cambios admisibles de la medida de referencia, $\mu \rightarrow \nu$ tal que $\mu \ll \nu$, para cualquier espacio medible $(\mathcal{X}, \mathcal{A})$, si y sólo si

$$\Phi(y) = \frac{k}{\lambda'(y)^2 y} \quad y > 0 \quad , \quad \Phi(0) = 0$$

Demostración. Vamos a seguir un desarrollo análogo al utilizado en la demostración de la Proposición 2.2. La nueva medida induce un cambio en las funciones de densidad

$$p = \frac{dP}{d\mu} \rightarrow \bar{p} = \frac{dP}{d\nu}$$

$$\bar{p} = \frac{dP}{d\nu} = \frac{dP}{d\mu} \frac{d\mu}{d\nu} = ph$$

donde $h \in \mathcal{L}^1(\nu)$.

La invariancia en el producto escalar se logra si:

$$\int_{\mathbf{x}} \frac{\partial \lambda(p(\mathbf{x}|\theta))}{\partial \theta_i} \frac{\partial \lambda(p(\mathbf{x}|\theta))}{\partial \theta_j} \Phi(p(\mathbf{x}|\theta)) d\mu(\mathbf{x}) = \int_{\mathbf{x}} \frac{\partial \lambda(h(x)p(\mathbf{x}|\theta))}{\partial \theta_i} \frac{\partial \lambda(h(x)p(\mathbf{x}|\theta))}{\partial \theta_j} \Phi(h(x)p(\mathbf{x}|\theta)) d\nu(\mathbf{x})$$

$$i, j = 1, \dots, n$$

Demostremos la condición de suficiencia, si $\Phi(y) = \frac{k}{\lambda'(y)^2 y}$, entonces

$$\int_{\mathbf{x}} \frac{k}{\Phi(p(\mathbf{x}|\theta))p(\mathbf{x}|\theta)} \partial_{\mu} p(\mathbf{x}|\theta) \partial_{\nu} p(\mathbf{x}|\theta) \Phi(p(\mathbf{x}|\theta)) h(x) d\nu(\mathbf{x}) =$$

$$= \int_{\mathbf{x}} \frac{k}{\Phi(h(x)p(\mathbf{x}|\theta))h(x)p(\mathbf{x}|\theta)} h(x)^2 \partial_{\mu} p(\mathbf{x}|\theta) \partial_{\nu} p(\mathbf{x}|\theta) \Phi(h(x)p(\mathbf{x}|\theta)) d\nu(\mathbf{x})$$

con lo que queda demostrada la condición de suficiencia.

Veamos ahora la condición necesaria. Consideremos $h(x) = \alpha \in \mathbf{R}^+$ para todo $x \in \mathcal{X}$, y dada la arbitrariedad del espacio de probabilidad, podemos trabajar con el caso particular en que tenemos dos conjuntos disjuntos $M, N \in \mathcal{A}$ tales que $\mu(M) = \mu(N) = 1$ y

$$p(\mathbf{x}|\theta) = \begin{cases} \theta_1 & x \in M \\ \theta_2 & x \in N \\ 0 & \text{en caso contrario} \end{cases}$$

$\theta_2 = 1 - \theta_1$. La condición a verificar se transforma en:

$$\lambda'(\theta_1)^2 \Phi(\theta_1) + \lambda'(\theta_2)^2 \Phi(\theta_2) = \alpha [\lambda'(\alpha\theta_1)^2 \Phi(\alpha\theta_1) + \lambda'(\alpha\theta_2)^2 \Phi(\alpha\theta_2)]$$

donde tomando en particular $\theta_1 = \theta_2 = \frac{1}{2}$, resulta:

$$\lambda'(1/2)^2 \Phi(1/2) = \alpha (\lambda'(\alpha/2)^2 \Phi(\alpha/2))$$

o lo que es lo mismo:

$$\Phi(y) = \frac{k}{\lambda'(y)^2} \quad \blacksquare$$

Por tanto el producto escalar en el espacio tangente en cada punto de la variedad define una métrica riemanniana en esta última que viene especificada por un tensor métrico de la forma

$$g_{ij}(\theta) = k E\left(\frac{\partial \log(p(\mathbf{x}|\theta))}{\partial \theta_i} \frac{\partial \log(p(\mathbf{x}|\theta))}{\partial \theta_j}\right) \quad i, j = 1, \dots, n \quad [3.3]$$

Señalemos en este punto la semejanza entre el tensor métrico definido en [3.3] con áquel obtenido en [2.4] tomando $k = 1$. Además el producto escalar obtenido es independiente de λ , con lo que podemos considerar equivalentes desde este punto de vista todas las λ -representaciones.

A partir de ahora consideraremos $E_0 \equiv E_0^\lambda$.

Recordemos que las condiciones de regularidad expuestas en el capítulo II hacen que $E(\partial_i \log p(\mathbf{x}|\theta)) = 0$ y por tanto el producto escalar coincide con la covarianza debido a que $E(XY) = \text{cov}(X, Y)$.

3.1.2. Representación del Espacio tangente dual.

Denominaremos **Espacio tangente dual** M_q^* de una variedad diferenciable M en un punto $q \in M$ al dual del espacio tangente en q .

Es bien sabido que M_q^* es a su vez un espacio vectorial de dimensión igual a la de M_q . Sea $(\frac{\partial}{\partial x_i})_q, i = 1, \dots, n$ una base de M_q , para cada $i = 1, \dots, n$ existe una forma lineal que notamos por $(\frac{\partial}{\partial x_i})_q^*$ tal que:

$$(\frac{\partial}{\partial x_i})_q^* : M_q \rightarrow \mathbb{R}$$

$$(\frac{\partial}{\partial x_i})_q^* (\frac{\partial}{\partial x_j})_q = \delta_{ij} \quad i, j = 1, \dots, n$$

donde $\delta_{ij} = 1$ si $i = j$ y $\delta_{ij} = 0$ si $i \neq j$. Es inmediato comprobar que la familia de formas lineales $(\frac{\partial}{\partial x_i})_q^*$ es una base de M_q^* . Dicha base es conocida como base dual de la base $(\frac{\partial}{\partial x_i})_q$.

En nuestro caso notaremos por E_0^* al dual de la representación definida en el apartado anterior E_0 del espacio tangente a la variedad D^λ en un punto $q = \xi^{-1}(0)$.

La base dual de la base $\frac{\partial \log p(\cdot|0)}{\partial \theta_i}, i = 1, \dots, n$ la notaremos por $(\frac{\partial \log p(\cdot|0)}{\partial \theta_i})^*, i = 1, \dots, n$. Por tanto y en la forma abreviada $\partial_i^* \log p(\cdot|0) = (\frac{\partial \log p(\cdot|0)}{\partial \theta_i})^*$

$$\partial_i^* \log p(\partial_j \log p(\cdot|0)) = \delta_{ij} \quad i, j = 1, \dots, n$$

donde δ_{ij} son las deltas de Kronecker.

Todo elemento $Y^* \in E_0^*$ se podrá expresar como una combinación lineal de la forma

$$Y^* = y_1 \partial_1^* \log p + \dots + y_n \partial_n^* \log p \quad [3.4]$$

Si $G(\theta) = [g_{ij}(\theta)]$ es la métrica definida por el producto escalar en E_θ , se define la métrica dual $G^{-1}(\theta) = [g^{ij}(\theta)]$ por $g_{ij} g^{jk} = g^{kj} g_{ji} = \delta_i^k$, siendo δ_i^k un tensor mixto definido por:

$$\delta_i^k = \begin{cases} 1 & \text{si } i = k \\ 0 & \text{si } i \neq k \end{cases}$$

Una mayor información sobre los conceptos de geometría diferencial presentados hasta el momento puede encontrarse en los clásicos textos de Spivak (1979) o Hicks (1965).

3.2. Distancia inmediata.

Sea (χ, A) un espacio medible sobre el que tenemos definido un modelo estadístico representado por la familia paramétrica $F_{(f, \Theta)} = \{p = f(\cdot, \theta) : \theta \in \Theta\}$, donde $\theta = (\theta_1, \dots, \theta_n)$ son los parámetros de la densidad y Θ un abierto de \mathbf{R}^n que representa el conjunto de parámetros para el cual la densidad está definida. Los elementos de χ son los que denominaremos individuos estadísticos.

Consideremos la siguiente aplicación entre el espacio muestral χ y el espacio E_0^* , dual de la representación E_0 del espacio tangente dual a $F_{(f, \Theta)}$ el punto de coordenadas $\theta = (\theta_1, \dots, \theta_n)$:

$$\begin{aligned} \delta: \chi &\rightarrow E_0^* \\ \mathbf{x} &\rightarrow \delta(\mathbf{x}) = Y^* \end{aligned} \quad [3.5]$$

tal que $Y^*(Y) = Y(\mathbf{x})$ para todo $Y \in E_0$.

Fijado θ , todo elemento $\mathbf{x} \in \chi$, se identifica mediante δ con una forma lineal perteneciente a E_0^* que aplica cada vector tangente de E_0 al valor que toma dicho vector sobre el individuo \mathbf{x} . Si consideramos el caso general sin fijar θ , cada individuo se identifica con un campo tensorial covariante de primer orden.

En estas condiciones se cumple lo siguiente:

$$Y^*\left(\frac{\partial \log p(\mathbf{x}|\theta)}{\partial \theta_i}\right) = \frac{\partial \log p(\mathbf{x}|\theta)}{\partial \theta_i} \equiv \partial_i \log p(\mathbf{x}) \quad i = 1, \dots, n$$

y por tanto de la representación [3.5] se sigue que todo individuo estadístico puede ser representado en E_0^* mediante unas coordenadas y_i $i = 1, \dots, n$ que son los valores que toman las n variables aleatorias $y_i = \partial_i \log p(\mathbf{x}|\theta)$ $i = 1, \dots, n$ sobre los diferentes individuos

$$x \rightarrow (\partial_1 \log p(x|0), \dots, \partial_n \log p(x|0)) \quad [3.6]$$

La distancia entre dos individuos estadísticos $x, \bar{x} \in \chi$ la definiremos como la distancia euclídea en E_0^* entre sus respectivas formas lineales asociadas:

$$\begin{aligned} d_x^2(x, \bar{x}) &= d_{E_0^*}^2(Y^*, \bar{Y}^*) = \langle Y^*, \bar{Y}^* \rangle_{E_0^*} = \\ &= (\partial_\theta \log p(\bar{x}|0) - \partial_\theta \log p(x|0))' G^{-1}(\theta) (\partial_\theta \log p(\bar{x}|0) - \partial_\theta \log p(x|0)) \end{aligned} \quad [3.7]$$

donde hemos utilizado la siguiente notación :

$$\partial_\theta \log p(x|0) \equiv \left(\frac{\partial \log p(x|0)}{\partial \theta_1}, \dots, \frac{\partial \log p(x|0)}{\partial \theta_n} \right)$$

y $G^{-1}(\theta)$ es la matriz del producto escalar en el espacio tangente dual, que corresponde a la inversa de la matriz de información de Fisher.

Queremos hacer notar que la distancia obtenida, no es una distancia intrínseca entre los individuos, en el sentido utilizado por Rao (1982), sino que depende de la población a la que pertenezcan los individuos, que nos determina el punto $p = \varphi^{-1}(\theta)$ en el cual está construido el espacio tangente dual.

Para que la distancia d sea una verdadera distancia métrica y no solamente una distancia pseudométrica o pseudodistancia, es necesario que se cumpla $d^2(x, \bar{x}) = 0$ si y sólo si $x = \bar{x}$. En nuestro caso no será cierto en general, ya que $d^2(x, \bar{x}) = 0$ cuando

$$\frac{\partial \log p(x|0)}{\partial \theta_i} = \frac{\partial \log p(\bar{x}|0)}{\partial \theta_i} \quad i = 1, \dots, n.$$

Si queremos obtener una auténtica distancia siempre podemos definir una relación de equivalencia en χ . Diremos que $x \sim \bar{x}$ si y sólo si $d^2(x, \bar{x}) = 0$, y ésto es una relación de equivalencia en χ que particiona χ en clases de equivalencia. Cada clase

consta de todos los individuos equivalentes a uno dado. Si X_1 y X_2 son dos clases de equivalencia, tomemos $x \in X_1$ y $\bar{x} \in X_2$ y definamos $d^2(X_1, X_2) = d^2(x, \bar{x})$. Nótese que hemos utilizado el mismo símbolo para la distancia aun cuando son claramente diferentes, al actuar una sobre χ y la otra sobre las clases de equivalencia.

Cuando consideramos el conjunto de las clases de equivalencia, d se convierte en una auténtica distancia métrica. A pesar de todo, para simplificar y haciendo un abuso del lenguaje continuaremos hablando de distancia sobre los individuos de χ en vez de sobre las clases de equivalencia.

Otra construcción es posible si consideramos el caso general de las variedades $(\varphi_{D^\lambda}, D^\lambda)$, e interpretando a los individuos como funciones escalares sobre D^λ

$$x \rightarrow f_x : q \rightarrow f_x(q) = q(x) = \lambda(p(x|\theta(q))) \quad \text{para todo } q \in D^\lambda$$

Por otra parte cada f_x estaría representada por un campo tensorial covariante

$$Y^* \in E_0^{\lambda*} \quad Y^*(Y) = Yf_x \quad \text{para todo } Y \in E_0^\lambda$$

y los individuos por tanto por:

$$x \rightarrow (\partial_1 \lambda(p(x|\theta(q))), \dots, \partial_n \lambda(p(x|\theta(q))))$$

Respecto la base dual de la base canónica asociada al sistema de coordenadas.

Si imponemos la condición de invarianza frente a transformaciones de la medida de referencia, tal y como hemos comprobado en apartados precedentes hemos de tomar $\lambda(x) = \log(x)$, coincidiendo con los resultados anteriores.

De forma análoga es posible la siguiente interpretación para las variables aleatorias. Sea V el espacio vectorial de todas las variables aleatorias definidas sobre

un espacio de probabilidad (χ, A, P) . Sea Z el subespacio vectorial de V formado por las variables aleatorias con los dos primeros momentos finitos, y Z_0 el subespacio de Z formado por las variables aleatorias centradas. La aplicación:

$$\begin{aligned} Z_0 \times Z_0 &\rightarrow \mathbf{R} \\ (X', Y') &\rightarrow \text{Cov}(X', Y') \end{aligned}$$

es un producto escalar no degenerado en Z_0 . Z_0 es entonces un espacio de Hilbert y para cada valor de θ tenemos un subespacio vectorial E_θ del mismo. Toda variable aleatoria $X \in Z$ puede identificarse con un elemento de E_θ mediante la aplicación $\Pi: Z \rightarrow E_\theta$, donde $\Pi(X)$ es la proyección ortogonal de $X - E(X)$ sobre E_θ . Si la dimensión de V es finita, no es necesaria la condición de existencia de los primeros momentos, al existir siempre un subespacio suplementario de E_θ de forma que $V = E_\theta \oplus H$ y todo $X \in V$ puede escribirse de forma única como $X = u + v$, donde $u \in E_\theta$ y $v \in H$. Identificaremos toda variable aleatoria con el elemento de E_θ que denominaremos proyección de X , $\Pi(X)$, y que será $\Pi(X) = u$.

Por lo expuesto anteriormente las variables aleatorias pueden identificarse con tensores contravariantes de orden 1 en E_θ , o en el caso general, sin fijar θ , con un campo tensorial contravariante de primer orden en la variedad.

Previamente a ver algunos ejemplos de construcción de la distancia, examinemos algunas propiedades importantes. En las siguientes proposiciones supondremos válidas las condiciones de regularidad asumidas generalmente y ya mencionadas en apartados precedentes.

Proposición 3.3. La distancia [3.7] es invariante frente a cambios admisibles de la medida de referencia $\mu \rightarrow \nu$ tal que $\mu \ll \nu$.

Demostración. Inmediato a partir de la invariancia de la matriz de información de Fisher y de las coordenadas $y_i = \partial_i \log p(\mathbf{x}|\theta)$ ■

Sea T una función medible que transforma el espacio medible (χ, A) en otro espacio medible (χ', A') . T induce un cambio en los espacios de probabilidad a través de los cambios en las medidas probabilísticas de la forma siguiente $P' = PT^{-1}$.

Proposición 3.4. Supuesta la existencia de una medida ν positiva y acotada sobre A' y que la función de densidad resultante sea de tipo paramétrico $g(t, \theta) \in \mathcal{L}^1(\nu)$ $P'(B) = \int_B g(t, \theta) d\nu(t)$ para todo $B \in A'$. Entonces si T es un estadístico suficiente:

$$d_{\chi}^2(\mathbf{x}, \bar{\mathbf{x}}) = d_{\chi'}^2(t, \bar{t}) \quad [3.8]$$

donde $t = T(\mathbf{x})$ y $\bar{t} = T(\bar{\mathbf{x}})$.

Demostración. Tal y como hemos indicado en el apartado 4 del capítulo II, la matriz de información de Fisher es invariante frente a transformaciones por estadísticos suficientes. Además y debido al Teorema de factorización de Neyman-Fisher, si T es un estadístico suficiente, bajo las suposiciones del enunciado:

$$p(\mathbf{x}|\theta) = g(t|\theta) h(\mathbf{x})$$

Por tanto y a partir de la definición [3.7] se sigue inmediatamente [3.8] ■

Proposición 3.5. La distancia [3.7] es no decreciente al aumentar el número de coordenadas.

Demostración. En efecto, [3.7] puede ser representada mediante la forma cuadrática $\mathbf{u}'G^{-1}\mathbf{u}$ donde $\mathbf{u} = (\partial_{\theta} \log p(\bar{\mathbf{x}}|\theta) - \partial_{\theta} \log p(\mathbf{x}|\theta))$, y recordemos que G es simétrica y definida positiva. Consideremos $\mathbf{u} = (u_1, u_2)$ y

$$G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}$$

la afirmación del enunciado quedará demostrada si conseguimos demostrar que $\mathbf{u}'G^{-1}\mathbf{u} - \mathbf{u}'_1G_{11}^{-1}\mathbf{u}_1 \geq 0$, donde:

$$G^{-1} = \begin{bmatrix} G^{11} & G^{12} \\ G^{21} & G^{22} \end{bmatrix}$$

Descompongamos $\mathbf{u} = \mathbf{v} + \mathbf{w}$ donde $\mathbf{v} = (G_{11}G_{11}^{-1}\mathbf{u}_1, G_{21}G_{11}^{-1}\mathbf{u}_1)'$ y $\mathbf{w} = (0, \mathbf{u}_2 - G_{21}G_{11}^{-1}\mathbf{u}_1)'$, entonces $\mathbf{u}'G^{-1}\mathbf{u} = \mathbf{w}'G^{-1}\mathbf{w} + \mathbf{v}'G^{-1}\mathbf{v} + 2\mathbf{v}'G^{-1}\mathbf{w}$. Es inmediato comprobar que el primer término $\mathbf{w}'G^{-1}\mathbf{w} \geq 0$, puesto que G^{-1} es también definida positiva, igualmente se comprueba que

$$\begin{aligned} \mathbf{v}'G^{-1}\mathbf{v} &= (\mathbf{u}'_1G_{11}^{-1}G_{11}, \mathbf{u}_1G_{11}^{-1}G_{21}) \begin{bmatrix} (G^{11}G_{11} + G^{12}G_{21})G_{11}^{-1}\mathbf{u}_1 \\ (G^{21}G_{11} + G^{22}G_{21})G_{11}^{-1}\mathbf{u}_1 \end{bmatrix} = \\ &= (\mathbf{u}'_1G_{11}^{-1}G_{11}, \mathbf{u}_1G_{11}^{-1}G_{21}) \begin{bmatrix} G_{11}^{-1}\mathbf{u}_1 \\ 0 \end{bmatrix} = \mathbf{u}'_1G_{11}^{-1}\mathbf{u}_1 \end{aligned}$$

además:

$$\begin{aligned} \mathbf{v}'G^{-1}\mathbf{w} &= \mathbf{v}' \begin{bmatrix} G^{12}(\mathbf{u}_2 - G_{21}G_{11}^{-1}\mathbf{u}_1) \\ G^{22}(\mathbf{u}_2 - G_{21}G_{11}^{-1}\mathbf{u}_1) \end{bmatrix} = \\ &= \mathbf{u}'_1G_{11}^{-1}(G_{11}G^{12} + G_{12}G^{22})(\mathbf{u}_2 - G_{21}G_{11}^{-1}\mathbf{u}_1) = 0 \end{aligned}$$

Por tanto $\mathbf{u}'G^{-1}\mathbf{u} - \mathbf{u}'_1G_{11}^{-1}\mathbf{u}_1 = \mathbf{w}'G^{-1}\mathbf{w} \geq 0$ ■

Proposición 3.6. Sean $\theta_1^1, \dots, \theta_{n_1}^1, \theta_1^2, \dots, \theta_{n_2}^2, \dots, \theta_1^k, \dots, \theta_{n_k}^k$ los parámetros de la densidad correspondiente a un modelo estadístico, donde $n_1 + \dots + n_k = n$. Si el tensor métrico es de la forma:

$$G(\theta) = \begin{bmatrix} A_1 & 0 & \dots & 0 \\ 0 & A_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & A_k \end{bmatrix}$$

donde A_i son matrices $n_i \times n_i$. El cuadrado de la distancia entre dos individuos en el espacio tangente dual de la variedad $F_{(f, \Theta)} = \{p = f(\cdot, \theta) : \theta \in \Theta\}$ donde $\theta = (\theta_1, \dots, \theta_n)$, es igual a la suma de las distancias al cuadrado entre los individuos en cada uno de los espacios tangentes duales de las k variedades $F_{(f_i, \Theta_i)} = \{p = f_i(\cdot, \theta_i) : \theta_i \in \Theta_i\}$, donde $\theta_i = (\theta_1^i, \dots, \theta_{n_i}^i) \quad i = 1, \dots, k$.

Demostración. Evidente a partir de la definición de la distancia en [3.7] y del hecho de que la inversa del tensor métrico es de la forma

$$G^{-1}(\theta) = \begin{bmatrix} A_1^{-1} & 0 & \dots & 0 \\ 0 & A_2^{-1} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & A_k^{-1} \end{bmatrix} \quad \blacksquare$$

3.3. Algunos ejemplos de distancias.

3.3.1. Distribuciones uniparamétricas.

En este caso E_0^* es un espacio vectorial de dimensión uno, y todo individuo $x \in \chi$ vendrá representado por una única coordenada $y = \frac{\partial \log p(x|\theta)}{\partial \theta}$. La métrica de E_0^* vendrá dada por $G^{-1}(\theta) = g^{11}(\theta) = 1/E_0((\frac{\partial \log p(x|\theta)}{\partial \theta})^2)$. Por tanto el cuadrado de la distancia entre dos individuos x e y será:

$$d^2(x, y) = g^{11}(\frac{\partial \log p(x|\theta)}{\partial \theta} - \frac{\partial \log p(y|\theta)}{\partial \theta})^2$$

Presentaremos a continuación resultados correspondientes a casos en que los individuos son:

i) Muestras de tamaño m de una distribución de Poisson de parámetro λ .

$$p(x_1, \dots, x_m | \lambda) = \exp(-m\lambda) \frac{\lambda^{m\bar{x}}}{x_1! \dots x_m!}$$

$$\lambda \in \mathbf{R}^+ \quad (x_1, \dots, x_m) \in (\mathbf{Z}^+)^m$$

donde $\mathbf{Z}^+ = \{0, 1, 2, \dots\}$, y $m\bar{x} = \sum_{i=1}^m x_i$.

Las coordenadas de un individuo vienen dadas por:

$$\frac{\partial \log p(x_1, \dots, x_m | \lambda)}{\partial \lambda} = -m + \frac{m\bar{x}}{\lambda}$$

El tensor métrico es de la forma:

$$g^{11}(\lambda) = \frac{\lambda}{m}$$

La distancia entre dos individuos $\mathbf{x} = (x_1, \dots, x_m)$ e $\mathbf{y} = (y_1, \dots, y_m)$ resulta:

$$d^2(\mathbf{x}, \mathbf{y}) = \frac{m}{\lambda} (\bar{x} - \bar{y})^2 \quad [3.9]$$

donde $\bar{x} = \frac{1}{m} \sum_{i=1}^m x_i$.

ii) Muestras de tamaño m de una distribución Weibull de parámetro λ .

$$p(x_1, \dots, x_m | \lambda) = r^m \lambda^m \exp\left(-\lambda \sum_{i=1}^m x_i^r\right) \prod_{i=1}^m x_i^{r-1}$$

$$\lambda \in \mathbf{R}^+ \quad (x_1, \dots, x_m) \in (\mathbf{R}^+)^m \quad r > 0$$

Las coordenadas de un individuo vienen dadas por:

$$\frac{\partial \log p(x_1, \dots, x_m | \lambda)}{\partial \lambda} = \frac{m}{\lambda} - \sum_{i=1}^m x_i^r$$

El tensor métrico es de la forma:

$$g^{11}(\lambda) = \frac{\lambda^2}{m}$$

La distancia entre dos individuos $\mathbf{x} = (x_1, \dots, x_m)$, $\mathbf{y} = (y_1, \dots, y_m)$ resulta:

$$d^2(\mathbf{x}, \mathbf{y}) = \frac{\lambda^2}{m} \left(\sum_{i=1}^m x_i^r - \sum_{i=1}^m y_i^r \right)^2 \quad [3.10]$$

iii) Muestras de tamaño m de una distribución Gamma de parámetro α conocido el valor de p .

$$p(x_1, \dots, x_m | \alpha, p_0) = \frac{\alpha^{mp_0}}{(\Gamma(p_0))^m} \exp\left(-\alpha \sum_{i=1}^m x_i\right) \prod_{i=1}^m x_i^{p_0-1}$$

$$\alpha \in \mathbf{R}^+ \quad (x_1, \dots, x_m) \in (\mathbf{R}^+)^m$$

Las coordenadas de un individuo vienen dadas por:

$$\frac{\partial \log p(x_1, \dots, x_m | \alpha, p_0)}{\partial \alpha} = \frac{p_0 m}{\alpha} - \sum_{i=1}^m x_i$$

El tensor métrico es de la forma:

$$g^{11}(\alpha) = \frac{\alpha^2}{p_0 m}$$

La distancia entre dos individuos $\mathbf{x} = (x_1, \dots, x_m)$, $\mathbf{y} = (y_1, \dots, y_m)$ es:

$$d^2(\mathbf{x}, \mathbf{y}) = \frac{m \alpha^2}{p_0} (\bar{x} - \bar{y})^2 \quad [3.11]$$

iv) Muestras de tamaño m de una distribución exponencial de parámetro λ .

Es un caso particular del anterior para $p = 1$, por tanto, la distancia entre dos individuos $\mathbf{x} = (x_1, \dots, x_m)$, $\mathbf{y} = (y_1, \dots, y_m)$ es:

$$d^2(\mathbf{x}, \mathbf{y}) = m \lambda^2 (\bar{x} - \bar{y})^2 \quad [3.12]$$

v) Muestras de tamaño m de una distribución Binomial de parámetro p .

$$p(x|p) = \binom{m}{x} p^x (1-p)^{m-x}$$

$$m, x \in \mathbb{Z}^+, \quad 0 < p < 1$$

Las coordenadas de un individuo vienen dadas por:

$$\frac{\partial \log p(x|p)}{\partial p} = \frac{x - mp}{p(1-p)}$$

El tensor métrico es de la forma:

$$g^{11}(p) = \frac{p(1-p)}{m}$$

La distancia entre dos individuos x e y es:

$$d^2(x, y) = \frac{(x - y)^2}{mp(1-p)} \quad [3.13]$$

vi) Muestra correspondiente a una distribución Binomial negativa.

$$p(x|p) = \binom{k+x-1}{x} p^k (1-p)^x$$

$$k, x \in \mathbb{Z}^+, \quad 0 < p < 1$$

Las coordenadas de un individuo vienen dadas por:

$$\frac{\partial \log p(x|p)}{\partial p} = \frac{k}{p} - \frac{x}{1-p}$$

El tensor métrico es de la forma:

$$g^{11}(p) = \frac{p^2(1-p)}{k}$$

La distancia entre dos individuos x e y es:

$$d^2(x, y) = \frac{(x - y)^2 p^2}{k(1-p)} \quad [3.14]$$

vii) Muestras de tamaño m de una distribución Normal univariante con varianza conocida.

$$p(x_1, \dots, x_m | \mu, \sigma_0^2) = \left(\frac{1}{\sqrt{2\pi\sigma_0^2}} \right)^m \exp\left(- \frac{\sum_{i=1}^m (x_i - \mu)^2}{2\sigma_0^2} \right)$$

$$\mu \in \mathbf{R} \quad (x_1, \dots, x_m) \in \mathbf{R}^m \quad \sigma_0 \in \mathbf{R}^+$$

Las coordenadas de un individuo vienen dadas por:

$$\frac{\partial \log p(x_1, \dots, x_m | \mu, \sigma_0^2)}{\partial \mu} = \frac{\sum_{i=1}^m (x_i - \mu)}{\sigma_0^2}$$

El tensor métrico es de la forma:

$$g^{11}(\mu) = \frac{\sigma_0^2}{m}$$

La distancia entre dos individuos $\mathbf{x} = (x_1, \dots, x_m)$, $\mathbf{y} = (y_1, \dots, y_m)$ es:

$$d^2(\mathbf{x}, \mathbf{y}) = \frac{m}{\sigma_0^2} (\bar{x} - \bar{y})^2 \quad [3.15]$$

viii) Muestras de tamaño m de una distribución Normal univariante con media conocida.

$$p(x_1, \dots, x_m | \mu_0, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^m \exp\left(- \frac{\sum_{i=1}^m (x_i - \mu_0)^2}{2\sigma^2} \right)$$

$$\mu_0 \in \mathbf{R} \quad (x_1, \dots, x_m) \in \mathbf{R}^m \quad \sigma \in \mathbf{R}^+$$

Las coordenadas de un individuo vienen dadas por:

$$\frac{\partial \log p(x_1, \dots, x_m | \mu_0, \sigma^2)}{\partial \sigma} = \frac{\sum_{i=1}^m (x_i - \mu_0)^2 - m\sigma^2}{\sigma^3}$$

El tensor métrico es de la forma:

$$g^{11}(\sigma) = \frac{\sigma^2}{2m}$$

La distancia entre dos individuos $\mathbf{x} = (x_1, \dots, x_m)$, $\mathbf{y} = (y_1, \dots, y_m)$ es:

$$d^2(\mathbf{x}, \mathbf{y}) = \frac{m}{2\sigma^4}(S^2(\mathbf{x}) - S^2(\mathbf{y}))^2 \quad [3.16]$$

donde $S^2(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^m (x_i - \mu_0)^2$.

3.3.2. Distribuciones multiparamétricas.

El espacio E_{θ}^* tendrá una dimensión igual al número de parámetros, los individuos vendrán representados por las n coordenadas [3.6] y la distancia entre dos individuos se obtendrá a partir de la expresión [3.7].

Veamos algunos ejemplos:

i) Muestras de tamaño m de una distribución multinomial.

$$p(x_1, \dots, x_{n+1} | p_1, \dots, p_n) = \frac{m!}{\mathbf{x}!} (p_1)^{x_1} \dots (p_{n+1})^{x_{n+1}}$$

$$(x_1, \dots, x_{n+1}) \in (\mathbb{Z}^+)^{n+1} \quad \sum_{i=1}^{n+1} x_i = m \quad \sum_{i=1}^{n+1} p_i = 1$$

donde $\mathbf{x}! = x_1! \dots x_{n+1}!$.

Las coordenadas de un individuo vienen dadas por:

$$\frac{\partial \log p(x_1, \dots, x_{n+1} | p_1, \dots, p_n)}{\partial p_i} = \frac{x_i}{p_i} - \frac{m - x_1 - \dots - x_n}{1 - p_1 - \dots - p_n} \quad i = 1, \dots, n$$

El tensor métrico toma la forma:

$$g_{ij} = m \left(\frac{\delta_{ij}}{p_i} + \frac{1}{1 - p_1 - \dots - p_n} \right)$$

y su inverso es $g^{ij} = \frac{1}{m} (\delta_{ij} p_i - p_i p_j) \quad i, j = 1, \dots, n$.

La distancia entre dos individuos $\mathbf{x} = (x_1, \dots, x_n)$, $\mathbf{y} = (y_1, \dots, y_n)$ viene expresada por:

$$d^2(\mathbf{x}, \mathbf{y}) = \frac{1}{m} \sum_{i=1}^{n+1} \frac{(x_i - y_i)^2}{p_i} \quad [3.17]$$

En el caso de muestras de tamaño uno obtenemos que la distancia entre los individuos $\mathbf{x} = (0, \dots, 0, \frac{1}{i}, 0, \dots, 0)$ e $\mathbf{y} = (0, \dots, 0, \frac{1}{j}, 0, \dots, 0)$ resulta:

$$d^2(\mathbf{x}, \mathbf{y}) = (1 - \delta_{ij})\left(\frac{1}{p_i} + \frac{1}{p_j}\right) \quad i, j = 1, \dots, n + 1 \quad [3.18]$$

ii) Muestras correspondientes a una distribución multinomial negativa.

$$p(x_1, \dots, x_n | p_1, \dots, p_n) = \frac{(x_1 + x_2 + \dots + x_n + r - 1)!}{\mathbf{x}! (r - 1)!} (p_1)^{x_1} \dots (p_n)^{x_n} (1 - |\mathbf{p}|)^r$$

$$(x_1, \dots, x_n, r) \in (\mathbb{Z}^+)^{n+1} \quad \sum_{i=1}^n p_i < 1$$

Las coordenadas de un individuo vienen dadas por:

$$\frac{\partial \log p(x_1, \dots, x_n | p_1, \dots, p_n)}{\partial p_i} = \frac{x_i}{p_i} - \frac{r}{1 - |\mathbf{p}|} \quad i = 1, \dots, n$$

donde $|\mathbf{p}| = p_1 + \dots + p_n$.

En Oller (1982) encontramos:

$$g_{ij} = \frac{r}{1 - |\mathbf{p}|} \left[\frac{1}{p_i} \delta_{ij} + \frac{1}{1 - |\mathbf{p}|} \right]$$

$$g^{ij} = \frac{1 - |\mathbf{p}|}{r} [\delta^{ij} p_i - p_i p_j] \quad i, j = 1, \dots, n$$

y la distancia entre dos individuos $\mathbf{x} = (x_1, \dots, x_n)$, $\mathbf{y} = (y_1, \dots, y_n)$ viene dada por:

$$\begin{aligned} d^2(\mathbf{x}, \mathbf{y}) &= \frac{1 - |\mathbf{p}|}{r} \sum_{i,j=1}^n \left[\frac{\delta^{ij}}{p_j} - 1 \right] (x_i - y_i)(x_j - y_j) = \\ &= \frac{1 - |\mathbf{p}|}{r} \left[\sum_{i=1}^n \frac{(x_i - y_i)^2}{p_i} - \left(\sum_{i=1}^n (x_i - y_i) \right)^2 \right] \end{aligned} \quad [3.19]$$

iii) Muestras de tamaño m de una distribución normal univariante

$$p(x_1, \dots, x_m | \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^m \exp \left(- \frac{\sum_{i=1}^m (x_i - \mu)^2}{2\sigma^2} \right)$$

$$\sigma \in \mathbf{R}^+ \quad \mu \in \mathbf{R} \quad (x_1, \dots, x_m) \in \mathbf{R}^m$$

Las coordenadas de un individuo vienen dadas por:

$$\frac{\partial \log p(x_1, \dots, x_m | \mu, \sigma^2)}{\partial \mu} = \frac{\sum_{i=1}^m (x_i - \mu)}{\sigma^2}$$

$$\frac{\partial \log p(x_1, \dots, x_m | \mu, \sigma^2)}{\partial \sigma} = \frac{\sum_{i=1}^m (x_i - \mu)^2 - m\sigma^2}{\sigma^3}$$

Teniendo presente que el tensor métrico es de la forma:

$$G = \begin{bmatrix} \frac{m}{\sigma^2} & 0 \\ 0 & \frac{2m}{\sigma^2} \end{bmatrix}$$

a partir de la proposición 3.6. y de [3.15] y [3.16], la distancia entre dos individuos

$\mathbf{x} = (x_1, \dots, x_m)$, $\mathbf{y} = (y_1, \dots, y_m)$ es:

$$d^2(\mathbf{x}, \mathbf{y}) = \frac{m}{2\sigma^4} (2\sigma^2(\bar{x} - \bar{y})^2 + (S^2(\mathbf{x}) - S^2(\mathbf{y}))^2) \quad [3.20]$$

donde $\bar{x} = \frac{1}{m} \sum_{i=1}^m x_i$ y $S^2(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^m (x_i - \mu)^2$.

iv) Muestras de tamaño m de una distribución de Wald (inversa gaussiana).

$$p(x_1, \dots, x_m | \lambda, \mu) = \prod_{i=1}^m \left(\frac{\lambda}{2\pi x_i^3} \right) \frac{1}{2} \exp - \left(\frac{\lambda(x_i - \mu)^2}{2\mu^2 x_i} \right)$$

$$(x_1, \dots, x_m) \in (\mathbb{R}^+)^m \quad \mu, \lambda \in \mathbb{R}^+$$

Las coordenadas de un individuo vienen dadas por:

$$\frac{\partial \log p(x_1, \dots, x_m | \lambda, \mu)}{\partial \mu} = \frac{\lambda}{\mu^2} \left(\frac{\sum_{i=1}^m x_i}{\mu} - m \right)$$

$$\frac{\partial \log p(x_1, \dots, x_m | \lambda, \mu)}{\partial \lambda} = \frac{1}{2} \left(\frac{m}{\lambda} - \frac{\sum_{i=1}^m x_i}{\mu^2} - \frac{m}{\sum_{i=1}^m x_i} \right) + \frac{m}{\mu}$$

El tensor métrico es de la forma:

$$G = \begin{bmatrix} \frac{m}{2\lambda^2} & 0 \\ 0 & \frac{m\lambda}{\mu^3} \end{bmatrix}$$

y la distancia entre dos individuos $\mathbf{x} = (x_1, \dots, x_m)$, $\mathbf{y} = (y_1, \dots, y_m)$ resulta:

$$d^2(\mathbf{x}, \mathbf{y}) = \frac{\lambda^2}{2m\mu^4} (m(\bar{y} - \bar{x}) + \mu^2 \sum_{i=1}^m \frac{x_i - y_i}{x_i y_i})^2 + \frac{\lambda m}{\mu^3} (\bar{x} - \bar{y})^2 \quad [3.21]$$

v) Muestras de tamaño m de una distribución logística.

$$p(x_1, \dots, x_m | \alpha, \beta) = \prod_{i=1}^m \frac{1}{4\beta} \operatorname{sech}^2((x_i - \alpha) / 2\beta)$$

$$(x_1, \dots, x_m) \in \mathbb{R}^m \quad \alpha \in \mathbb{R} \quad \beta \in \mathbb{R}^+$$

Las coordenadas de un individuo vienen dadas por:

$$\frac{\partial \log p(x_1, \dots, x_m | \alpha, \beta)}{\partial \alpha} = \frac{1}{\beta} \sum_{i=1}^m \operatorname{tgh} \left(\frac{x_i - \alpha}{2\beta} \right)$$

$$\frac{\partial \log p(x_1, \dots, x_m | \alpha, \beta)}{\partial \beta} = -\frac{m}{\beta} - \sum_{i=1}^m \frac{x_i - \alpha}{\beta^2} \operatorname{tgh} \left(\frac{x_i - \alpha}{2\beta} \right)$$

El tensor métrico es de la forma:

$$G = \begin{bmatrix} \frac{m}{3\beta^2} & 0 \\ 0 & \frac{m(\pi^2 + 3)}{9\beta^2} \end{bmatrix}$$

y la distancia entre dos individuos $\mathbf{x} = (x_1, \dots, x_m)$ e $\mathbf{y} = (y_1, \dots, y_m)$ resulta:

$$d^2(\mathbf{x}, \mathbf{y}) = 3 \left[\sum_{i=1}^m (T(x_i) - T(y_i)) \right]^2 + \frac{9}{\beta^2 (\pi^2 + 3)} \left[\sum_{i=1}^m \{x_i T(x_i) - y_i T(y_i) + \alpha(T(x_i) - T(y_i))\} \right]^2$$

[3.22]

donde $T(x_i) = \operatorname{tgh}\left(\frac{x_i - \alpha}{2\beta}\right)$.

vi) Muestras de tamaño 1 de una distribución normal multivariante con matriz de covarianzas conocida.

$$p(X|\mu, \Sigma_0) = (2\pi)^{-\frac{n}{2}} |\Sigma_0|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(X - \mu)' \Sigma_0^{-1} (X - \mu)\right)$$

$$X, \mu \in \mathbf{R}^n, \quad \Sigma_0 \in P_n(\mathbf{R})$$

donde $P_n(\mathbf{R})$ es el grupo de matrices regulares simétricas y definidas positivas a coeficientes reales.

Las coordenadas de un individuo las expresamos en notación matricial:

$$\frac{\partial \log p(X|\mu, \Sigma_0)}{\partial \mu} = \Sigma_0^{-1} (X - \mu)$$

El tensor métrico es:

$$G^{-1}(\mathbf{0}) = \Sigma_0$$

y la distancia entre dos individuos X e Y es:

$$d^2(X, Y) = (X - Y)' \Sigma_0^{-1} (X - Y) \quad [3.23]$$

expresión que coincide formalmente, con la conocida distancia de Mahalanobis, aunque esta última definida entre poblaciones, Mahalanobis (1936).

vii) Muestras de tamaño 1 de una distribución normal multivariante con vector de medias conocido.

$$p(X|\mu_0, \Sigma) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(X - \mu_0)' \Sigma^{-1} (X - \mu_0)\right)$$

$$X, \mu_0 \in \mathbf{R}^n \quad \Sigma \in P_n(\mathbf{R})$$

donde $P_n(\mathbf{R})$ es el grupo de matrices regulares simétricas y definidas positivas a coeficientes reales.

Las coordenadas de un individuo vienen dadas por:

$$\frac{\partial \log p}{\partial \theta_\alpha} = -\frac{1}{2} \left[\frac{1}{|\Sigma|} \frac{\partial |\Sigma|}{\partial \sigma_{ij}} + \sum_{p=1}^n \sum_{q=1}^n \left(\frac{\delta_{ij}}{2} - 1 \right) (\sigma^{ip} \sigma^{qj} + \sigma^{pj} \sigma^{qi}) (x_p - m_p)(x_q - m_q) \right]$$

donde $\theta_\alpha = \sigma_{ij}$ $\alpha = 1, \dots, \frac{n(n+1)}{2}$ con $\alpha = (i-1)(n - \frac{i}{2}) + j$ y $i \leq j$ y donde σ^{ij} son los elementos de Σ^{-1} .

El tensor métrico resulta:

$$G^{-1}(0) = H$$

donde $H \in P_{\frac{n(n+1)}{2}}$ definida por $h_{\alpha\beta} = \sigma_{ik} \sigma_{jl} + \sigma_{il} \sigma_{jk}$, donde $\alpha, \beta, i, j, k, l = 1, \dots, n$ con $i \leq j$ y $k \leq l$.

La distancia entre dos individuos $X = (x_1, \dots, x_n)$ e $Y = (y_1, \dots, y_n)$ resulta finalmente:

$$d^2(X, Y) = \frac{1}{4} \left\{ \sum_{\substack{i, j, k, l=1 \\ i \leq j \quad k \leq l}}^n (\sigma_{ik} \sigma_{jl} + \sigma_{il} \sigma_{jk}) \left(\frac{\delta_{ij}}{2} - 1 \right) \left(\frac{\delta_{kl}}{2} - 1 \right) \right. \\ \left. \left[\sum_{p, q=1}^n (\sigma^{ip} \sigma^{qj} + \sigma^{pj} \sigma^{qi}) (y_p - m_p)(y_q - m_q) - (x_p - m_p)(x_q - m_q) \right] \right. \\ \left. \left[\sum_{p, q=1}^n (\sigma^{pk} \sigma^{ql} + \sigma^{pl} \sigma^{qk}) (y_p - m_p)(y_q - m_q) - (x_p - m_p)(x_q - m_q) \right] \right\} \quad [3.24]$$

donde $\mu_0 = (m_1, \dots, m_n)$.

En notación matricial llegamos a la siguiente expresión:

$$d^2(X, Y) = \frac{1}{2} [(\Delta Y)' \Sigma^{-1} \Delta Y)^2 + (\Delta X)' \Sigma^{-1} \Delta X)^2 - 2 (\Delta X)' \Sigma^{-1} \Delta Y)^2] \quad [3.25]$$

donde $\Delta X = X - \mu_0$ y $\Delta Y = Y - \mu_0$.

viii) Muestras de tamaño 1 de una distribución normal multivariante.

$$p(X|\mu, \Sigma) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(X - \mu)' \Sigma^{-1} (X - \mu)\right)$$

$$X, M \in \mathbb{R}^n \quad \Sigma \in P_n(\mathbb{R})$$

donde $P_n(\mathbb{R})$ es el grupo de matrices regulares simétricas y definidas positivas a coeficientes reales.

Teniendo en cuenta que el tensor métrico es de la forma:

$$G = \begin{bmatrix} \Sigma^{-1} & 0 \\ 0 & H^{-1} \end{bmatrix} \in P_{\frac{n(n+1)}{2}}(\mathbb{R})$$

III - Distancias entre individuos estadísticos

donde H tiene el mismo significado que en el apartado anterior, y considerando la proposición 3.6. la distancia entre dos individuos $X = (x_1, \dots, x_n)$ e $Y = (y_1, \dots, y_n)$ resulta ser la suma de las expresiones [3.23] y [3.25].

$$d^2(X, Y) = (X - Y)' \Sigma^{-1} (X - Y) + \frac{1}{2} [(\Delta Y' \Sigma^{-1} \Delta Y)^2 + (\Delta X' \Sigma^{-1} \Delta X)^2 - 2(\Delta X' \Sigma^{-1} \Delta Y)^2] \quad [3.26]$$

donde $\Delta X = X - \mu$ y $\Delta Y = Y - \mu$.

3.4. Distancia estructural

En el desarrollo del presente apartado, (χ, A) será un espacio medible, con la suposición adicional de que χ tiene una estructura de espacio de Banach real $(\chi, \|\cdot\|_\chi)$ de dimensión $m < \infty$, donde $\|\cdot\|_\chi$ es una norma en χ que tendrá asociada una métrica d .

La distancia al cuadrado entre dos puntos próximos $\mathbf{x}, \mathbf{x} + d\mathbf{x} \in \chi$, vendrá dada utilizando dicha métrica por:

$$ds^2 = \|d\mathbf{x}\|_\chi^2$$

La definición de la aplicación [3.5] $\delta: \chi \rightarrow E_0^*$, nos permite la introducción de una métrica en χ definiendo la distancia entre dos puntos $\mathbf{x}, \mathbf{x} + d\mathbf{x} \in \chi$, como la distancia en S entre sus respectivas formas lineales asociadas, donde S es el conjunto imagen de χ por la aplicación δ , $\delta(\chi) = S$.

$$\begin{aligned} \delta(\mathbf{x}) &= Y^* \\ \delta(\mathbf{x} + d\mathbf{x}) &= Y^* + dY^* \end{aligned}$$

Bajo condiciones generales, S es una variedad, donde introduciendo la métrica $G^{-1}(\theta)$ el elemento de línea resulta:

$$ds^2 = g^{ij} dy_i dy_j \quad ij = 1, \dots, n \quad [3.27]$$

donde y_i son las coordenadas en E_0^* .

Continúa siendo válido el comentario efectuado en el apartado 3.2. en el sentido de que es sobre las clases de equivalencia de χ sobre las cuales está realmente definida la distancia, a pesar de referirnos a elementos de χ .

III - Distancias entre individuos estadísticos

Tal y como desarrollaremos a continuación, es posible, bajo condiciones generales, trasladar el cálculo de la distancia al espacio χ introduciendo en este último una forma cuadrática diferencial mediante una transformación de las coordenadas.

Impongamos como condición de regularidad la existencia de las derivadas parciales $\frac{\partial y_i}{\partial x_\alpha}$, es posible entonces efectuar la transformación siguiente:

$$dy_i = \frac{\partial^2 \log p(\mathbf{x}|\theta)}{\partial \theta_i \partial x_\alpha} dx_\alpha \quad \alpha = 1, \dots, m \quad i = 1, \dots, n \quad [3.28]$$

resultando:

$$\begin{aligned} ds^2 &= g^{ij} \frac{\partial^2 \log p(\mathbf{x}|\theta)}{\partial \theta_i \partial x_\alpha} dx_\alpha \frac{\partial^2 \log p(\mathbf{x}|\theta)}{\partial \theta_j \partial x_\beta} dx_\beta = \\ &= g^{ij} \frac{\partial^2 \log p(\mathbf{x}|\theta)}{\partial \theta_i \partial x_\alpha} \frac{\partial^2 \log p(\mathbf{x}|\theta)}{\partial \theta_j \partial x_\beta} dx_\alpha dx_\beta \end{aligned}$$

donde llamando:

$$g_{\alpha\beta} = g^{ij} \frac{\partial^2 \log p(\mathbf{x}|\theta)}{\partial \theta_i \partial x_\alpha} \frac{\partial^2 \log p(\mathbf{x}|\theta)}{\partial \theta_j \partial x_\beta} \quad [3.29]$$

podemos escribir:

$$ds^2 = g_{\alpha\beta} dx_\alpha dx_\beta \quad [3.30]$$

La matriz $G = [g_{ij}]$ son las componentes de un tensor de segundo orden, covariante simétrico y en general semidefinido positivo. La variedad resultante será por tanto en general una variedad semi-Riemanniana.

Siguiendo el criterio utilizado hasta este momento, parece más conveniente, en aquellos casos en que sea posible, la utilización de la distancia en el conjunto $\delta(\chi) = S$, en lugar de la distancia euclídea ordinaria en el espacio tangente dual, puesto que supone en cierta manera aprovechar más la información proporcionada por la organización del "espacio" en el cual se encuentran los individuos.

Tengamos en cuenta los siguientes resultados:

Proposición 3.7. Si el subconjunto del espacio tangente dual que corresponde a la imagen de χ por la aplicación δ definida en [3.5], $\delta(\chi) = S$, es un conjunto convexo, la distancia estructural entre dos formas lineales pertenecientes a S , coincide con su distancia inmediata en E_0^* .

Demostración. Evidente a partir de las definiciones proporcionadas en el presente capítulo ■

Corolario. La distancia inmediata [3.7] basada en la estructura euclídea del espacio tangente dual coincide con la distancia estructural si los individuos pertenecen a distribuciones uniparamétricas.

Demostración. Evidente a partir de la Proposición 3.7., al ser en este caso la dimensión de E_0^* igual a 1. ■

Interesará estudiar la convexidad del conjunto S para decidir la utilización de una u otra distancia. Veamos el siguiente resultado:

Proposición 3.8. Sea la dimensión del espacio muestral igual a 1. Si las derivadas parciales $\partial_i \log p(x|\theta)$ $i = 1, \dots, n$ son linealmente independientes, el conjunto S no es convexo.

Demostración. En efecto, si la dimensión de χ es 1, S representa una curva en E_0^* , y será convexo únicamente si dicha curva es una recta continua. Las coordenadas de un individuo x serán $\partial_i \log p(x|\theta)$ $i = 1, \dots, n$. Si S es una recta las coordenadas serán una función lineal

$$\partial_i \log p(x|\theta) = \lambda(x)v_i + t_i \quad i = 1, \dots, n$$

donde $\lambda(x)$ es una función continua.

Debido a que $E(\partial_i \log p(x|\theta)) = 0$ $i = 1, \dots, n$, obtenemos que $t_i = -v_i E(\lambda(x))$ $i = 1, \dots, n$, y por tanto:

$$\partial_i \log p(x|\theta) = v_i(\lambda(x) - E(\lambda(x))) = v_i \mu(x)$$

con lo que dadas dos coordenadas:

$$\partial_i \log p(x|\theta) = v_i \mu(x)$$

$$\partial_j \log p(x|\theta) = v_j \mu(x)$$

se comprueba inmediatamente su dependencia lineal.

Si S es convexo las derivadas parciales del logaritmo de la densidad respecto a los parámetros son linealmente dependientes dos a dos ■

Como un caso particular de la proposición anterior consideremos que ocurre cuando los individuos son muestras de tamaño 1 de una distribución normal univariante.

Proposición 3.9. La imagen del espacio muestral χ por la aplicación [3.5] cuando los individuos son muestras de tamaño m de una distribución normal univariante, es un subconjunto convexo de E_0^* si y sólo si $m \geq 2$.

Demostración. Las coordenadas son:

$$\frac{\partial \log p}{\partial \mu} = \frac{\sum_{i=1}^m (x_i - \mu)}{\sigma^2}$$

$$\frac{\partial \log p}{\partial \sigma} = \frac{\sum_{i=1}^m (x_i - \mu)^2 - m\sigma^2}{\sigma^3}$$

El subconjunto $S = \delta(\chi)$ será convexo si para cualquier $x, y \in S$ y para $0 \leq t \leq 1$ se cumple que $tx + (1-t)y \in S$. Se debe cumplir por tanto la existencia de unos valores z_1, \dots, z_m tales que:

$$t \frac{\sum_{i=1}^m (x_i - \mu)}{\sigma^2} + (1-t) \frac{\sum_{i=1}^m (y_i - \mu)}{\sigma^2} = \frac{\sum_{i=1}^m (z_i - \mu)}{\sigma^2}$$

$$t \frac{\sum_{i=1}^m (x_i - \mu)^2 - m\sigma^2}{\sigma^3} + (1-t) \frac{\sum_{i=1}^m (y_i - \mu)^2 - m\sigma^2}{\sigma^3} = \frac{\sum_{i=1}^m (z_i - \mu)^2 - m\sigma^2}{\sigma^3}$$

llegándose finalmente a las igualdades:

$$\bar{z} = t\bar{x} + (1-t)\bar{y}$$

$$\overline{(z^2)} = t\overline{(x^2)} + (1-t)\overline{(y^2)} \quad [3.31]$$

donde $m\bar{x} = \sum_{i=1}^m x_i$ y $m\overline{(x^2)} = \sum_{i=1}^m x_i^2$.

Este sistema es incompatible para muestras de tamaño uno, con lo cual el conjunto S no será convexo, tal y como ya sabíamos debido a la proposición 3.8. Vamos a

comprobar ahora que el sistema tiene solución para muestras de tamaño mayor o igual a dos.

Llamemos $a = \bar{z}$ y $b = \overline{(z^2)}$ y tomemos soluciones de la forma $z_i = a + h_i$ $i = 1, \dots, m$, resulta entonces que $a = \bar{z} = a + \bar{h}$ y por tanto $\bar{h} = 0$. Además $b = \overline{(z^2)} = a^2 + 2a\bar{h} + \overline{(h^2)} = a^2 + \overline{(h^2)}$. Las soluciones se obtendrán de:

$$\begin{aligned} \bar{h} &= 0 \\ a^2 + \overline{(h^2)} &= b \end{aligned}$$

tomemos $h_j = 0$ $j = 3, \dots, m$ y $h_1 = -h_2$, en estas condiciones $h_1 = (\frac{b - a^2}{2})^{1/2}$. Existirá solución siempre que $b - a^2 \geq 0$, comprobemos que ésto último es cierto siempre. Recordemos que $\overline{(x^2)} \geq (\bar{x})^2$, por tanto :

$$\begin{aligned} b - a^2 &= \overline{(x^2)} + (1-t)\overline{(y^2)} - t^2(\bar{x})^2 - 2t(1-t)\bar{x}\bar{y} - (1-t)^2(\bar{y})^2 \geq \\ &\geq \overline{(x^2)} + (1-t)\overline{(y^2)} - t^2x^2ba - 2t(1-t)\bar{x}\bar{y} - (1-t)^2\overline{(y^2)} = \\ &= t(1-t)\overline{(x^2)} + t(1-t)\overline{(y^2)} - 2t(1-t)\bar{x}\bar{y} = t(1-t)(\overline{(x^2)} + \overline{(y^2)} - 2\bar{x}\bar{y}) \geq \\ &\geq t(1-t)((\bar{x})^2 + (\bar{y})^2 - 2\bar{x}\bar{y}) = t(1-t)(\bar{x} - \bar{y})^2 \geq 0 \quad \blacksquare \end{aligned}$$

La distancia entre individuos que corresponden a muestras de tamaño uno de una distribución Normal univariante, será por tanto conveniente calcularla utilizando la distancia geodésica entre individuos. Procederemos a continuación a su obtención.

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

$$x, \mu \in \mathbf{R} \quad \sigma \in \mathbf{R}^+$$

La matriz inversa del tensor métrico es igual a:

$$g^{11} = \sigma^2 \quad g^{22} = \frac{\sigma^2}{2} \quad g^{12} = g^{21} = 0$$

$$\frac{\partial^2 \log p(x|\mu, \sigma^2)}{\partial \mu \partial x} = \frac{1}{\sigma^2} \frac{\partial^2 \log p(x|\mu, \sigma^2)}{\partial \sigma \partial x} = \frac{2(x - \mu)}{\sigma^3}$$

El elemento de línea resulta:

$$ds^2 = \frac{2(x - \mu)^2 + \sigma^2}{\sigma^4} (dx)^2 \quad [3.32]$$

La distancia entre dos puntos x^1 y x^2 será:

$$s = \left| \int_{x^1}^{x^2} \left(\frac{2(x - \mu)^2 + \sigma^2}{\sigma^4} \right)^{\frac{1}{2}} dx \right| \quad [3.33]$$

Efectuando el cambio $\sqrt{2} \frac{x - \mu}{\sigma} = y$, obtenemos:

$$s = \frac{1}{\sqrt{2}} \int_{y_1}^{y_2} \sqrt{y^2 + 1} \, dy$$

que mediante el nuevo cambio $y = sh \, t = \frac{e^t - e^{-t}}{2}$, resulta:

$$s = \frac{1}{4\sqrt{2}} \left(\frac{1}{2} e^{2t} - \frac{1}{2} e^{-2t} + 2t \right) \Big|_{t_1}^{t_2}$$

o bien:

$$s = \frac{1}{4\sqrt{2}} (sh \, 2 \, t_2 - sh \, 2 \, t_1 + 2t_2 - 2t_1) \quad [3.34]$$

Un cambio alternativo:

$$\frac{x - \mu}{\sigma} = y \quad [3.35]$$

nos lleva a:

$$s = \int_{y_1}^{y_2} \sqrt{2y^2 + 1} \, dy$$

resultando:

$$s = -\frac{\sqrt{2}}{4} \left[\frac{u_1^4 - 1}{4u_1^2} - \frac{u_2^4 - 1}{4u_2^2} + \ln \frac{u_1}{u_2} \right] \quad [3.36]$$

donde $u = \sqrt{1 + 2y^2} - \sqrt{2}y$.

IV. Aplicaciones de las Distancias entre individuos.

En este capítulo presentamos algunas aplicaciones de las distancias entre individuos a la estimación de parámetros, utilizando tanto la distancia inmediata como la estructural, a la verificación de hipótesis estadísticas y a los problemas de discriminación de individuos.

- 4.1. Estimación de parámetros.
- 4.2. Contraste de hipótesis.
- 4.3. Clasificación y discriminación.

4.1. Estimación de Parámetros.

4.1.1. Procedimiento general.

Dentro de las posibles aplicaciones de las distancias entre individuos para enfocar desde un punto de vista geométrico diversos aspectos de la estadística, destaquemos la definición de un método general de estimación basado en consideraciones geométricas.

Sea (χ, A) un espacio medible sobre el que tenemos definido un modelo estadístico representado por $F_{(f, \Theta)} = \{p = f(\cdot, \theta) : \theta \in \Theta\}$, donde $\theta = (\theta_1, \dots, \theta_n)$ son los parámetros de la densidad y Θ el espacio paramétrico.

Sea $\bar{x} = (\bar{x}_1, \dots, \bar{x}_n)$ una muestra aleatoria que suponemos proviene de una población teórica, entendida como un elemento de la familia paramétrica $F_{(f, \Theta)}$ caracterizado por un determinado valor de los parámetros. El problema de la estimación es determinar de alguna forma el valor de dichos parámetros a partir de la información facilitada por la muestra.

El procedimiento propuesto por nosotros, se basa en determinar la distancia entre la muestra obtenida \bar{x} y otra hipotética muestra x suponiendo ambas procedentes de una población estadística caracterizada por un valor $\theta = (\theta_1, \dots, \theta_n)$ de los parámetros, utilizando para ello la distancia en el espacio tangente dual E_0^* entre sus respectivas formas lineales asociadas.

$$d_{\chi}^2(\bar{x}, x) = d_{E_0^*}^2(\delta(\bar{x}), \delta(x)) \quad [4.1]$$

donde $\delta(x) = Y^*$ es la forma lineal asociada con el individuo x a través de la aplicación δ definida en [3.5].

IV - Aplicaciones de las distancias entre individuos

Una forma razonable para estimar el valor real de los parámetros de la población de la cual proviene la muestra \bar{x} puede ser tomar aquel valor de los parámetros $\hat{\theta}$ que minimice la esperanza de la distancia al cuadrado entre \bar{x} y la muestra x .

$$E_{\theta}(d^2(\bar{x}, x)) = \int_{\mathcal{X}} p(x|\theta) d^2(\bar{x}, x) d\mu(x) \quad [4.2]$$

Por tanto, y supuesta la existencia del mínimo para la esperanza de la distancia al cuadrado, la estimación $\hat{\theta}$ verificará:

$$E_{\hat{\theta}}(d^2(\bar{x}, x)) = \min_{\theta \in \Theta} E_{\theta}(d^2(\bar{x}, x)) \quad [4.3]$$

Se trata de un criterio meramente geométrico puesto que dada una muestra asignamos como parámetros aquellos que definen la población estadística que tiene sus individuos, más cercanos, en promedio, a la muestra obtenida.

Debemos hacer notar que el mínimo exigido en [4.3] no tiene porque existir necesariamente, así mismo resaltemos que el resultado de la estimación dependerá de la definición de la distancia entre individuos que utilizemos. Al poseer dos diferentes definiciones para la distancia entre elementos del espacio tangente dual o entre individuos, la distancia inmediata y la distancia estructural, las estimaciones obtenidas diferirán en tanto que las distancias también difieran.

4.1.2. Estimación utilizando la distancia inmediata.

La distancia entre la muestra \bar{x} y una muestra hipotética x , utilizando la distancia euclídea en el espacio tangente dual, resulta, utilizando una notación matricial:

$$d^2(\bar{x}, x) = d_{E_0}^2(\delta(\bar{x}), \delta(x)) =$$

IV - Aplicaciones de las distancias entre individuos

$$= [\partial_0 \log p(\tilde{\mathbf{x}}|\theta) - \partial_0 \log p(\mathbf{x}|\theta)]' G^{-1}(\theta) [\partial_0 \log p(\tilde{\mathbf{x}}|\theta) - \partial_0 \log p(\mathbf{x}|\theta)] \quad [4.4]$$

donde $G^{-1}(\theta)$ es la matriz asociada al producto escalar del espacio tangente dual.

La esperanza de la distancia al cuadrado es en este caso:

$$\begin{aligned} E_0(d^2(\tilde{\mathbf{x}}, \mathbf{x})) &= (\partial_0 \log p(\tilde{\mathbf{x}}|\theta))' G^{-1}(\theta) (\partial_0 \log p(\tilde{\mathbf{x}}|\theta)) \\ &\quad - 2(\partial_0 \log p(\tilde{\mathbf{x}}|\theta))' G^{-1}(\theta) E(\partial_0 \log p(\mathbf{x}|\theta)) + E((\partial_0 \log p(\mathbf{x}|\theta))' G^{-1}(\theta) (\partial_0 \log p(\mathbf{x}|\theta))) = \\ &= (\partial_0 \log p(\tilde{\mathbf{x}}|\theta))' G^{-1}(\theta) (\partial_0 \log p(\tilde{\mathbf{x}}|\theta)) + E(\text{traza}(G^{-1}(\theta) (\partial_0 \log p(\mathbf{x}|\theta)) (\partial_0 \log p(\mathbf{x}|\theta))')) = \\ &= (\partial_0 \log p(\tilde{\mathbf{x}}|\theta))' G^{-1}(\theta) (\partial_0 \log p(\tilde{\mathbf{x}}|\theta)) + \text{traza}(G^{-1}(\theta) E((\partial_0 \log p(\mathbf{x}|\theta)) (\partial_0 \log p(\mathbf{x}|\theta))')) = \\ &= (\partial_0 \log p(\tilde{\mathbf{x}}|\theta))' G^{-1}(\theta) (\partial_0 \log p(\tilde{\mathbf{x}}|\theta)) + n \end{aligned} \quad [4.5]$$

donde hemos teniendo en cuenta que $E(\partial_0 \log p(\mathbf{x}|\theta)) = \mathbf{0}$.

Resulta igual a la suma del número de parámetros y la norma al cuadrado del vector de coordenadas de la muestra $\tilde{\mathbf{x}}$ en el espacio tangente dual.

El mínimo se alcanza, caso de existir, como solución de:

$$\| \partial_0 \log p(\tilde{\mathbf{x}}|\theta) \| = 0 \quad [4.6]$$

Dicha solución, bajo las habituales condiciones de regularidad, se corresponde con la anulación de las n ecuaciones de verosimilitud

$$\frac{\partial \log p(\mathbf{x}|\theta)}{\partial \theta_i} = 0 \quad i = 1, \dots, n \quad [4.7]$$

IV - Aplicaciones de las distancias entre individuos

Los resultados de estimación coincidirán con los obtenidos por el método clásico de maximizar la función de verosimilitud de la muestra, caso de existir un máximo para tal función. Sin embargo el método de minimizar la esperanza del cuadrado de la distancia no impone ninguna restricción sobre los ceros de las ecuaciones de verosimilitud, no exige buscar un máximo para la verosimilitud. En este sentido podríamos considerar que la obtención de estimadores máximo verosímiles implica la obtención de distancia mínima, pero no así a la inversa. No es la primera ocasión en que se plantea la utilización de soluciones de las ecuaciones de verosimilitud que no se corresponden con el máximo de la misma para la obtención de estimaciones, Duda y Hart (1973), intentan evaluar la función de verosimilitud correspondiente a una mezcla de distribuciones normales donde todos los parámetros son desconocidos, y encuentran que la verosimilitud puede hacerse arbitrariamente grande si la varianza tiende a cero, por tanto la solución máximo verosímil es singular. Sin embargo aducen que empíricamente se obtienen estimaciones razonables utilizando el mayor de los máximos locales de la función de verosimilitud. Nosotros iríamos más lejos proponiendo como resultado de la estimación cualquier raíz consistente de las ecuaciones de verosimilitud.

Presentamos a continuación algunos resultados ya conocidos sobre la existencia de soluciones de las ecuaciones de verosimilitud.

Bajo las condiciones de regularidad del capítulo 2 , a las que añadimos:

B1. Las observaciones x_1, \dots, x_m son independientes e idénticamente distribuidas con función de densidad $p(\cdot|\theta)$ $\theta \in \Theta$.

B2. Las distribuciones correspondientes poseen soporte común y son identificables, es decir $\theta \neq \theta'$ implica que $p(\cdot|\theta) \neq p(\cdot|\theta')$.

IV - Aplicaciones de las distancias entre individuos

B3. Existe un abierto $W \subset \Theta$ que contiene el verdadero punto de los parámetros θ_0 tal que para casi todo x existen las derivadas terceras

$$\frac{\partial^3}{\partial \theta_i \partial \theta_j \partial \theta_k} p(x|\theta) \quad \text{para todo } \theta \in W$$

B4. Existen funciones M_{ijk} tales que para todo $i, j, k = 1, \dots, n$

$$\left| \frac{\partial^3}{\partial \theta_i \partial \theta_j \partial \theta_k} \log p(x|\theta) \right| \leq M_{ijk}(x) \quad \text{para todo } \theta \in \Theta$$

donde $E_{\theta_0}(M_{ijk}(x)) < \infty$.

Entonces:

Proposición 4.1. (Teorema de Consistencia de Cramer) Con probabilidad tendiendo a uno cuando $n \rightarrow \infty$ existen soluciones de las ecuaciones de verosimilitud

$$\frac{\partial \log p(x|\theta)}{\partial \theta_i} = 0 \quad i = 1, \dots, n$$

tales que convergen a θ_0 en probabilidad, es decir son consistentes.

Demostración. Puede encontrarse en Lehman (1983).

Destaquemos que el estimador máximo verosímil no coincide necesariamente con la raíz consistente garantizada por la proposición anterior. Por ejemplo Kraft y Lecam (1956) proporcionan un ejemplo en el que el estimador máximo verosímil existe, es único y satisface las ecuaciones de verosimilitud, aunque no es consistente.

En Bandorff-Nielsen (1978) o Makelainen et al. (1981) pueden encontrarse también condiciones suficientes para la existencia o existencia y unicidad de raíces de

las ecuaciones de verosimilitud. Destaquemos que si $p(x|\theta)$ es una familia exponencial multiparamétrica no degenerada, entonces existe como máximo una solución.

4.1.3. Estimación utilizando la distancia estructural.

La posibilidad de realizar la estimación utilizando como distancia entre individuos la distancia geodésica definida en 3.4 nos impone una reflexión sobre la conveniencia de emplear una u otra estimación.

Basándonos en el corolario de la Proposición 3.7., en caso de que los individuos pertenezcan a distribuciones uniparamétricas, la la distancia estructural coincide con la distancia inmediata, y por tanto la estimación utilizando la distancia estructural se reduce a lo comentado en el apartado anterior.

En la Proposición 3.8., hemos demostrado que si el tamaño muestral es $m = 1$, y la dimensión del espacio paramétrico es $n \geq 2$, la subvariedad S no es convexa, las distancias inmediata y estructural difieren, y por tanto los resultados de la estimación también deberán diferir.

Veamos como ejemplo la familia paramétrica definida por la función de densidad:

$$p(x|\mu, \beta) = \frac{1}{\beta} F\left(\left(\frac{x - \mu}{\beta}\right)^2\right)$$

donde $\mu \in \mathbf{R}$, $\beta > 0$ y F es una función $F: \mathbf{R}^+ \cup \{0\} \rightarrow \mathbf{R}^+ \cup \{0\}$ que satisface:

$$\int_{-\infty}^{\infty} F(u^2) du = 1$$

Asumiendo que:

IV - Aplicaciones de las distancias entre individuos

$$a = 4 \int_0^{\infty} t^{1/2} (\mathcal{L}F)^2(t) F(t) dt < \infty$$

y

$$b = \int_{-\infty}^{\infty} (1 + 2(\mathcal{L}F)(u^2)u^2)^2 F(u^2) du < \infty$$

donde $\mathcal{L}F \equiv \frac{F}{F}$. Puede comprobarse fácilmente que el inverso de la matriz de información de Fisher es:

$$g^{11} = \frac{\beta^2}{a} \quad g^{22} = \frac{\beta^2}{b} \quad g^{12} = g^{21} = 0$$

Además suponiendo $\mathcal{L}F$ derivable respecto x :

$$\frac{\partial^2 \log p}{\partial \mu \partial x} = \frac{1}{\beta^2} f \left(\left(\frac{x - \mu}{\beta} \right)^2 \right)$$

$$\frac{\partial^2 \log p}{\partial \beta \partial x} = \frac{1}{\beta^2} g \left(\left(\frac{x - \mu}{\beta} \right)^2 \right) \left(\frac{x - \mu}{\beta} \right)$$

donde $f(u) = -2 \{ \mathcal{L}F(u) + 2(\mathcal{L}F)'(u)u \}$ y $g(u) = -4 \{ (\mathcal{L}F)'(u)u^2 + \mathcal{L}F(u) \}$. Por tanto a partir de [3.29] y [3.30] tenemos:

$$ds^2 = \frac{1}{\beta^2} h^2 \left(\left(\frac{x - \mu}{\beta} \right)^2 \right) (dx)^2$$

donde $h(u) = \left(\frac{1}{a} f^2(u) + \frac{1}{b} g^2(u)u \right)^{1/2}$.

Tendremos que:

$$E(d^2(\tilde{x}, x)) = \int_{-\infty}^{\infty} \left[\int_x^x \frac{1}{\beta} h \left(\left(\frac{y - \mu}{\beta} \right)^2 \right) dy \right]^2 \frac{1}{\beta} F \left(\left(\frac{x - \mu}{\beta} \right)^2 \right) dx$$

IV - Aplicaciones de las distancias entre individuos

donde haciendo los cambios $z = \frac{x - \mu}{\beta}$ y $t = \frac{y - \mu}{\beta}$, resulta:

$$\begin{aligned} E(d^2(\bar{x}, x)) &= \int_{-\infty}^{\infty} \left[\int_z^z h(t^2) dt \right]^2 F(z^2) dz = \\ &= \int_{-\infty}^{\infty} \left[\int_z^0 h(t^2) dt \right]^2 F(z^2) dz + \int_{-\infty}^{\infty} \left[\int_0^z h(t^2) dt \right]^2 F(z^2) dz + \\ &+ 2 \int_{-\infty}^{\infty} \left[\int_z^0 h(t^2) dt \right] \left[\int_0^z h(t^2) dt \right] F(z^2) dz \end{aligned}$$

que al ser el segundo miembro independiente de la muestra y anularse el tercero, resulta

$$E(d^2(\bar{x}, x)) = \left[\int_z^0 h(t^2) dt \right]^2 + K$$

donde K es una constante independiente de la muestra. La estimación que minimiza la esperanza del cuadrado de la distancia será por tanto:

$$\bar{z} = \frac{\bar{x} - \mu}{\beta} = 0$$

es decir $\hat{\mu} = \bar{x}$, quedando $\hat{\beta}$ indeterminada.

Como un caso particular del ejemplo anterior consideremos la distribución normal univariante, donde $F(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$. Por la Proposición 3.9, si el tamaño muestral es superior o igual a 2, S es convexa y por tanto la estimación de los parámetros se realizará buscando raíces de las ecuaciones de verosimilitud, sin embargo en caso de trabajar con muestras de tamaño $m = 1$, la imagen del espacio muestral por la aplicación [3.5], no es un conjunto convexo, y teniendo en cuenta el resultado anterior, la estimación resultante será $\hat{\mu} = \bar{x}$ quedando $\hat{\sigma}$ indeterminado.

El resultado presentado contrasta con el que obtenemos aplicando directamente el método de la máxima verosimilitud, puesto que obtendríamos como estimación

IV - Aplicaciones de las distancias entre individuos

$\hat{\mu} = \tilde{x}$, $\hat{\sigma} = 0$. Creemos más coherente el resultado proporcionado por el método de minimizar la esperanza del cuadrado de la distancia, ya que al disponer de una muestra de tamaño uno, no podemos conjeturar nada sobre el parámetro de dispersión de la distribución y por tanto la estimación que nos permite tomar un valor arbitrario es más razonable que aquella que le asigna el valor 0.

4.2. Contraste de Hipótesis.

4.2.1. Consideraciones generales.

Otra posible aplicación de la distancia entre individuos es la verificación de hipótesis estadísticas paramétricas sobre los resultados de un experimento. Consideremos un cierto contraste de hipótesis definido sobre los parámetros de la densidad, donde la hipótesis nula genera una cierta restricción sobre el espacio paramétrico original Θ , generándose una subvariedad que denominaremos Θ_H , podemos representar dicho contraste por:

$$H_0: \theta \in \Theta_H$$

$$H_1: \theta \in \Theta$$

Para verificar la veracidad de la hipótesis nula dada una muestra $\bar{x} \in \chi$, consideremos los siguientes estadísticos:

$$\min_{\theta \in \Theta_H} E_{\theta}(d^2(\bar{x}, x)) = E_{\theta_1}(d^2(\bar{x}, x)) = E_1(\bar{x})$$

$$\min_{\theta \in \Theta} E_{\theta}(d^2(\bar{x}, x)) = E_{\theta_2}(d^2(\bar{x}, x)) = E_2(\bar{x})$$

Obtenemos E_1 y E_2 sustituyendo en $E_{\theta}(d^2(\bar{x}, x))$ los parámetros $\theta = (\theta_1, \dots, \theta_n)$ por sus estimaciones de distancia mínima dentro de cada una de las variedades paramétricas correspondientes a la hipótesis nula H_0 o a la hipótesis alternativa H_1 .

Podríamos adoptar como criterio de decisión, rechazar la hipótesis nula H_0 , cuando $E_1(\bar{x})$ sea razonablemente mayor que $E_2(\bar{x})$, al no lograrse un mínimo aceptable para la $E(d^2(\bar{x}, x))$. Rechazaríamos por tanto la hipótesis nula cuando el cociente:

$$\lambda = \frac{\bar{E}_1(\bar{\mathbf{x}})}{\bar{E}_2(\bar{\mathbf{x}})}$$

fuese suficientemente grande.

La región crítica del test vendrá dada por tanto por

$$W = \{(\bar{x}_1, \dots, \bar{x}_m) / \lambda \geq \lambda_\varepsilon\} \quad [4.8]$$

donde λ_ε es una constante que vendrá dada por la elección de un coeficiente de significación ε tal que $P(\lambda \geq \lambda_\varepsilon | H_0) \leq \varepsilon$.

4.2.2. Hipótesis nula simple.

4.2.2.1. Distribución asintótica bajo la hipótesis nula.

Supongamos que deseamos contrastar una hipótesis nula del tipo $H_0 : \theta = \theta_0$.

En apartados precedentes hemos comprobado que bajo las condiciones de regularidad enunciadas en el capítulo II, las variables aleatorias $\partial_i \log p(\mathbf{x}|\theta_0)$ cumplen lo siguiente:

$$E(\partial_i \log p(\mathbf{x}|\theta_0)) = 0$$

$$\text{cov}(\partial_i \log p(\mathbf{x}|\theta_0), \partial_j \log p(\mathbf{x}|\theta_0)) = g_{ij}(\theta_0)$$

donde $g_{ij}(\theta_0)$ son los elementos de la matriz de información de Fisher.

Tomemos una muestra aleatoria de tamaño m , $\bar{\mathbf{x}} = (\bar{x}_1, \dots, \bar{x}_m)$ procedente de una población, elemento de la familia paramétrica $F_{(\mathcal{I}, \Theta)}$ caracterizado por un determinado valor de los parámetros, sobre la cual quremos efectuar el contraste definido

anteriormente. Entonces $\partial_l \log p(\bar{x}|\theta_0) = \partial_l \log p(x_1|\theta_0) + \dots + \partial_l \log p(x_m|\theta_0)$ y por consiguiente, obtenemos las expresiones $E(\partial_l \log p(\bar{x}|\theta_0)) = 0$ y $\text{cov}(\partial_l \log p(\bar{x}|\theta_0), \partial_j \log p(\bar{x}|\theta_0)) = m g_{lj}(\theta_0)$.

Para valores de m elevados, $m \rightarrow \infty$, el vector aleatorio $\partial_0 \log p(\bar{x}|\theta_0) = \left(\frac{\partial \log p(\bar{x}|\theta_0)}{\partial \theta_1}, \dots, \frac{\partial \log p(\bar{x}|\theta_0)}{\partial \theta_n} \right)' \xrightarrow{\mathcal{L}} X \sim N(0, G)$, es decir, converge en ley a X donde X es un vector aleatorio n -dimensional distribuido según una normal multivariante con vector de medias cero y matriz de covarianzas igual a la matriz de información de Fisher.

Como consecuencia de lo anterior y debido a las propiedades de la normalidad multivariante:

$$(\partial_0 \log p(\bar{x}|\theta_0))' G^{-1}(0) (\partial_0 \log p(\bar{x}|\theta_0)) \xrightarrow{\mathcal{L}} Y \sim X_n^2 \quad [4.9]$$

es decir, converge en ley a Y , donde Y se distribuye según una ji-cuadrado con n grados de libertad.

La región crítica del test:

$$\begin{aligned} H_0 : \theta &= \theta_0 \\ H_1 : \theta &\neq \theta_0 \end{aligned} \quad [4.10]$$

tal y como la hemos definido en [4.8], teniendo en cuenta el resultado del apartado 4.1.2. , vendrá dada por:

$$W = \{ \bar{x} : (\partial_0 \log p(\bar{x}|\theta_0))' G^{-1}(0) (\partial_0 \log p(\bar{x}|\theta_0)) \geq c_\alpha \} \quad [4.11]$$

El test obtenido en [4.11] ha sido también formulado independientemente bajo la denominación de test de los Multiplicadores de Lagrange, Aitchison y Silvey (1958),

o test de los scores, Tarone (1988), también denominado del gradiente, y utilizado por vez primera por Fisher (1935).

Entre sus propiedades destaca el ser asintóticamente equivalente al test de Wald basado en estimadores máximo verosímiles y al test de la Razón de Verosimilitud, Tarone (1988).

Una de las aplicaciones desarrolladas para el test de los "scores", es en contrastes del tipo $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$, donde generalmente no existe un test localmente de potencia máxima (LMP), si embargo utilizando el estadístico

$$Z^2 = g_{11}^{-1} (\partial_0 \log p(\mathbf{x}|\theta))^2$$

se obtiene el test asintótico insesgado localmente de potencia máxima (LMPU), Wald (1941). Del mismo modo para contrastes con vectores paramétricos n-dimensionales $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$, se considera un estadístico idéntico al definido en [4.11] con una distribución asintótica ji-cuadrado con n grados de libertad. Y para contrastes sobre la igualdad de parámetros entre K poblaciones, se ha desarrollado un test basado en el estadístico:

$$\Xi_H^2 = \sum_{k=1}^K \{S_k(\hat{\theta})\}' \{G_k(\hat{\theta})\}^{-1} \{S_k(\hat{\theta})\}$$

donde $S(\theta) = \partial_0 \log p(\tilde{\mathbf{x}}|\theta)$, y $\hat{\theta}$ es el estimador máximo verosímil basado en los datos de las K poblaciones. Bajo la hipótesis nula de un valor común de los parámetros para las K poblaciones, Ξ_H^2 se distribuye asintóticamente según una ji-cuadrado con $n(K-1)$ grados de libertad.

4.2.2.2. Ejemplos.: i) Distribución normal univariante con varianza conocida.

Deseamos contrastar la hipótesis:

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

Siguiendo el desarrollo precedente es inmediato que:

$$\lambda = \frac{E_1}{E_2} = 1 + \frac{m}{\sigma_0^2} (\mu_0 - \bar{x})^2 \quad [4.12]$$

por tanto si H_0 es cierta: $\frac{m}{\sigma_0^2} (\mu_0 - \bar{x})^2$ sigue una distribución χ^2 con un grado de libertad.

Nótese que en este caso al ser la distribución de

$$\frac{(\bar{x} - \mu_0)}{\sigma_0 / \sqrt{m}}$$

una $N(0,1)$, la distribución del estadístico es exactamente una ji-cuadrado de un grado de libertad, independientemente del tamaño muestral m .

ii) Distribución de Poisson de parámetro λ .

Deseamos contrastar la hipótesis:

$$H_0: \lambda = \lambda_0$$

$$H_1: \lambda \neq \lambda_0$$

Llegamos a obtener:

IV - Aplicaciones de las distancias entre individuos

$$\frac{E_1}{E_2} = 1 + \frac{m}{\lambda_0} (\lambda_0 - \bar{x})^2 \quad [4.13]$$

por tanto si H_0 es cierta $\frac{m}{\lambda_0} (\lambda_0 - \bar{x})^2$ sigue asintóticamente una distribución χ^2 con un grado de libertad.

iii) Distribución Multinomial.

Deseamos contrastar la hipótesis:

$$H_0: \mathbf{p} = \mathbf{p}_0$$

$$H_1: \mathbf{p} \neq \mathbf{p}_0$$

A partir de los resultados obtenidos en 3.3.2. es fácil comprobar que el test definido en [4.11] toma la forma

$$W = \left\{ \mathbf{x} : \sum_{i=1}^{n+1} \frac{x_i^2}{m p_i} - m \geq c_\epsilon \right\} \quad [4.14]$$

y el estadístico $\sum_{i=1}^{n+1} \frac{x_i^2}{m p_i} - m$ si H_0 es cierta sigue asintóticamente una distribución χ^2 con n grados de libertad.

Como puede comprobarse el resultado obtenido en [4.14] coincide con el conocido estadístico χ^2 de Pearson.

4.3. Discriminación y Clasificación.

El problema de asignar un individuo ω a una de varias poblaciones Π_1, \dots, Π_k caracterizada por unos determinados valores de los parámetros $\theta_1, \dots, \theta_n$, puede ser tratada también con ayuda de la distancia entre individuos.

Sea \bar{x} una observación a clasificar entre un conjunto de poblaciones Π_1, \dots, Π_k . Podemos definir la siguiente función discriminante:

$$f_i(\bar{x}) = E_{(\Pi_i)}(d^2(\bar{x}, \mathbf{x})) \quad [4.15]$$

donde \mathbf{x} es al igual que en el caso de la estimación, una hipotética muestra procedente de la población Π_i .

La regla de decisión consistirá en asignar \bar{x} a la población Π_i si:

$$f_i(\bar{x}) = \min \{f_1(\bar{x}), \dots, f_k(\bar{x})\} \quad [4.16]$$

Una posible variante a la función discriminante propuesta también en Cuadras (1989,a), consiste en tomar:

$$f_i(\bar{x}) = d_{\Pi_i}^2(\bar{x}, 0) \quad [4.17]$$

puesto que $E(\partial_i \log p(\mathbf{x}|\theta)) = 0$, pudiéndose considerar la distancia al cuadrado entre la muestra \bar{x} y el individuo "medio" de la población.

Posteriores comentarios a la aplicación de la distancia entre individuos en problemas de clasificación y discriminación, incluyendo la estimación de la distancia entre individuos cuando las poblaciones vienen caracterizadas por muestras aleatorias, puede encontrarse en Cuadras (1989 a,b) y Sanchez (1989).

V. Sobre una clase de funciones de densidad de probabilidad

En este capítulo introducimos una clase de funciones de densidad de probabilidad basadas en la representación en un espacio de Hilbert de las medidas probabilísticas. Las funciones que se obtienen pueden ser caracterizadas en una variedad paramétrica de dimensión finita estudiándose la geometría informacional del modelo. Las variedades resultantes tienen curvatura constante y positiva, y ha sido posible la obtención de una expresión para las geodésicas y la distancia de Rao entre dos distribuciones. Se introducen varios ejemplos con el correspondiente estudio probabilístico y finalmente efectuamos un pequeño comentario sobre la aplicación de dichas familias a la estadística no paramétrica.

- 5.1. Construcción y consideraciones básicas.
- 5.2. Geometría informacional del modelo.
- 5.3. Ejemplos de familias.
- 5.4. Consideraciones finales.

5.1. Construcción y consideraciones básicas.

En ciertas ocasiones, la geometría de algunos modelos estadísticos es bastante complicada, por ejemplo la geometría del modelo normal multivariante, del cual todavía no ha sido posible obtener una expresión explícita para la distancia entre dos distribuciones, vease por ejemplo Eriksen (1986) o Calvo (1988). Este hecho sugiere que quizá sería interesante considerar la construcción de modelos con propiedades geométricas sencillas y con la capacidad de ajustarse de forma adecuada a una amplia clase de muestras, teniendo en cuenta además que en muchas ocasiones no existen razones teóricas para optar entre un modelo u otro.

Dentro de este contexto podemos establecer alguna relación entre los aspectos geométricos del análisis estadístico y la estadística no paramétrica, mediante la representación de las distribuciones en un espacio de Hilbert a partir de un desarrollo en serie utilizando sistemas ortonormales de funciones, donde los parámetros de la serie podrán ser estimados a partir de los datos. Mostraremos que las variedades resultantes serán de curvatura constante y positiva, la geometría será la geometría esférica y obtendremos la fórmula explícita de la distancia entre dos distribuciones.

Sea μ una σ -finita medida aditiva de base numerable, definida sobre una σ -álgebra de los subconjuntos de un espacio medible (χ, A) . M designa el espacio de todas las funciones reales μ -medibles sobre χ y $\mathcal{L}^2(\chi, \mu) \equiv \mathcal{L}^2$ es el espacio de todas las $f \in M$ tales que

$$\int_{\chi} f^2(x) d\mu(x) < \infty$$

\mathcal{L}^2 es un espacio de Hilbert real con el producto escalar y la norma

$$\langle f, g \rangle = \int_{\chi} f(x)g(x) d\mu(x) \quad , \quad \|f\|^2 = \int_{\chi} f^2(x) d\mu(x) \quad f, g \in \mathcal{L}^2$$

V - Sobre una clase de funciones de densidad de probabilidad.

Bajo las suposiciones anteriores, \mathcal{L}^2 es un espacio separable y consecuentemente existen sistemas ortogonales completos de funciones, en particular también ortonormales, de forma que cualquier $f \in \mathcal{L}^2$ puede ser expresado como $f = \sum_{l=1}^{\infty} c_l \varphi_l(x)$ donde $\{\varphi_l(x)\}_{l \in \mathbb{N}}$ es un sistema ortogonal completo y c_l los coeficientes de Fourier de f con respecto a $\{\varphi_l(x)\}_{l \in \mathbb{N}}$, que toman la forma:

$$c_l = \frac{\langle f, \varphi_l \rangle}{\|\varphi_l(x)\|^2} = \frac{1}{\|\varphi_l(x)\|^2} \int_{\mathcal{X}} f(x) \varphi_l(x) d\mu(x)$$

Sea $l^2 = \{(\theta_1, \theta_2, \dots) \in \mathbb{R}^{\infty} : \sum_{l=1}^{\infty} \theta_l^2 < \infty, \}$, l^2 es un espacio de Hilbert con el producto escalar $\langle \theta, \eta \rangle = \sum_{l=1}^{\infty} \theta_l \eta_l$, $\theta, \eta \in l^2$. Podemos también considerar los subconjuntos $S_{\infty}(r) = \{(\theta_1, \theta_2, \dots) \in \mathbb{R}^{\infty} : \sum_{l=1}^{\infty} \theta_l^2 = r^2, \}$, para $0 < r < \infty$, esferas de radio r en l^2 , y $S_n(r) = \{(\theta_1, \dots, \theta_n) \in \mathbb{R}^n : \sum_{l=1}^n \theta_l^2 = r^2, \}$, para $0 < r < \infty$ la esfera de radio r usual en \mathbb{R}^n . Una vez fijado un sistema ortonormal $\{\varphi_l(x)\}_{l \in \mathbb{N}}$, existe un isomorfismo natural entre \mathcal{L}^2 y l^2 de la forma siguiente:

$$\begin{aligned} \xi : \mathcal{L}^2 &\rightarrow l^2 \\ f &\rightarrow \xi(f) = (\theta_1, \theta_2, \dots) \end{aligned} \quad [5.1]$$

donde $(\theta_1, \theta_2, \dots)$ son los coeficientes de Fourier de f con respecto a $\{\varphi_l(x)\}_{l \in \mathbb{N}}$. Los parámetros $(\theta_1, \theta_2, \dots)$ pueden ser considerados como las coordenadas de f con respecto a $\{\varphi_l(x)\}_{l \in \mathbb{N}}$. Cuando por otra parte consideramos $f \in M$ tal que $f = \sum_{l=1}^{\infty} \theta_l \varphi_l(x)$ donde $(\theta_1, \theta_2, \dots) \in S_{\infty}(r)$, debido a que ξ es también una isometría obtenemos un subconjunto de \mathcal{L}^2 que es una esfera de radio r en \mathcal{L}^2 , la llamaremos $\mathcal{L}^2(r)$, $\xi^{-1}(S_{\infty}(r)) = \mathcal{L}^2(r)$, también se verifica que $\xi(\mathcal{L}^2(r)) = S_{\infty}(r)$.

Sea ahora Π el conjunto de medidas probabilísticas sobre (\mathcal{X}, A) tal que $P \ll \mu$ para todo $P \in \Pi$. Consideremos la inmersión de P en \mathcal{L}^2 definida por $P \rightarrow \alpha \left[\frac{dP}{d\mu} \right]^{1/2}$, donde $\alpha \in \mathbb{R}$. Es fácil verificar que la distancia en Π inducida por la distancia en \mathcal{L}^2 permanece invariante bajo cambios de la medida de referencia μ .

Si consideramos la siguiente representación:

$$\alpha \left[\frac{dP}{d\mu} \right]^{1/2} = \sum_{i=1}^{\infty} \theta_i \varphi_i(x) \quad [5.2]$$

donde $\{\varphi_i(x)\}_{i \in \mathbb{N}}$ son un sistema ortonormal completo, entonces $\theta = (\theta_1, \theta_2, \dots) \in S_{\infty}(\alpha)$, puesto que $\alpha \left[\frac{dP}{d\mu} \right]^{1/2} \in \mathcal{L}^2(\alpha)$. La expresión [5.2] nos sugiere una forma de definir una clase de familias de funciones de densidad de probabilidad de dimensión finita de la manera siguiente:

$$p(x|\theta) = \left[\sum_{i=1}^{n+1} \theta_i \varphi_i(x) \right]^2 \quad [5.3]$$

donde $\{\varphi_i(x)\}_{i=1, \dots, n+1}$ son miembros de un sistema ortonormal completo en \mathcal{L}^2 y $\theta = (\theta_1, \dots, \theta_{n+1}) \in S_{n+1}(1)$.

Tomemos ahora la variedad paramétrica n-dimensional $S_{n+1}(1) = \{(\theta_1, \dots, \theta_{n+1}) \in \mathbb{R}^{n+1} : \sum_{i=1}^{n+1} \theta_i^2 = 1\}$, claramente una variedad real n-dimensional C^{∞} conexa y compacta. S_{n+1} puede ser considerada como el espacio paramétrico que represente el conjunto de parámetros para el cual las densidades están definidas $\Theta \equiv S_{n+1}(1)$. Nótese que la condición de identificabilidad, es decir si $\theta \neq \theta'$ y $\theta, \theta' \in \Theta$ entonces $p(x|\theta) \neq p(x|\theta')$, no se cumple en este caso, puesto que obtendríamos la misma densidad si consideráramos los conjuntos $\theta = (\theta_1, \dots, \theta_{n+1})$ y $\theta' = (-\theta_1, \dots, -\theta_{n+1})$. Es conveniente introducir sobre la familia de densidades de probabilidad la condición de identificabilidad, que puede conseguirse introduciendo ciertas restricciones sobre el espacio paramétrico Θ , dando lugar a nuevos espacios paramétricos restringidos Θ_c .

Una condición suficiente de identificabilidad, puede ser dada en caso de que χ sea un espacio topológico, con A la correspondiente álgebra de Borel generada por la topología, a través de la siguiente

Proposición 5.1. Sea $p(x|\theta) = \left[\sum_{i=1}^{n+1} \theta_i \varphi_i(x) \right]^2$ una familia de densidades de probabilidad definidas como en [5.3], y consideremos el siguiente espacio paramétrico $\Theta_c = \{S_{n+1} : v'\theta > 0\}$, donde fijado $v \in \mathbb{R}^{n+1}$, $v'\theta > 0$ para toda función de la familia. Supongamos que toda $\varphi_i(x)$ $i = 1, \dots, n+1$ es una función continua en χ y que $\varphi_1(x), \dots, \varphi_{n+1}(x)$ son linealmente independientes en todo abierto no vacío de χ . Sea $\gamma(x) = \theta_1 \varphi_1(x) + \dots + \theta_{n+1} \varphi_{n+1}(x)$ y $\beta(x) = \bar{\theta}_1 \varphi_1(x) + \dots + \bar{\theta}_{n+1} \varphi_{n+1}(x)$ con $\gamma^2(x) = \beta^2(x)$ entonces $\theta_i = \bar{\theta}_i$ $i = 1, \dots, n+1$, y la familia antes mencionada es por tanto identificable.

Demostración. Debido a que $\int_{\chi} \gamma^2 d\mu = \int_{\chi} \beta^2 d\mu = 1$ existe un punto $x \in \chi$ tal que $\gamma^2(x) = \beta^2(x) \neq 0$, y consecuentemente existe un abierto V con $x \in V$ tal que en cada punto $c \in V$, $\gamma(x)$ y $\gamma(c)$ tienen el mismo signo. De idéntica manera y debido a que β es también una función continua, existe un abierto W con $x \in W$ tal que en todo punto $z \in W$, $\beta(x)$ y $\beta(z)$ tienen el mismo signo. Esto significa que $\gamma(x) = \beta(x)$ o $\gamma(x) = -\beta(x)$ para todo $x \in W \cap V$. Se sigue que $\sum_{i=1}^{n+1} (\theta_i - \bar{\theta}_i) \varphi_i(x) = 0$ o $\sum_{i=1}^{n+1} (\theta_i + \bar{\theta}_i) \varphi_i(x) = 0$, y debido a la independencia lineal de φ_i en $W \cap V$ concluimos que $\theta_i = \bar{\theta}_i$ o $\theta_i = -\bar{\theta}_i$ para todo $i = 1, \dots, n+1$, pero debido a que $v'\theta > 0$ y $v'\bar{\theta} > 0$, se sigue inmediatamente la proposición ■

El ejemplo 1 de la sección 5.3 ofrece un ejemplo de una familia de densidades de probabilidad discretas, donde la identificabilidad se logra porque los parámetros se identifican con la raíz cuadrada positiva de medidas probabilísticas, dando lugar a un nuevo espacio paramétrico restringido diferente del anterior.

5.2. Geometría informacional del modelo.

Para obtener la distancia entre dos distribuciones, podemos seguir el esquema usual de la geometría riemanniana definiendo en cualquiera de las variedades Θ o Θ_c un campo tensorial métrico a partir de la matriz de información de Fisher, Rao (1945), Burbea y Rao (1982), y consecuentemente el espacio paramétrico se convierte en una variedad Riemanniana n-dimensional, como veremos a continuación.

En primer lugar debemos definir un entorno coordenado de todo punto θ de la variedad. Dado un punto $\theta \in \Theta \equiv S_{n+1}(1)$ consideremos un entorno abierto U de θ . Puesto que $\theta \in S_{n+1}(1)$ $\theta_j \neq 0$ para algún $1 \leq j \leq n+1$, podemos asumir sin pérdida de generalidad que $\theta_{n+1} \neq 0$. La carta local (U, ψ) donde $\psi : U \rightarrow \mathbb{R}^n$ está definida por ejemplo, por $\psi(\theta) = (\theta_1, \dots, \theta_n)$ $\theta \in U$, induce un familia de funciones coordenadas reales definida por $x_i = u_i \circ \psi$ $i = 1, \dots, n$ donde u_i son las coordenadas naturales en \mathbb{R}^n .

A partir de ahora consideramos como variedad paramétrica la variedad n-dimensional $\Theta = S_{n+1}$. Los resultados son esencialmente los mismos para ciertas variedades paramétricas restringidas, como por ejemplo la definida en la **Proposición 5.1** y la correspondiente al ejemplo 1 de la sección 5.3. .

El logaritmo de la función de densidad resulta:

$$\log p(x|\theta) = 2 \log \left(\sum_{k=1}^n \theta_k \varphi_k(x) + (1 - \theta_1^2 - \dots - \theta_n^2)^{1/2} \varphi_{n+1}(x) \right) \quad [5.4]$$

derivando parcialmente respecto θ_i obtenemos:

$$\partial_i \log p(x|\theta) = \frac{2 \left[\varphi_i(x) - \frac{\theta_i \varphi_{n+1}(x)}{(1 - \theta_1^2 - \dots - \theta_n^2)^{1/2}} \right]}{\sum_{k=1}^n \theta_k \varphi_k(x) + (1 - \theta_1^2 - \dots - \theta_n^2)^{1/2} \varphi_{n+1}(x)} \quad [5.5]$$

V - Sobre una clase de funciones de densidad de probabilidad.

resultando por tanto las componentes del campo tensorial métrico:

$$g_{ij}(\theta) = E(\partial_i \log p(x|\theta) \partial_j \log p(x|\theta)) = 4 \left[\delta_{ij} + \frac{\theta_i \theta_j}{1 - \theta_1^2 - \dots - \theta_n^2} \right] \quad i, j = 1, \dots, n \quad [5.6]$$

o en notación matricial:

$$G(\theta) = 4 \left(I + \frac{\theta \theta'}{1 - \theta' \theta} \right)$$

Puede comprobarse que es C^∞ , simétrico, definido positivo y su determinante es igual a:

$$|g_{ij}(\theta)| = \frac{4^n}{1 - \theta_1^2 - \dots - \theta_n^2} \quad [5.7]$$

definiendo por tanto una métrica Riemanniana sobre la variedad.

Pasemos ahora a calcular el inverso del tensor métrico, resultando:

$$g^{ij}(\theta) = \frac{1}{4} (\delta^{ij} - \theta_i \theta_j) \quad i, j = 1, \dots, n \quad [5.8]$$

en notación matricial:

$$G^{-1}(\theta) = \frac{1}{4} (I - \theta \theta')$$

En efecto, puede comprobarse de forma inmediata que:

$$4 \left(I + \frac{\theta \theta'}{1 - \theta' \theta} \right) \frac{1}{4} (I - \theta \theta') = I \quad [5.9]$$

Procedamos al cálculo de los símbolos de Christoffel de primera y segunda especie. Obtenemos en primer lugar:

V - Sobre una clase de funciones de densidad de probabilidad.

$$\frac{\partial g_{ik}}{\partial \theta_j} = \frac{(\delta_{ij}\theta_k + \delta_{jk}\theta_i)(1 - \theta_1^2 - \dots - \theta_n^2) + 2\theta_i\theta_j\theta_k}{(1 - \theta_1^2 - \dots - \theta_n^2)^2} \quad [5.10]$$

resultando por tanto los símbolos de Christoffel de primera especie:

$$[ij,k] = 4 \left[\frac{\delta_{ij}\theta_k}{1 - \theta_1^2 - \dots - \theta_n^2} + \frac{\theta_i\theta_j\theta_k}{(1 - \theta_1^2 - \dots - \theta_n^2)^2} \right] \quad [5.11]$$

$$i,j,k = 1, \dots, n$$

A continuación obtengamos los de segunda especie:

$$\begin{aligned} \Gamma_{jk}^i &= \frac{1}{4} (\delta^{ii} - \theta_i\theta_i) 4 \left[\frac{\delta_{jk}\theta_i}{1 - \theta_1^2 - \dots - \theta_n^2} + \frac{\theta_j\theta_k\theta_i}{(1 - \theta_1^2 - \dots - \theta_n^2)^2} \right] = \\ &= \left[\frac{\delta^{ii}\delta_{jk}\theta_i}{1 - \theta_1^2 - \dots - \theta_n^2} - \frac{\delta_{jk}\theta_i\theta_i^2}{1 - \theta_1^2 - \dots - \theta_n^2} \right] + \\ &+ \left[\frac{\delta^{ii}\theta_j\theta_k\theta_i}{(1 - \theta_1^2 - \dots - \theta_n^2)^2} - \frac{\theta_i\theta_j\theta_k\theta_i^2}{(1 - \theta_1^2 - \dots - \theta_n^2)^2} \right] = \\ &= \frac{\delta_{jk}\theta_i}{1 - \theta_1^2 - \dots - \theta_n^2} - \frac{\delta_{jk}\theta_i(1 - (1 - \theta_1^2 - \dots - \theta_n^2))}{1 - \theta_1^2 - \dots - \theta_n^2} + \\ &+ \frac{\theta_j\theta_k\theta_i}{(1 - \theta_1^2 - \dots - \theta_n^2)^2} - \frac{\theta_i\theta_j\theta_k(1 - (1 - \theta_1^2 - \dots - \theta_n^2))}{(1 - \theta_1^2 - \dots - \theta_n^2)^2} = \\ &= \delta_{jk}\theta_i + \frac{\theta_i\theta_j\theta_k}{1 - \theta_1^2 - \dots - \theta_n^2} \quad [5.12] \end{aligned}$$

$$i,j,k = 1, \dots, n$$

Para obtener la curvatura riemanniana de la variedad necesitamos obtener el tensor de Riemann-Christoffel de primera especie, y previamente el de segunda especie.

V - Sobre una clase de funciones de densidad de probabilidad.

$$\begin{aligned}
 \Gamma_{ik}^{\beta} \Gamma_{\beta j}^{\alpha} &= \left\{ \left[\delta_{ik} \theta_{\beta} + \frac{\theta_i \theta_k \theta_{\beta}}{1 - \theta_1^2 - \dots - \theta_n^2} \right] \left[\delta_{\beta j} \theta_{\alpha} + \frac{\theta_{\beta} \theta_j \theta_{\alpha}}{1 - \theta_1^2 - \dots - \theta_n^2} \right] \right\} = \\
 &= \left[\delta_{ik} \theta_{\alpha} \theta_j + \frac{\theta_{\alpha} \theta_i \theta_k \theta_j + \delta_{ik} \theta_j \theta_{\alpha} (1 - (1 - \theta_1^2 - \dots - \theta_n^2))}{1 - \theta_1^2 - \dots - \theta_n^2} + \frac{\theta_i \theta_k \theta_j \theta_{\alpha} (1 - (1 - \theta_1^2 - \dots - \theta_n^2))}{(1 - \theta_1^2 - \dots - \theta_n^2)^2} \right] = \\
 &= \frac{\delta_{ik} \theta_j \theta_{\alpha}}{1 - \theta_1^2 - \dots - \theta_n^2} + \frac{\theta_i \theta_k \theta_j \theta_{\alpha}}{(1 - \theta_1^2 - \dots - \theta_n^2)^2} \quad [5.13]
 \end{aligned}$$

Obtengamos ahora:

$$\Gamma_{ik}^{\beta} \Gamma_{\beta j}^{\alpha} - \Gamma_{ij}^{\beta} \Gamma_{\beta k}^{\alpha} = \frac{\delta_{ik} \theta_j \theta_{\alpha} - \delta_{ij} \theta_k \theta_{\alpha}}{1 - \theta_1^2 - \dots - \theta_n^2} \quad [5.14]$$

$$\frac{\partial \Gamma_{ik}^{\alpha}}{\partial \theta_j} = \delta_{ik} \delta_{aj} + \frac{(\delta_{aj} \theta_i \theta_k + \delta_{ij} \theta_{\alpha} \theta_k + \delta_{kj} \theta_{\alpha} \theta_i) (1 - \theta_1^2 - \dots - \theta_n^2) + 2 \theta_j \theta_{\alpha} \theta_i \theta_k}{(1 - \theta_1^2 - \dots - \theta_n^2)^2} \quad [5.15]$$

El tensor de Riemann-Christoffel de segunda especie resulta finalmente:

$$\begin{aligned}
 R_{ijk}^{\alpha} &= \frac{\partial \Gamma_{ik}^{\alpha}}{\partial \theta_j} - \frac{\partial \Gamma_{ij}^{\alpha}}{\partial \theta_k} + \Gamma_{ik}^m \Gamma_{mj}^{\alpha} - \Gamma_{ij}^m \Gamma_{mk}^{\alpha} = \\
 &= \delta_{ik} \delta_{aj} - \delta_{ij} \delta_{ak} + \frac{\delta_{aj} \theta_i \theta_k - \delta_{ak} \theta_i \theta_j}{1 - \theta_1^2 - \dots - \theta_n^2} \\
 & \quad a, i, j, k = 1, \dots, n
 \end{aligned} \quad [5.16]$$

El tensor de Riemann-Christoffel de primera especie se obtiene a partir de:

$$\begin{aligned}
 R_{hijk} &= R_{ijk}^{\alpha} g_{ah} = \\
 &= \sum_{\alpha=1}^n \left\{ \left[\delta_{ik} \delta_{aj} - \delta_{ij} \delta_{ak} + \frac{\delta_{aj} \theta_i \theta_k - \delta_{ak} \theta_i \theta_j}{1 - \theta_1^2 - \dots - \theta_n^2} \right] 4 \left[\delta_{ak} + \frac{\theta_{\alpha} \theta_h}{1 - \theta_1^2 - \dots - \theta_n^2} \right] \right\} =
 \end{aligned}$$

V - Sobre una clase de funciones de densidad de probabilidad.

$$= 4 \left[\delta_{ik} \delta_{kj} - \delta_{ij} \delta_{kh} + \frac{\delta_{jh} \theta_i \theta_k - \delta_{hk} \theta_i \theta_j}{1 - \theta_1^2 - \dots - \theta_n^2} + \frac{\delta_{ik} \theta_j \theta_h - \delta_{ij} \theta_k \theta_h}{1 - \theta_1^2 - \dots - \theta_n^2} \right] \quad [5.17]$$

$$h, i, j, k = 1, \dots, n$$

Por otra parte:

$$g_{hj} g_{ik} = 15 \left[\delta_{hj} \delta_{ik} + \frac{\delta_{hj} \theta_i \theta_k + \delta_{ik} \theta_h \theta_j}{1 - \theta_1^2 - \dots - \theta_n^2} + \frac{\theta_h \theta_j \theta_i \theta_k}{(1 - \theta_1^2 - \dots - \theta_n^2)^2} \right] \quad [5.18]$$

y

$$\begin{aligned} g_{hj} g_{ik} - g_{hk} g_{ij} &= \\ &= 15 \left[\delta_{hj} \delta_{ik} - \delta_{hk} \delta_{ij} + \frac{\delta_{hj} \theta_i \theta_k + \delta_{ik} \theta_h \theta_j - \delta_{hk} \theta_i \theta_j - \delta_{ij} \theta_h \theta_k}{1 - \theta_1^2 - \dots - \theta_n^2} \right] \quad [5.19] \end{aligned}$$

Resultando por tanto:

$$R_{hijk} = \frac{1}{4} (g_{hj} g_{ik} - g_{hk} g_{ij}) \quad [5.20]$$

$$h, i, j, k = 1, \dots, n$$

y la curvatura riemanniana es:

$$\kappa = \frac{1}{4} \quad [5.21]$$

positiva y constante en todo el espacio e independiente de la orientación del plano que consideremos, es decir el espacio es isótropo.

Nuestro objetivo es ahora hallar la distancia entre dos puntos de la variedad paramétrica Θ . Para ello debemos hallar la curva geodésica que conecta a ambos. Las ecuaciones diferenciales de las geodésicas son en nuestro caso:

$$\ddot{\theta}_i + \left[\delta_{jk} \theta_l + \frac{\theta_l \theta_j \theta_k}{1 - \theta_1^2 - \dots - \theta_n^2} \right] \dot{\theta}_j \dot{\theta}_k = 0 \quad i = 1, \dots, n \quad [5.22]$$

donde el punto indica derivación respecto el parámetro de la curva s , y donde, al igual que en el resto del capítulo, hacemos uso del convenio de sumación de los índices repetidos.

Deben ser resueltas con la condición de norma unitaria:

$$4 \left[\delta_{\mu\nu} + \frac{\theta_\mu \theta_\nu}{1 - \theta_1^2 - \dots - \theta_n^2} \right] \dot{\theta}_\mu \dot{\theta}_\nu = 1 \quad [5.23]$$

El sistema [5.22] puede escribirse como:

$$\ddot{\theta}_i + \theta_l \sum_{j=1}^n (\dot{\theta}_j)^2 + \frac{\theta_l}{1 - \theta_1^2 - \dots - \theta_n^2} \theta_j \theta_k \dot{\theta}_j \dot{\theta}_k = 0 \quad i = 1, \dots, n \quad [5.24]$$

y de [5.23] obtenemos:

$$\sum_{\mu=1}^n (\dot{\theta}_\mu)^2 + \frac{1}{1 - \theta_1^2 - \dots - \theta_n^2} \theta_\mu \theta_\nu \dot{\theta}_\mu \dot{\theta}_\nu = \frac{1}{4} \quad [5.25]$$

sustituyendo [5.25] en [5.24] obtenemos:

$$4\ddot{\theta}_i + \theta_i = 0 \quad i = 1, \dots, n \quad [5.26]$$

Multipliquemos cada ecuación por θ_i y sumemos las n ecuaciones resultantes, obtenemos:

V - Sobre una clase de funciones de densidad de probabilidad.

$$4\ddot{\theta}_i\theta_i + 1 - (1 - \theta_1^2 - \dots - \theta_n^2) = 0 \quad [5.27]$$

a partir de:

$$\theta_{n+1}^2 = 1 - \theta_1^2 - \dots - \theta_n^2 \quad [5.28]$$

derivemos dos veces respecto el parámetro de la curva, obtenemos:

$$(\dot{\theta}_{n+1})^2 + \ddot{\theta}_{n+1}\theta_{n+1} = - \sum_{i=1}^n [(\dot{\theta}_i)^2 + \ddot{\theta}_i\theta_i] = - \sum_{i=1}^n (\dot{\theta}_i)^2 + \frac{1 - \theta_{n+1}^2}{4} \quad [5.29]$$

a partir de [5.28] derivemos una vez y elevemos al cuadrado,

$$(\theta_{n+1}\dot{\theta}_{n+1})^2 = \theta_i\theta_j\dot{\theta}_i\dot{\theta}_j \quad [5.30]$$

de [5.25] y [5.30] obtenemos:

$$4 \sum_{i=1}^n (\dot{\theta}_i)^2 = 1 - 4(\dot{\theta}_{n+1})^2 \quad [5.31]$$

y sustituyendo en [5.29]:

$$(\dot{\theta}_{n+1})^2 + \ddot{\theta}_{n+1}\theta_{n+1} = (\dot{\theta}_{n+1})^2 - \frac{\theta_{n+1}^2}{4} \quad [5.32]$$

o lo que es lo mismo:

$$\ddot{\theta}_{n+1}\theta_{n+1} + \frac{\theta_{n+1}^2}{4} = 0 \rightarrow 4\ddot{\theta}_{n+1} + \theta_{n+1} = 0 \quad [5.33]$$

por lo que el sistema [5.26] puede escribirse teniendo en cuenta todos los parámetros como:

$$4\ddot{\theta}_i + \theta_i = 0 \quad i = 1, \dots, n + 1 \quad [5.34]$$

V - Sobre una clase de funciones de densidad de probabilidad.

que es un sistema de ecuaciones diferenciales lineales homogéneas de segundo orden cuya solución general viene dada por:

$$\theta_i = A_i \cos \frac{s}{2} + B_i \sin \frac{s}{2} \quad i = 1, \dots, n + 1 \quad [5.35]$$

donde s es la distancia a lo largo de la curva geodésica, y A_i y B_i son constantes de integración que hemos de elegir convenientemente, de forma que dados dos individuos $\mathbf{a} = (a_1, \dots, a_{n+1})$ y $\mathbf{b} = (b_1, \dots, b_{n+1})$ debe suceder que para $s = 0$, $\theta_i = a_i$ y para $s = \rho$ $\theta_i = b_i$, donde ρ es la distancia geodésica. De forma inmediata se obtiene $A_i = a_i$ y a partir de:

$$\sum_{i=1}^n (A_i \cos \frac{s}{2} + B_i \sin \frac{s}{2})^2 = 1 \quad [5.36]$$

siguiendo un desarrollo ya conocido llegamos a $B_i = (b_i - a_i \cos \frac{s}{2})(\sin \frac{s}{2})^{-1}$, y finalmente, teniendo en cuenta la condición de norma unitaria, la distancia geodésica entre dos distribuciones resulta:

$$s(P, P') = 2 \arccos \left(\sum_{i=1}^{n+1} a_i b_i \right) \quad [5.37]$$

donde $\frac{dP}{d\mu} = \left(\sum_{i=1}^{n+1} a_i \varphi_i(x) \right)^2$ con $\sum_{i=1}^{n+1} a_i^2 = 1$ y $\frac{dP'}{d\mu} = \left(\sum_{i=1}^{n+1} b_i \varphi_i(x) \right)^2$ con $\sum_{i=1}^{n+1} b_i^2 = 1$.

La distancia está claramente acotada por $0 \leq s \leq 2\pi$.

Nótese la relación entre nuestra construcción y la distancia de Bhattacharyya entre poblaciones, Bhattacharyya (1943) o posteriores desarrollos Rao (1949), donde siendo Π el conjunto de medidas probabilísticas introducido en el apartado 5.1., las funciones de densidad $p = \frac{dP}{d\mu}$, $P \in \Pi$, son representadas en una porción $P(\alpha)$ sobre

V - Sobre una clase de funciones de densidad de probabilidad.

la superficie de una hiperesfera por medio de las raices cuadradas positivas $p \rightarrow +\alpha\sqrt{p} \in P(\alpha)$, es decir

$$P(\alpha) = \{f \in \mathcal{L}^2 : f = +\alpha\sqrt{p} \quad p = \frac{dP}{d\mu} \quad P \in \Pi\} \quad [5.38]$$

obviamente $f(x) \in \mathbf{R}^+$, $\|f\| = \alpha$.

Sea ahora T la clase de densidades de probabilidad definida en la sección 5.1. , podemos representar cada función de densidad $p \in T$ con $p = g^2$, por g , donde $g(x) = \sum_{i=1}^{n+1} \theta_i \varphi_i(x)$, bajo las suposiciones mencionadas en la sección 5.1. . Nótese que g es una raíz cuadrada de p pero no necesariamente la positiva. Obtenemos una porcion de esfera Q de \mathcal{L}^2 , representada por:

$$Q = \{g \in \mathcal{L}^2 : g(x) = \sum_{i=1}^{n+1} \theta_i \varphi_i(x)\} \quad [5.39]$$

diferente de $P(\alpha)$, puesto que $g(x)$ no pertenece necesariamente a \mathbf{R}^+ . Consecuentemente, se obtendrán diferentes distancias, aunque ambas definidas sobre la superficie de una hiperesfera. Ambas distancias coincidirán si los signos de las funciones g en [5.39] son los mismos. En particular, si el signo es positivo ambas superficies coincidirán.

5.3. Algunos ejemplos de familias.

El primero de los ejemplos presentados corresponde a un caso en el cual trabajamos con una medida discreta y el resto al caso continuo.

5.3.1. Distribuciones multinomiales.

Sean π_1, \dots, π_{n+1} $n+1$ clases formando una partición de una determinada población, cada una de ellas de probabilidad p_i . Definamos una medida discreta μ sobre el espacio medible (χ, A) donde $\chi = \mathbf{R}^{n+1}$ y A es una σ -álgebra de conjuntos sobre χ , de la forma siguiente. Sea $E = \{\mathbf{e}_1, \dots, \mathbf{e}_{n+1}\}$ donde $\mathbf{e}_i = (0, \dots, 1, \dots, 0) \in \mathbf{R}_{n+1}$, entonces $\mu(\{\mathbf{e}_i\}) = 1$ $i = 1, \dots, n+1$ y $\mu(B) = 0$ si $B \cap E = \emptyset$. Sea $f_i: \mathbf{R}^{n+1} \rightarrow \mathbf{R}$ $i = 1, \dots, n+1$ una familia de funciones reales definidas del modo siguiente:

$$f_i(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} = \mathbf{e}_i \\ 0 & \mathbf{x} \neq \mathbf{e}_i \end{cases} \quad i = 1, \dots, n+1$$

Obviamente $f_i(\mathbf{x})$ son un sistema ortonormal en \mathcal{L}^2 y por tanto podemos construir una familia de funciones de densidad de probabilidad como:

$$p(\mathbf{x}|\theta) = (\theta_1 f_1(\mathbf{x}) + \dots + \theta_{n+1} f_{n+1}(\mathbf{x}))^2 \quad [5.40]$$

tomando $\theta_i = +\sqrt{p_i}$ obtenemos que la familia es identificable, y la distancia entre dos poblaciones caracterizadas por (p_1, \dots, p_{n+1}) y (q_1, \dots, q_{n+1}) resulta:

$$d = 2 \cos^{-1} \left(\sum_{i=1}^{n+1} \sqrt{p_i q_i} \right) \quad [5.41]$$

V - Sobre una clase de funciones de densidad de probabilidad.

distancia de Rao y también de Bhattacharyya entre dos distribuciones multinomiales, tal y como se puede ver por ejemplo en Atkinson y Mitchell (1981) u Oller (1982). En este caso la distancia está acotada por $0 \leq d \leq \pi$.

En general, para cualquier miembro de la clase de funciones de densidad de probabilidad definida en [5.3] cualquier momento, supuesta su existencia, se puede obtener a partir de: $E(X^k) = \int_{\mathcal{X}} x^k \left(\sum_{l=1}^{n+1} \theta_l \varphi_l(x) \right)^2 d\mu$ también expresable a través de la forma cuadrática $E(X^k) = \theta' B_k \theta$, donde $\theta = (\theta_1, \dots, \theta_{n+1})'$ y B_k es una matriz simétrica de orden $n+1$ de la forma:

$$b_{ij}^k = \int_{\mathcal{X}} x^k \varphi_i(x) \varphi_j(x) d\mu \quad [5.42]$$

resultando, en este caso, el conocido resultado:

$$\begin{aligned} E(\mathbf{x}) &= (E(x_1), \dots, E(x_{n+1})) = \sum_{l,j,k=1}^{n+1} \mathbf{e}_k \sqrt{p_l} \sqrt{p_j} f_l(\mathbf{e}_k) f_j(\mathbf{e}_k) = \\ &= \sum_{l=1}^{n+1} \mathbf{e}_l p_l f_l^2(\mathbf{e}_l) = (p_1, \dots, p_{n+1}) \end{aligned} \quad [5.43]$$

La covarianza se obtiene a partir de:

$$E(x_i x_j) = \sum_{\alpha, \beta, \gamma=1}^{n+1} \delta_{ij} \delta_{\alpha\gamma} \delta_{\beta\gamma} \sqrt{p_\alpha} \sqrt{p_\beta} = \delta_{ij} p_i$$

donde δ_{ij} son las delta de Kronecker, resultando como es natural:

$$\text{cov}(x_i, x_j) = \begin{cases} p_i(1 - p_i) & \text{si } i = j \\ -p_i p_j & \text{si } i \neq j \end{cases} \quad [5.44]$$

En los siguientes tres ejemplos, los espacios paramétricos, teniendo en cuenta la proposición 5.1 para lograr identificabilidad vendrán dados por $\Theta = \{(\theta_0, \dots, \theta_n) \in \mathbf{R}^{n+1} : \sum_{l=0}^n \theta_l^2 = 1, \theta_j > 0\}$ para un parámetro j fijo.

5.3.2. Sistema de funciones de Hermite.

Sea μ la medida de Lebesgue sobre la recta real $(-\infty, +\infty)$ y tomemos el sistema de funciones de Hermite normalizado

$$h_i(x) = H_i(x) e^{-x^2/2} 2^{-i/2} (i!)^{-1/2} \pi^{-1/4} \quad i = 0, 1, 2, \dots \quad [5.45]$$

donde $H_i(x) = (-1)^i e^{x^2} \frac{d^i}{dx^i} (e^{-x^2})$. La correspondiente familia de funciones de densidad resulta:

$$p(x|\theta) = \left(\sum_{i=0}^n \theta_i h_i(x) \right)^2 \quad [5.46]$$

con las convenientes restricciones sobre los parámetros θ_i

Los polinomios de Hermite normalizados satisfacen las siguientes fórmulas de recurrencia, Courant y Hilbert (1953):

$$x h_i(x) = \left(\frac{i}{2}\right)^{1/2} h_{i-1}(x) + \left(\frac{i+1}{2}\right)^{1/2} h_{i+1}(x) \quad i = 1, 2, \dots \quad [5.47]$$

a partir de [5.42] la esperanza puede ser expresada únicamente respecto a los parámetros, resultando tras algunos cálculos:

$$E(X) = \sqrt{2} \sum_{i=0}^{n-1} \theta_i \theta_{i+1} (i+1)^{1/2} \quad [5.48]$$

De la misma forma podemos obtener:

$$\begin{aligned} \text{var}(X) &= \sum_{i=0}^n \left(i + \frac{1}{2}\right) \theta_i^2 + \sum_{i=0}^{n-2} \theta_i \theta_{i+2} \left\{ (i+1)(i+2) \right\}^{1/2} - \\ &\quad - \left[\sqrt{2} \sum_{i=0}^{n-1} \theta_i \theta_{i+1} (i+1)^{1/2} \right]^2 \end{aligned} \quad [5.49]$$

V - Sobre una clase de funciones de densidad de probabilidad.

Momentos de orden mayor serían más complicados, pero todos ellos se pueden hallar en función únicamente de los parámetros a partir de [5.47], que da lugar a :

$$\int_x x^k h_j h_i dx = \left(\frac{j}{2}\right)^{1/2} \int_x x^{k-1} h_{j-1} h_i dx + \left(\frac{j+1}{2}\right)^{1/2} \int_x x^{k-1} h_{j+1} h_i dx \quad [5.50]$$

La función de distribución puede ser así mismo obtenida a partir de una forma cuadrática del modo siguiente: $F(x) = \theta' C(x) \theta$, donde $\theta = (\theta_1, \dots, \theta_{n+1})'$ y C es una matriz simétrica de orden $n+1$ de la forma:

$$c_{ij} = \int_{-\infty}^x h_i(t) h_j(t) dt \quad [5.51]$$

Para la obtención de los coeficientes de la matriz hemos de considerar dos casos separadamente:

a) $i < j$

$$\int_{-\infty}^x h_i(t) h_j(t) dt = \int_{-\infty}^x \frac{H_i(t) e^{-t^2/2}}{(2^i i! \pi^{1/2})^{1/2}} \frac{H_j(t) e^{-t^2/2}}{(2^j j! \pi^{1/2})^{1/2}} dt = \frac{1}{(2^{i+j} i! j! \pi)^{1/2}} \int_{-\infty}^x H_i(t) H_j(t) e^{-t^2} dt$$

podemos escribir la integral de la forma siguiente:

$$\int_{-\infty}^x H_i(t) H_j(t) e^{-t^2} dt = (-1)^j \int_{-\infty}^x H_i(t) \frac{d^j}{dt^j} e^{-t^2} dt$$

sabemos que $H'_i = 2iH_{i-1}$, por lo tanto haciendo una integración por partes:

$$\begin{aligned} \int_{-\infty}^x H_i(t) \frac{d^j}{dt^j} e^{-t^2} dt &= \left[H_i(t) \frac{d^{j-1}}{dt^{j-1}} e^{-t^2} \right]_{-\infty}^x - \int_{-\infty}^x 2i H_{i-1}(t) \frac{d^{j-1}}{dt^{j-1}} e^{-t^2} dt = \\ &= H_i(x) H_{j-1}(x) (-1)^{j-1} e^{-x^2} - 2i \int_{-\infty}^x H_{i-1}(t) \frac{d^{j-1}}{dt^{j-1}} e^{-t^2} dt \end{aligned}$$

teniendo en cuenta que $\frac{d^j}{dt^j} e^{-t^2} = H_j(t) (-1)^j e^{-t^2}$.

Por repetida integración por partes llegamos finalmente a:

$$\int_{-\infty}^x H_i(t) \frac{d^j}{dt^j} e^{-t^2} dt = \sum_{n=0}^i 2^n \frac{i!}{(i-n)!} H_{i-n}(x) H_{j-n-1}(x) (-1)^{j-1} e^{-x^2}$$

resultando finalmente:

$$\int_{-\infty}^x h_i(t) h_j(t) dt = \left(\frac{-1}{\sqrt{2}}\right) \left(\frac{i!}{j!}\right)^{1/2} \sum_{n=0}^i h_{i-n}(x) h_{j-n-1}(x) \left(\frac{(j-n-1)!}{(i-n)!}\right)^{1/2} \quad [5.52]$$

b) $i = j$

Mediante un desarrollo análogo al anterior llegamos a:

$$\int_{-\infty}^x h_i^2(t) dt = \left(\frac{-1}{\sqrt{2}}\right) \sum_{n=0}^{i-1} \frac{h_{i-n}(x) h_{i-n-1}(x)}{(i-n)^{1/2}} + \pi^{-1/2} \int_{-\infty}^x e^{-t^2} dt \quad [5.53]$$

5.3.3. Sistema de funciones de Laguerre.

Sobre el intervalo $(0, +\infty)$ podemos considerar el sistema formado por los polinomios de Laguerre, el cual en la versión normalizada toma la forma:

$$l_i(x) = \frac{1}{i!} e^{x/2} \frac{d^i}{dx^i} (x^i e^{-x}) \quad i = 0, 1, 2, \dots \quad [5.54]$$

Los polinomios de Laguerre normalizados satisfacen las siguientes fórmulas de recurrencia, Courant y Hilbert (1953):

$$(1 + 2j - x) l_j(x) = (j + 1) l_{j+1}(x) + j l_{j-1}(x) \quad [5.55]$$

a partir de [5.55] puede ser obtenido cualquier momento de orden k , teniendo en cuenta que la matriz [5.42] toma la forma:

V - Sobre una clase de funciones de densidad de probabilidad.

$$\int_x x^k l_j dx = (1 + 2i) \int_x x^{k-1} l_j dx - (i + 1) \int_x x^{k-1} l_{i+1} dx - i \int_x x^{k-1} l_{i-1} dx \quad [5.56]$$

En concreto la esperanza de la función de densidad $p(x|0) = (\sum_{i=0}^n \theta_i l_i(x))^2$ resulta:

$$E(X) = \sum_{i=0}^n \theta_i^2 (1 + 2i) - 2 \sum_{i=0}^{n-1} \theta_i \theta_{i+1} (i + 1) \quad [5.57]$$

y la varianza

$$\begin{aligned} \text{var}(X) = & \sum_{i=0}^n \theta_i^2 (6i^2 + 6i + 2) - 2 \sum_{i=0}^{n-1} \theta_i \theta_{i+1} (4i^2 + 8i + 4) + 2 \sum_{i=0}^{n-2} \theta_i \theta_{i+2} (i^2 + 3i + 2) - \\ & - \left[\sum_{i=0}^n \theta_i^2 (1 + 2i) - 2 \sum_{i=0}^{n-1} \theta_i \theta_{i+1} (i + 1) \right]^2 \end{aligned} \quad [5.58]$$

5.3.4. Sistema de funciones de Legendre.

Sobre el intervalo $(-1, 1)$ podemos trabajar con los polinomios de Legendre normalizados:

$$p_i(x) = \frac{P_i(x)(2i + 1)^{1/2}}{2^{1/2}} \quad i = 0, 1, 2, \dots \quad [5.59]$$

donde $P_i(x) = \frac{1}{2^i i!} \frac{d^i}{dx^i} (x^2 - 1)^i \quad i = 0, 1, \dots$

Satisfacen las siguientes fórmulas de recurrencia, Courant y Hilbert (1953):

$$x p_j(x) = \frac{j + 1}{2j + 1} \left[\frac{2j + 1}{2j + 3} \right]^{1/2} p_{j+1}(x) + \frac{j}{2j + 1} \left[\frac{2j + 1}{2j - 1} \right]^{1/2} p_{j-1}(x) \quad [5.60]$$

La matriz [5.42] que nos permite obtener cualquier momento toma la forma:

V - Sobre una clase de funciones de densidad de probabilidad.

$$\int_x x^k p_i p_j dx = \frac{i+1}{((2i+1)(2i+3))^{1/2}} \int_x x^{k-1} p_{i+1} p_j dx + \frac{i}{((2i+1)(2i-1))^{1/2}} \int_x x^{k-1} p_{i-1} p_j dx \quad [5.61]$$

obtenemos que la esperanza de la función de densidad $p(x|0) = (\sum_{i=0}^n \theta_i p_i(x))^2$ resulta:

$$E(X) = 2 \sum_{i=0}^{n-1} \theta_i \theta_{i+1} \frac{i+1}{((2i+1)(2i+3))^{1/2}} \quad [5.62]$$

y la varianza es:

$$\begin{aligned} \text{var}(X) = \sum_{i=0}^n \theta_i^2 \left[\frac{i^2}{(2i+1)(2i-1)} + \frac{(i+1)^2}{(2i+3)(2i+1)} \right] + 2 \sum_{i=0}^{n-2} \theta_i \theta_{i+2} \left[\frac{(i+1)(i+2)}{(2i+3)((2i+1)(2i+5))^{1/2}} \right] - \\ - \left[2 \sum_{i=0}^{n-1} \theta_i \theta_{i+1} \frac{i+1}{((2i+1)(2i+3))^{1/2}} \right]^2 \quad [5.63] \end{aligned}$$

Es inmediata la generalización del caso anterior al trabajo en un intervalo de longitud arbitraria $[-a, a]$.

5.4. Consideraciones finales.

Un problema fundamental en estadística es proponer modelos que se ajusten a una muestra de observaciones, de forma que posteriores análisis, por ejemplo contraste de hipótesis o clasificaciones, pueden ser llevados a cabo teniendo en cuenta el modelo desarrollado. A veces los datos poseen tal estructura que no existe un modelo sencillo que se ajuste a ellos. Al no poder efectuar ninguna suposición de la distribución de probabilidad del modelo, se debe hacer uso de modelos que puedan emplearse con cualquier tipo de datos y que poseen un número ilimitado de parámetros, éstos son los llamados modelos no paramétricos.

La clase de funciones de densidad de probabilidad que hemos considerado en las secciones previas puede ajustarse con cualquier clase de observaciones. Los parámetros de la función pueden ser estimados a partir de los datos por alguno de los conocidos métodos de estimación. Uno de los más ampliamente utilizados es el método de la máxima verosimilitud (MLE), sin embargo este método presenta algunos problemas cuando tratamos con funciones que poseen un gran número de parámetros y una gran capacidad de ajuste a una muestra. En el siguiente capítulo comprobamos que la estimación máximo verosimil conduce a soluciones de poca utilidad, tal y como es sobradamente conocido y puede encontrarse una abundante bibliografía al respecto. Por tanto es conveniente la utilización de técnicas suplementarias o alternativas, a las cuales se adaptan las funciones estudiadas en el presente capítulo.

Además de obtener una razonable estimación para los parámetros, podemos considerar el problema de efectuar análisis estadísticos. Las técnicas no paramétricas usuales pueden ser complementadas con otras técnicas basadas en consideraciones geométricas, principalmente en las distancias entre funciones de densidad. Debido a que conocemos la geometría informacional del modelo y poseemos una expresión para

V - Sobre una clase de funciones de densidad de probabilidad.

la distancia de Rao entre dos funciones pertenecientes a una familia, es posible aplicar técnicas estadísticas basadas en esa distancia. Señalemos la posibilidad de construir contrastes de hipótesis para comparar dos o más poblaciones basados en la distancia de Rao. Burbea y Oller (1989) proponen varios tests bajo ciertas condiciones de regularidad. También es posible la representación gráfica de las poblaciones en dimensión reducida mediante el análisis de coordenadas principales, técnicas de taxonomía numérica o técnicas de multidimensional scaling, Mardia et al. (1979).

VI. Estimación no paramétrica de la densidad.

Consideramos en el presente capítulo el problema de la estimación de las funciones de densidad introducidas en el capítulo anterior. Comprobamos los inconvenientes de la estimación máximo verosímil y proponemos un algoritmo tipo stepwise basado en ella, que intenta paliar tales inconvenientes. También presentamos algunos resultados desarrollando el procedimiento de estimación propuesto en el capítulo 4, utilizando una aproximación a la distancia estructural.

- 6.1. Estimación no paramétrica de la densidad
- 6.2. Estimación máximo verosímil.
- 6.3. Un algoritmo stepwise de estimación no paramétrica.
- 6.4. Estimación minimizando la esperanza del cuadrado de la distancia.

6.1. Estimación no paramétrica de la densidad.

Comenzaremos desarrollando un procedimiento de estimación en el que suponemos conocida la verdadera función de densidad de probabilidad de la población y deseamos aproximar la misma mediante un elemento de la clase de funciones definida en [5.3].

Supongamos que una población está caracterizada por la función de densidad de probabilidad $p(x)$. Sea $\hat{p}(x)$ una estimación de $p(x)$ de la forma definida en [5.3].

$$\hat{p}(x) \equiv p(x|\hat{\theta}) = \left[\sum_{l=1}^{n+1} \hat{\theta}_l \varphi_l(x) \right]^2 \quad [6.1]$$

Nosotros deseamos que esta estimación minimice la función de error definida como:

$$R(\theta) = \|\sqrt{p(x)} - \sum_{l=1}^{n+1} \theta_l \varphi_l(x)\|^2 \quad [6.2]$$

Queremos por tanto, determinar los coeficientes θ_l que hagan mínimo:

$$\begin{aligned} R(\theta) &= \int_{\mathcal{X}} p(x) - 2\sqrt{p(x)} \sum_{l=1}^{n+1} \theta_l \varphi_l(x) + \left(\sum_{l=1}^{n+1} \theta_l \varphi_l(x) \right)^2 dx = \\ &= \int_{\mathcal{X}} p(x) dx - 2 \sum_{l=1}^{n+1} \theta_l \int_{\mathcal{X}} \sqrt{p(x)} \varphi_l(x) dx + \sum_{l=1}^{n+1} \theta_l^2 \end{aligned}$$

Derivando parcialmente respecto θ_k asumiendo condiciones de regularidad suficientes, obtenemos:

$$\frac{\partial R}{\partial \theta_k} = -2 \int_{\mathcal{X}} \sqrt{p(x)} \varphi_k(x) dx + 2\theta_k \quad k = 1, \dots, n+1$$

y por tanto obtenemos la estimación:

$$\hat{\theta}_k = \int_{\mathcal{X}} \sqrt{p(x)} \varphi_k(x) dx \quad k = 1, \dots, n+1 \quad [6.3]$$

donde normalizando resulta:

$$\hat{\theta}_k = \frac{1}{\|\theta\|} \int_x \sqrt{p(x)} \varphi_k(x) dx \quad [6.4]$$

Nótese que [6.3] es la definición del valor esperado de la función $\frac{\varphi_k(x)}{\sqrt{p(x)}}$, valor que puede ser aproximado a partir de la media muestral:

$$\int_x \sqrt{p(x)} \varphi_k(x) dx \approx \frac{1}{m} \sum_{j=1}^m \frac{\varphi_k(x_j)}{\sqrt{p(x_j)}} \quad [6.5]$$

[6.5] admite de forma inmediata una forma recursiva al añadir una muestra a las existentes. Si llamamos

$$\hat{\theta}_k(m) = \frac{1}{m} \sum_{j=1}^m \frac{\varphi_k(x_j)}{\sqrt{p(x_j)}}$$

es inmediato que:

$$\hat{\theta}_k(m+1) = \frac{1}{m+1} \left[m \hat{\theta}_k(m) + \frac{\varphi_k(x_{m+1})}{\sqrt{p(x_{m+1})}} \right] \quad [6.6]$$

Una vez determinado el valor de los diferentes parámetros podemos formular la función de densidad aplicando [6.1]. Si poseemos la verdadera densidad $p(x)$, el número $n+1$ de términos que debemos escoger, del cual depende evidentemente la calidad de la aproximación, puede ser determinado por comparación directa entre la estimación y la verdadera densidad mediante la distancia

$$d = 2 \cos^{-1} \int_x \sqrt{p(x)} \sum_{i=1}^{n+1} \hat{\theta}_i \varphi_i(x) dx \quad [6.7]$$

que puede ser evaluada de forma numérica. Añadiremos términos hasta que la adición de nuevos términos no provoque una variación significativa en la distancia o se supere cierta tolerancia predeterminada.

6.2. Estimación máximo verosímil.

6.2.1. Planteamiento.

Un enfoque alternativo al expuesto en el apartado precedente y utilizable en el caso frecuente de que la única información que poseamos sobre la función de densidad $p(x)$ venga dada por una muestra aleatoria de tamaño m , x_1, \dots, x_m , es considerar la función de densidad de probabilidad de la población como miembro de una de las familias definidas en [5.3], y estimando los parámetros por el método de la máxima verosimilitud.

El problema de estimar una función de densidad de probabilidad $p \in \mathcal{L}^1(\mu)$ a partir de una muestra aleatoria x_1, \dots, x_m puede ser tratado a partir de los trabajos de Fisher (1922), donde se introduce la llamada función de verosimilitud. Por la verosimilitud de que $p(x) \in \mathcal{L}^1(\mu)$ de lugar a la muestra aleatoria x_1, \dots, x_m entendemos:

$$L(p) = \prod_{i=1}^m p(x_i) \quad [6.8]$$

El procedimiento de estimación por máxima verosimilitud (MLE) consiste en maximizar $L(p)$ sujeto a las restricciones

$$p \in F, \quad \int_{\mathcal{X}} p d\mu = 1, \quad p \geq 0$$

donde F es una determinada clase de funciones de densidad a la cual p pertenece, y que puede considerarse un subespacio de dimensión finita o infinita de $\mathcal{L}^1(\mu)$, $F \subset \mathcal{L}^1(\mu)$.

6.2.2. Resultados generales.

Dada una muestra aleatoria simple x_1, \dots, x_m a partir de la cual queremos estimar los parámetros de un miembro de la clase de funciones definidas en [5.3], la función de verosimilitud es:

$$L = \prod_{j=1}^m \left[\sum_{l=1}^{n+1} \theta_l \varphi_l(x_j) \right]^2 \quad [6.9]$$

y su logaritmo:

$$\log L = 2 \sum_{j=1}^m \log \left(\sum_{l=1}^{n+1} \theta_l \varphi_l(x_j) \right) \quad [6.10]$$

El problema es ahora maximizar el $\log L$ con la condición $\sum_{l=1}^{n+1} \theta_l^2 = 1$.

Resulta el sistema de ecuaciones:

$$2 \sum_{j=1}^m \frac{\varphi_k(x_j)}{\sum_{l=1}^{n+1} \theta_l \varphi_l(x_j)} + 2 \lambda \theta_k = 0 \quad k = 1, \dots, n+1 \quad [6.11]$$

$$\sum_{l=1}^{n+1} \theta_l^2 = 1 \quad [6.12]$$

Multiplicando cada una de las $n+1$ primeras ecuaciones por θ_k y sumando obtenemos:

$$2 \sum_{k=1}^{n+1} \sum_{j=1}^m \frac{\varphi_k(x_j) \theta_k}{\sum_{l=1}^{n+1} \theta_l \varphi_l(x_j)} + 2 \lambda \sum_{k=1}^{n+1} \theta_k^2 = 0 \quad [6.13]$$

sustituyendo [6.12] en [6.13] resulta:

$$2m + 2\lambda = 0 \quad [6.14]$$

por tanto finalmente obtenemos el sistema:

$$\sum_{j=1}^m \frac{\varphi_k(x_j)}{m \sum_{l=1}^{n+1} \theta_l \varphi_l(x_j)} - m \theta_k = 0 \quad k = 1, \dots, n+1 \quad [6.15]$$

sistema de ecuaciones que puede ser resuelto mediante técnicas de cálculo numérico.

El resultado obtenido por máxima verosimilitud es análogo al obtenido en [6.5] considerando $p(x)$ como miembro de las familias [5.3].

Nótese que en el caso $m = 1$, la estimación resulta:

$$\hat{\theta}_k = \frac{\varphi_k(x)}{(\varphi_1^2(x) + \dots + \varphi_{n+1}^2(x))^{1/2}} \quad [6.16]$$

Al no conocer la forma explícita completa de $p(x)$, no podemos determinar por comparación directa el número de términos a utilizar. Además, es un resultado conocido la no existencia en general de un máximo para la función de verosimilitud si el subespacio $F \subset \mathcal{L}^1(\mu)$ al cual pertenece la función de densidad es de dimensión infinita, Tapia y Thompson (1978).

La no existencia de solución máximo verosímil en el caso infinito dimensional se traduce en estimaciones que presentan un fenómeno de **inestabilidad dimensional** cuando se aplican a variedades con un número grande, aunque finito, de parámetros, tal y como es el caso de nuestras familias. Por inestabilidad dimensional entendemos la aparición de irregularidades, "picos", en la estimación de las funciones de densidad al aumentar el número de dimensiones. Este hecho se produce al tender la estimación máximo verosímil en el caso de dimensión infinita, a una combinación de funcionales **delta de Dirac** en los puntos de la muestra.

$$\hat{p}(x) \rightarrow p^*(x) = \frac{1}{m} \sum_{l=1}^m \delta(x - x_l) \quad [6.17]$$

donde δ es la función delta de Dirac, definida por:

$$\int_{-\infty}^{\infty} \delta(y) dy = 1$$

$$\delta(y) = 0 \quad \text{si } y \neq 0$$
[6.18]

Para la estimación [6.17] el valor de la verosimilitud sería $+\infty$.

Cuando el número de dimensiones es pequeño o tenemos un conocimiento previo y preciso de la familia a la cual pertenece la función de densidad, aportando una mayor rigidez al modelo, el problema de inestabilidad dimensional no interfiere para obtener un resultado satisfactorio.

En el siguiente apartado observamos, presentando una serie de ejemplos, como el fenómeno de inestabilidad dimensional aparece al trabajar con las familias definidas en el capítulo anterior y los problemas que esto plantea.

6.2.3. Resultados con las familias del apartado 5.3.2.

Para la ejecución de los ejemplos numéricos, hemos tomado como familia de funciones de densidad la definida en el apartado 5.3.2., utilizando como funciones ortogonales los polinomios de Hermite normalizados. Entre las razones que podemos aducir para la selección de esta familia en concreto, citaremos que su dominio de definición comprende toda la recta real y que hemos obtenido la expresión explícita para la función de distribución, tal y como figura en [5.52] y [5.53]. Esto último nos permitirá la obtención de muestras simuladas correspondientes a diferentes elementos de la familia, que utilizaremos en el apartado siguiente.

En los ejemplos numéricos correspondientes a las figuras 1 a 8, hemos trabajado siempre con datos tales que tienen media muestral 0 y varianza muestral corregida 1/2.

VI - Estimación no paramétrica de la densidad

Adoptamos este convenio debido a que la función de densidad correspondiente a una distribución normal de media 0 y varianza 1/2 coincide con el elemento de la familia:

$$p(x|0) = h_0^2(x) \quad [6.19]$$

y nos simplifica de esta forma la resolución numérica de las ecuaciones de verosimilitud [6.15] al poder facilitar a los métodos numéricos una aproximación inicial razonable al algoritmo numérico:

$$\theta_0 = 1 \quad , \quad \theta_1 = \dots = \theta_n = 0 \quad [6.20]$$

La resolución de las ecuaciones de verosimilitud [6.15] se ha efectuado aplicando el algoritmo de Newton-Raphson, Cohen (1977).

En las **Figs. 1 y 2** vemos como para una muestra de tamaño $m = 1$ $x = 0$, al aumentar el número de parámetros, 20 y 50 respectivamente, la solución máximo verosímil converge a una función delta de Dirac en el punto $x = 0$. El mismo efecto se comprueba en las **Figs. 3, 4, 5 y 6** para una muestra de tamaño $m = 2$ $x_1 = -0.5$, $x_2 = +0.5$ donde consideramos 10, 15, 20 y 50 parámetros, y en las **Figs. 7 y 8** con $m = 3$ $x_1 = -1/\sqrt{2}$, $x_2 = 0.0$, $x_3 = +1/\sqrt{2}$ con 20 y 50 parámetros.

En las **Figs. 9a, 9b** se presenta el resultado correspondiente a una simulación de tamaño $m=30$ de una distribución $N(0,1)$ donde hemos considerado 15 y 30 parámetros respectivamente. Para los datos simulados se ha obtenido $\bar{x} = 0.11470$ y $s = 0.70483$. Se comprueba como el resultado difiere notablemente de la distribución real. La distancia entre la real y la estimada, a partir de la fórmula $d = 2 \cos^{-1} \int_x \sqrt{p} \sqrt{q} dx$, es de 0.674556 y 0.768904 respectivamente. En las **Figs. 10a y 10b** hemos procedido del mismo modo, utilizando una muestra de tamaño $m=50$ de una distribución $N(0,1)$, también estimando con 15 y 30 parámetros. En este caso

VI - Estimación no paramétrica de la densidad.

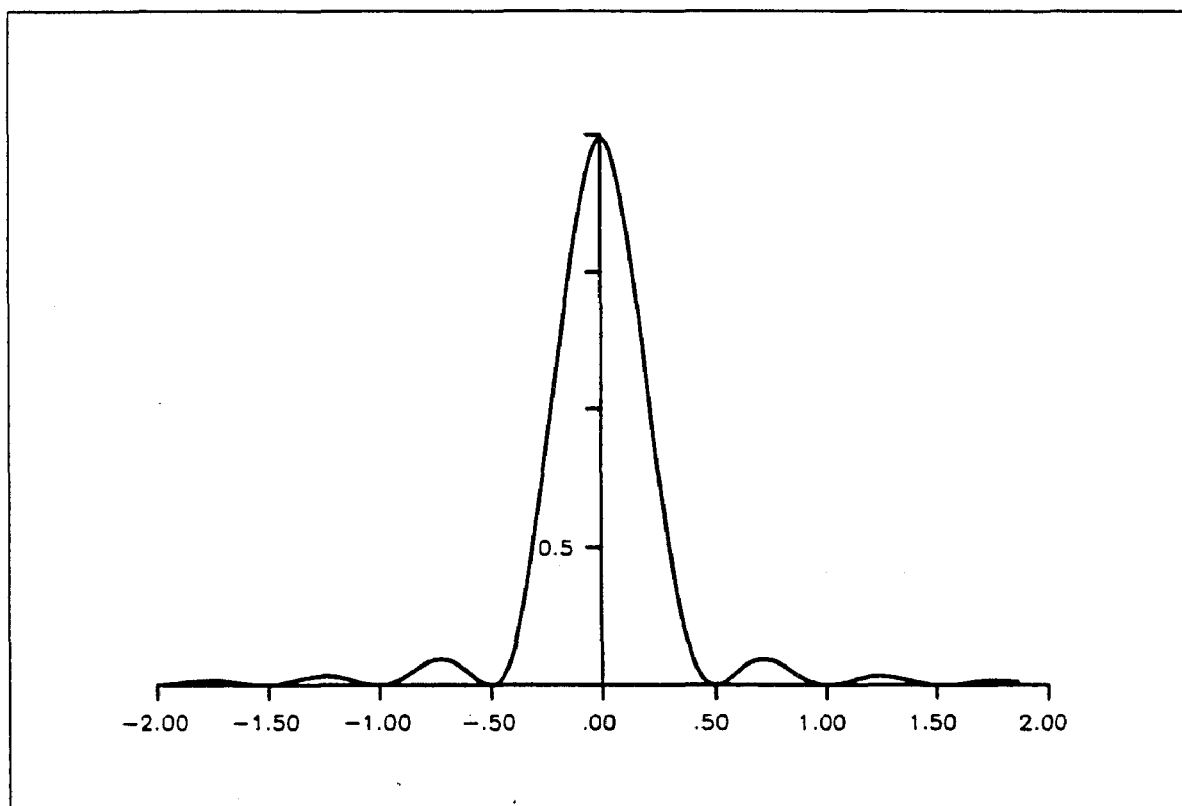


Figura 1. Estimación máximo verosímil con una muestra de tamaño $l = 0.0$ y $n = 20$ parámetros.

Valores obtenidos para los parámetros.

$$\begin{aligned}\theta_0 &= 0.53270435 \\ \theta_2 &= -0.37667912 \\ \theta_4 &= 0.32621378 \\ \theta_6 &= -0.29779083 \\ \theta_8 &= 0.27855760 \\ \theta_{10} &= -0.26426286 \\ \theta_{12} &= 0.25301236 \\ \theta_{14} &= -0.24380869 \\ \theta_{16} &= 0.23606664 \\ \theta_{18} &= -0.22941542\end{aligned}$$

$$\theta_{2k+1} = 0.0 \quad k = 0, \dots, 9$$

VI - Estimación no paramétrica de la densidad.

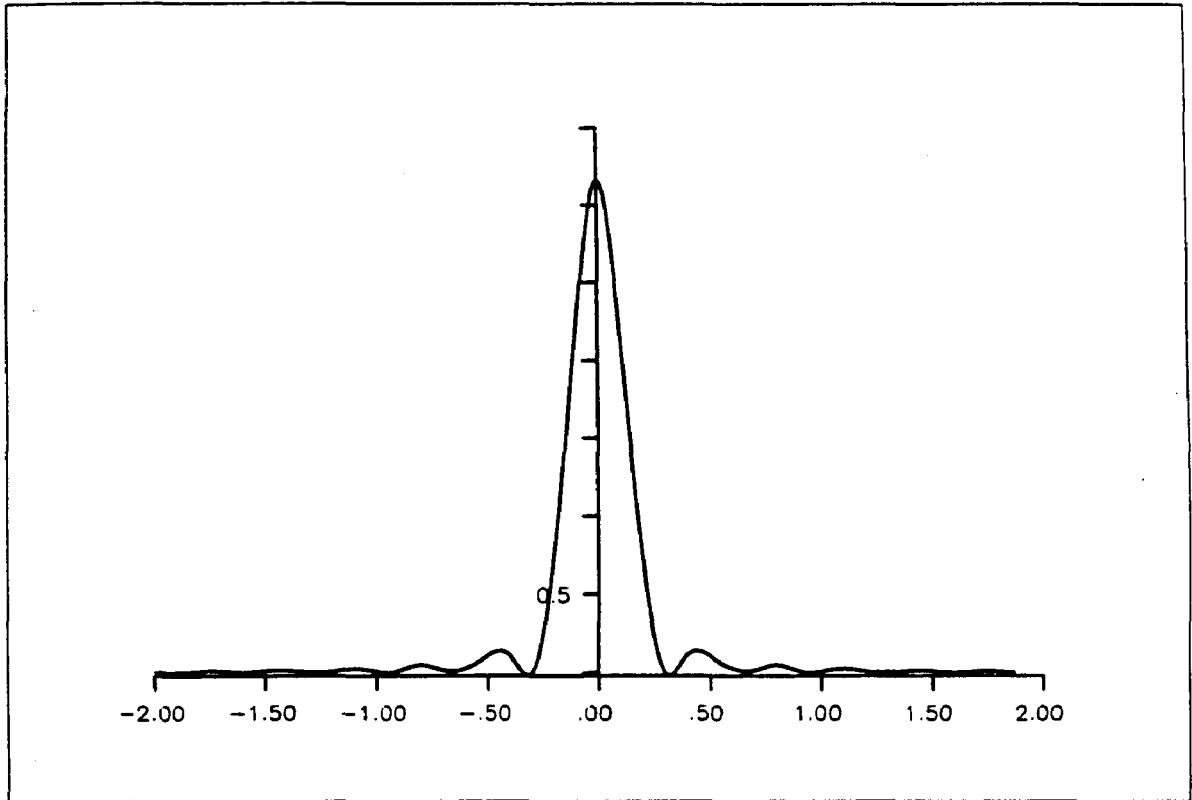


Figura 2. Estimación máximo verosímil con una muestra de tamaño $l = 1$ $x = 0.0$ y $n = 50$ parámetros.

Valores obtenidos para los parámetros.

$\theta_0 = 0.42206043$	$\theta_{20} = 0.17716289$	$\theta_{40} = 0.14944089$
$\theta_2 = -0.29844189$	$\theta_{22} = -0.17308950$	$\theta_{42} = -0.14765102$
$\theta_4 = 0.25845838$	$\theta_{24} = 0.16944498$	$\theta_{44} = 0.14596349$
$\theta_6 = -0.23593891$	$\theta_{26} = -0.16615439$	$\theta_{46} = -0.14436817$
$\theta_8 = 0.22070044$	$\theta_{28} = 0.16316032$	$\theta_{48} = 0.14285642$
$\theta_{10} = -0.20937479$	$\theta_{30} = -0.16041791$	
$\theta_{12} = 0.20046109$	$\theta_{32} = 0.15789139$	
$\theta_{14} = -0.19316900$	$\theta_{34} = -0.15555209$	
$\theta_{16} = 0.18703497$	$\theta_{36} = 0.15337634$	
$\theta_{18} = -0.18176526$	$\theta_{38} = -0.15134478$	

$$\theta_{2k+1} = 0.0 \quad k = 0, \dots, 24$$

VI - Estimación no paramétrica de la densidad.

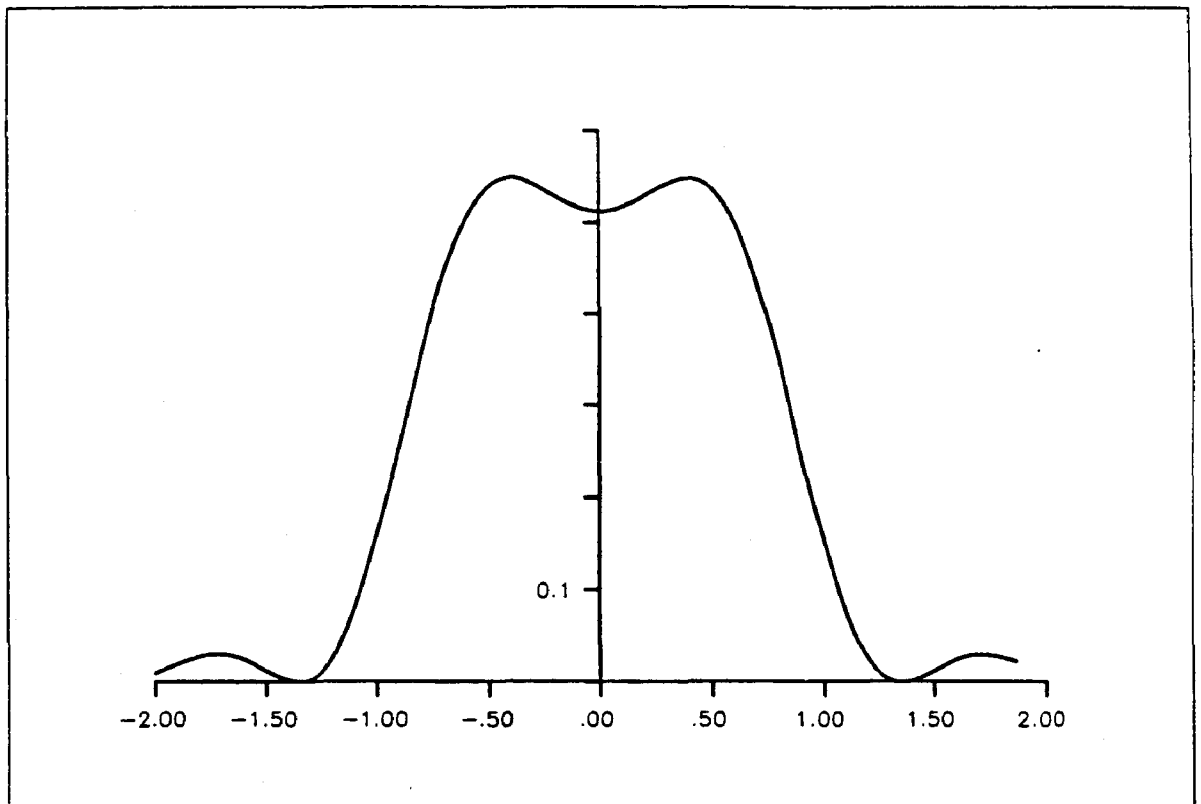


Figura 3. *Estimación máximo verosímil con una muestra de tamaño 2*
 $x_1 = -0.5$, $x_2 = +0.5$ y $n = 10$ parámetros.

Valores obtenidos para los parámetros.

$$\begin{aligned}\theta_0 &= 0.90311569 \\ \theta_2 &= -0.31930006 \\ \theta_4 &= 0.46087399E - 01 \\ \theta_6 &= 0.13042092 \\ \theta_8 &= -0.25158531\end{aligned}$$

$$\theta_{2k+1} = 0.0 \quad k = 0, \dots, 4$$

VI - Estimación no paramétrica de la densidad.

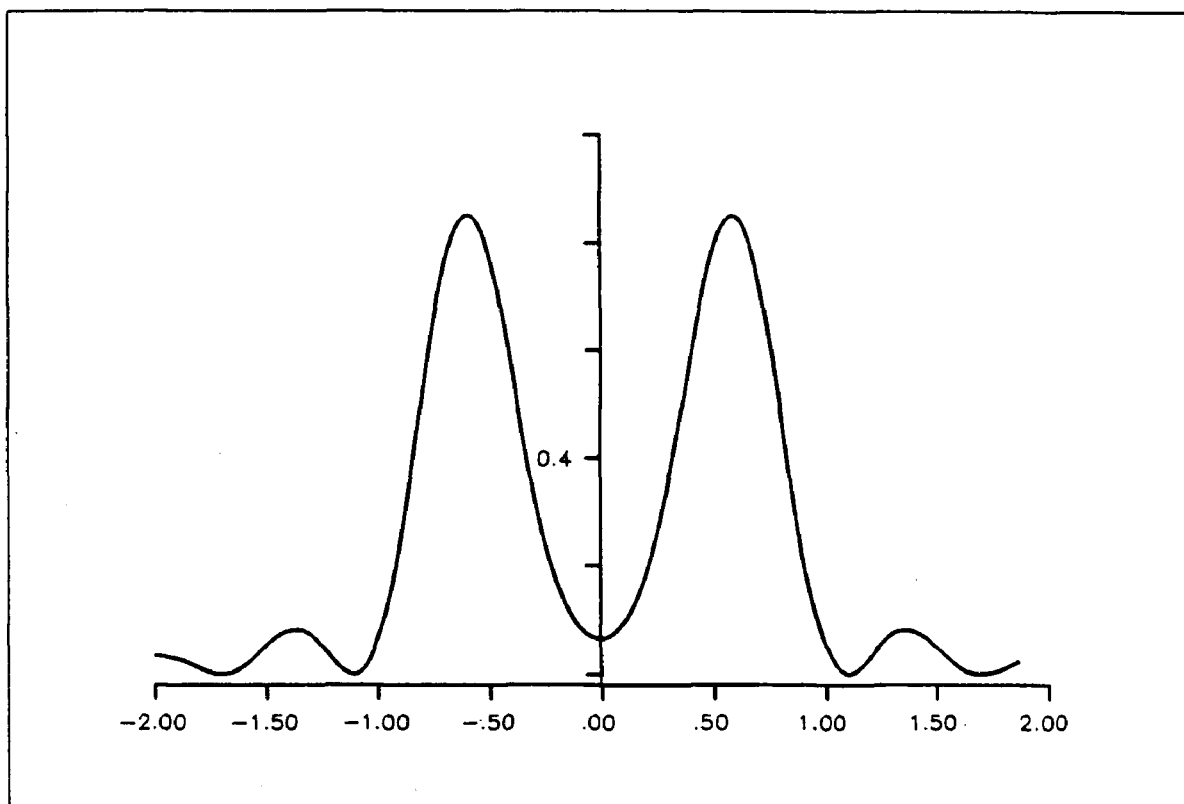


Figura 4. Estimación máximo verosímil con una muestra de tamaño 2
 $x_1 = -0.5$, $x_2 = +0.5$ y $n = 15$ parámetros.

Valores obtenidos para los parámetros.

$$\begin{aligned}\theta_0 &= 0.75208128 \\ \theta_2 &= -0.26590151 \\ \theta_4 &= 0.38379878E - 01 \\ \theta_6 &= 0.10860974 \\ \theta_8 &= -0.20951092 \\ \theta_{10} &= 0.27871990 \\ \theta_{12} &= -0.32432085 \\ \theta_{14} &= 0.35161954\end{aligned}$$

$$\theta_{2k+1} = 0.0 \quad k = 0, \dots, 6$$

VI - Estimación no paramétrica de la densidad.

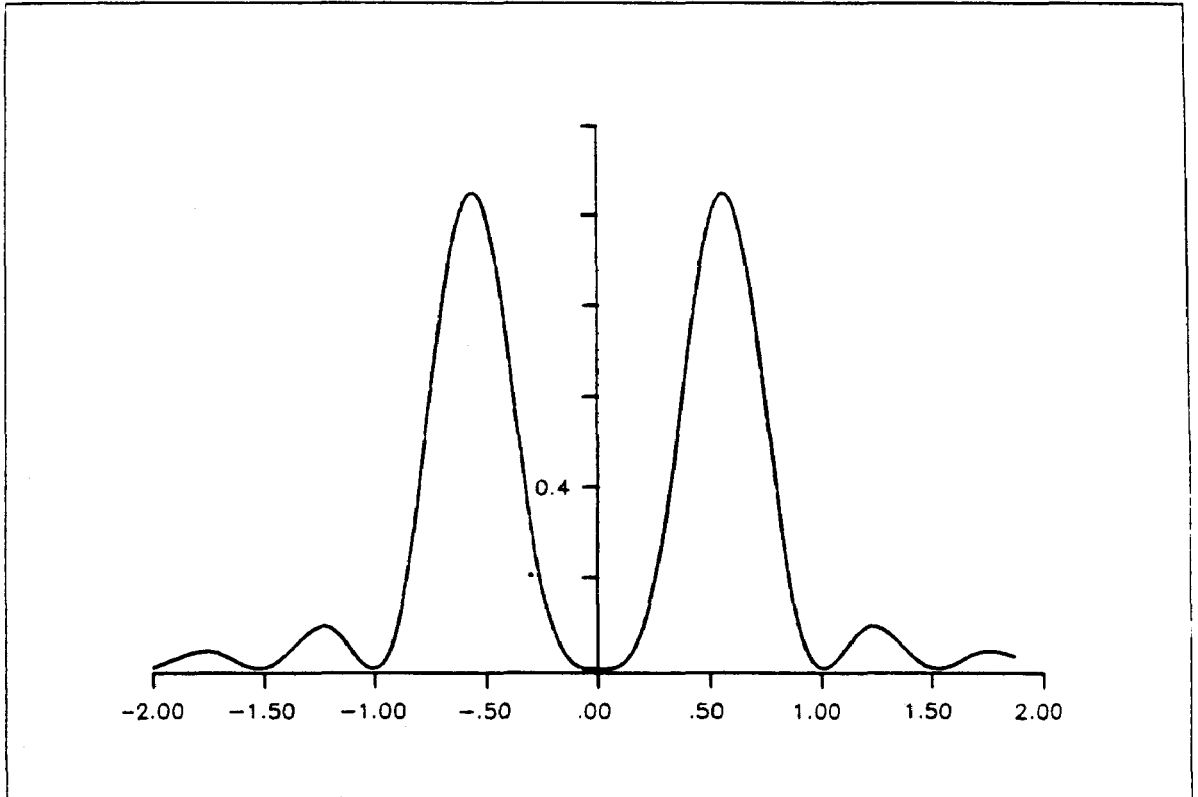


Figura 5. Estimación máximo verosímil con una muestra de tamaño 2 $x_1 = -0.5$, $x_2 = +0.5$ y $n = 20$ parámetros.

Valores obtenidos para los parámetros.

$$\begin{aligned}\theta_0 &= 0.668275 \\ \theta_2 &= -0.23627120 \\ \theta_4 &= 0.34103114E - 01 \\ \theta_6 &= 0.96507013E - 01 \\ \theta_8 &= -0.18616456 \\ \theta_{10} &= 0.24766141 \\ \theta_{12} &= -0.28818089 \\ \theta_{14} &= 0.31243742 \\ \theta_{16} &= -0.32382667 \\ \theta_{18} &= 0.32493872\end{aligned}$$

$$\theta_{2k+1} = 0.0 \quad k = 0, \dots, 9$$

VI - Estimación no paramétrica de la densidad.

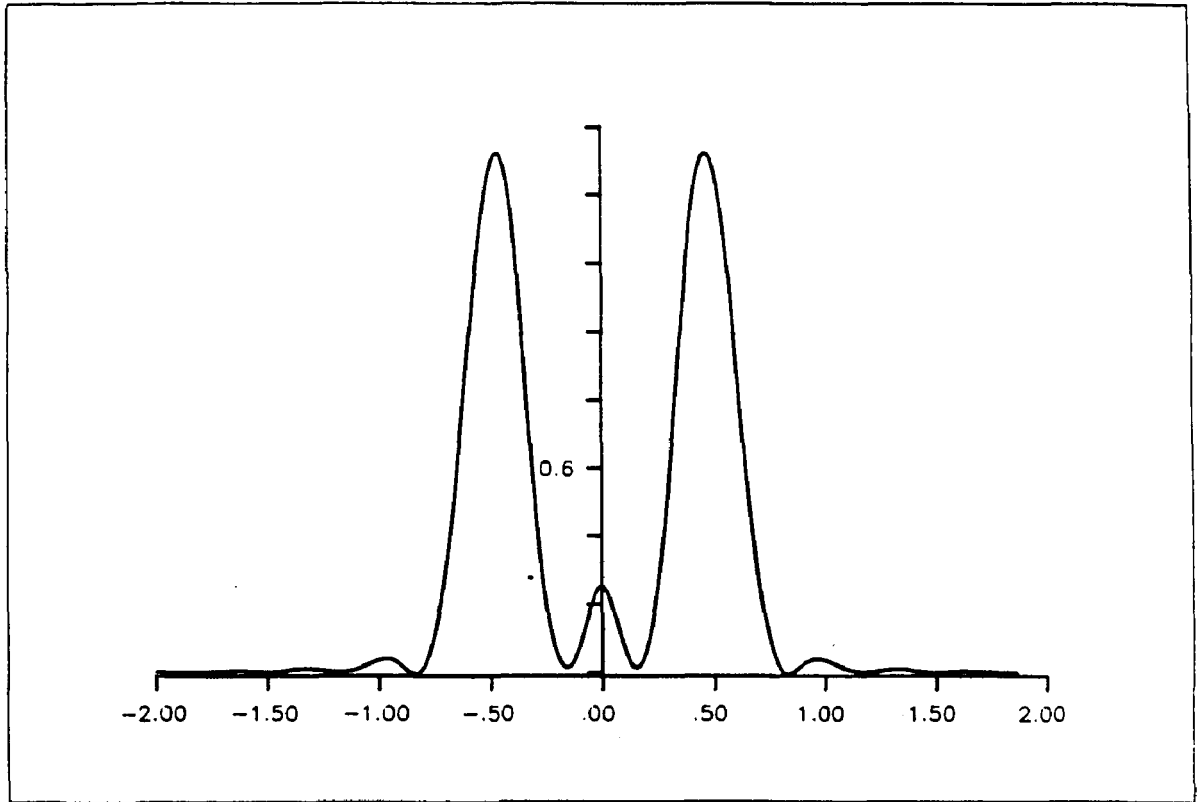


Figura 6. Estimación máximo verosímil con una muestra de tamaño 2 $x_1 = -0.5$, $x_2 = +0.5$ y $n = 50$ parámetros.

Valores obtenidos para los parámetros.

θ_0	=	0.54083997	θ_{20}	=	-0.25722069	θ_{40}	=	-0.46232384E-01
θ_2	=	-0.19121599	θ_{22}	=	0.24616414	θ_{42}	=	0.22427015E-01
θ_4	=	0.27599897E-01	θ_{24}	=	-0.23092723	θ_{44}	=	0.53845951E-03
θ_6	=	0.78103840E-01	θ_{26}	=	0.21244943	θ_{46}	=	-0.22488456E-01
θ_8	=	-0.15066427	θ_{28}	=	-0.19152576	θ_{48}	=	0.43279905E-01
θ_{10}	=	0.20043415	θ_{30}	=	0.16883129			
θ_{12}	=	-0.23322690	θ_{32}	=	-0.14494193			
θ_{14}	=	0.25285810	θ_{34}	=	0.12034905			
θ_{16}	=	-0.26207525	θ_{36}	=	-0.95471382E-01			
θ_{18}	=	0.26297534	θ_{38}	=	0.70665181E-01			

$$\theta_{2k+1} = 0.0 \quad k = 0, \dots, 24$$

VI - Estimación no paramétrica de la densidad.

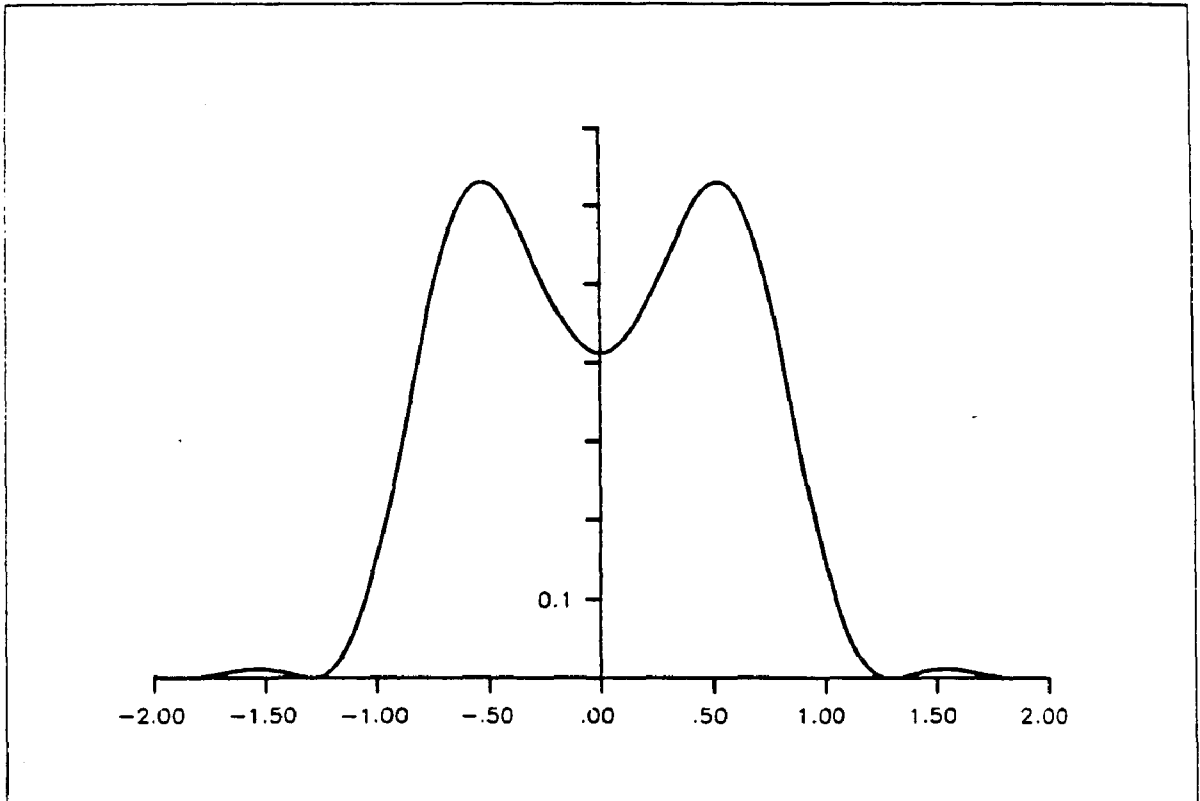


Figura 7. Estimación máximo verosímil con una muestra de tamaño 3 $x_1 = -1/\sqrt{2}$, $x_2 = 0.0$, $x_3 = +1/\sqrt{2}$ y $n = 20$ parámetros.

Valores obtenidos para los parámetros.

$$\begin{aligned}\theta_0 &= 0.92203099 \\ \theta_2 &= -0.27597427 \\ \theta_4 &= 0.21917030E - 01 \\ \theta_6 &= 0.98894596E - 01 \\ \theta_8 &= -0.14547068 \\ \theta_{10} &= 0.14582139 \\ \theta_{12} &= -0.11687243 \\ \theta_{14} &= 0.69989562E - 01 \\ \theta_{16} &= -0.13192762E - 01 \\ \theta_{18} &= -0.47736742E - 01\end{aligned}$$

$$\theta_{2k+1} = 0.0 \quad k = 0, \dots, 9$$

VI - Estimación no paramétrica de la densidad.

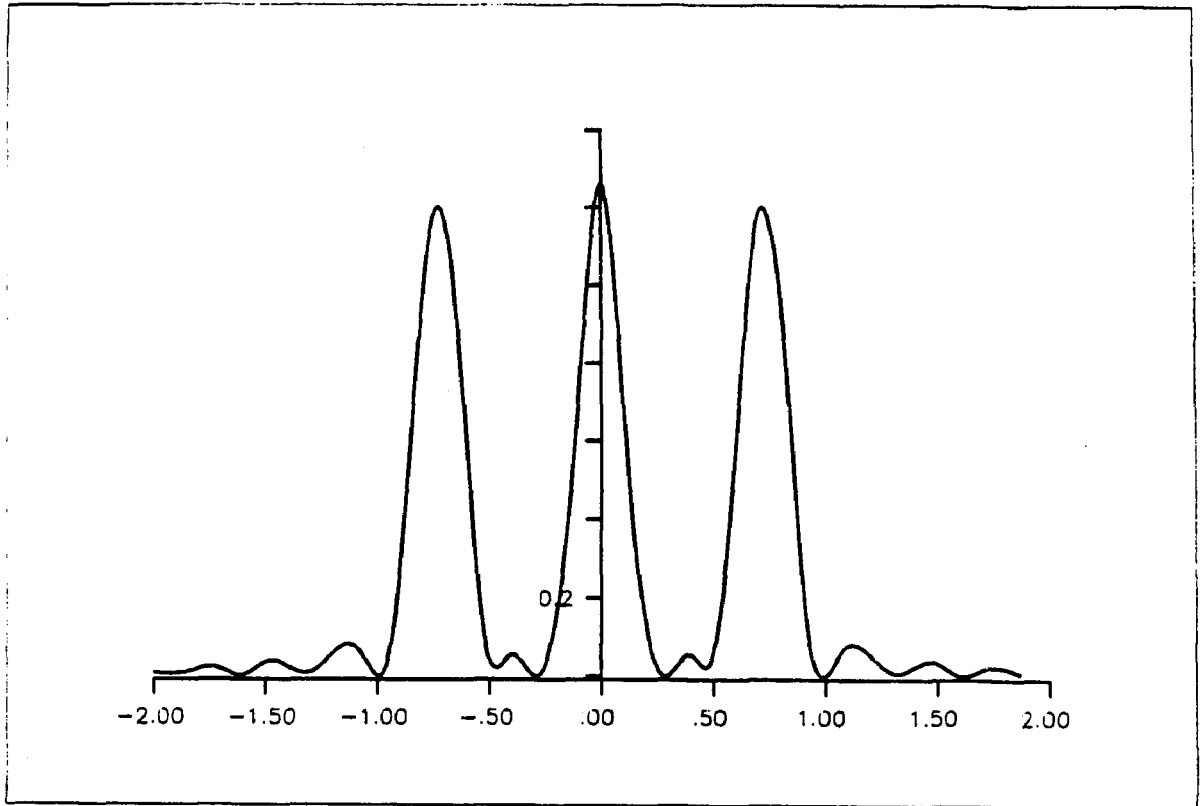


Figura 8. Estimación máximo verosímil con una muestra de tamaño 3 $x_1 = -1/\sqrt{2}$, $x_2 = 0.0$, $x_3 = +1/\sqrt{2}$ y $n = 50$ parámetros.

Valores obtenidos para los parámetros.

$\theta_0 = 0.57461506$	$\theta_{20} = 0.57008199E-01$	$\theta_{40} = 0.23863804$
$\theta_2 = -0.15755957$	$\theta_{22} = -0.95619559E-01$	$\theta_{42} = -0.23204494$
$\theta_4 = -0.71681105E-02$	$\theta_{24} = 0.13080823$	$\theta_{44} = 0.22224397$
$\theta_6 = 0.85206807E-01$	$\theta_{26} = -0.16166627$	$\theta_{46} = -0.20971042$
$\theta_8 = -0.11474329$	$\theta_{28} = 0.18764556$	$\theta_{48} = 0.19491756$
$\theta_{10} = 0.11402583$	$\theta_{30} = -0.20847744$	
$\theta_{12} = -0.94126523E-01$	$\theta_{32} = 0.22411019$	
$\theta_{14} = 0.62498227E-01$	$\theta_{34} = -0.23465908$	
$\theta_{16} = -0.24408344E-01$	$\theta_{36} = 0.24036402$	
$\theta_{18} = -0.16343396E-01$	$\theta_{38} = -0.24155790$	

$$\theta_{2k+1} = 0.0 \quad k = 0, \dots, 24$$

VI - Estimación no paramétrica de la densidad.

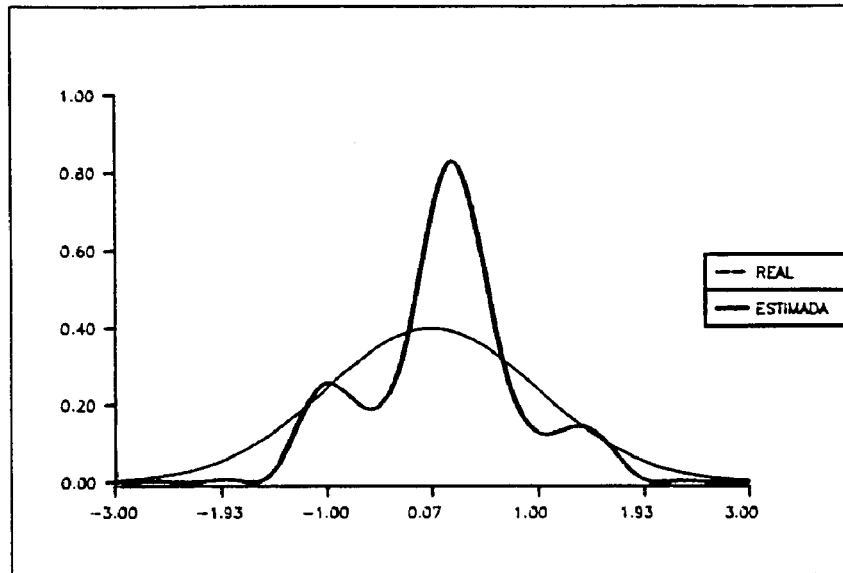


Figura 9a. Estimación MLE de una muestra de tamaño $m=30$ de una $N(0,1)$ con 15 parámetros.

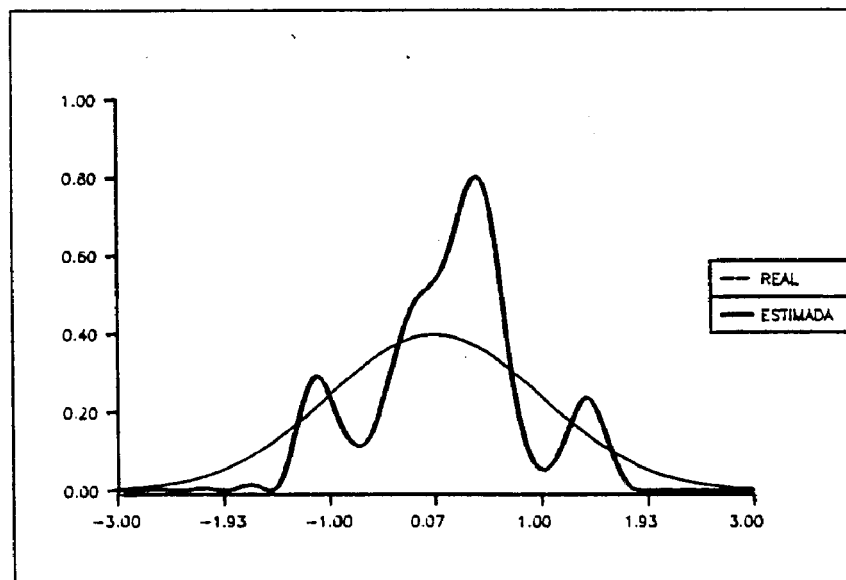


Figura 9b. Estimación MLE de una muestra de tamaño $m=30$ de una $N(0,1)$ con 30 parámetros.

VI - Estimación no paramétrica de la densidad.

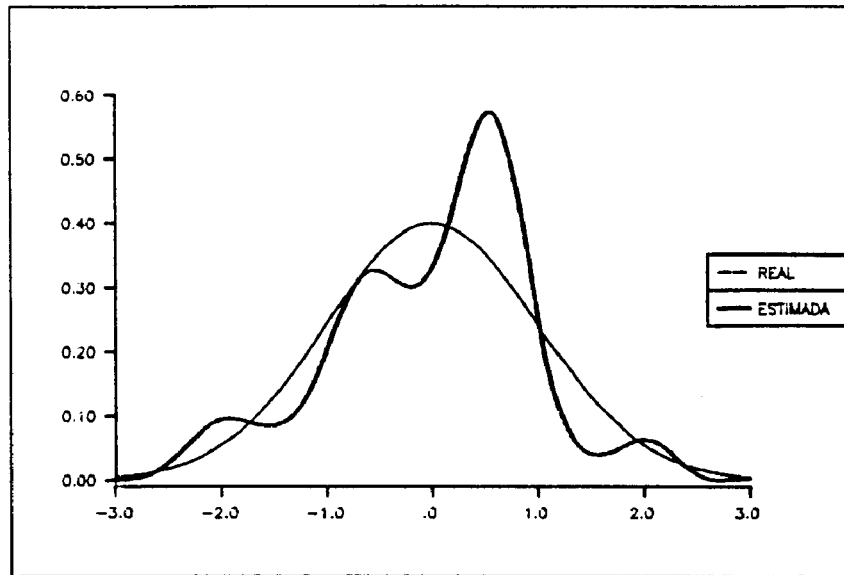


Figura 10a. Estimación MLE de una muestra de tamaño $m=50$ de una $N(0,1)$ con 15 parámetros.

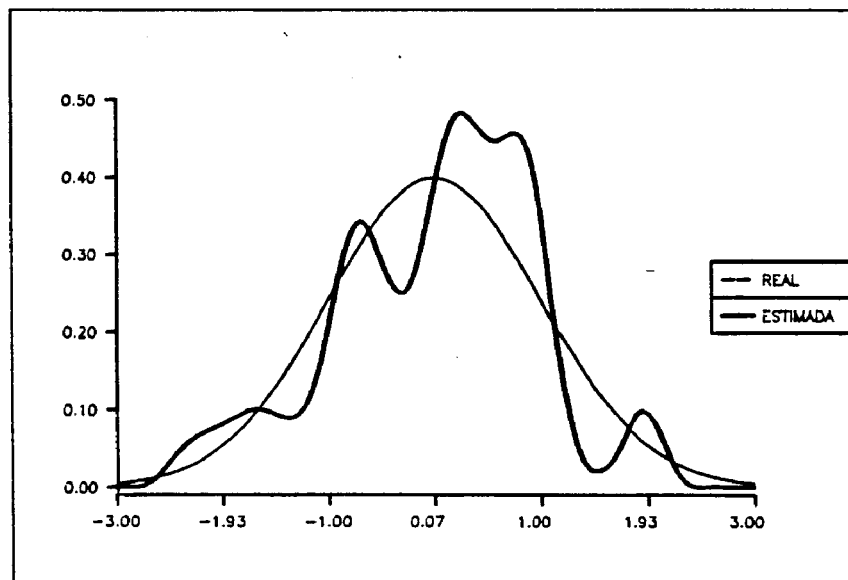


Figura 10b. Estimación MLE de una muestra de tamaño $m=50$ de una $N(0,1)$ con 30 parámetros.

VI - Estimación no paramétrica de la densidad

la media muestral era de $\bar{x} = 0.01014$ y la desviación típica muestral de $s = 0.95971$. Las distancias entre las funciones simuladas y la real es de 0.404867 y de 0.436314 para 15 y 30 parámetros respectivamente. En ambos casos se observa como al aumentar el número de parámetros que intervienen en la estimación, aumenta la distorsión de las funciones.

Por lo visto hasta el momento se desprende que la utilización de la estimación máximo verosímil sin ninguna consideración adicional no es plenamente satisfactoria en los modelos que aquí consideramos. Se observa que la estimación depende enormemente del número de parámetros considerado y que así como la elección de un número de parámetros excesivamente bajo produce estimaciones alejadas de la realidad de los datos debido a la rigidez de las funciones, incrementar de forma ilimitada los mismos produce una inestabilidad en las estimaciones que no podemos evaluar ni solucionar sin un conocimiento previo de la distribución real.

En los siguientes apartados proponemos algunas soluciones para resolver el problema anterior. En el apartado 6.3 presentamos un algoritmo stepwise que selecciona automáticamente el orden de las estimaciones y las funciones que intervienen en las mismas. Esto se realiza mediante comparaciones de las verosimilitudes para sucesivas estimaciones al aumentar o disminuir el número de parámetros. En el apartado 6.4 utilizamos un método diferente de estimación utilizando las distancias entre individuos definidas en el capítulo 3 y el proceso de estimación introducido en el capítulo 4.

6.3. Un algoritmo stepwise de estimación no paramétrica.

6.3.1. Desarrollo del algoritmo.

El algoritmo consta de tres partes básicas, una primera que comprende las etapas 1 a 4, en la cual tras una adecuación de los datos a los métodos numéricos, efectuamos un proceso de estimación ascendente, aumentando progresivamente el número de parámetros empleados hasta que no se produzca un incremento significativo en el máximo de la función de verosimilitud. En la segunda etapa, que corresponde al número 5, a partir de la estimación anterior, buscamos los parámetros correspondientes a la función de densidad con los datos sin transformar. Finalmente en el paso 6, efectuamos un proceso descendente suprimiendo de la estimación aquellos parámetros cuya eliminación no provoca un descenso significativo del máximo de la función de verosimilitud.

1. En primer lugar efectuamos una transformación lineal de los datos del tipo $y_i = \frac{x_i - a}{b}$ $i = 1, \dots, m$. El motivo de tal transformación es solventar problemas numéricos que se pueden presentar al trabajar con la familia ortonormal, en particular nos permitirá proporcionar una buena aproximación inicial a la solución de las ecuaciones de verosimilitud. Una alternativa es trabajar con $a =$ mediana muestral, y $b =$ recorrido intercuartil muestral. Al trabajar con los polinomios de Hermite hemos tomado $a = \bar{x}$ y $b = \sqrt{2} S \sqrt{\frac{m-1}{m}}$, con lo cual los datos transformados pasan a tener media muestral 0 y varianza muestral insesgada 1/2, pudiendo tomar como aproximación inicial la definida en [6.20].

2. Dada una muestra y_1, \dots, y_m , y fijados unos coeficientes $\alpha_1, \dots, \alpha_k \in \mathbb{N}$ definimos:

$$\hat{l}_{\alpha_1, \dots, \alpha_k} = \sup_{\theta_{\alpha_1}, \dots, \theta_{\alpha_k} \in S_k(1)} L(\theta_{\alpha_1}, \dots, \theta_{\alpha_k}) = \sup_{\theta_{\alpha_1}, \dots, \theta_{\alpha_k} \in S_k(1)} \prod_{j=1}^m \left[\sum_{i=1}^k \theta_{\alpha_i} \varphi_{\alpha_i}(y_j) \right]^2 \quad [6.21]$$

3. Seleccionamos un número arbitrario N de parámetros a explorar y definimos el conjunto de índices $I = \{1, \dots, N\}$. La elección del número N vendrá motivada por criterios de tiempo computacional. Es razonable la selección de los N primeros términos, ya que al ser la media muestral de los datos 0 y la varianza muestral insesgada $1/2$ es de esperar que $\theta_j \rightarrow 0$ al incrementarse j debido a [6.19].

4. Desde $k = 1$ y hasta un máximo de N , supuesto que en el paso k tenemos fijados $\alpha_1, \dots, \alpha_{k-1}$, buscamos para $i \in I - \{\alpha_1, \dots, \alpha_{k-1}\}$, aquel i tal que

$$\hat{l}_{\alpha_1, \dots, \alpha_{k-1}, i} \geq \hat{l}_{\alpha_1, \dots, \alpha_{k-1}, j} \quad [6.22]$$

para todo $j \in I - \{\alpha_1, \dots, \alpha_{k-1}\}$. Si $k \geq 2$ efectuamos una prueba de significación del término i , hallando la razón de verosimilitud

$$\Lambda = \frac{\hat{l}_{\alpha_1, \dots, \alpha_{k-1}}}{\hat{l}_{\alpha_1, \dots, \alpha_{k-1}, i}} \quad [6.23]$$

Por el teorema de Wilks, bajo la hipótesis de igualdad de ambos espacios paramétricos, $U = -2 \ln \Lambda$ converge a una distribución χ^2 con un grado de libertad. Si aplicando el test rechazamos la hipótesis y consideramos que existen diferencias significativas entre las verosimilitudes, fijamos $\alpha_k = i$, e incrementando k en una unidad volvemos a 4. En caso contrario si el test no detecta diferencias significativas, detenemos el proceso de adición de términos y pasamos a la siguiente etapa del algoritmo.

VI - Estimación no paramétrica de la densidad

5. Con los parámetros obtenidos en el paso anterior formamos según [6.1] la función de densidad estimada correspondiente a los datos transformados

$$\hat{p}_Y(y) = \left[\sum_{i=1}^{k-1} \theta_{\alpha_i} \varphi_{\alpha_i} \right]^2 \quad [6.24]$$

Para obtener los coeficientes correspondientes a la función con datos no transformados, realizamos numéricamente la proyección

$$\theta_i = \int_x \sqrt{\hat{p}_X(x)} \varphi_i dx \quad i = 1, \dots, r \quad [6.25]$$

donde $\hat{p}_X(x) = \hat{p}_Y\left(\frac{x-a}{b}\right) \frac{1}{b}$ y r es tal que para una tolerancia arbitraria ϵ , $1 - \sum_{j=1}^r \theta_j^2 \leq \epsilon$.

6. Hacemos un filtrado final con los datos sin transformar. Definimos $I' = \{\alpha_1, \dots, \alpha_r\}$ donde $\alpha_1, \dots, \alpha_r \in 1, \dots, r$, y buscamos aquel $i \in I'$ tal que

$$\hat{l}_{\alpha_1, \dots, \alpha_{i-1}, \alpha_{i+1}, \dots, \alpha_r} \leq \hat{l}_{\alpha_1, \dots, \alpha_{j-1}, \alpha_{j+1}, \dots, \alpha_r} \quad [6.26]$$

para todo $j \in I'$. Si

$$\Lambda = \frac{\hat{l}_{\alpha_1, \dots, \alpha_{i-1}, \alpha_{i+1}, \dots, \alpha_r}}{\hat{l}_{\alpha_1, \dots, \alpha_r}} \quad [6.27]$$

no proporciona diferencias significativas entre las verosimilitudes, entonces eliminamos el término i , redefinimos $I' = \{\alpha_1, \dots, \alpha_{i-1}, \alpha_{i+1}, \dots, \alpha_r\}$ y volvemos a ejecutar la búsqueda definida en 6. En caso contrario el algoritmo finaliza. El cociente [6.27] lo calculamos siempre entre la verosimilitud después de haber eliminado el parámetro conveniente y la verosimilitud inicial con los r parámetros originales, de forma que los grados

de libertad de la χ^2 son en cada paso por [6.27] el número de parámetros que llevamos eliminados respecto al número inicial r .

6.3.2. Resultados.

Para verificar la eficacia del algoritmo, hemos optado en primer lugar por la siguiente estrategia. Fijado un elemento de las familias del capítulo 5 utilizando los polinomios de Hermite normalizados $h_t(x)$ como familia ortonormal, simulamos una muestra aleatoria simple de tamaño m de una población que tenga áquel elemento como función de densidad de probabilidad. La simulación se ha llevado a cabo por el Método de Montecarlo utilizando la función de distribución de probabilidad hallada por nosotros en [5.52] y [5.53]. A continuación aplicamos el algoritmo de estimación a la muestra simulada y evaluamos la bondad del resultado comparando la estimación con la función original mediante la distancia [5.37]. En general hemos comprobado que es suficiente trabajar con una tolerancia ε de 0,01 y $N = 15$.

Ejemplo 6.3.2.1.

Corresponde a una función de densidad unimodal, tal y como se refleja en la Fig. 11 . Los parámetros originales son: $\theta_1 = 0.9$, $\theta_2 = 0.4$, $\theta_3 = 0.17$, $\theta_4 = 0.0$ en otro caso. En la Fig. 12 presentamos para una simulación en particular, la gráfica comparativa de las funciones de distribución real y empírica, comprobando el ajuste entre las mismas. Al igual que en los ejemplos que presentaremos a continuación, se han obtenido muestras simuladas de tamaños 30 y 200, sobre las que hemos aplicado el algoritmo. También se han considerado dos diferentes niveles de significación para los contrastes del algoritmo, 5 % y 10 %. Los resultados obtenidos son presentados en las Tablas 1 a 4 .

Ejemplo 6.3.2.2.

La gráfica de la función de densidad se presenta en la Fig. 13 , y tal y como se aprecia, es una distribución multimodal y simétrica. Los parámetros originales son: $\theta_1 = 0.5$, $\theta_3 = 0.5$, $\theta_5 = 0.5$, $\theta_7 = 0.5$, $\theta_l = 0.0$ en caso contrario. En la Fig. 14 presentamos la gráfica del ajuste para una simulación particular, y los resultados se presentan en las Tablas 5 a 8 .

Ejemplo 6.3.2.3.

Corresponde a una función de densidad multimodal y asimétrica, de la forma que se observa en la Fig. 15 . Los parámetros originales son: $\theta_3 = 0.9$, $\theta_4 = 0.4$, $\theta_6 = 0.17$, $\theta_l = 0.0$ en otro caso. En la Fig. 16 presentamos la gráfica del ajuste para una simulación particular, y los resultados se presentan en las Tablas 9 a 12 .

Entre las conclusiones que se pueden extraer de los resultados, destaquemos:

- * El ajuste mejora notablemente al aumentar el tamaño muestral.
- * Con menores tamaños muestrales, intervienen menos aquellos parámetros con menor peso en la función de densidad.
- * No se aprecian diferencias significativas entre los resultados obtenidos con los niveles de significación del 5% y del 10%.
- * No se ha encontrado una correlación significativa entre las distancias finales y el ajuste de las simulaciones.

VI - Estimación no paramétrica de la densidad.

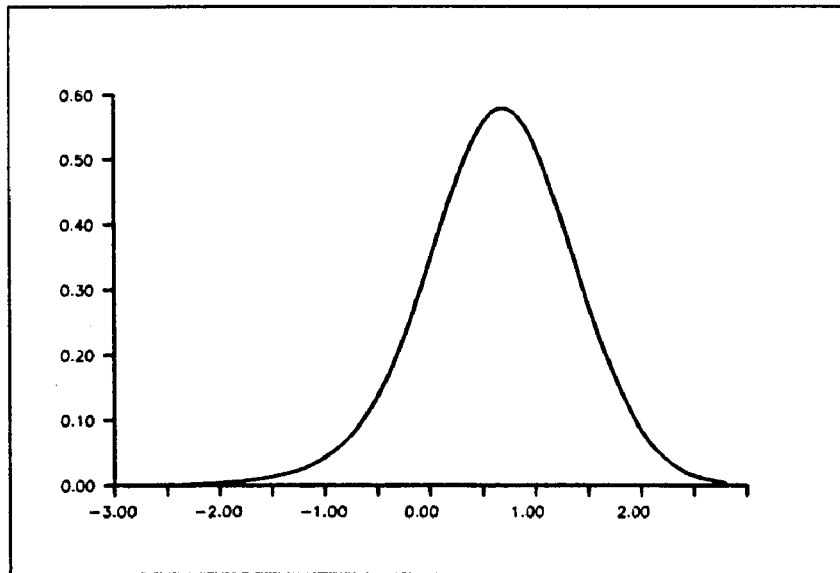


Figura 11. Función de densidad correspondiente a los parámetros del ejemplo 6.3.2.1.

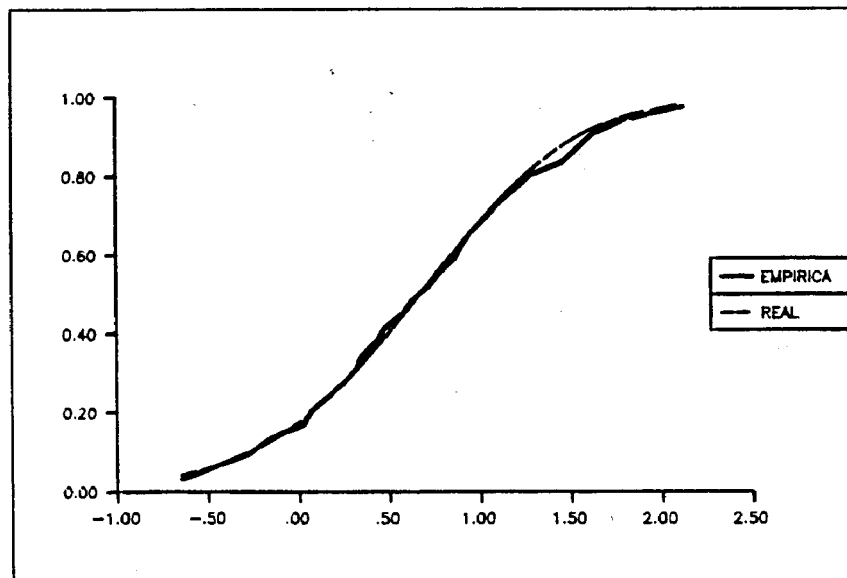


Figura 12. Funciones de distribución real y empírica correspondiente a una simulación de tamaño 200 a partir de los parámetros del ejemplo 6.3.2.1. Máxima diferencia en valor absoluto para esta simulación 0.0447.

VI - Estimación no paramétrica de la densidad.

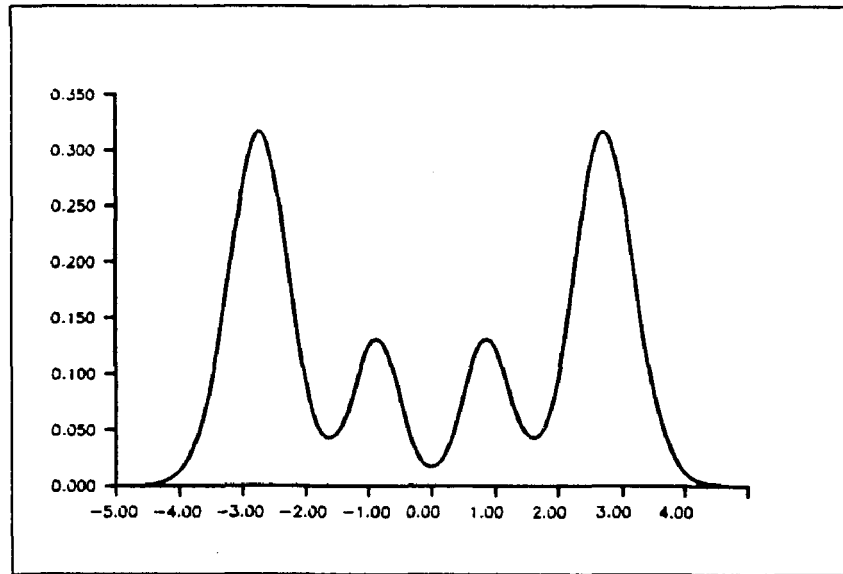


Figura 13. Función de densidad correspondiente a los parámetros del ejemplo 6.3.2.2.

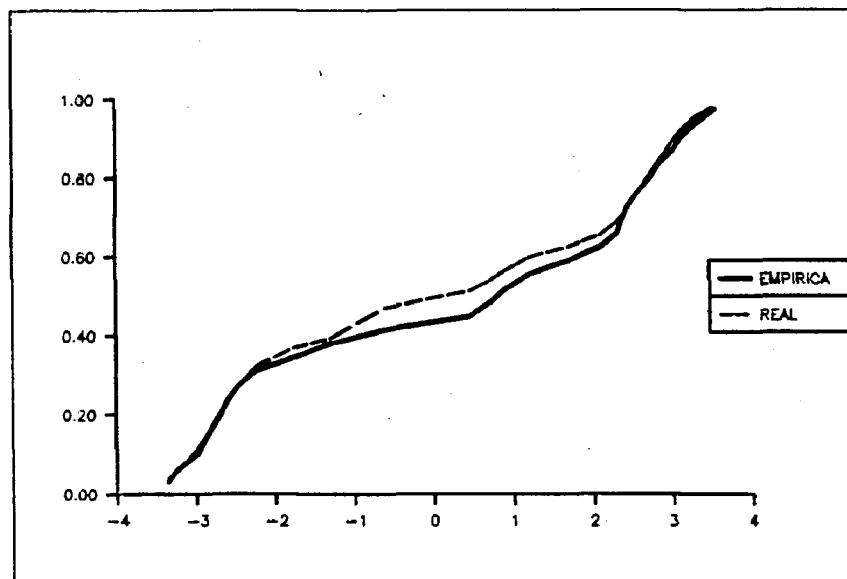


Figura 14. Funciones de distribución real y empírica correspondiente a una simulación de tamaño 200 a partir de los parámetros del ejemplo 6.3.2.2. Máxima diferencia en valor absoluto para esta simulación 0.06478.

VI - Estimación no paramétrica de la densidad.

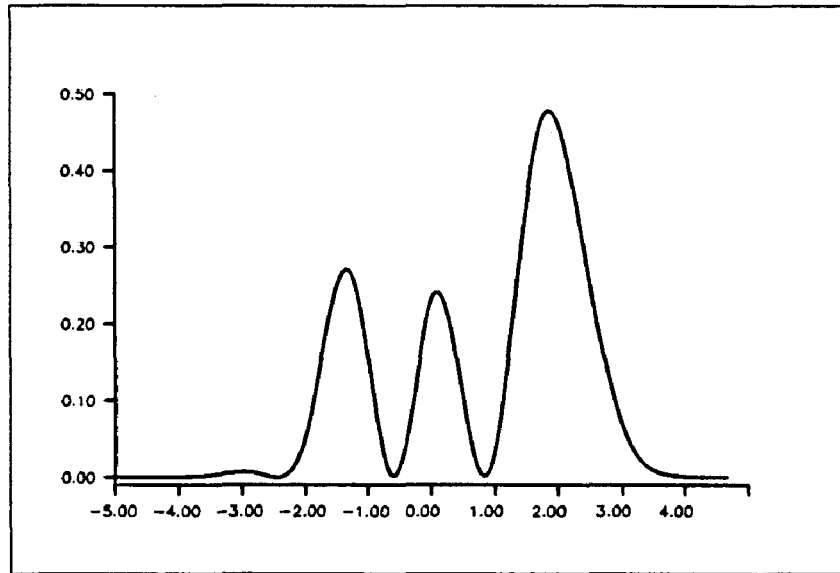


Figura 15. Función de densidad correspondiente a los parámetros del ejemplo 6.3.2.3.

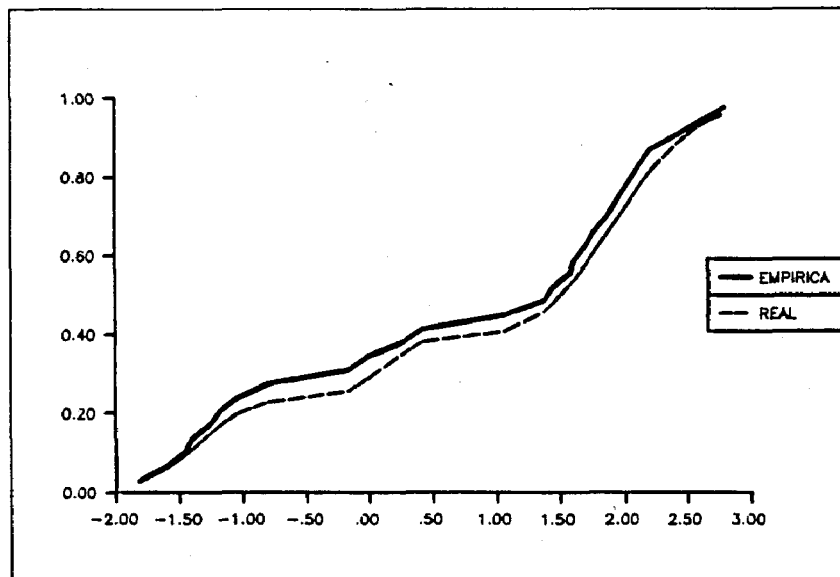


Figura 16. Funciones de distribución real y empírica correspondiente a una simulación de tamaño 200 a partir de los parámetros del ejemplo 6.3.2.3. Máxima diferencia en valor absoluto para esta simulación 0.04882.

VI - Estimación no paramétrica de la densidad.

Muestras procedentes de una simulación de parámetros

$$\theta_1 = 0.9, \quad \theta_2 = 0.4, \quad \theta_3 = 0.17$$

$$\theta_i = 0.0 \quad \text{en caso contrario}$$

Nivel de significación: 5 %
 Tamaño muestral simulado: 30 Número de simulaciones: 70

Ajuste de las simulaciones. $D_n = \sup |S_n(x) - F(x)|$

Valor máximo: 0.287 Valor mínimo: 0.0680
 Media: 0.144 Desviación típica: 0.0470
 Superan Kolmogorov-Smirnov: 68 Porcentaje: 97.14 %

Combinaciones de parámetros resultantes.

Combinación	Número	Porcentaje
1-2	52	74.29 %
1-2-3	14	20.00 %
1-2-9	2	2.86 %
1-2-15	1	1.43 %
1-2-3-4	1	1.43 %

Parámetro aislado	Número	Media	Desviación típica
1	70	0.885	0.049
2	70	0.433	0.102
3	15	0.255	0.032
9	2	0.126	0.040
4	1	-0.281	0.000
15	1	0.237	0.000

Distancias entre la función real y la estimada

Valor máximo: 0.8120 Valor mínimo: 0.1350
 Media: 0.3960 Desviación típica: 0.1100

Tabla 1

VI - Estimación no paramétrica de la densidad.

Muestras procedentes de una simulación de parámetros

$$\theta_1 = 0.9 \quad , \quad \theta_2 = 0.4 \quad , \quad \theta_3 = 0.17$$

$$\theta_i = 0.0 \quad \text{en caso contrario}$$

Nivel de significación: 5 %
 Tamaño muestral simulado: 200 Número de simulaciones: 64

Ajuste de las simulaciones. $D_n = \sup |S_n(x) - F(x)|$

Valor máximo: 0.0985 Valor mínimo: 0.0307
 Media: 0.0573 Desviación típica: 0.0161
 Superan Kolmogorov-Smirnov: 63 Porcentaje: 98.44 %

Combinaciones de parámetros resultantes.

Combinación	Número	Porcentaje
1-2-3	46	71.88 %
1-2	18	28.13 %

Parámetro aislado	Número	Media	Desviación típica
1	64	0.898	0.022
2	64	0.407	0.047
3	46	0.186	0.031

Distancias entre la función real y la estimada

Valor máximo: 0.5235 Valor mínimo: 0.0676
 Media: 0.1877 Desviación típica: 0.1274

Tabla 2

VI - Estimación no paramétrica de la densidad.

Muestras procedentes de una simulación de parámetros

$$\theta_1 = 0.9, \quad \theta_2 = 0.4, \quad \theta_3 = 0.17$$

$$\theta_i = 0.0 \quad \text{en caso contrario}$$

Nivel de significación: 10 %
 Tamaño muestral simulado: 30 Número de simulaciones: 80

Ajuste de las simulaciones. $D_n = \sup |S_n(x) - F(x)|$

Valor máximo: 0.277 Valor mínimo: 0.0637
 Media: 0.140 Desviación típica: 0.0462
 Superan Kolmogorov-Smirnov: 79 Porcentaje: 98.750 %

Combinaciones de parámetros resultantes.

Combinación	Número	Porcentaje
1-2	50	62.50 %
1-2-3	24	30.00 %
1-2-13	2	2.50 %
1-2-4	1	1.25 %
1-2-5	1	1.25 %
1-2-3-5	1	1.25 %
1-2-3-9	1	1.25 %

Parámetro aislado	Número	Media	Desviación típica
1	80	0.884	0.055
2	80	0.421	0.108
3	26	0.269	0.039
5	2	-0.058	0.297
13	2	0.200	0.126
4	1	0.240	0.000
9	1	0.115	0.000

Distancias entre la función real y la estimada

Valor máximo: 0.906 Valor mínimo: 0.106
 Media: 0.390 Desviación típica: 0.140

Tabla 3

VI - Estimación no paramétrica de la densidad.

Muestras procedentes de una simulación de parámetros

$$\theta_1 = 0.9 \quad , \quad \theta_2 = 0.4 \quad , \quad \theta_3 = 0.17$$

$$\theta_i = 0.0 \quad \text{en caso contrario}$$

Nivel de significación: 10 %
 Tamaño muestral simulado: 200 Número de simulaciones: 65

Ajuste de las simulaciones. $D_n = \sup |S_n(x) - F(x)|$

Valor máximo: 0.0885 Valor mínimo: 0.0296
 Media: 0.0579 Desviación típica: 0.0156
 Superan Kolmogorov-Smirnov: 65 Porcentaje: 100.00 %

Combinaciones de parámetros resultantes.

Combinación	Número	Porcentaje
1-2-3	50	76.92 %
1-2	14	21.54 %
1-2-3-9	1	1.54 %

Parámetro aislado	Número	Media	Desviación típica
1	65	0.894	0.019
2	65	0.416	0.043
3	51	0.175	0.027
9	1	0.121	0.000

Distancias entre la función real y la estimada

Valor máximo: 0.4623 Valor mínimo: 0.0693
 Media: 0.1678 Desviación típica: 0.1204

Tabla 4

VI - Estimación no paramétrica de la densidad.

Muestras procedentes de una simulación de parámetros

$$\theta_1 = 0.5, \quad \theta_3 = 0.5, \quad \theta_5 = 0.5, \quad \theta_7 = 0.5$$

$$\theta_i = 0.0 \quad \text{en caso contrario}$$

Nivel de significación: 5 %
 Tamaño muestral simulado: 30 Número de simulaciones: 44

Ajuste de las simulaciones. $D_n = \sup |S_n(x) - F(x)|$

Valor máximo: 0.2412 Valor mínimo: 0.0627
 Media: 0.1467 Desviación típica: 0.0389
 Superan Kolmogorov-Smirnov: 44 Porcentaje: 100.00 %

Combinaciones de parámetros resultantes.

Combinación	Número	Porcentaje
1-3-5	13	29.55 %
1-2-3-5	11	25.00 %
1-3-5-7	9	20.45 %
1-2-5	3	6.82 %
1-5-9	2	4.55 %
1-2-5-7	2	4.55 %
5	1	2.27 %
3-9	1	2.27 %
1-5-7	1	2.27 %
1-3-5-9	1	2.27 %

Parámetro aislado	Número	Media	Desviación típica
5	43	0.622	0.095
1	42	0.596	0.087
3	35	0.432	0.092
2	16	0.414	0.054
7	12	0.318	0.074
9	4	-0.444	0.150

Distancias entre la función real y la estimada

Valor máximo: 3.1416 Valor mínimo: 1.9250
 Media: 2.3991 Desviación típica: 0.2544

Tabla 5

VI - Estimación no paramétrica de la densidad.

Muestras procedentes de una simulación de parámetros

$$\theta_1 = 0.5, \quad \theta_3 = 0.5, \quad \theta_5 = 0.5, \quad \theta_7 = 0.5$$

$$\theta_i = 0.0 \quad \text{en caso contrario}$$

Nivel de significación: 5 %
 Tamaño muestral simulado: 200 Número de simulaciones: 52

Ajuste de las simulaciones. $D_n = \sup |S_n(x) - F(x)|$

Valor máximo: 0.1111 Valor mínimo: 0.0300
 Media: 0.0562 Desviación típica: 0.0175
 Superan Kolmogorov-Smimov: 51 Porcentaje: 98.08 %

Combinaciones de parámetros resultantes.

Combinación	Número	Porcentaje
1-3-5-7	52	100.00 %

Parámetro aislado	Número	Media	Desviación típica
1	52	0.498	0.032
3	52	0.495	0.035
5	52	0.501	0.028
7	52	0.501	0.032

Distancias entre la función real y la estimada

Valor máximo: 0.2409 Valor mínimo: 0.0309
 Media: 0.1171 Desviación típica: 0.0484

Tabla 6

VI - Estimación no paramétrica de la densidad.

Muestras procedentes de una simulación de parámetros

$$\theta_1 = 0.5, \quad \theta_3 = 0.5, \quad \theta_5 = 0.5, \quad \theta_7 = 0.5$$

$$\theta_i = 0.0 \quad \text{en caso contrario}$$

Nivel de significación: 10 %
 Tamaño muestral simulado: 30 Número de simulaciones: 31

Ajuste de las simulaciones. $D_n = \sup |S_n(x) - F(x)|$

Valor máximo: 0.2413 Valor mínimo: 0.0680
 Media: 0.1416 Desviación típica: 0.0450
 Superan Kolmogorov-Smirnov: 31 Porcentaje: 100.00 %

Combinaciones de parámetros resultantes.

Combinación	Número	Porcentaje
1-3-5	12	38.76 %
1-2-3-5	5	16.15 %
1-6	3	9.69 %
1-2-5-7	3	9.69 %
1-2-3-5-7	2	6.45 %
1-4	1	3.23 %
1-5	1	3.23 %
1-5-8	1	3.23 %
1-7-9	1	3.23 %
1-2-5-9	1	3.23 %
1-3-5-9	1	3.23 %

Parámetro aislado	Número	Media	Desviación típica
1	31	0.674	0.137
5	26	0.586	0.080
3	20	0.426	0.073
2	11	0.414	0.074
7	6	0.334	0.045
6	3	0.332	0.012
9	3	-0.351	0.118
4	1	-0.234	0.000
8	1	-0.380	0.000

Distancias entre la función real y la estimada

Valor máximo: 2.5599 Valor mínimo: 1.7470
 Media: 2.2621 Desviación típica: 0.2329

Tabla 7

VI - Estimación no paramétrica de la densidad.

Muestras procedentes de una simulación de parámetros

$$\theta_1 = 0.5, \quad \theta_3 = 0.5, \quad \theta_5 = 0.5, \quad \theta_7 = 0.5$$

$$\theta_i = 0.0 \quad \text{en caso contrario}$$

Nivel de significación: 10 %
 Tamaño muestral simulado: 200 Número de simulaciones: 66

Ajuste de las simulaciones. $D_n = \sup |S_n(x) - F(x)|$

Valor máximo: 0.1088 Valor mínimo: 0.0294
 Media: 0.0565 Desviación típica: 0.0183
 Superan Kolmogorov-Smirnov: 63 Porcentaje: 95.46 %

Combinaciones de parámetros resultantes.

Combinación	Número	Porcentaje
1-3-5-7	64	96.97 %
1-3-5-7-9	2	3.03 %

Parámetro aislado	Número	Media	Desviación típica
1	66	0.500	0.033
3	66	0.500	0.032
5	66	0.499	0.032
7	66	0.497	0.034
9	2	0.501	0.106

Distancias entre la función real y la estimada

Valor máximo: 0.2507 Valor mínimo: 0.0173
 Media: 0.1236 Desviación típica: 0.0490

VI - Estimación no paramétrica de la densidad.

Muestras procedentes de una simulación de parámetros

$$\theta_3 = 0.9 \quad , \quad \theta_4 = 0.4 \quad , \quad \theta_6 = 0.17$$

$$\theta_i = 0.0 \quad \text{en caso contrario}$$

Nivel de significación: 5 %
 Tamaño muestral simulado: 30 Número de simulaciones: 56

Ajuste de las simulaciones. $D_n = \sup |S_n(x) - F(x)|$

Valor máximo: 0.2922 Valor mínimo: 0.0512
 Media: 0.1380 Desviación típica: 0.0535
 Superan Kolmogorov-Smirnov: 52 Porcentaje: 92.86 %

Combinaciones de parámetros resultantes.

Combinación	Número	Porcentaje
3-4	30	53.57 %
3	14	25.00 %
1-3-4-5	5	8.93 %
1-3-5	3	5.36 %
3-6	2	3.57 %
2-4-5	1	1.79 %
1-3-4-6	1	1.79 %

Parámetro aislado	Número	Media	Desviación típica
3	55	0.870	0.170
4	37	0.412	0.077
5	9	0.349	0.053
1	8	0.723	0.084
6	3	0.367	0.099
2	1	0.726	0.000

Distancias entre la función real y la estimada

Valor máximo: 2.6633 Valor mínimo: 0.1936
 Media: 0.7978 Desviación típica: 0.6260

Tabla 9

VI - Estimación no paramétrica de la densidad.

Muestras procedentes de una simulación de parámetros

$$\theta_3 = 0.9, \quad \theta_4 = 0.4, \quad \theta_6 = 0.17$$

$$\theta_i = 0.0 \quad \text{en caso contrario}$$

Nivel de significación: 5 %
 Tamaño muestral simulado: 200 Número de simulaciones: 49

Ajuste de las simulaciones. $D_n = \sup |S_n(x) - F(x)|$

Valor máximo: 0.1108 Valor mínimo: 0.0255
 Media: 0.0598 Desviación típica: 0.0208
 Superan Kolmogorov-Smirnov: 46 Porcentaje: 93.88 %

Combinaciones de parámetros resultantes.

Combinación	Número	Porcentaje
3-4-6	32	65.31 %
3-4	8	16.33 %
3-5-4	4	8.16 %
3-4-7	1	2.04 %
3-4-6-1	1	2.04 %
3-4-5-7	1	2.04 %
1-3-4-5-6-7	1	2.04 %
1-2-3-4-5-6-7	1	2.04 %

Parámetro aislado	Número	Media	Desviación típica
3	49	0.892	0.073
4	49	0.388	0.051
6	34	0.170	0.039
5	7	0.157	0.084
7	4	-0.002	0.409
1	3	-0.402	0.463
2	1	0.132	0.000

Distancias entre la función real y la estimada

Valor máximo: 1.7767 Valor mínimo: 0.0665
 Media: 0.2711 Desviación típica: 0.3371

Tabla 10

VI - Estimación no paramétrica de la densidad.

Muestras procedentes de una simulación de parámetros

$$\theta_3 = 0.9 \quad , \quad \theta_4 = 0.4 \quad , \quad \theta_6 = 0.17$$

$$\theta_i = 0.0 \quad \text{en caso contrario}$$

Nivel de significación: 10 %
 Tamaño muestral simulado: 30 Número de simulaciones: 48

Ajuste de las simulaciones. $D_n = \sup |S_n(x) - F(x)|$

Valor máximo: 0.2639 Valor mínimo: 0.0643
 Media: 0.1389 Desviación típica: 0.0480
 Superan Kolmogorov-Smirnov: 47 Porcentaje: 97.92 %

Combinaciones de parámetros resultantes.

Combinación	Número	Porcentaje
3-4	26	55.08 %
3	11	22.88 %
1-3-4-5	3	6.24 %
1-4-2-5	2	4.17 %
3-6	1	2.08 %
1-3-5	1	2.08 %
2-4-5	1	2.08 %
3-4-5	1	2.08 %
3-4-6	1	2.08 %
1-3-5-6-9	1	2.08 %

Parámetro aislado	Número	Media	Desviación típica
3	45	0.876	0.154
4	34	0.434	0.108
5	9	0.365	0.080
1	7	0.679	0.088
2	3	0.501	0.086
6	3	0.254	0.075
9	1	0.323	0.000

Distancias entre la función real y la estimada

Valor máximo: 2.6889 Valor mínimo: 0.1348
 Media: 0.8391 Desviación típica: 0.6980

Tabla 11

VI - Estimación no paramétrica de la densidad.

Muestras procedentes de una simulación de parámetros

$$\theta_3 = 0.9, \quad \theta_4 = 0.4, \quad \theta_6 = 0.17$$

$$\theta_i = 0.0 \quad \text{en caso contrario}$$

Nivel de significación: 10 %
 Tamaño muestral simulado: 200 Número de simulaciones: 54

Ajuste de las simulaciones. $D_n = \sup |S_n(x) - F(x)|$

Valor máximo: 0.1108 Valor mínimo: 0.0255
 Media: 0.0595 Desviación típica: 0.0208
 Superan Kolmogorov-Smirnov: 46 Porcentaje: 92.00 %

Combinaciones de parámetros resultantes.

Combinación	Número	Porcentaje
3-4-6	32	64.00 %
3-4	8	16.00 %
3-5-4	4	8.00 %
3-4-7	1	2.00 %
3-4-6-1	1	2.00 %
3-4-5-7	1	2.00 %
1-3-4-5-6	1	2.00 %
1-3-4-5-6-7	1	2.00 %
1-2-3-4-5-6-7	1	2.00 %

Parámetro aislado	Número	Media	Desviación típica
3	50	0.884	0.090
4	50	0.386	0.052
6	35	0.169	0.039
5	8	0.182	0.105
1	4	0.480	0.409
7	4	-0.002	0.180
2	1	0.132	0.000

Distancias entre la función real y la estimada

Valor máximo: 1.8560 Valor mínimo: 0.0665
 Media: 0.3027 Desviación típica: 0.4020

Tabla 12

VI - Estimación no paramétrica de la densidad.

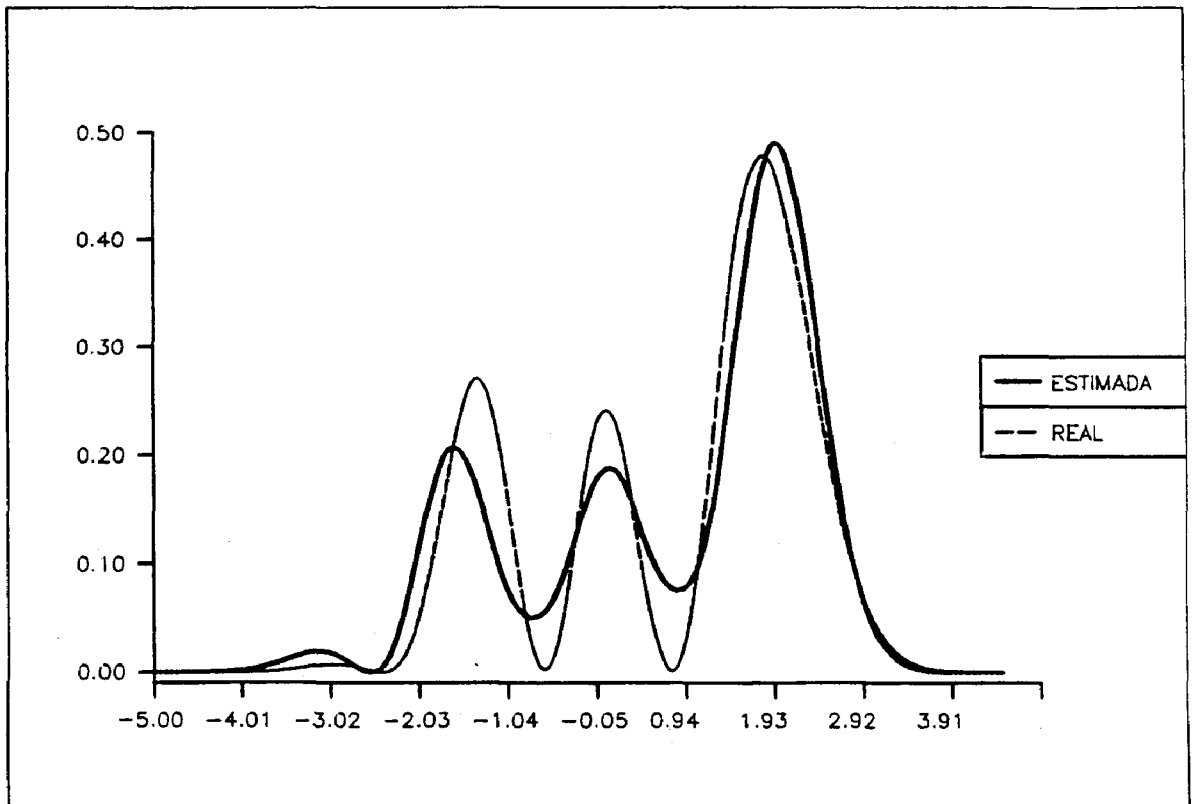


Figura 17. *Función de densidad correspondiente a los parámetros del ejemplo 6.3.2.2. frente a la estimación obtenida de la misma de parámetros*

$$\begin{aligned}\theta_1 &= 0.670 \\ \theta_2 &= 0.132 \\ \theta_3 &= 0.537 \\ \theta_4 &= 0.257 \\ \theta_5 &= 0.278 \\ \theta_6 &= 0.264 \\ \theta_7 &= -0.181\end{aligned}$$

En la Fig. 17 presentamos un resultado correspondiente al ejemplo 6.3.2.3. con una muestra de tamaño 200 y un nivel de significación del 5 %, que corresponde al resultado que difiere más del original de los presentados en la Tabla 10 . Se observa que la moda más importante es recuperada con un grado aceptable de precisión, mientras que en las otras modas la ubicación es la correcta aun cuando en la función estimada son menos acusadas. El logaritmo de la verosimilitud para la muestra simulada con los parámetros originales es de -305.3454, mientras que con los parámetros obtenidos por el algoritmo es de -309.275. La distancia entre la función real y la estimada es de 1.7767.

Ejemplo 6.3.2.4.

En las Figs. 18a y 18b presentamos la estimación obtenida al aplicar el algoritmo sobre los datos utilizados en las Figs. 9a, 9b, 10a y 10b . La simulación corresponde a una distribución $N(0,1)$ con tamaños muestrales de 30 y 50 respectivamente. Comparando los resultados entonces obtenidos aplicando de forma directa la estimación máximo verosímil con 15 y 30 parámetros, con los resultados obtenidos ahora, comprobamos como desaparece la distorsión. Con la muestra de tamaño 30 el único parámetro significativo es el primero, con la de tamaño 50 son significativos el primero y tercero con valores 0.9793 y 0.2023 respectivamente. Las distancias entre la función de densidad estimada y la real son de 0.539275 para la muestra de tamaño 30, y de 0.176686 para la de tamaño 50. En la Fig. 18a se aprecia una desviación notable de la $N(0,1)$. Sin embargo el resultado es lógico ya que, tal y como indicamos en el apartado 6.2.3. , la desviación típica muestral obtenida en esta simulación es de $s = 0.70483$, y el resultado coincide con una distribución $N(0, 1/2)$. En el presente ejemplo así como en todos los siguientes, hemos trabajado con un nivel de significación del 5 % para todos los contrastes.

VI - Estimación no paramétrica de la densidad.

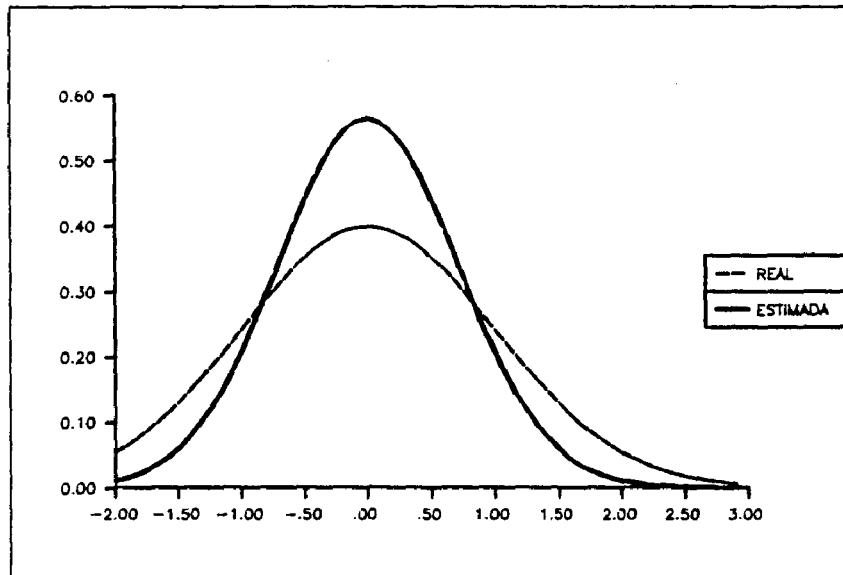


Figura 18a. Estimación stepwise obtenida a partir de una muestra simulada de tamaño 30 de una distribución $N(0,1)$. Los datos son los mismos que los utilizados en las Figs. 9a y 9b.

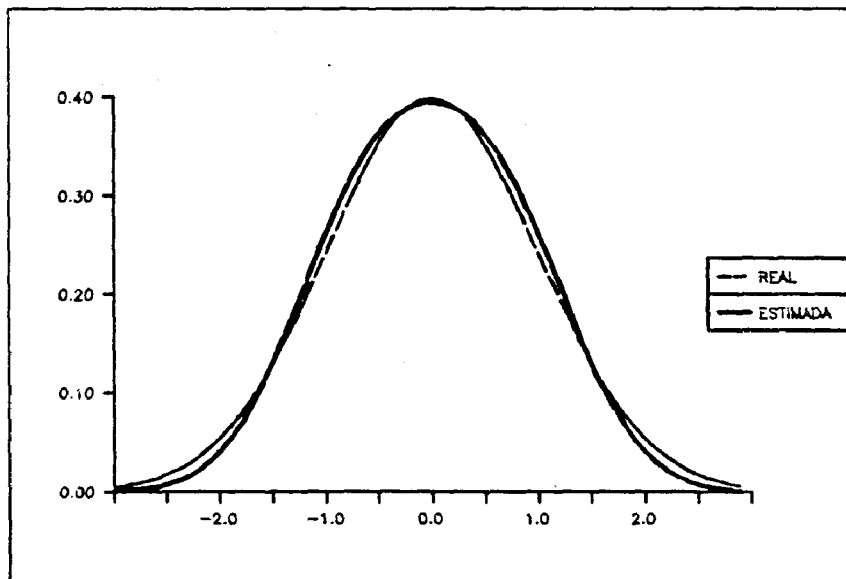


Figura 18b. Estimación stepwise obtenida a partir de una muestra simulada de tamaño 50 de una distribución $N(0,1)$. Los datos son los mismos que los utilizados en las Figs. 10a y 10b.

VI - Estimación no paramétrica de la densidad.

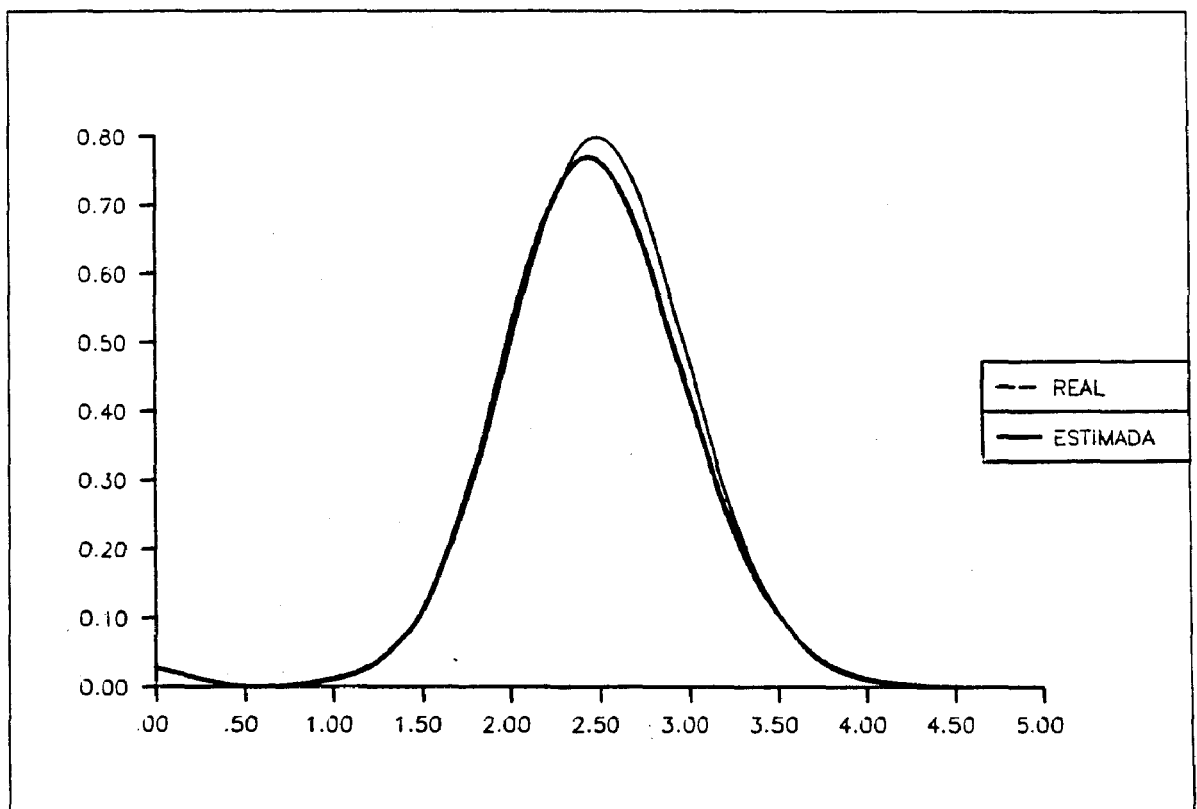


Figura 19. Estimación stepwise obtenida a partir de una muestra simulada de tamaño 100 de una distribución $N(5/2 , 1/4)$.

Valores obtenidos para los parámetros.

$$\begin{aligned}\theta_2 &= 0.3275 \\ \theta_3 &= 0.4869 \\ \theta_4 &= 0.5247 \\ \theta_5 &= 0.4434 \\ \theta_6 &= 0.3350 \\ \theta_7 &= 0.2674\end{aligned}$$

VI - Estimación no paramétrica de la densidad.

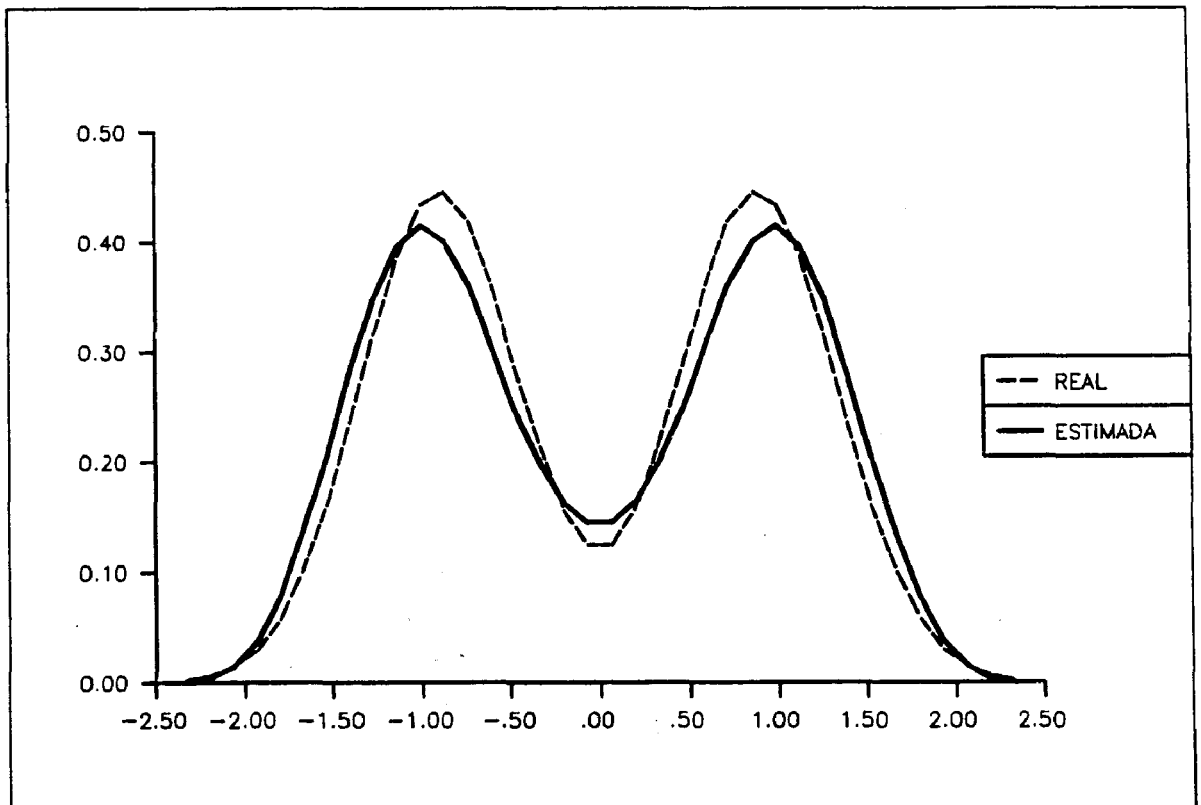


Figura 20. Estimación stepwise obtenida a partir de una muestra simulada de tamaño 100 de una distribución $0.5 N(- 2/\sqrt{5} , 1/\sqrt{5}) + 0.5 N(+ 2/\sqrt{5} , 1/\sqrt{5})$.

Valores obtenidos para los parámetros.

$$\theta_1 = 0.9035$$

$$\theta_3 = 0.3351$$

$$\theta_5 = -0.2672$$

VI - Estimación no paramétrica de la densidad.

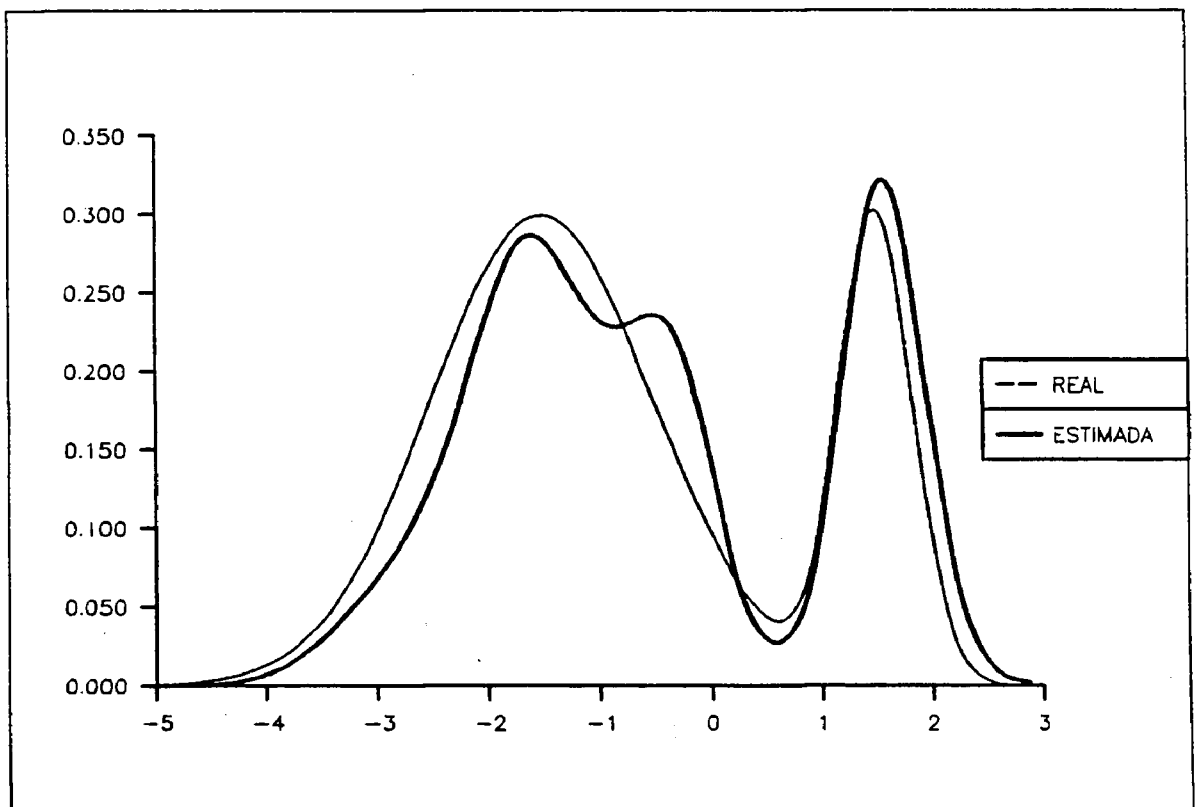


Figura 21. Estimación stepwise obtenida a partir de una muestra simulada de tamaño 200 de una distribución $0.75 N(-3/2, 1) + 0.25 N(+3/2, 1/3)$.

Valores obtenidos para los parámetros.

$$\begin{aligned}\theta_1 &= 0.7299 \\ \theta_2 &= -0.1481 \\ \theta_3 &= 0.5905 \\ \theta_5 &= 0.2017 \\ \theta_6 &= -0.1987 \\ \theta_9 &= 0.1284\end{aligned}$$

Ejemplo 6.3.2.5.

En la Fig. 19 presentamos los resultados obtenidos a partir de una simulación de tamaño 100 de una distribución $N(5/2, 1/4)$. La distancia entre la distribución real y la estimación es de 0.349931. Se han encontrado 6 parámetros significativos.

Ejemplo 6.3.2.6.

Simulada una muestra de tamaño 100 de una distribución bimodal con función de densidad mezcla de dos distribuciones normales del tipo $0.5 N(-2/\sqrt{5}, 1/\sqrt{5}) + 0.5 N(+2/\sqrt{5}, 1/\sqrt{5})$, el resultado obtenido es presentado en la Fig. 20. La distancia entre la estimación y la densidad real es de 0.214552, con 3 parámetros significativos.

Ejemplo 6.3.2.7.

Finalmente en la Fig. 21 presentamos el resultado obtenido al realizar una simulación de tamaño 200 de una mezcla de normales $0.75 N(-3/2, 1) + 0.25 N(+3/2, 1/3)$. La distancia entre la estimación y la densidad real es de 0.297978, y se han obtenido 6 parámetros significativos.

6.3.3. Comparación con otros métodos.

Uno de los métodos más ampliamente utilizados en estimación no paramétrica, es el método de los Kernel, Rosenblatt (1956), Parzen(1962). Dada una muestra aleatoria x_1, \dots, x_m , consideramos como estimador de $p(x)$:

$$\hat{p}(x) = \frac{1}{m h_m} \sum_{j=1}^m K\left(\frac{x - x_j}{h_m}\right) \quad [6.28]$$

VI - Estimación no paramétrica de la densidad

donde K es una función Kernel que verifica:

$$\int_{-\infty}^{\infty} |K(y)| dy < \infty$$

$$\sup_{-\infty < y < \infty} |K(y)| < \infty$$

$$\lim_{y \rightarrow \infty} y|K(y)| = 0$$

$$K(y) \geq 0$$

$$\int_{-\infty}^{\infty} K(y) dy = 1$$

y h_m es un parámetro de la amplitud de los intervalos de estimación.

Existen varias funciones Kernel de uso general, nosotros para comparar resultados con el algoritmo stepwise hemos optado por una de las más utilizadas, la función Kernel de Gauss, definida por:

$$K(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \quad |y| < \infty. \quad [6.29]$$

La comparación la hemos realizado simulando 25 muestras de tamaño 25 de una distribución $N(0,1)$, aplicando a cada simulación la estimación Kernel y la stepwise y evaluando posteriormente la distancia $d = 2 \cos^{-1} \int_x \sqrt{p} \sqrt{q} dx$. La estimación Kernel depende en gran medida del valor h_m que se considere. Nosotros hemos utilizado $h_m = 0.56$, al ser el óptimo para las simulaciones utilizadas según se desprende de Tapia y Thompson (1978). Los resultados se muestran en la Tabla 13.

ALGORITMO STEPWISE

Valor máximo:	0.688	Valor mínimo:	0.174
Media:	0.419	Desviación típica:	0.162

ESTIMACION KERNEL

Valor máximo:	0.819	Valor mínimo:	0.189
Media:	0.416	Desviación típica:	0.146

Tabla 13

Como puede observarse los resultados son prácticamente coincidentes con ambos métodos.

Existen otros métodos explícitamente desarrollados con el objetivo de eliminar el problema de la inestabilidad dimensional, destaquemos el concepto de máxima verosimilitud penalizada (MPLE). Debido en su origen a Good (1971) y desarrollado posteriormente por él mismo, Good y Gaskins (1971,1980) y por otros autores, de Montricher et al. (1975), Tapia y Thompson (1978), el método se basa en las siguientes consideraciones.

Sea F una clase de funciones de densidad. Una función penalizadora $\varphi: F \rightarrow \mathbf{R}$ es un funcional real definido sobre F . Dada una muestra aleatoria x_1, \dots, x_m , se define la verosimilitud φ -penalizada de que $p \in F$ dé lugar a la muestra, como el funcional:

$$\hat{L}(p) = \prod_{i=1}^n p(x_i) \exp(-\varphi(p)) \quad [6.30]$$

Obtener la solución MPLE consiste en maximizar [6.49] condicionado a:

$$p \in F \quad , \quad \int_{\mathcal{X}} p \, d\mu = 1 \quad , \quad p \geq 0 \quad [6.31]$$

Diversas funciones penalizadoras han sido utilizadas, destaquemos la propuesta por Good (1971), donde:

$$\varphi(p) = \alpha \int_{-\infty}^{+\infty} (p^{(1)}(x))^2 \, dx + \beta \int_{-\infty}^{+\infty} (p^{(2)}(x))^2 \, dx \quad [6.32]$$

Good y Gaskins (1971) sugieren, por criterios de simplicidad, trabajar con funciones $\hat{p}(x)$ de la forma definida en [5.3], coincidentes con las funciones estudiadas por nosotros en el capítulo 5 . En los trabajos citados anteriormente pueden encontrarse estudios teóricos referentes a la existencia de soluciones MPLE así como ejemplos numéricos de la aplicación de la técnica.

6.4. Estimación minimizando la esperanza del cuadrado de la distancia.

6.4.1. Consideraciones generales.

Una posible alternativa al empleo de las técnicas MLE surge con el método de estimación presentado en la sección 4.1, basado en la distancia entre individuos. Tal y como comprobamos en el capítulo 4 la minimización de la esperanza de la distancia al cuadrado entre la muestra real y otra muestra hipotética conduce a la estimación MLE si utilizamos la distancia inmediata [3.6], pero no así utilizando la distancia estructural [3.27]. Nuestro objetivo es comprobar que la utilización de la distancia [3.27] proporciona estimaciones que corrigen en parte el defecto de inestabilidad dimensional presentado por la estimación MLE.

Tomando como base las funciones definidas en el apartado 5.3.2 y una muestra x_1, \dots, x_m , el tensor métrico [3.29] respecto a las coordenadas x_i resulta:

$$g_{\alpha\beta} = \frac{2}{m} \sum_{k=0}^n \frac{\sum_{i,j=0}^n \theta_j \theta_i [k^{1/2} h_{k-1}(x_\alpha) h_j(x_\alpha) - j^{1/2} h_{j-1}(x_\alpha) h_k(x_\alpha)] [k^{1/2} h_{k-1}(x_\beta) h_i(x_\beta) - i^{1/2} h_{i-1}(x_\beta) h_k(x_\beta)]}{[\sum_{l=0}^n \theta_l h_l(x_\alpha)]^2 [\sum_{l=0}^n \theta_l h_l(x_\beta)]^2}$$

$$\alpha, \beta = 1, \dots, m \quad [6.33]$$

La complejidad de las funciones empleadas no ha permitido por el momento la obtención de la fórmula explícita de la distancia entre dos individuos estadísticos. En el siguiente apartado proponemos una aproximación a la distancia riemanniana entre dos puntos de una variedad paramétrica y posteriormente utilizaremos dicha aproximación para llevar a cabo la estimación de los parámetros.

6.4.2. Una aproximación a la distancia riemanniana.

Aun cuando una variedad paramétrica no sea euclídea, siempre es posible encontrar un sistema de coordenadas en el cual los símbolos de Christoffel se anulen en un punto P del espacio. Tales coordenadas se denominan coordenadas geodésicas para ese punto particular o localmente cartesianas en P. Por tanto el tensor métrico puede ser considerado constante en un entorno restringido del punto P y podemos calcular la distancia entre dos puntos cualesquiera de la variedad como si fuera un espacio euclídeo.

Las leyes de transformación de los símbolos de Christoffel pueden ser expresadas como:

$$\frac{\partial^2 x_\epsilon}{\partial y_i \partial y_j} + x^\alpha \Gamma_{\alpha\beta}^\epsilon \frac{\partial x_\alpha}{\partial y_i} \frac{\partial x_\beta}{\partial y_j} = y^\gamma \Gamma_{ij}^\gamma \frac{\partial x_\epsilon}{\partial y_\gamma} \quad [6.34]$$

$$\epsilon, i, j = 1, \dots, m$$

donde $x^\alpha \Gamma_{\alpha\beta}^\epsilon$ es el valor de los símbolos de Christoffel en las coordenadas originales, y $y^\gamma \Gamma_{ij}^\gamma$ es el valor en las nuevas coordenadas, pero puesto que $y^\gamma \Gamma_{ij}^\gamma = 0$ resulta:

$$\frac{\partial^2 x_\epsilon}{\partial y_i \partial y_j} + \Gamma_{\alpha\beta}^\epsilon \frac{\partial x_\alpha}{\partial y_i} \frac{\partial x_\beta}{\partial y_j} \quad \epsilon, i, j = 1, \dots, m \quad [6.35]$$

Sokolnikoff (1971) propone una solución particular para este sistema en forma de polinomio de segundo grado:

$$x_\alpha = x_\alpha^p + y_\alpha - \frac{1}{2} \Gamma_{\beta\epsilon}^\alpha y_\beta y_\epsilon \quad \alpha = 1, \dots, m \quad [6.36]$$

donde x_α^p es el valor de x_α en p y $\Gamma_{\beta\epsilon}^\alpha$ el valor de los símbolos de Christoffel en p .

De [6.36] podemos ver que las coordenadas del punto p en el nuevo sistema de referencia serán $(0, \dots, 0)$. Además en el nuevo sistema:

$$\frac{\partial x_\alpha}{\partial y_\varepsilon} = \delta_\varepsilon^\alpha - (\Gamma_{\beta\varepsilon}^\alpha)_p y_\beta \quad [6.37]$$

por tanto el tensor métrico vendrá dado en el nuevo sistema por:

$$\bar{g}_{\varepsilon\nu} = (\delta_\varepsilon^\alpha - (\Gamma_{\mu\varepsilon}^\alpha)_p y_\mu) (\delta_\nu^\beta - (\Gamma_{\pi\nu}^\beta)_p y_\pi) g_{\alpha\beta} \quad [6.38]$$

y como en el punto p $\frac{\partial x_\alpha}{\partial y_\mu} = \delta_\mu^\alpha$, el tensor métrico en el punto p será $\bar{g}_{\mu\nu}(0, \dots, 0) = g_{\mu\nu}(x_1^p, \dots, x_m^p)$

Introduzcamos en [6.36] un parámetro perturbador λ , resulta

$$\lambda \Delta_\alpha = + y_\alpha - \frac{1}{2} \Gamma_{\beta\varepsilon}^\alpha y_\beta y_\varepsilon \quad \alpha = 1, \dots, m \quad [6.39]$$

donde $\Delta_\alpha = x_\alpha - x_\alpha^p$, evidentemente para $\lambda = 0$ tenemos $y_\alpha = 0 \quad \alpha = 1, \dots, m$.

A continuación vamos a expresar las nuevas coordenadas $y_\alpha(\lambda)$ mediante un desarrollo en serie de Taylor en el punto $\lambda = 0$, la solución que nos interesa es la correspondiente a $\lambda = 1$. Derivemos [6.39] respecto λ

$$\Delta_\alpha = \frac{\partial y_\alpha}{\partial \lambda} - (\Gamma_{\beta\varepsilon}^\alpha)_p y_\varepsilon \frac{\partial y_\beta}{\partial \lambda} \quad \alpha = 1, \dots, m \quad [6.40]$$

derivemos una segunda vez y resulta:

$$0 = \frac{\partial^2 y_\alpha}{\partial \lambda^2} - (\Gamma_{\beta\varepsilon}^\alpha)_p \left[\frac{\partial y^\varepsilon}{\partial \lambda} \frac{\partial y^\beta}{\partial \lambda} + y_\varepsilon \frac{\partial^2 y_\beta}{\partial \lambda^2} \right] \quad [6.41]$$

Las expresiones [6.40] y [6.41] en el punto $\lambda = 0$ dan lugar a:

VI - Estimación no paramétrica de la densidad

$$\frac{\partial y_\alpha}{\partial \lambda} = \Delta_\alpha \quad , \quad \frac{\partial^2 y_\alpha}{\partial \lambda^2} = \Gamma_{\beta\epsilon}^\alpha \Delta_\beta \Delta_\epsilon \quad [6.42]$$

por lo tanto a partir del desarrollo de Taylor de grado 2 de las nuevas coordenadas en función de λ , podemos aproximar el valor de $y_\alpha(1)$ mediante:

$$y_\alpha(1) = \Delta_\alpha + \frac{1}{2} \Gamma_{\beta\epsilon}^\alpha \Delta_\beta \Delta_\epsilon \quad \alpha = 1, \dots, m \quad [6.43]$$

La distancia entre dos individuos $\mathbf{a} = (x_1^a, \dots, x_m^a)$ y $\mathbf{b} = (x_1^b, \dots, x_m^b)$ puede calcularse dado un punto p como:

$$d^2(\mathbf{a}, \mathbf{b}) \approx (g_{\alpha\beta})_p (y_\alpha^b - y_\alpha^a) (y_\beta^b - y_\beta^a) \quad [6.44]$$

y tomando como valor de las coordenadas y_α la primera aproximación proporcionada por [6.42], $y_\alpha \approx \Delta_\alpha = x_\alpha - x_\alpha^p$, obtenemos una sencilla expresión que nos proporciona una aproximación a la distancia:

$$d^2(\mathbf{a}, \mathbf{b}) \approx (g_{\alpha\beta})_p (x_\alpha^b - x_\alpha^a) (x_\beta^b - x_\beta^a) \quad [6.45]$$

Tomando como aproximación a las nuevas coordenadas la expresión [6.43] completa, obtendríamos finalmente como aproximación a la distancia:

$$\begin{aligned} d^2(\mathbf{a}, \mathbf{b}) \approx & (g_{\alpha\beta})_p (x_\alpha^b - x_\alpha^a) (x_\beta^b - x_\beta^a) + \frac{1}{2} \left(\frac{\partial g_{\mu\nu}}{\partial x_\beta} \right)_p (x_\mu^b - x_\mu^a) (x_\nu^b - x_\nu^a) (x_\beta^b - x_\beta^a) + \\ & + \frac{1}{4} [\mu\nu, \beta]_p (\Gamma_{\epsilon\tau}^\beta)_p (x_\mu^b - x_\mu^a) (x_\nu^b - x_\nu^a) (x_\epsilon^b - x_\epsilon^a) (x_\tau^b - x_\tau^a) \end{aligned} \quad [6.46]$$

Una posibilidad para obtener una evaluación aproximada de la esperanza de la distancia al cuadrado entre la muestra que se posee y cualquier muestra hipotética procedente de una distribución concreta sería utilizar las expresiones [6.45] ó [6.46]. La elección del punto P es en cierta manera arbitraria, la opción más favorable para

obtener una buena aproximación a la distancia sería tomar un punto intermedio entre las dos muestras para las cuales estamos calculando la distancia. Sin embargo debido a la complejidad del tensor métrico [6.28] hemos optado por tomar como punto P el correspondiente a la muestra real, por lo tanto a la hora de calcular la esperanza, [6.28] es constante y la complejidad del cálculo queda reducida enormemente.

6.4.3. Algoritmo de minimización numérica.

La obtención de los valores de los parámetros que minimizan la esperanza del cuadrado de la distancia, requiere la solución de un problema de minimización numérica de funciones de varias variables con restricciones. Hemos optado por la utilización del algoritmo de minimización debido a Rosenbrok (1960).

Sea $g(\mathbf{x} = (x_1, \dots, x_m))$ la función a minimizar, el método parte de una aproximación \mathbf{x}_0 al mínimo, y de un conjunto de vectores unidad y direcciones mutuamente ortogonales $\xi_{10}, \dots, \xi_{m0}$. Inicialmente tomamos como vectores ortonormales los de la base canónica. En la etapa $k + 1$, tomamos un paso d_1 en la dirección ξ_{1k} , y evaluamos la función en el punto $\mathbf{x}_k = d_1 \xi_{1k}$. Si, para un $\alpha > 1$ dado

$$\begin{aligned} g(\mathbf{x}_k + d_1 \xi_{1k}) &\leq g(\mathbf{x}_k) \\ \text{y} \quad g(\mathbf{x}_k + \alpha d_1 \xi_{1k}) &> g(\mathbf{x}_k) \end{aligned} \quad [6.47]$$

el paso se considera provechoso. Si por el contrario

$$g(\mathbf{x}_k + \alpha d_1 \xi_{1k}) \leq g(\mathbf{x}_k) \quad [6.48]$$

sustituimos d_1 por αd_1 , y se intenta satisfacer [6.47] de nuevo. Si inicialmente tuvieramos que:

$$g(\mathbf{x}_k + d_1 \xi_{1k}) > g(\mathbf{x}_k) \quad [6.49]$$

sustituiríamos d_1 por $-\beta d_1$ ($0 < \beta < 1$) e intentaríamos satisfacer de nuevo [6.47]. A continuación avanzamos un paso d_2 a partir del punto $x_k + d_1 \xi_{1k}$ en la dirección ξ_{2k} repitiendo las comprobaciones [6.47] a [6.49]. Análogamente obtendríamos d_3, \dots, d_m . Llegando al final de la $k+1$ ésima etapa, obteniendo una nueva aproximación x_{k+1} al mínimo de g , donde

$$x_{k+1} = x_k + d_1 \xi_{1k} + \dots + d_m \xi_{mk} \quad [6.50]$$

El siguiente paso consiste en establecer un nuevo conjunto de vectores direccionales ortogonales, $\xi_{1,k+1}, \dots, \xi_{m,k+1}$ mediante el procedimiento de ortogonalización de Gram-Schmidt sobre el conjunto de vectores:

$$\begin{aligned} A_1 &= d_1 \xi_{1k} + \dots + d_m \xi_{mk} \\ A_2 &= d_2 \xi_{2k} + \dots + d_m \xi_{mk} \\ &\dots\dots\dots \\ A_m &= d_m \xi_{mk} \end{aligned} \quad [6.51]$$

El efecto de aplicar [6.51] y posteriormente Gram-Schmidt es asegurar que $\xi_{1,k+1}$ esté en la dirección de avance más rápida hacia el mínimo. $\xi_{2,k+1}$ es la mejor dirección normal a $\xi_{1,k+1}$, y así sucesivamente. Las iteraciones finalizan tras un número finito de etapas o, alternativamente, cuando $|x_{k+1} - x_k|$ es menor que una tolerancia prevista.

El problema de introducir la restricción [6.12] puede ser resuelto mediante la técnica de superficies de respuesta creadas, Carroll (1961), que para el caso de restricciones de la forma

$$k_i(x) = 0 \quad i = 1, \dots, p \quad [6.52]$$

consiste en la creación de superficies de respuesta $\varphi(\mathbf{x}, \alpha_r)$, resultando

$$\varphi(\mathbf{x}, \alpha_r) = g(\mathbf{x}) + \left(\frac{1}{\alpha_r}\right) \sum_{i=1}^p k_i^2(\mathbf{x}) \quad r = 1, 2, \dots \quad [6.53]$$

Esta técnica asegura que la función toma valores muy grandes en la proximidad de una restricción y se crea un mínimo dentro de la región permitida. Partiendo de un valor de α_1 buscamos un mínimo para $\varphi(\mathbf{x}, \alpha_1)$ y buscamos un α_2 tal que $0 < \alpha_2 < \alpha_1$. Obtenemos una sucesión de puntos que bajo condiciones generales, tiende al verdadero mínimo de $g(\mathbf{x})$ para $\alpha_r \rightarrow 0$. Una completa exposición del método, del tratamiento de restricciones más generales así como ejemplos de aplicación puede encontrarse en Cohen (1977).

6.4.4. Resultados.

Presentamos a continuación algunos resultados comparativos de la aplicación de los métodos de estimación máximo verosímil (MLE) y de la minimización de la esperanza de la distancia entre individuos al cuadrado (MESD). Los ejemplos considerados son los mismos que algunos de los comentados en el apartado 6.2.3.

En la Fig. 22 con los datos de la Fig. 1, una muestra de tamaño $m = 1$, $x = 0$, y una estimación de 20 parámetros, se observa que la estimación MESD está más suavizada que la MLE.

En las Figs. 23 y 24 comparamos los resultados con los obtenidos en las Figs. 3 y 5. La muestra es de tamaño 2, $x_1 = -0.5$, $x_2 = 0.5$, y las estimaciones de 10 y 20 parámetros. Nuevamente se observa como las estimaciones MESD son sensiblemente más suaves que la MLE.

VI - Estimación no paramétrica de la densidad.

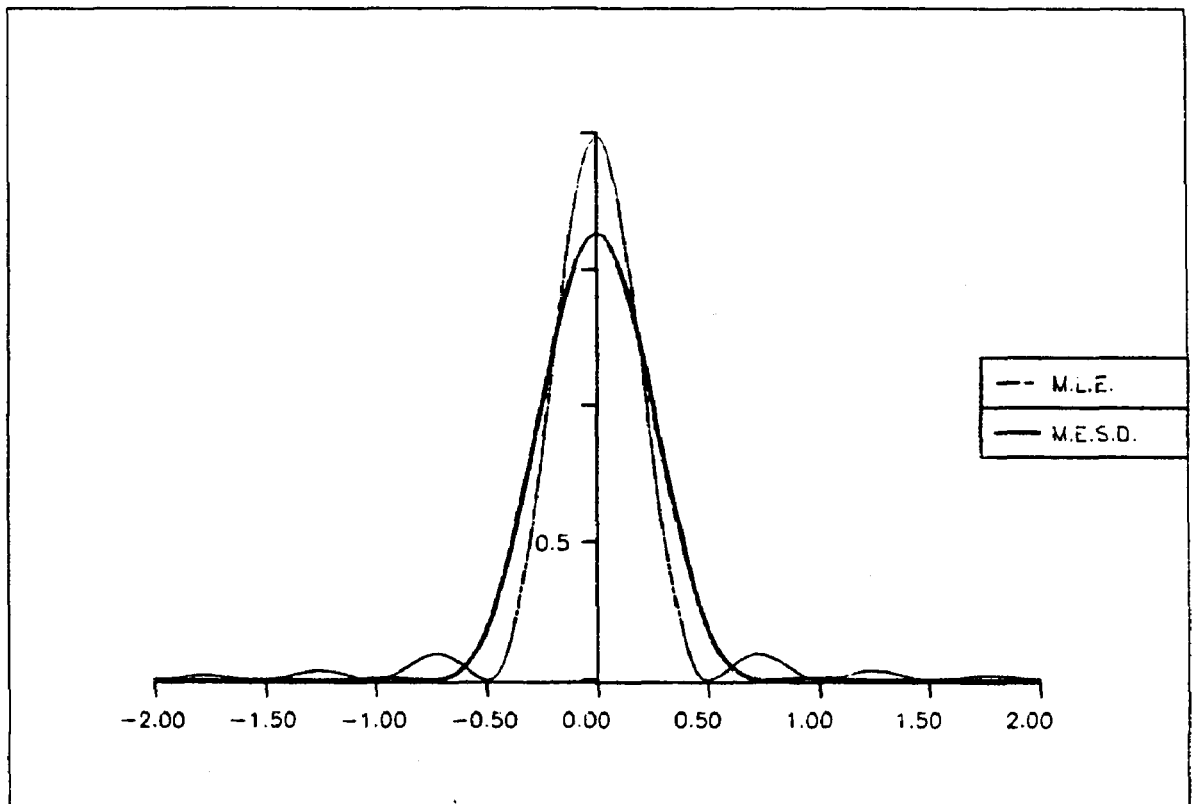


Figura 22. Estimaciones MLE y MESD con una muestra de tamaño $l = 20$ y $n = 20$ parámetros.

Valores obtenidos para los parámetros en la estimación MESD.

$$\begin{aligned}\theta_0 &= 0.7338459 \\ \theta_2 &= -0.44893847 \\ \theta_4 &= 0.33528298 \\ \theta_6 &= -0.25568469 \\ \theta_8 &= 0.19961080 \\ \theta_{10} &= -0.15284868 \\ \theta_{12} &= 0.11085614 \\ \theta_{14} &= -0.07919429 \\ \theta_{16} &= 0.05034236 \\ \theta_{18} &= -0.0246061\end{aligned}$$

$$\theta_{2k+1} = 0.0 \quad k = 0, \dots, 9$$

VI - Estimación no paramétrica de la densidad.

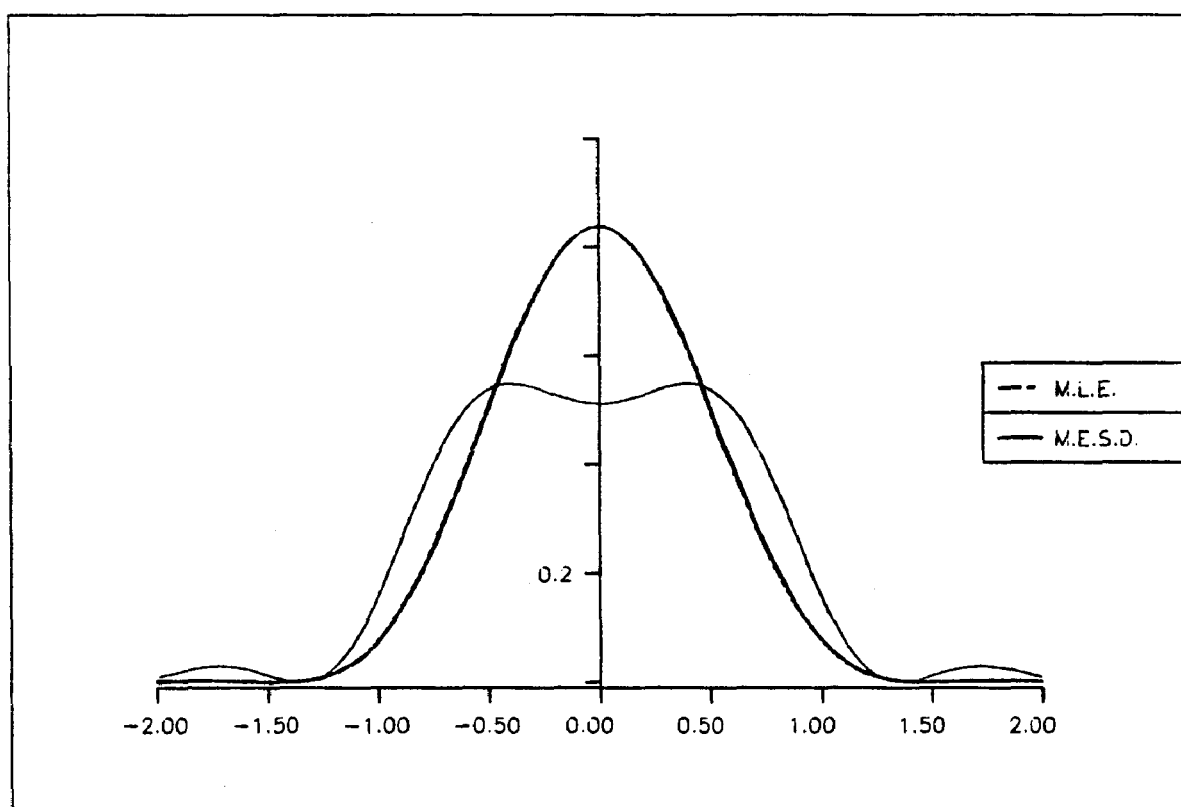


Figura 23. Estimaciones MLE y MESD con una muestra de tamaño 2 $x_1 = -0.5$, $x_2 = +0.5$ y $n = 10$ parámetros.

Valores obtenidos para los parámetros en la estimación MESD.

$$\theta_0 = 0.93427048$$

$$\theta_2 = -0.33978577$$

$$\theta_4 = 0.10416150$$

$$\theta_6 = 0.00405403$$

$$\theta_8 = -0.03206056$$

$$\theta_{2k+1} = 0.0 \quad k = 0, \dots, 4$$

VI - Estimación no paramétrica de la densidad.

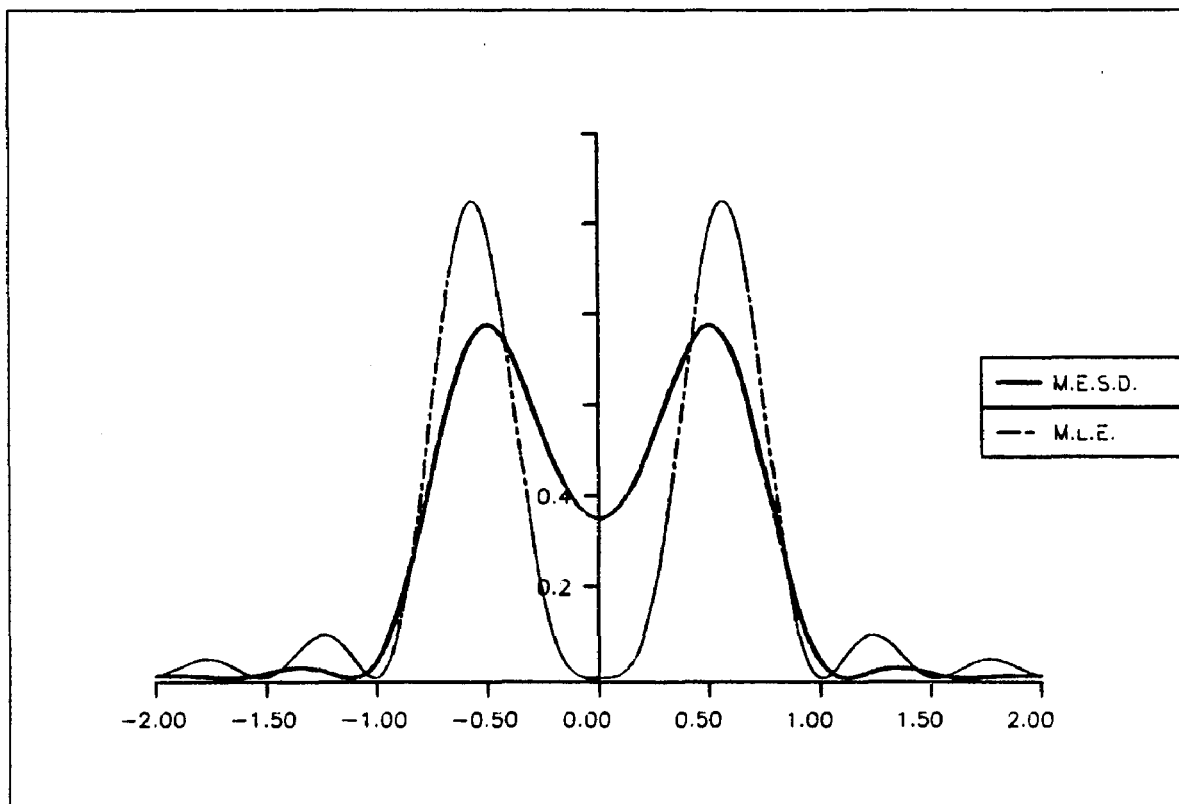


Figura 24. Estimaciones MLE y MESD con una muestra de tamaño $n = 20$ y $x_1 = -0.5$, $x_2 = +0.5$ y $n = 20$ parámetros.

Valores obtenidos para los parámetros en la estimación MESD.

$$\begin{aligned} \theta_0 &= 0.8731400 \\ \theta_2 &= -0.34151663 \\ \theta_4 &= 0.10378813 \\ \theta_6 &= 0.03673594 \\ \theta_8 &= -0.11868716 \\ \theta_{10} &= 0.15917754 \\ \theta_{12} &= -0.16823010 \\ \theta_{14} &= 0.15251627 \\ \theta_{16} &= -0.11687457 \\ \theta_{18} &= 0.06504072 \end{aligned}$$

$$\theta_{2k+1} = 0.0 \quad k = 0, \dots, 9$$

VI - Estimación no paramétrica de la densidad.

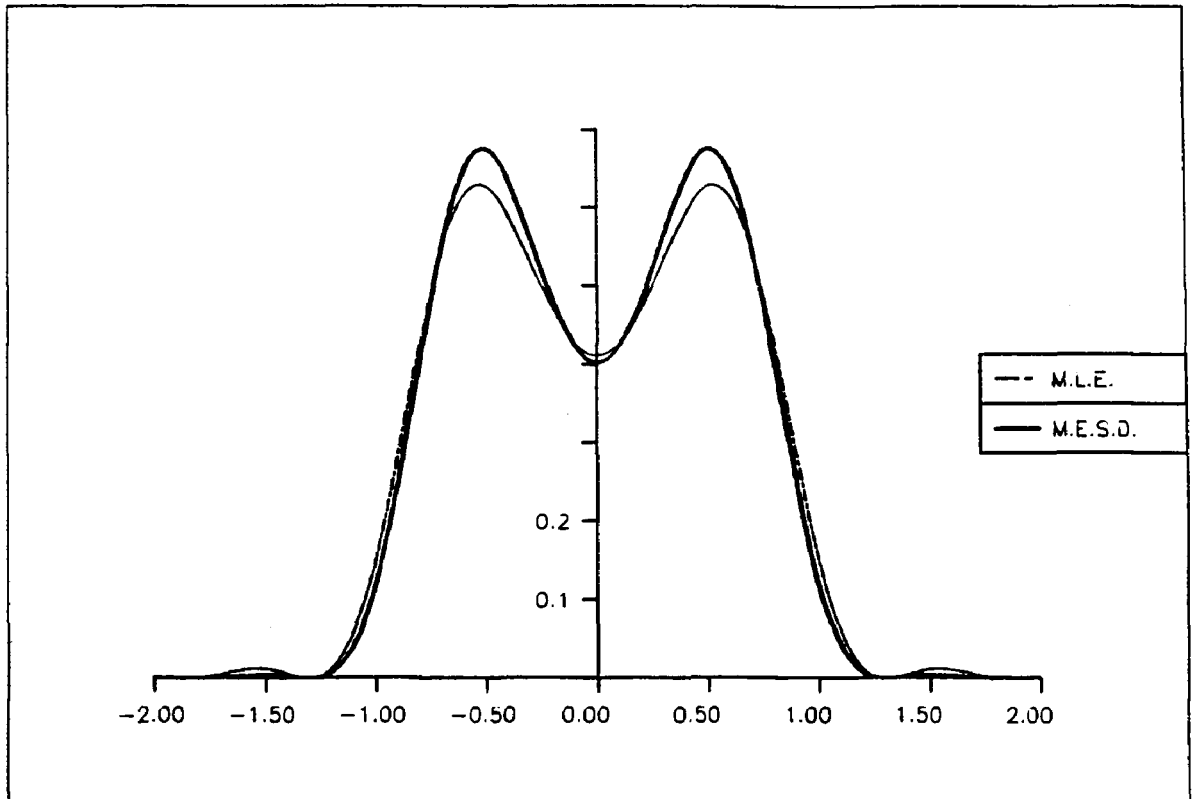


Figura 25. Estimaciones MLE y MESD con una muestra de tamaño 3 $x_1 = -1/\sqrt{2}$, $x_2 = 0.0$, $x_3 = +1/\sqrt{2}$ y $n = 20$ parámetros.

Valores obtenidos para los parámetros.

$$\begin{aligned} \theta_0 &= 0.92715721 \\ \theta_2 &= -0.27855287 \\ \theta_4 &= 0.02928407 \\ \theta_6 &= 0.08542500 \\ \theta_8 &= -0.12854033 \\ \theta_{10} &= 0.13065706 \\ \theta_{12} &= -0.11018162 \\ \theta_{14} &= 0.07944685 \\ \theta_{16} &= -0.047096878 \\ \theta_{18} &= 0.019217292 \end{aligned}$$

$$\theta_{2k+1} = 0.0 \quad k = 0, \dots, 9$$

VI - Estimación no paramétrica de la densidad

Finalmente en la Fig. 25 presentamos el resultado comparativo frente a la Fig. 7 , donde con una muestra de tamaño 3, $x_1 = -1/\sqrt{2}$, $x_2 = 0.$, $x_3 = +1/\sqrt{2}$, y una estimación de 20 parámetros se observa que en este ejemplo la estimación MESD resulta ser más bimodal que la MLE.

Debido al elevado tiempo de ejecución requerido por el algoritmo de minimización numérica al trabajar con un elevado número de variables, no ha sido posible la comparación de ambos métodos para un número mayor de parámetros.

VII. Aplicaciones a la Biología.

Presentamos algunas aplicaciones a la Biología de varias técnicas expuestas en capítulos precedentes. En un primer ejemplo utilizamos el algoritmo stepwise para estimar la función de densidad y distribución de una medida biométrica a partir de unos datos en forma de histograma donde se aprecia una bimodalidad acusada. En un segundo ejemplo realizamos una clasificación de diversos patrones electroforéticos, asimilándolos a funciones de densidad.

7.1. Estimación del tamaño en *Octopus vulgaris*.

7.2. Clasificación de patrones electroforéticos.

7.1. Estimación del tamaño en *Octopus vulgaris*.

7.1.1. Introducción.

La exploración previa de los datos con vistas a determinar o confirmar el tipo de distribución de los mismos, es un paso importante en todo análisis. Efectuar descripciones mediante histogramas o diagramas de barras es un adecuado primer paso que puede dar indicios de desviación de hipótesis previas o de la presencia de una mezcla de varias poblaciones. Cuando se presenten tales circunstancias y deseemos una representación funcional de la función de densidad o verificar la significación de irregularidades aparecidas en los histogramas, la aplicación de un método de estimación no paramétrico de la densidad del tipo del algoritmo stepwise, es claramente aconsejable.

Como ejemplo del funcionamiento del algoritmo stepwise de estimación no paramétrica aplicado a datos reales, hemos realizado el siguiente estudio basado en unos datos obtenidos de Alonso et al. (1976). En el texto mencionado se comentan algunos métodos para la discriminación entre varias poblaciones mediante el estudio de la función de distribución empírica. Se ilustran dichos comentarios con unos datos obtenidos en individuos de la especie *Octopus vulgaris* sobre los cuales se evaluaba la variable aleatoria T (tamaño). En total se consideraron 505 individuos extraídos del Oceano Atlántico, a una profundidad que va de 15 a 35 metros, durante una campaña que se desarrolló en los meses de Abril y Mayo de 1974. Los datos eran presentados en forma de histograma, agrupando los diferentes valores de la variable T en 24 intervalos y proporcionando la frecuencia observada en cada uno de ellos. De la visión directa del histograma ya se desprende la presencia de dos modas y la posible mezcla de dos poblaciones de *Octopus vulgaris*. Nuestro objetivo es comprobar que la

aplicación del algoritmo stepwise confirma la presencia de ambas modas en la función de densidad estimada.

Hemos realizado una previa transformación del histograma centrando los intervalos en el origen de coordenadas y reduciendo la longitud de cada uno de ellos a 0,1 unidades. El histograma transformado puede observarse en la Fig. 26 .

Al utilizar el algoritmo datos continuos y presentarse éstos en forma agrupada, hemos optado por la estrategia consistente en repartir las diferentes observaciones de forma uniforme dentro del intervalo. De manera que cuando la frecuencia de un intervalo es n , las observaciones se presentan en el centro de los n subintervalos iguales en los que podría ser dividido el intervalo original.

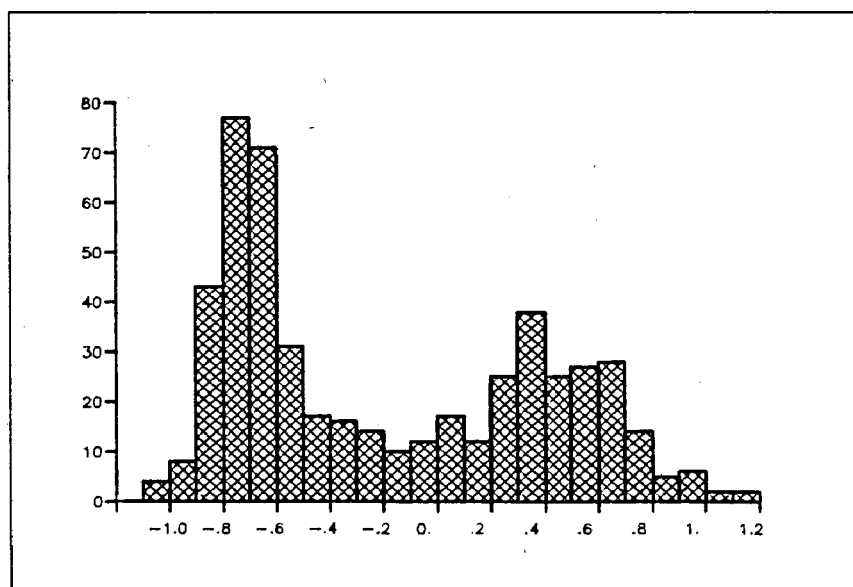


Figura 26. *Histograma correspondiente a los datos utilizados en el ejemplo 7.1. Alonso et al. (1976).*

7.1.2. Resultados.

Tras la aplicación del algoritmo stepwise hemos encontrado, trabajando como en el resto de ejemplos con un nivel de significación en los contrastes del 5% , un total de 9 parámetros significativos, que son

$$\begin{array}{ll}
 \theta_1 = 0.90349 & \theta_{11} = 0.15720 \\
 \theta_2 = -0.14180 & \theta_{12} = -0.11380 \\
 \theta_3 = -0.17587 & \theta_{13} = -0.12335 \\
 \theta_7 = 0.15970 & \theta_{16} = -0.15258 \\
 \theta_9 = -0.17608 &
 \end{array}$$

En la **Fig. 27** presentamos la función de densidad estimada y en la **Fig. 28** la función de distribución obtenida utilizando los nueve parámetros significativos a partir de las expresiones [5.52] y [5.53].

Puede comprobarse que efectivamente en la función de densidad estimada se recuperan las dos modas principales que presentaba el histograma, siendo ambas por tanto significativas según el algoritmo stepwise. Podría sugerirse como conclusión la posible presencia, para la variable tamaño, de dos poblaciones de *Octopus vulgaris* en la muestra estudiada.

Creemos que puede considerarse el ejemplo precedente una validación de los resultados proporcionados por el algoritmo stepwise al trabajar con datos reales.

VII - Aplicaciones a la Biología.

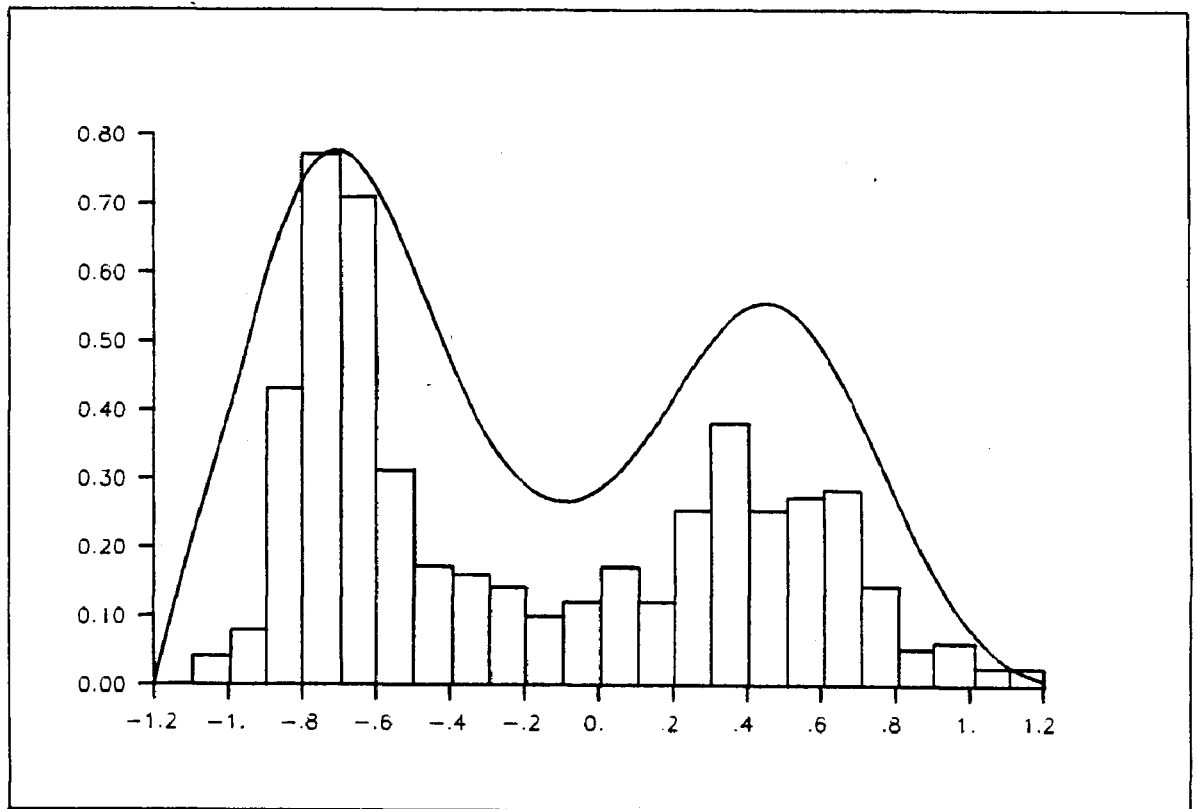


Figura 27. Estimación stepwise de la densidad correspondiente a los datos del ejemplo 7.1. Se ha superpuesto el histograma a la misma escala.

VII - Aplicaciones a la Biología.

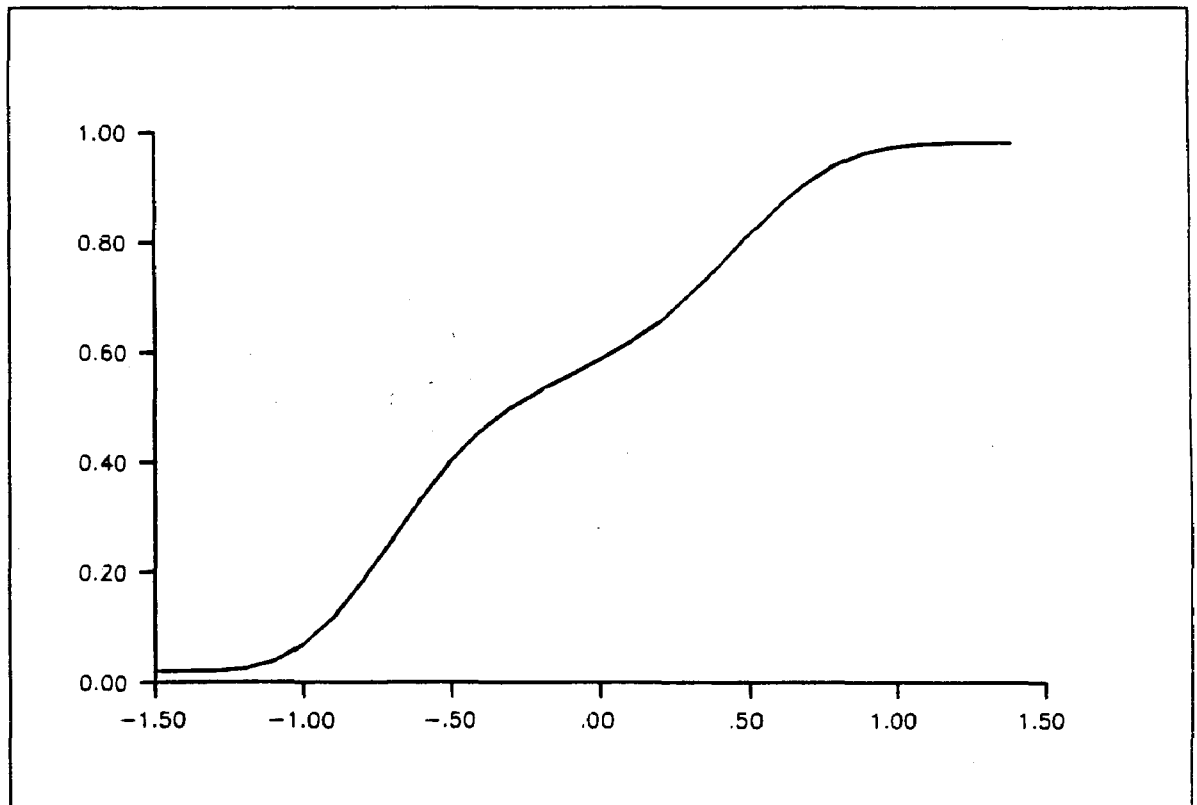


Figura 28. *Función de distribución obtenida a partir de los parámetros significativos de la estimación stepwise de la densidad con los datos del ejemplo 7.1.*

7.2. Clasificación de patrones electroforéticos.

7.2.1. Introducción.

Con frecuencia nos encontramos con la necesidad de analizar datos donde las observaciones sobre cada individuo vienen dadas en forma de función continua sobre un eje. Por ejemplo podemos citar, espectros de absorción a diferentes longitudes de onda, densidades a lo largo de un gradiente continuo, etc. La metodología que sugerimos se basa en la identificación de cada gráfica obtenida con una función de densidad de probabilidad y a continuación aplicar sobre las mismas las técnicas de análisis basadas en las funciones de densidad.

Para identificar cada gráfica con una función de densidad se procederá en primer lugar a una normalización de la función. Sea $f_i(x)$ $x \in \chi$ la función observada sobre el individuo i , que supondremos continua y positiva, y sea $K_i = \int_{\chi} f_i(x) dx$. La función normalizada será por tanto $p_i(x) = \frac{f_i(x)}{K_i}$.

Una vez que definida sobre cada individuo una función de densidad $p_i(x)$, la representación de la misma en forma analítica se llevará a cabo utilizando el procedimiento de estimación desarrollado en el apartado 6.1. Como resultado tendremos para cada individuo una representación funcional tal como se ha definido en [5.3] y que nos permite introducir una distancia entre los diferentes individuos utilizando la expresión [5.37] y posteriormente aplicar técnicas de clasificación y representación que nos proporcionarán una idea de las relaciones entre los individuos.

Como ejemplo práctico hemos utilizado una serie de patrones electroforéticos de proteínas en suero, obtenidos en pacientes con mieloma múltiple. El mieloma múltiple, según Miale (1982), es una enfermedad tumoral en la que se produce una anormal

proliferación de células plasmáticas de origen B-linfocítico. Entre las características que presenta la enfermedad destaquemos la destrucción del hueso, la sustitución de la médula osea por células plasmáticas, hiperproteinemia e hiperglobulinemia. El anormal metabolismo de las proteínas y la alta concentración de inmunoglobulina produce una hiperviscosidad con un flujo sanguíneo defectuoso y como consecuencia una insuficiencia vascular. Una de las evidencias de estas anormalidades proteínicas, utilizada para el diagnóstico de la enfermedad, es una electroforesis del suero que muestre cantidades anómalas en las bandas correspondientes a las globulinas. Existen diferentes tipos de anormalidades proteínicas asociadas al mieloma múltiple, en la mayoría de casos se detecta un notable incremento de la Ig G (alrededor del 60 % de los casos) o de la Ig A (alrededor del 18 %) que se traduce a la hora de realizar el patrón electroforético en irregularidades en las bandas correspondientes a la gamma y beta globulinas, al ser los principales componentes de las inmunoglobulinas. También existen casos, alrededor del 1 % de los mielomas, en los cuales no se aprecia ninguna anormalidad proteínica, denominándose mieloma hiposecretorio.

Los patrones electroforéticos utilizados en el presente ejemplo han sido obtenidos de Miale (1982) y se presentan en la Fig. 29 .

La experiencia aleatoria asociada estaría formada por un espacio muestral continuo que representaría los diferentes valores de migración sobre la placa electroforética, de las diferentes proteínas, y la función de densidad representaría la distribución de probabilidad de las mismas.

VII - Aplicaciones a la Biología.

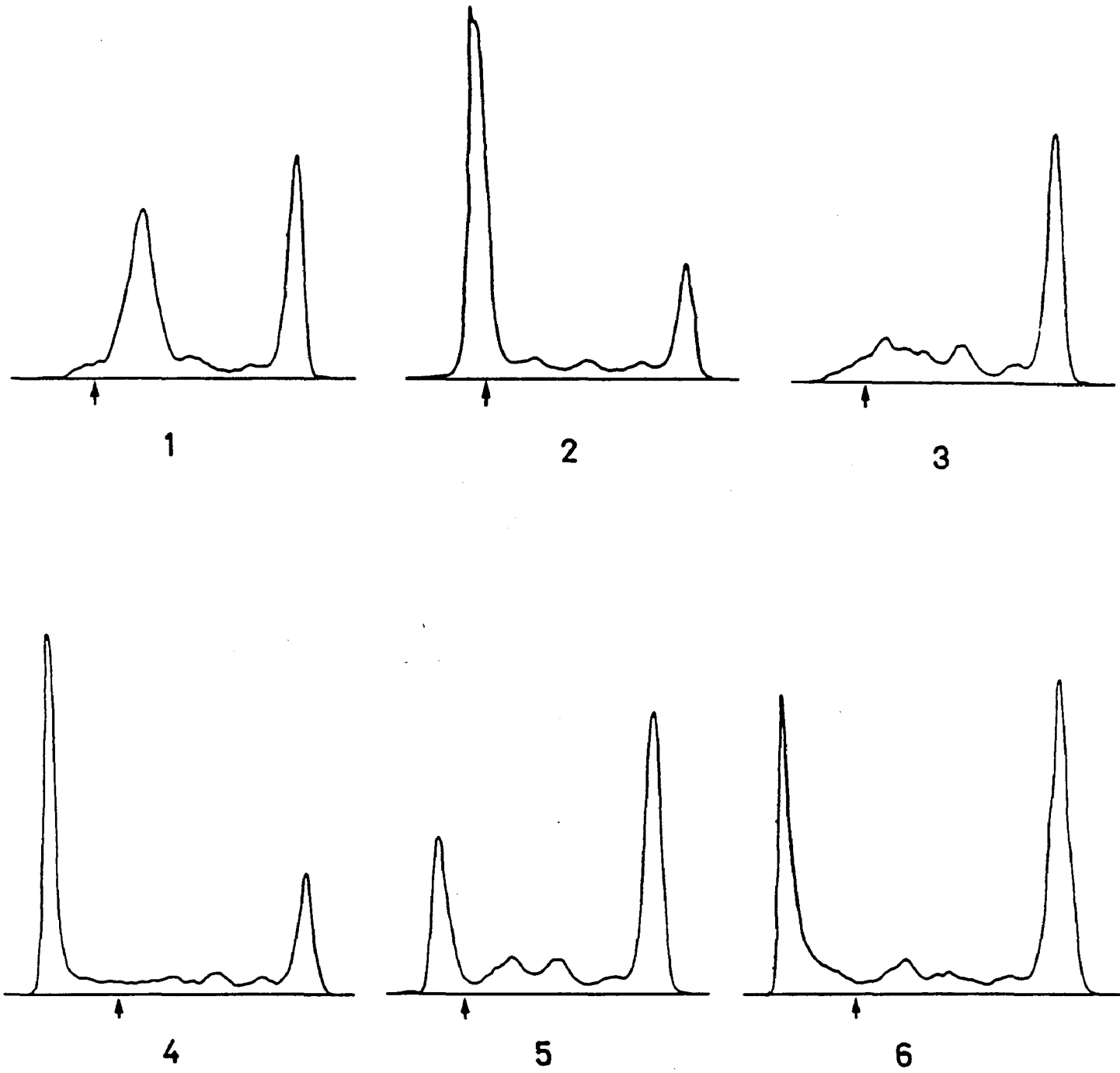


Figura 29. Patrones electroforéticos estudiados en el ejemplo 7.2.

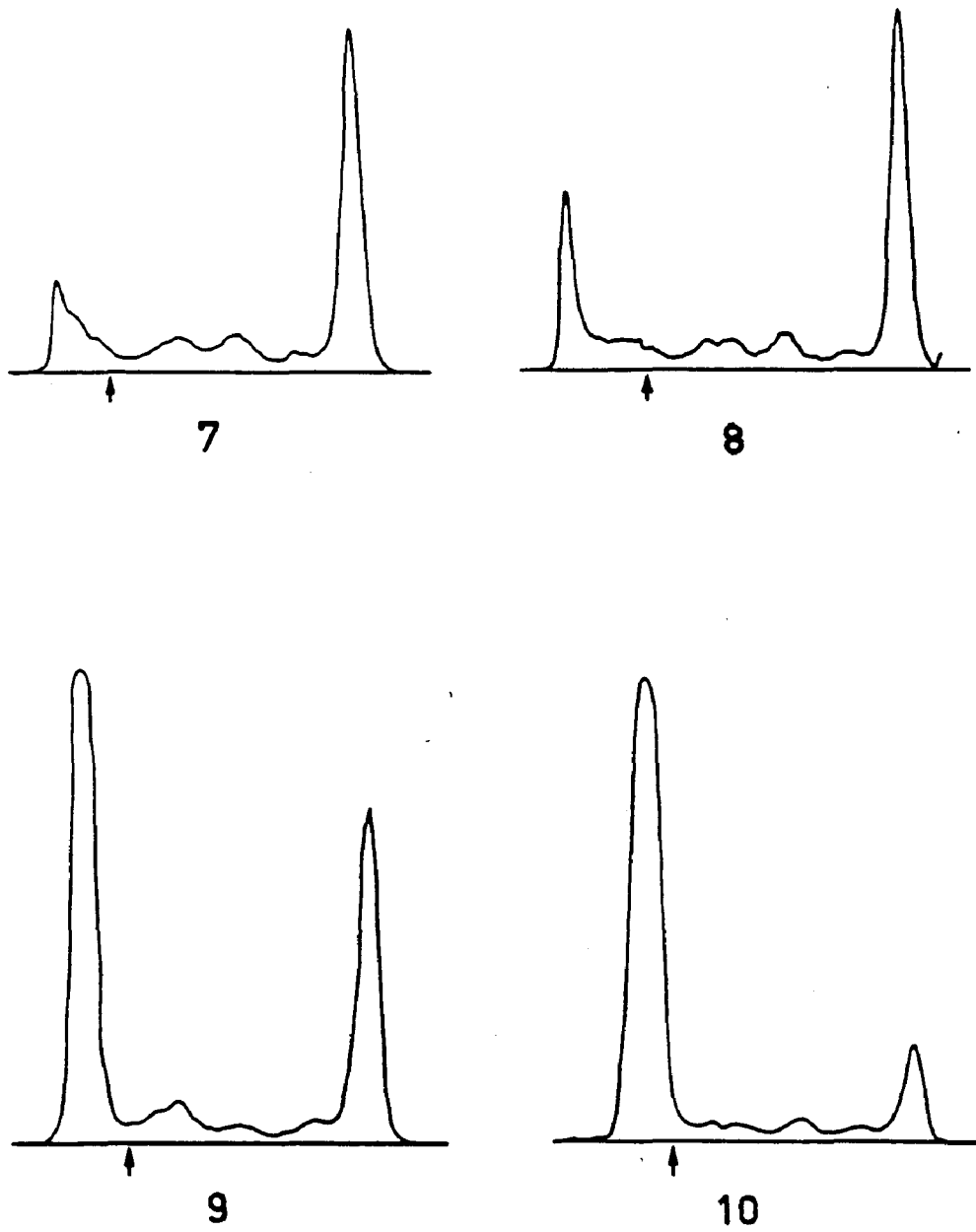


Figura 29. (continuación). Patrones electroforéticos estudiados en el ejemplo 7.2.

7.2.2. Resultados.

Cada patrón electroforético ha sido transformado en valores numéricos para permitir llevar a cabo el proceso de estimación de las funciones de densidad. La estimación se ha obtenido aplicando [6.4] y tras un análisis de la bondad de las estimaciones, hemos considerado funciones con un total de 50 parámetros. En la **Tabla 14** presentamos la matriz de interdistancias entre los diez individuos prescindiendo de la diagonal principal.

1.34644									
0.72795	1.09438								
1.76739	1.98122	1.63299							
1.14743	1.01632	0.79540	1.61076						
1.41473	1.72876	1.24756	0.68683	1.40740					
1.10852	1.11068	0.71575	1.47115	0.25018	1.22691				
1.29274	1.47144	1.04431	0.86684	0.89969	0.67151	0.70696			
1.44774	1.38018	1.17908	1.39413	0.66257	1.19457	0.60130	0.66801		
1.75375	0.96964	1.52867	1.95548	0.92032	1.77798	1.02093	1.34078		
0.94203									

Tabla 14. *Matriz de interdistancias entre los diez patrones electroforéticos.*

Posteriormente se ha realizado un Análisis de Coordenadas Principales, Cuadras (1981), sobre la matriz de interdistancias anterior, con objeto de representar los individuos en dimensión reducida y permitir una clara visualización de las relaciones entre ambos. El Análisis ha sido realizado con el programa VCOOR cedido por el Dr. Angel Villarroya. En la **Fig. 30** presentamos la representación de los diez patrones según las dos primeras coordenadas principales. El porcentaje de variabilidad explicado es del 50.62 % con la primera coordenada y del 79.17 % con las dos primeras.

También se ha llevado a cabo una clasificación jerárquica utilizando técnicas de Taxonomía Numérica. Utilizando el paquete CLUSTAN hemos obtenido una

disimilaridad ultramétrica entre los patrones mediante el método UPGMA, Cuadras (1981), y la posterior representación en forma de dendrograma se muestra en la Fig. 31. La correlación cofenética obtenida es de 0.69.

7.2.3. Discusión.

De los resultados tanto del Análisis de Coordenadas Principales como de la Taxonomía Numérica, se desprende la aparición de tres grupos diferenciados. La discriminación entre los tres grupos se basa principalmente en la ubicación de los picos irregulares del patrón electroforético. En el primer grupo, formado por los individuos 4, 6 y 8 aparece una clara irregularidad en el extremo izquierdo de las gráficas, correspondiente a la banda de la gamma-globulina. Los individuos 2, 5, 7, 9 y 10 presentan un pico irregular más desplazado a la derecha, afectando igualmente a la banda de la gamma-globulina, pero parcialmente también afecta a la de la beta-globulina. Finalmente los individuos 1 y 3 no presentan ninguna irregularidad en la zona correspondiente a la gamma-globulina y sí en la zona de la beta-globulina, mucho más pronunciada en el individuo 1 que en el 3. Este último corresponde a un individuo con niveles prácticamente normales de globulinas, excepto un ligero exceso en la beta-globulina, tras una terapia médica y nos sirve para apreciar las irregularidades presentes en los demás.

De forma global podemos considerar que existe una adecuada correspondencia entre los resultados obtenidos con los métodos de representación a partir de la matriz de interdistancias, con los que cabría esperar conociendo las particularidades de cada uno de los individuos, y por tanto podemos considerar válida y adecuada la representación de los individuos mediante las funciones de densidad estimadas.

VII - Aplicaciones a la Biología.

Además de las representaciones gráficas y clasificaciones jerárquicas, el poseer una representación funcional y una distancia entre funciones de densidad, permite desarrollar otras técnicas, como por ejemplo el diagnóstico y clasificación automática de los individuos, siguiendo alguno de los algoritmos de diagnóstico desarrollados en diferentes trabajos, por ejemplo en Oller (1982).

VII - Aplicaciones a la Biología.

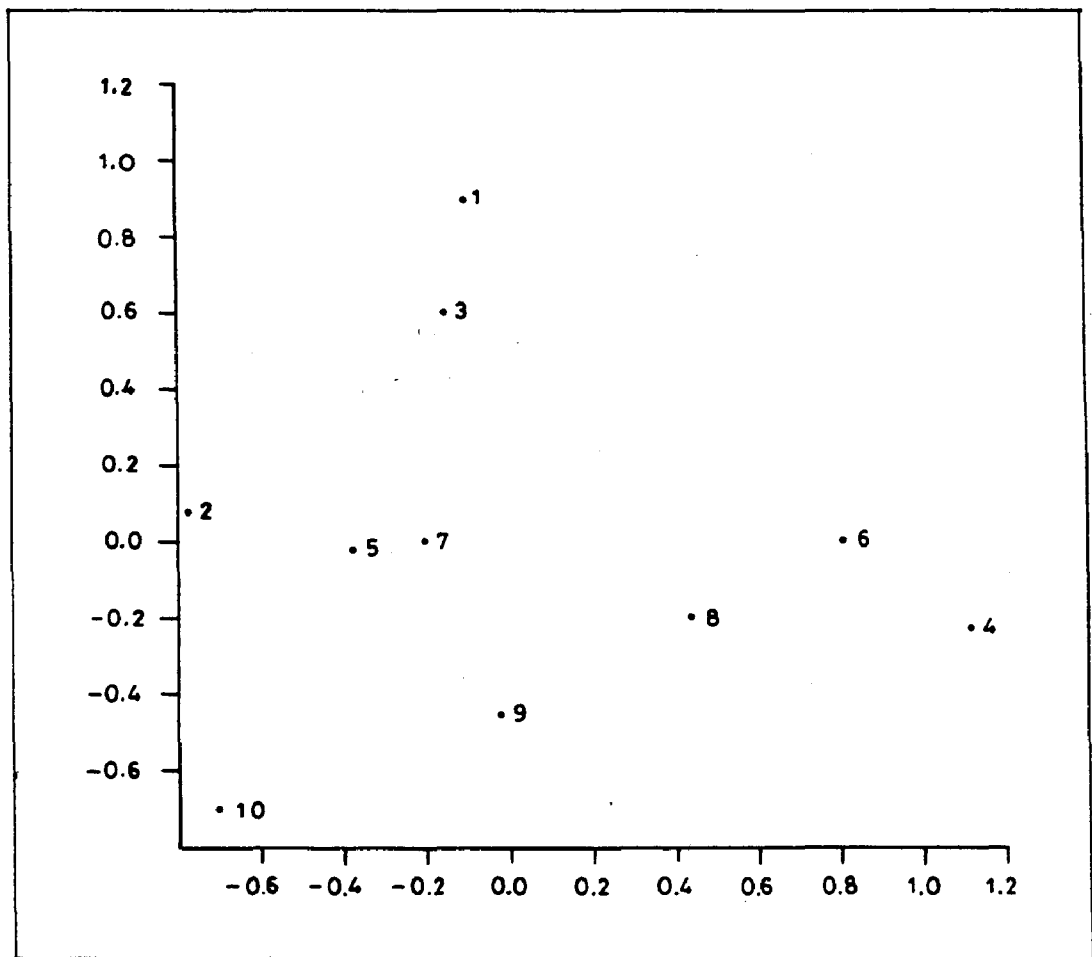


Figura 30. Representación de los patrones electroforéticos mediante las dos primeras coordenadas principales.

VII - Aplicaciones a la Biología.

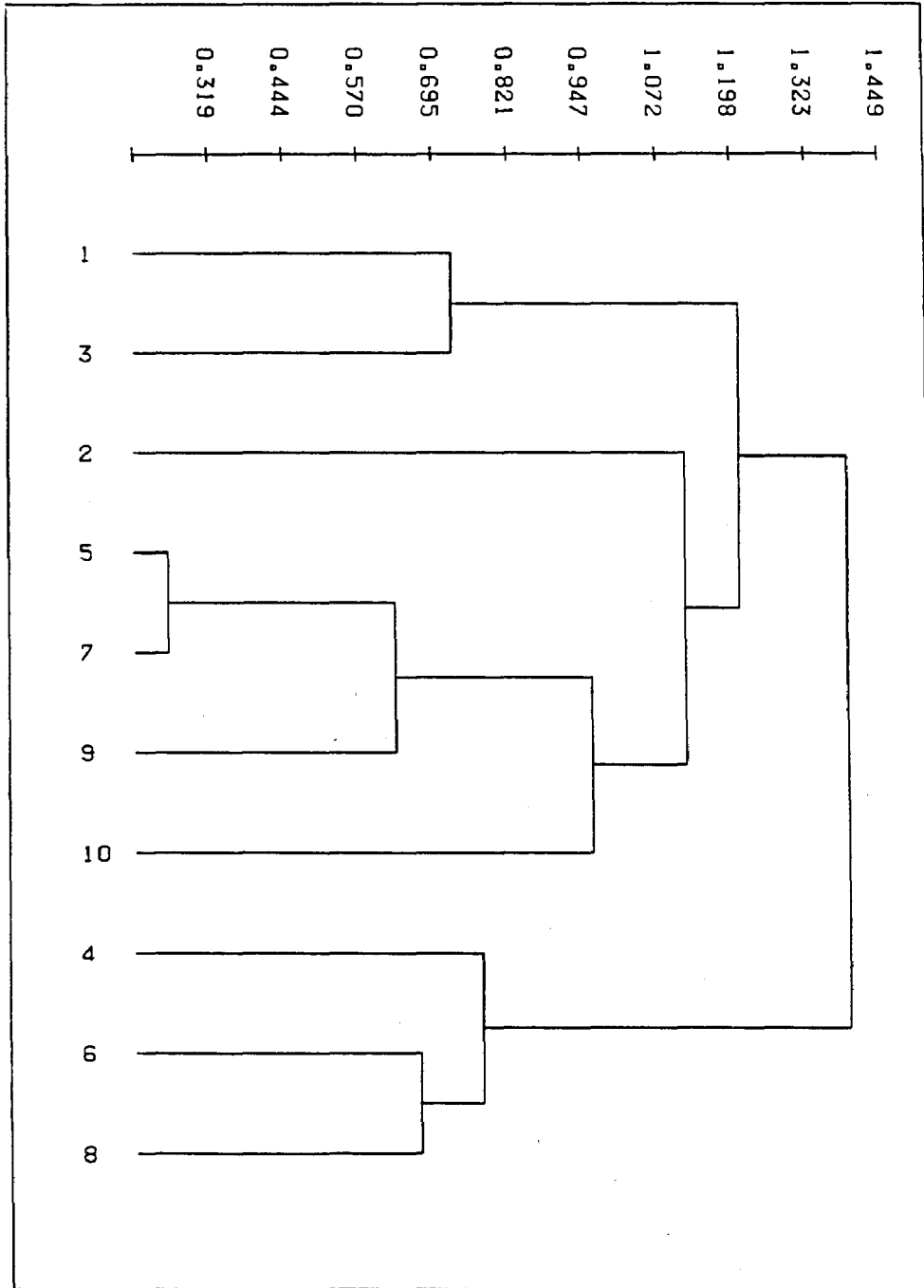


Figura 31. Dendrograma obtenido a partir de las distancias entre los patrones electroforéticos.

VIII. Resumen de Resultados.

Resumen de Resultados.

Presentamos a continuación un resumen de los principales resultados aportados por la presente memoria.

Comenzamos realizando una aproximación al concepto de modelo estadístico desde el punto de vista geométrico, centrándonos principalmente en consideraciones sobre la introducción de distancias, y en particular estudiando la métrica informacional y sus propiedades.

Dada una variedad paramétrica correspondiente a un modelo estadístico, hemos efectuado un estudio del espacio tangente y del espacio tangente dual en un punto a la variedad, introduciendo representaciones adecuadas de los mismos. Tales representaciones han permitido identificar a los elementos del espacio muestral con campos tensoriales covariantes de primer orden en la variedad, mientras que las variables aleatorias pueden ser identificadas con campos tensoriales contravariantes también de primer orden.

Hemos introducido dos definiciones de distancias, en sentido estricto pseudodistancias, entre valores muestrales basadas ambas en distancias en el espacio tangente dual entre sus respectivas formas lineales asociadas según la aplicación

$$\begin{aligned} \delta: \chi &\rightarrow E_0^* \\ \mathbf{x} &\rightarrow \delta(\mathbf{x}) = Y^* \end{aligned}$$

tal que $Y^*(Y) = Y(\mathbf{x})$ para todo $Y \in E_0$.

La primera, a la que denominamos distancia inmediata, es definida a partir de la distancia euclídea en el espacio tangente dual. Se han obtenido expresiones explícitas

para la distancia cuando los individuos estadísticos son muestras correspondientes a las distribuciones Poisson, Weibull, Gamma, Exponencial, Binomial, Binomial negativa, Multinomial, Multinomial negativa, Wald, Logística, Normal univariante y Normal multivariante. Se han estudiado ciertas propiedades relacionadas con la distancia inmediata, entre las que destacamos su invarianza frente a cambios de la medida de referencia y transformaciones por estadísticos suficientes, y su no decrecimiento al aumentar el número de parámetros de las variedades.

La distancia estructural es definida a partir de la distancia sobre el conjunto imagen del espacio muestral por la aplicación anterior, al que denominaremos S . Se demuestra que coincide con la distancia inmediata si S es un conjunto convexo y también que S no es convexo si la dimensión del espacio muestral es uno y el número de parámetros de la variedad mayor o igual a dos. Se ha obtenido la expresión explícita para la distancia estructural entre muestras de tamaño uno correspondientes a una distribución Normal univariante.

Se han estudiado las aplicaciones de las distancias entre individuos a técnicas clásicas de inferencia estadística, definiendo nuevos procedimientos de estimación de parámetros y contraste de hipótesis desde un punto de vista geométrico. Se comprueba como utilizando la distancia inmediata se recuperan gran parte de los resultados clásicos, en particular las ecuaciones de verosimilitud y el contraste de hipótesis mediante el test de los multiplicadores de Lagrange. Hemos comprobado también como utilizando en estimación de parámetros la distancia estructural en un ejemplo en que ésta difiere de la inmediata, se obtienen resultados que difieren respecto a la máxima verosimilitud clásica y que podemos considerar más acordes con resultados intuitivos al dejar indeterminada la estimación de la varianza trabajando con muestras de tamaño uno de una distribución Normal univariante.

Se ha introducido una clase de funciones de densidad de probabilidad que pueden ser caracterizadas en una variedad paramétrica de dimensión finita. Se comprueba que las variedades resultantes son de curvatura constante y positiva. Se han obtenido las expresiones para las geodésicas y la distancia de Rao entre dos distribuciones. Hemos efectuado un estudio probabilístico en varios ejemplos y finalmente consideramos la aplicación de tales familias a la estimación no paramétrica de funciones de densidad gracias a su capacidad de adaptación.

Se ha abordado el problema de la estimación de parámetros en las familias anteriormente citadas. Comprobamos los inconvenientes de la estimación máximo verosímil y para subsanarlos hemos propuesto un algoritmo tipo stepwise que toma en cuenta la significación de los incrementos de la verosimilitud al modificar el número de parámetros de las familias. Utilizamos diversas simulaciones para comprobar la bondad del algoritmo, obteniendo resultados satisfactorios tanto al trabajar con distribuciones clásicas como con las nuevas familias. Se han comparado los resultados con otros métodos clásicos de estimación no paramétrica, en particular con el método de los Kernel.

También se ha estudiado el método de minimizar la esperanza del cuadrado de la distancia estructural entre individuos (MESD). Para poder llevar a cabo tal estudio se ha desarrollado una aproximación a la distancia Riemanniana y se han utilizado técnicas de minimización numérica de funciones de varias variables con restricciones. Se han obtenido algunos ejemplos que muestran un mejor comportamiento de la estimación MESD frente a la MLE.

VIII - Resumen de Resultados.

Finalmente se han considerado dos ejemplos prácticos consistentes en la estimación de una función de densidad bimodal a partir de unos datos en forma de histograma y en la clasificación de diversos patrones electroforéticos asimilándolos a funciones de densidad. En ambos ejemplos los resultados parecen validar completamente la metodología empleada.

Bibliografía

- Aitchison, J. and Silvey, S.D. (1958) Maximum likelihood estimation of parameters subject to restraints. *Ann. Math. Statist.*, 29, 813-828.
- Alonso, G., Ocaña, J. y Cuadras, C.M. (1976) *Fundamentos de Probabilidad en Bioestadística*. Eunibar, Barcelona.
- Amari, S. (1968) Theory of Information Space: a Geometrical Foundation of the Analysis of Communication Systems. *RAAG Memoirs, Vol 4* . , 373-418.
- Amari, S. (1980) Theory of Information Space: a Differential-Geometrical Foundation of Statistics. *RAAG Reports, 106* . , 1-53.
- Amari, S. (1982) Geometrical theory of asymptotic ancillarity and conditional inference. *Biometrika*, 69(1), 1-17.
- Atkinson, C. and Mitchell, A.F.S. (1981) Rao's distance measure. *Sankhya*, 43 A, 345-365.
- Bandorff-Nielsen, O. (1978) *Information and Exponential Families in Statistical Theory*. John Wiley, New York.
- Benzecri, J.P. (1976) *L'analyse des donnees I. La toxiologie. L'analyse des donnees II. L'analyse des correspondances*. Dunod, Paris.
- Bhattacharyya, A. (1943) On discrimination and divergence. *Proc. 29th. Indian Sci. Cong. Part III*, 13.

- Bhattacharyya, A. (1946) On a measure of divergence between two multinomial populations. *Sankhya*, 7, 401-406.
- Bose, R.C. (1936) *Sankhya*., 2, 143-154.
- Burbea, J. (1986) Informative Geometry of probability spaces. *Expositiones Mathematicae*, 4, 347-378.
- Burbea, J. and Oller, J.M. (1988) The information metric for linear elliptic models. *Statistics and Decision*, vol. 6
- Burbea, J. and Oller, J.M. (1989) On Rao distance asymptotic distribution. *Mathematics Preprint Series No. 67. Universitat de Barcelona.*
- Burbea, J. and Rao, C.R. (1982) Entropy differential metric, distance and divergence measures in probability spaces: a unified approach. *J. Multivariate Anal.*, 12 , 575-596.
- Burbea, J. and Rao, C.R. (1984) Differential metrics in Probability Spaces. *Prob. and Math. Statistics*, 3(2), 241-258.
- Calvo, M. (1988) *Sobre la geometria informacional del modelo normal multivariante. Aplicaciones a la Biologia.* Tesis Doctoral. Universitat de Barcelona.
- Calvo, M. and Oller, J.M. (1990) A Distance between Multivariate Normal Distributions Based in an Embedding into the Siegel Group. *Journal of Multivariate Analysis*, 35(2) , 223-242.
- Carroll, C.W. (1961) The Created Response Surface Technique for Optimizing Non linear Restraint Systems. *Operations Research*, 9.
- Cavalli-Sforza, L.L. (1969) Human Diversity. en *Proceddings, XII Internat. Congr. Genetics, Tokyo*, Vol3, 405-416.

- Chentsov, N.N. (1972) *Statistical Decision Rules and Optimal Inference*. (en ruso) Nauka, Moscow. Traducido al inglés (1982) Math. Monographs, 53, Amer. Math. Soc., Providence.
- Cohen, A.M. (1977) *Análisis Numérico*. Reverté, Barcelona.
- Courant, R. and Hilbert, D. (1953) *Methods of Mathematical Physics, vol. I*. Interscience, New York.
- Cuadras, C.M. (1981) *Métodos de Análisis Multivariante*. Eunibar, Barcelona.
- Cuadras, C.M. (1989a) Distance analysis in discrimination and classification using both continuous and categorical variables. en *Statistical Data Analysis and Inference (Y. Dodge, Ed.)* North-Holland Pu. Co., Amsterdam.
- Cuadras, C.M. (1989b) Distancias estadísticas entre individuos y poblaciones con variables mixtas. *Actas XVIII Reunión Nac. Estad. e I. Oper., Univ. Santiago de Compostela.*, 143-148.
- Dawid, A.P. (1975) Discussion to Efron's Paper. *Annals of Statistics* , 3, 1189-1217.
- Dawid, A.P. (1977) Further Comments on a paper by Bradley Efron. *Annals of Statistics* , 3, 1249.
- Duda, R.O. and Hart, P.E. (1973) *Pattern Classification and Scene Analysis*. John Wiley and sons, New York.
- Efron, B. (1975) Defining the Curvature of a Statistical Problem (With Application to Second Order Efficiency) (with discussion). *Annals of Statistics* , 3, 1189-1242.

- Eriksen, P.S. (1986) Geodesic connected with the Fisher metric on the multivariate normal manifold. *Preprint No. R 86-13. Institute of Electronic Systems. Aalborg, University Center, Denmark.*
- Fisher, R.A. (1912) On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, 41, 155-160.
- Fisher, R.A. (1913) Application of Vector Analysis to Geometry. *Messenger of Mathematics*, 42, 161-178.
- Fisher, R.A. (1915) Frequency distribution of the values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika*, 10, 507-521.
- Fisher, R.A. (1922) The mathematical foundations of theoretical statistics. *Phil. Trans. A.*, 222, 309-368.
- Fisher, R.A. (1925) Theory of Statistical Estimation. *Proc. Camb. Phil. Soc.*, 22, 700-725.
- Fisher, R.A. (1936) The Fiducial Argument in Statistical Inference. *Annals of Eugenics*, 6(4).
- Good, I.J. (1971) Non-parametric Roughness Penalty for Probability Densities. *Nature Physical-Sciences*, 229, 29-30.
- Good, I.J. and Gaskins, R.A. (1971) Nonparametric Roughness Penalties for Probability Densities. *Biometrika*, 58(2), 255-277
- Good, I.J. and Gaskins, R.A. (1980) Density estimation and bumps hunting by the penalized likelihood method exemplified by scattering and meteorite data. *J. Amer. Statist. Assoc.*, 75, 42-73.

- Hadjisavvas, N. (1981) Distance Between states and statistical inference in quantum theory. *Ann. Inst. Henri Poincaré*, 35(4), 287-309.
- Halmos, P.R. (1950) *Measure Theory*. Van Nostrand.
- Hicks, N.J. (1965) *Notes on Differential Geometry*. Van Nostrand Reinhold Comp., London.
- Hotelling, H. (1933) Analysis of a complex of Statistical variables into principal components. *J. Educ. Psychol.*, 24(6), 417-441.
- Jauch, J.M. (1968) *Foundations of Quantum Mechanics*. Addison-Wesley, Mass.
- Jeffreys, H. (1946) An Invariant Form for the Prior Probability in Estimation Problems. *Proc. Roy. Soc., A*, 196, 453-461.
- Kass, R.E. (1987) *Differential Geometry in Statistical Inference. Cap. I Introduction*. Inst. of Math. Statist. Hayward, Cal.
- Kraft, C. and Le Cam, L. (1956) A remark on the roots of the maximum likelihood equation. *Ann. Math. Statist.*, 27, 1174-1177.
- Kullback, S. and Leibler, R.A. (1951) On Information and Sufficiency. *Ann. Math. Statist.*, 22, 79-86.
- Lehman, E.L. (1983) *Theory of Point Estimation*. Wiley, New York.
- Mackey, G.W. (1963) *Mathematical foundations of quantum mechanics*. W.A. Benjamin Inc., New York.
- Madsen, L.T. (1979) The geometry of statistical model - a generalization of curvature. *Res. Report 79-1 Statist. Res. Unit.*, Danish Medical Res. Council.

- Mahalanobis, P.C. (1936) On the Generalized Distance in Statistics. *Proc. Natl. Inst. Sci. India.*, 2(1), 49-55.
- Makelainen, J., Schmidt, K. and Styan, G. (1981) On the existence and uniqueness of the maximum likelihood estimate of a vector-valued parameter in fixed-size samples. *Ann. Statist.*, 9, 758-767.
- Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979) *Multivariate Analysis*. Academic Press, London.
- Matusita, K. (1964) Distance and Decision Rules. *Ann. Inst. Stat. Math.*, 16, 301-315.
- Miale, J.B. (1982) *Laboratory medicine Hematology*. The C.V. Mosby Company, St. Louis, Missouri.
- Miñarro, A. (1985) *Transformaciones de coordenadas en Análisis de Datos. Aplicaciones a la Biología*. Tesina de Licenciatura. Universitat de Barcelona.
- Miñarro, A. and Oller, J.M. (1990) An Algorithm for Geodesic Coordinates in Riemannian Manifolds. *Indian J. pure appl. Math.*, 21(7), 707-715.
- Misra, B. (1974) en *Physical reality and Mathematical description*. C. Enz and J. Mehra eds., D. Reidel publishing company, Dordrecht, Holland.
- de Montricher, G.F., Tapia, R.A. and Thompson, J.R. (1975) Nonparametric Maximum Likelihood Estimation of Probability Densities by Penalty Function Methods. *Annals of Statistics*, 3, 1329-1348.
- Nei, M. (1978) The Theory of genetic distance and evolution of human races. *Japan J. Human Genet.*, 23, 341-369.

Bibliografía.

- Oller, J.M. (1982) *Utilización de métricas riemannianas en análisis de datos multidimensionales y su aplicación a la Biología*. Tesis Doctoral. Universitat de Barcelona.
- Oller, J.M. (1987) Information metric for extreme value and logistic probability distributions. *Sankhya*, 49, A, 17-23.
- Oller, J.M. (1988) Comunicación personal.
- Oller, J.M. (1989) Some geometrical aspects of data analysis and statistics. en *Statistical Data Analysis and Inference (Y. Dodge, Ed.)* North-Holland Pu. Co., Amsterdam.
- Oller, J.M. and Cuadras, C.M. (1985a) Sobre ciertas condiciones que deben verificar las distancias entre espacios probabilísticos. *Actas de la XV Reunión Nacional de Est. e Inv. Op.*, tomo II, 503-509.
- Oller, J.M. and Cuadras, C.M. (1985b) Rao's distance for negative multinomial distributions. *Sankhya*, 47, A, 75-83.
- Parzen, E. (1962) On estimation of a Probability Density Function and Mode. *Ann. Math. Statist.*, 33, 1065-1076.
- Pearson, K. (1896) Mathematical Contributions to the Theory of Evolution. *Phil. Tran. Roy. Soc.*, 187.
- Pearson, K. (1901) On lines and Planes of Closest Fit to Systems of Points in Space. *Phil. Mag., Ser. 6*, 2 (11), 559-572.
- Pearson, K. (1926) On the coefficient of racial likeness. *Biometrika*, 18, 105-117.

- Prevosti, A., Ocaña, J. and Alonso, G. (1975) Distances between Populations of *Drosophila suboscuro* Based on Chromosome Arrangement Frequencies. *Theoretical and Applied Genetic*, 45, 231-241.
- Rao, C.R. (1945) Information and accuracy attainable in the estimation of parameters. *Bull. Calcutta Math. Soc.*, 37, 81-97.
- Rao, C.R. (1949) On the distance between two populations. *Sankhyā*, 9, 246-248.
- Rao, C.R. (1961) Asymptotic efficiency and limiting information. en *Proc. Fourth Berkeley Symp. Math. Statist. Prob. I (J. Neyman, Ed.)*, Univ. of California Press.
- Rao, C.R. (1962) Efficient estimates and optimum inference procedures in large samples. *J. Roy. Statist. Soc. B.*, 24, 46-72.
- Rao, C.R. (1973) *Linear Statistical Inference and its Applications*. Wiley, New York.
- Rao, C.R. (1982) Diversity and Dissimilarity Coefficients: A Unified Approach. *J. Theoretical Population Biology*, 21, 24-43.
- Reeds, J. (1977) Discussion on Professor Efron's paper (1975). *Annals of Statistics*, 5, 1234-1238.
- Rios, M. and Cuadras, C.M. (1986) Distancia entre modelos lineales normales. *Questiú*, 10(2), 83-92.
- Rosenblatt, M. (1956) Remarks on some nonparametric estimates of a density function. *Ann. of Math. Statist.*, 27, 832-835.
- Rosenbrok, H.H. (1960) An automatic method for finding the greatest or least value of a function. *Computer J.*, 3, 175-184.

- Rudin, W. (1966) *Real and Complex Analysis.*, Mac Graw-Hill, New York.
- Sanchez, P. (1989) Funciones discriminantes basadas en distancias estadísticas. *Actas XVIII Reunión Nac. Estad. e I. Oper., Univ. Santiago de Compostela.*, 468-472.
- Sato, Y., Sugawa, K . and Kawaguchi, M. (1979) A geometrical structure of the parameter space of the two-dimensional norma distribution. *Reports on Math. Phys.*, 16, 11-119.
- Skovgaard, L. T. (1981) A Riemannian Geometry of the Multivariate Normal Model. Research Report 81/3. Statistical Research Unit, Danish Medical Research Council.
- Skovgaard, L. T. (1984) A Riemannian Geometry of the Multivariate Normal Model. *Scand. J. Statist.*, 11, 211-223.
- Sokolnikoff, I.S. (1971) *Análisis Tensorial.* Index, Madrid.
- Spivak, M. (1979) *A Comprehensive Introduction to Differential Geometry.* Publish or Perish, Berkeley, CA.
- Takiyama, R. (1974) On geometrical structures of parameter spaces of one-dimensional distributions. *Trans. Inst. Electr. Comm. Eng. Japan*, 57-A, 67-69.
- Tapia, R.A. and Thompson, J.R. (1978) *Nonparametric Probability Density Estimation.* The John Hopkins University Press, Baltimore.
- Tarone, R.E. (1988) Score Statistics. en *Encyclopedia of Statistical Sciences.* Vol. 8 (S. Kotz and N.L. Eds.), John Wiley, New York.
- Wald, A. (1941) Some examples of asymptotically most powerful tests. *Ann. Math. Statist.*, 12, 396-408.

Bibliografia.

Yoshizawa, T. (1971) A geometrical interpretation of location and scale parameters.
Memo TYH-2, Harvard. Univ.