Modelos de mezcla de distribuciones α-estables

Aplicación a micromatrices de expresión genética

Tesis Doctoral

Diego Salas González

Departamento de Física Aplicada
Universidad de Granada -España-
dsalas@ugr.es

Editor: Editorial de la Universidad de Granada
Autor: Diego Salas González
D.L.: GR.1909-2008
ISBN: 978-84-691-5804-3

D. Diego Pablo Ruiz Padillo, Doctor en Ciencias Físicas y Profesor Titular del Departamento de Física Aplicada de la Facultad de Ciencias de la Universidad de Granada y

Dña. María del Carmen Carrión Pérez, Doctora en Ciencias Físicas y Profesora Titular del Departamento de Física Aplicada de la Facultad de Ciencias de la Universidad de Granada y

D. Ercan Engin Kuruoglu, Doctor por la Universidad de Cambridge en Inglaterra e Investigador del Istituto di Scienza e Tecnologie dell'Informazione del Consiglio Nazionale della Ricerca de Pisa en Italia,

MANIFIESTAN:

Que la presente Memoria titulada "Modelos de Mezcla de Distribuciones α-estables. Aplicación a Micromatrices de Expresión Genética", presentada por Diego Salas González para optar al grado de Doctor por la Universidad de Granada, ha sido realizada bajo nuestra dirección en el Departamento de Fisica Aplicada de la Universidad de Granada y el Instituto de Ciencia y Tecnologías de la información de Pisa en Italia. Con esta fecha, autorizamos la presentación de la misma.

Granada, a 10 de junio de 2008

Fdo: Diego Pablo Ruiz Padillo Fdo: Mª del Carmen Carrión Pérez

Fdo: Ercan Engin Kuruoglu

Memoria presentada por Diego Salas González para optar al Grado de Doctor por la Universidad de Granada.

Fdo: Diego Salas González

"Sería fantástico,
que la ciencia fuese neutral".

Joan Manuel Serrat
-Seria Fantàstic-

"Es mi barrio pobre, pero lo respeto porque es donde vivo.
Y además, bueno...
además lo quiero porque un día de enero me encontré contigo".

Tito Fernández
-Mi barrio-

"*Todos realizamos diariamente multitud de tareas que no serían posible sin el desarrollo de la tecnología. La tecnología nos permite despertarnos con el sonido de un despertador, asearnos con agua corriente o tener luz eléctrica en casa. Todas estas actividades y muchísimas, por otra parte normales en la mayoría de los hogares españoles, son posibles gracias a la labor de miles de científicos e investigadores que a lo largo de los siglos han aportado su granito de arena para comprender un poquito mejor la naturaleza.*

Formar parte de esta cadena de personas (los científicos) que dura ya miles de años (desde la invención del fuego o la rueda por poner un ejemplo) que edifican, a partir del conocimiento científico de cada época y mediante un esfuerzo intelectual considerable, nuevas teorías que ayudan a explicar los fenómenos que nos rodean fue mi principal motivación para comenzar la carrera de investigación en el Departamento de Física Aplicada de la Universidad de Granada.

La incertidumbre que siempre conlleva dedicarse a la investigación en un país caracterizado por la 'fuga de cerebros' como es España, hace de la investigación una profesión muy vocacional y muchas veces muy competitiva por el número escaso de becas y oportunidades. Sin embargo el descubrimiento de nuevas teorías después de meses o años de trabajo de investigación, el saber que estamos contribuyendo al desarrollo de la ciencia del mismo modo que otros científicos lo han estado haciendo durante siglos produce una sensación indescriptible, que sin duda es el motor de todos los que elegimos, por vocación, la carrera de investigación"

Diego Salas González, IDEAL, año 2005.

Agradecimientos

Esta Tesis no hubiera sido posible sin la colaboración de mis directores Diego Pablo Ruiz y María del Carmen Carrión. Les agradezco sinceramente que me hayan acogido en su grupo de investigación en estos cuatro años que han supuesto para mí un crecimiento tanto en el plano profesional como personal.

A Ercan Kuruoglu, por su inestimable ayuda en los diez meses vividos en Italia.

A todos los buenos profesores de los que he tenido la fortuna de ser alumno desde el colegio hasta el doctorado (María del Mar Camacho, Agustín Monzón, Juan López, Ángel Delgado, Javier de la Portilla, Joaquín Fdez-Valdivia, etc.) porque, de algún modo, aún hoy sigo aprendiendo de ellos.

A todas aquellas personas que han hecho que me sienta en Italia como en casa. Especialmente a Alessandra, Maria Grazia, Gabriele, Massimo, Sara y Davide.

En Granada, a mis compañeros de despacho, Pepe, Isa, David, Antonio, Susana y Otilia. Por todos los buenos ratos vividos.

A Juanma Górriz y Javier Ramírez por su confianza en mi trabajo.

A Conrado Ferrer por la música.

A mis alumnos de Ingeniería Acústica de 3º de Ingeniería de Telecomunicación, por aplaudirme en el momento más oportuno.

A toda mi familia por su ejemplo de trabajo y haberme enseñado a ser feliz.

A Laura, por acompañarme siempre en esta aventura y mostrarme que las cuestiones más importantes no pueden explicarse mediante ecuaciones. Afortunadamente.

Título de Doctor con Mención Europea

Con el fin de obtener la Mención Europea en el Título de Doctor (aprobada en Junta de Gobierno de la Universidad de Granada el 5 de Febrero de 2001), se han cumplido, en lo que atañe a esta Tesis Doctoral y a su Defensa, los siguientes requisitos:

- Durante su etapa de formación, el doctorando ha realizado una estancia superior a 3 meses fuera de España en una institución de enseñanza superior de otro Estado europeo cursando estudios o realizando trabajos de investigación que le han sido reconocidos por el órgano responsable del programa. En concreto el doctorando ha acumulado 10 meses en 3 estancias en el ISTI-CNR de Pisa en Italia.
- Parte de la Tesis se ha redactado y presentado en una de las lenguas oficiales de la Unión Europea distinta a alguna de las lenguas oficiales en España. En concreto en Inglés.
- La Tesis ha sido informada por más de dos expertos pertenecientes a alguna institución de educación superior o instituto de investigación de un Estado miembro de la Unión Europea distinto de España.
- Un experto perteneciente a alguna institución de educación superior o instituto de investigaciçón de un Estado miembro de la Unión Europea distinto de España, con el grado de doctor, y distinto de los mencionados antes, forme parte del tribunal evaluador de la Tesis.

Financiación

Las principales fuentes de financiación de esta Tesis Doctoral han sido las siguientes:

- Beca-contrato de Formación del Profesorado Universitario (FPU) del Ministerio de Educación y Ciencia, con número de referencia AP2003-4441. Desde el 1 de febrero de 2004 hasta el 31 de enero de 2008.
- Proyecto "Tratamiento de señal en ambientes no estacionarios usando técnicas de procesado conjunto ICA, HOS y PF. Aplicaciones en comunicaciones e imágenes", del Ministerio de Ciencia y Tecnología con referencia TEC 2004-06096-C03-02/TCM. Desde el año 2004 hasta 2007.
- Ayudas para estancias de la beca-contrato FPU del Ministerio de Educación y Ciencia para la realización de tres estancias breves en los años 2005, 2006 y 2007, con una duración total de 10 meses en el *Istituto di Scienza e Tecnologie dell'Informazione*, perteneciente al *Consiglio Nazionalle della Ricerca* de Italia, en la ciudad de Pisa.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

Esta Tesis Doctoral está organizada en dos grandes bloques. En el primero de ellos se desarrollan nuevos métodos estadísticos para el análisis de señales mediante un modelo Bayesiano de mezclas de distribuciones α-estables. Como se explicará con detalle en el siguiente capítulo, la distribución α-estable engloba una familia de distribuciones impulsivas y asimétricas que contiene a la distribución Gaussiana como caso particular. Además, una variable aleatoria con distribución α-estable posee una serie de propiedades tales como la propiedad de estabilidad o el Teorema Fundamental del Límite Generalizado, que justifican el desarrollo de nuevas técnicas de modelado y estimación paramétrica usando este tipo de distribuciones.

El desarrollo de los modelos de mezclas finitas ha propiciado la aparición de una herramienta de modelado estadístico de gran flexibilidad, capaz de encontrar aplicaciones en multitud de fenómenos en disciplinas dispares, como en la astronomía, física, biología, genética, medicina, psiquiatría, economía, ingeniería o ciencias sociales. El modelo de mezclas consiste en una distribución densidad de probabilidad que es una combinación convexa de varias distribuciones. Matemáticamente, si la variable aleatoria discreta Y es una mezcla de k componentes, entonces la función densidad de probabilidad de Y, $p_Y(y)$, es una suma de distribuciones con distintos pesos w_i de los componentes. Este modelo se representa del siguiente modo:

$$p_Y(y) = \sum_{j=1}^{k} w_j p(y|\theta_j) \tag{1}$$

$$0 \leq w_j \leq 1 \ \ (\forall j) \ \ ; \ \ \sum_{j=1}^{k} w_j = 1$$

donde w_j es la proporción del componente j en la mezcla y $p_Y(y)$ es la función densidad de probabilidad de la señal y de longitud N. $p(y|\theta_j)$ representa una distribución cualquiera con variable y y n parámetros que denotamos genéricamente como $\theta = [\theta_1, ..., \theta_n]$.

En la primera parte de esta Tesis Doctoral, desarrollamos un modelo Bayesiano de mezcla de distribuciones α-estable tanto simétricas como asimétricas. Esta distinción entre ambos casos, se debe al hecho de que la distribución simétrica posee algunas características propias que hacen que ambos algoritmos sean diferentes. Además de comparar los modelos de mezcla desarrollados con otros modelos de mezcla existentes en la literatura, aplicaremos este nuevo modelo a datos procedentes de distintas disciplinas, como la astronomía, geología, biología y economía. De este modo resaltaremos el alto rango de aplicación de los algoritmos presentados en esta memoria.

En la segunda parte de esta Tesis Doctoral, se aplican las propiedades de la distribución α-estable al estudio de la distribución de la expresión genética en micromatrices. Las micromatrices de ADN son un conjunto de celdas microscópicas de ADN. Cada una de estas celdas representa un gen determinado dispuesto en forma de matriz, es decir, formando distintas filas y columnas. Este dispositivo permite realizar medidas tanto cualitativas como cuantitativas de la expresión genética bajo distintas condiciones. Típicamente, los datos procedentes de micromatrices proporcionan información de la expresión de miles de genes simultáneamente.

Existen diversos métodos de estudio de la expresión genética mediante micromatrices. El caso estudiado en esta Tesis Doctoral es el que hace uso de la técnica de micromatrices de dos colores. Típicamente, esta técnica consiste en la comparación entre dos tipos de ADN complementario procedente de dos muestras distintas (por ejemplo, células sanas y células enfermas), estas dos muestras son diferenciadas mediante dos colores fluorescentes distintos, generalmente rojo y verde.

El par de muestras de ADN complementario, se mezcla en una misma micromatriz, la cual es escaneada con el fin de medir la fluorescencia tras la excitación con un haz láser. Finalmente, los valores del logaritmo de la intensidad relativa para cada uno de los colores, rojo y verde, convenientemente normalizados, proporcionan la información necesaria para identificar qué genes están expresados.

0.1. Mezcla de distribuciones α-estables

0.1.1. Modelo de mezcla α-estable simétrico

Planteamiento del problema

El modelo de mezcla de distribuciones α-estable simétricas viene descrito matemáticamente mediante la siguiente expresión:

$$p_Y(y) = \sum_{j=1}^{k} w_j f_{\alpha_j,0}(y|\gamma_j, \mu_j)$$

$$0 \leq w_j \leq 1(\forall j) \text{ y } \sum_{j=1}^{k} w_j = 1$$

donde w_j es el peso del componente j y $p_Y(y)$ es la función densidad de probabilidad del vector observación y. $f_{\alpha,\beta}(y|\gamma,\mu)$ denota una distribución α-estable de parámetros α, β, γ y μ. Este modelo considera que los valores del vector y han sido aleatoriamente extraídos de un número k de componentes, que denotamos como $j = 1, 2, ...k$.

En este primer acercamiento al modelado de mezcla de α-estables, el parámetro β, que es el que tiene en cuenta la asimetría de la distribución es igual a cero. Esto simplifica los cálculos, y nos permite realizar la inferencia de los parámetros desconocidos γ_j, μ_j, w_j mediante el algoritmo de Gibbs, obteniendo unas distribuciones a posteriori muy similares a las obtenidas mediante uso de la estadística Bayesiana y métodos Monte Carlo basados en cadenas de Markov para el caso de la mezcla de Gaussianas.

Trabajos anteriores

En la literatura existen multitud de trabajos sobre mezcla de distintas distribuciones: t-student [McLachlan & Peel, 1998], Gamma [M. Wiper & Ruggieri, 2001], Poisson [Fernandez & Green, 2002] o Weibull [Tsionas, 1999] entre otras. Los modelos de mezcla han encontrado un amplio rango de aplicaciones en diversas disciplinas tanto en física, ingeniería, biología o ciencias sociales [McLachlan & Peel, 2000]. De todos los modelos de mezcla estudiados, la mezcla de Gaussianas es el que ha sido aplicado en mayor número de problemas. Esto es debido a un doble motivo: en primer lugar, algunas de las propiedades de dicha distribución, como la propiedad de estabilidad y el Teorema Central del Límite,

permiten, en algunos casos, justificar el uso de la distribución Normal. Y, por otra parte, la expresión analítica de la distribución Gaussiana es sencilla y el significado de sus dos parámetros, la media y la desviación estándar, es muy intuitivo.

A pesar de, tal y como se ha apuntado, la existencia de una gran cantidad de modelos de mezcla de distintas distribuciones, el modelo de mezcla de distribuciones α-estable simétricas no se había resuelto con anterioridad al trabajo realizado en esta Tesis. Probablemente, debido al hecho de la no existencia de expresión analítica para expresar la función densidad de probabilidad estable.

Por este motivo, en el Capítulo 2, donde se presenta el modelo de mezcla de distribuciones α-estable simétricas comparamos el algoritmo propuesto con la mezcla Gaussianas.

Aportaciones en esta memoria

En el Capítulo 2 se presenta el modelo de mezclas de distribuciones α-estables simétricas. El problema provocado por la inexistencia de una expresión analítica para la pdf α-estable es solventado mediante el uso de la representación mediante mezcla escalada de distribuciones Normales. Esta representación es válida para varias familias de distribuciones subgaussianas y permite, mediante la introducción de una variable aleatoria auxiliar λ de dominio $\lambda \in [0, +\infty)]$, escribir la pdf α-estable como una distribución condicionalmente Gaussiana.

Por tanto, en el Capítulo 2 de esta memoria realizamos las siguientes aportaciones originales:

- Proponemos por vez primera en la literatura un modelo de mezcla de distribuciones α-estables simétricas, que entre otras características, es una generalización del ampliamente estudiado modelo de mezclas de Gaussianas.

- Usamos la representación mediante mezcla escalada de Gaussianas para obtener una expresión analítica de la función densidad de probabilidad α-estable simétrica. Esta representación, nunca había usada anteriormente en el contexto de la los modelos de mezcla.

- El uso de la representación mediante mezcla escalada de Gaussianas de la distribución simétrica α-estable, nos permite escribir la función densidad de probabilidad α-estable simétrica como una distribución Normal, condicionada a la variable auxiliar λ. Por tanto, aunque el algoritmo aquí propuesto resuelve el complejo modelo de mezclas α-estable simétricas, com-

parte la sencillez del modelo Bayesiano de mezcla de Gaussianas. Esto nos permite usar la distribución a priori conjugada y escribir una expresión analítica para las distribuciones a posteriori de algunos de los parámetros desconocidos del problema.

- Resolvemos la estimación de parámetros mediante un planteamiento estrictamente Bayesiano del problema. Usando métodos de muestreo Monte Carlo como el algoritmo de muestreo por rechazo, muestreo de Gibbs y Metropolis.

- En el contexto de mezcla de distribuciones α-estables, nunca con anterioridad se habían considerados métodos Monte Carlo de dimensión variable como el algoritmo Monte Carlo basado en cadenas de Markov con saltos reversibles (RJMCMC, por sus siglas en inglés). El cual nos permite calcular el número de componentes que componen la mezcla.

- Al ser este modelo una generalización del modelo de mezclas de Gaussianas, el rango de aplicación de éste se extiende a multitud de distintas disciplinas y materias. Por otro lado, presenta una ventaja bastante importante respecto al modelo Gaussiano, ya que la mezcla de distribuciones α-estables permite, además, modelar datos cuya distribución es una mezcla de componentes impulsivos.

- Este algoritmo es comparado con el modelo de mezcla Gaussiano. Se comprueba a partir del análisis de las simulaciones realizadas, que el modelo α-estable simétrico permite modelar datos como mezcla de distribuciones impulsivas de manera más compacta que lo hace la distribución Gaussiana ya que precisa un menor número de componentes.

- Por otro lado, el modelo de mezcla α-estable simétrico demuestra funcionar muy bien y estimar correctamente todos los parámetros del modelo, incluso cuando los datos son mezcla de distribuciones Normales. No es posible decir lo mismo en el caso contrario, es decir, para un vector de datos mezcla de distribuciones α-estables simétricas.

- El amplio rango de aplicación y distintas posibilidades que posee este modelo es mostrado mediante tres simulaciones con datos reales de disciplinas dispares como la biología, astrofísica y geología.

0.1.2. Modelo de mezcla α-estable asimétrico

Planteamiento del problema

El modelo de mezcla de distribuciones α-estable asimétricas viene descrito matemáticamente del siguiente modo:

$$p_Y(y) = \sum_{j=1}^{k} w_j f_{\alpha_j,\beta_j}(y|\gamma_j,\mu_j)$$

$$0 \leq w_j \leq 1(\forall j) \text{ y } \sum_{j=1}^{k} w_j = 1$$

donde w_j es el peso del componente j y $p_Y(y)$ es la función densidad de probabilidad del vector observación y.

Nuestro objetivo es, para un vector observación y dado, estimar el resto de parámetros desconocidos del problema, $\alpha_j, \beta_j, \gamma_j, \mu_j, w_j, k$ mediante el uso de la estadística Bayesiana y métodos Monte Carlo basados en cadenas de Markov. Aunque existen en la literatura multitud de modelos de mezcla para distintas familias de distribuciones, este objetivo nunca había sido alcanzado con anterioridad de manera satisfactoria para mezcla de α-estables asimétricas.

Trabajos anteriores

En la literatura, existe tan sólo un trabajo previo en mezcla de distribuciones α-estables [Casarin, 2004]. Se trata de un informe técnico que nunca fue publicado y que posee varios puntos débiles. En primer lugar, el problema de la inexistencia de función densidad de probabilidad lo resuelve mediante la representación de la densidad α-estable propuesta en [Buckle, 1995]. En dicho trabajo, la estimación de los parámetros de la α-estable tiene graves problemas de convergencia, por lo que cada parámetro es inicializado con valores muy cercanos a los valores reales. Además, es muy costoso computacionalmente ya que precisa de la introducción y actualización de dos variables aleatorias de dimensión igual al tamaño del vector observación considerado. Por otro lado, en las simulaciones presentadas en [Casarin, 2004], o bien no se estiman todos los parámetros, es decir, se consideran conocidos algunos de ellos con la consecuente simplificación que esto produce en el problema; o los distintos componentes de la mezcla están muy separados entre sí, (en un ejemplo de dos componentes, la moda de cada uno de ellos es $\mu_1 = 0$ y $\mu_2 = 30$. Para este caso, aunque

existen dos componentes, la mezcla de dichos componentes no se ha producido, por lo que el algoritmo propuesto en [Casarin, 2004] no está siendo aplicado a una mezcla de distribuciones sino a dos componentes individuales, como hacía el algoritmo anterior de [Buckle, 1995]. Además de todos estos inconvenientes, el modelo planteado en [Casarin, 2004] supone conocido el número de componentes de la mezcla, por lo que la flexibilidad de dicho algoritmo es menor que la del modelo planteado en el Capítulo 3 de esta memoria, en la que sí se calcula el número de componentes de la mezcla.

Aportaciones en esta memoria

En el Capítulo 3 se presenta el modelo de mezclas de distribuciones α-estables más general. El problema provocado por la inexistencia de una expresión analítica para la pdf α-estable es superado mediante la resolución numérica de la integral de la función característica α-estable. Este modelo, por lo tanto, además de ser una generalización del modelo de mezclas Gaussiano, es una generalización del modelo de mezcla α-estable simétrico presentado en el Capítulo 2.

A continuación se resumen algunas de las aportaciones del modelo Bayesiano de mezclas de distribuciones α-estable presentado con detalle en el Capítulo 3 de esta memoria:

- Proponemos por primera vez en la literatura un análisis Bayesiano del modelo de mezcla α-estable que permite la estimación de todos los parámetros del problema de manera exacta.
- El modelo es una generalización de la mezcla de Gaussianas. En el caso en que los datos son mezcla de componentes impulsivos, la mezcla de Gaussianas no converge, mientras que nuestro modelo es robusto frente a señales impulsivas.
- Por otro lado, en caso de datos impulsivos, la mezcla de α-estables requiere un menor número de componentes para ajustar la distribución de los datos que en el caso Gaussiano.
- Al igual que en la aportación del Capítulo 2, en el modelo de mezcla α-estable más general, el número de componentes en la mezcla es calculado satisfactoriamente mediante técnicas Monte Carlo basadas en cadenas de Markov de dimensión variable, en concreto el algoritmo de saltos reversibles (RJMCMC, *Reversible jump Markov chain Monte Carlo*). Esta Tesis Doctoral presenta las dos únicas ocasiones en que este algoritmo se ha usado en el contexto de mezclas α-estables.

- El algoritmo ha sido ampliamente comparado con el método propuesto por [Casarin, 2004], las ventajas del método presentado en esta memoria son claras: menor complejidad computacional debido a la integración numérica de la función característica de la distribución α-estable. El algoritmo es mucho más robusto y la convergencia del mismo no depende de los valores iniciales de los parámetros.

- Del mismo modo que en el caso de la mezcla simétrica de distribuciones α-estable, hemos querido mostrar el amplio rango de aplicación de este algoritmo mediante el estudio de datos reales de diversas disciplinas, como la economía y la biología.

0.2. Expresión genética y distribución α-estable

0.2.1. Modelado de la distribución de la expresión de genes mediante la distribución α-estable

Planteamiento del problema

En la segunda parte de esta Tesis Doctoral, se aplican las propiedades de la distribución α-estable al estudio de la distribución de la expresión genética en micromatrices. Las micromatrices de ADN son un conjunto de celdas microscópicas de ADN. Cada una de estas celdas representa un gen determinado, dispuesto en forma de matriz, es decir, formando distintas filas y columnas. Este dispositivo permite realizar medidas tanto cualitativas como cuantitativas de la expresión genética bajo distintas condiciones. Típicamente, los datos procedentes de micromatrices proporcionan información de la expresión de miles de genes simultáneamente.

Existen diversos métodos de estudio de la expresión genética mediante micromatrices, el caso estudiado en esta Tesis Doctoral es el que hace uso de la técnica de micromatrices de dos colores. Típicamente, esta técnica consiste en la comparación entre dos tipos de ADN complementario procedente de dos muestras distintas (por ejemplo, células sanas y células enfermas), estas dos muestras son diferenciadas mediante dos colores fluorescentes distintos, generalmente rojo y verde.

Las dos muestras de ADN complementario se mezclan en una misma micromatriz que es escaneada con el fin de medir la fluorescencia tras la excitación con un haz láser. Finalmente, los valores del logaritmo de la intensidad relativa para cada uno de los colores, rojo y verde, convenientemente normalizados,

proporcionan la información necesaria para identificar genes expresados.

En el Capítulo 5 estudiamos la distribución del logaritmo de las intensidades relativas de la expresión genética en datos de micromatrices usando la distribución α-estable. Además, comprobamos que la distribución de la expresión de genes comparte algunas de sus propiedades empíricas con la distribución α-estable.

Trabajos anteriores

Tras el proceso de normalización, la distribución de la expresión de genes para distintos organismos, posee una forma parecida independientemente . En general, presenta colas con peso superior a la distribución Gaussiana y cierto grado de asimetría. Esta distribución ha sido modelada en la literatura mediante el uso de distintas funciones densidad de probabilidad:

[Kuznetsov, 2001] modela la distribución de la expresión de genes mediante diferentes clases de densidades asimétricas como la distribución de Poisson, la distribución exponencial, series logarítmicas y la distribución de Pareto. En dicho trabajo, muestra el resultado obtenido tan sólo para la distribución Pareto, ya que es la que ofrece un mejor ajuste de la distribución de genes.

Por otro lado, [Hoyle *et al.*, 2002] analiza una gran cantidad de datos genéticos reales. Finalmente, la distribución de genes es aproximada mediante dos distribuciones: una Log-normal en el centro y una ley de potencias en las colas. Así que el comportamiento de la cola de la distribución no es exponencial. Además, en dicho artículo, se resalta que la varianza de las intensidades muestran una correlación positiva con el número de genes. Es decir, la varianza de los datos genéticos de micromatrices no se estabiliza conforme aumenta el número de datos estudiados.

En [Purdom & Holmes, 2005], la distribución de la expresión genética es ajustada mediante la distribución de Laplace Asimétrica, que tiene un comportamiento de las colas de tipo exponencial. La mejora de este método con respecto al uso de la distribución Gaussiana es notable debido a que esta distribución presenta mayor peso en las colas y asimetría. Por otra parte, aunque la distribución de Laplace Asimétrica posee colas de mayor peso que la Normal, el comportamiento asintótico de dichas colas es exponencial, en lugar de potencial, y no sigue un comportamiento tipo ley de Pareto.

En [Khondoker *et al.*, 2006], se presenta un modelo estadístico para estimar la expresión genética usando datos procedente de múltiples escaneados mediante láser. La distribución de la expresión de genes en dicho trabajo se modela mediante una distribución de Cauchy.

Aportaciones en esta memoria

En esta memoria, se propone un modelado de la distribución de la expresión de genes usando la distribución α-estable. Este modelado mejora a otros trabajos anteriores existentes en la literatura. Además, la distribución de la expresión de genes comparte propiedades empíricas con la distribución α-estable. Las principales aportaciones del Capítulo 5 de esta Tesis Doctoral se resumen a continuación:

- En [Khondoker *et al.*, 2006], se modela la distribución de la expresión de genes mediante distintas familias de distribuciones asimétricas. Finalmente, la distribución que ofrece mejor ajuste fue la distribución Pareto aunque para ello tuvo que introducirse un parámetro posición adicional para generalizar dicha distribución. La distribución α-estable ya cuenta con dicho parámetro posición y proporciona un buen ajuste tanto en el centro de la distribución como en las colas. Además, la distribución α-estable también posee comportamiento asintótico de tipo Pareto (ley de potencias) en las colas cuando $\alpha < 2$.

- Mandelbrot hizo hincapié en los primeros trabajos de aplicación de la distribución α-estable, en el hecho de que el uso de dicha distribución para el estudio y descripción de datos biológicos era preferible al uso de distribuciones de tipo Zipf-Pareto debido a motivos tanto teóricos como prácticos.

- La distribución α-estable permite el modelado de la distribución de la expresión de genes de manera más compacta, mediante el uso de una sola distribución. Al contrario de lo que sucede en [Hoyle *et al.*, 2002], donde la aproximación se realiza mediante una distribución Log-normal en el centro de la distribución y una ley de potencias o ley Zipf en las colas.

- Además, en [Hoyle *et al.*, 2002], se apunta que la varianza σ^2 de las intensidades logarítmicas aumenta conforme el número de genes estudiado aumenta. Este resultado está en completo acuerdo con las propiedades de la distribución α-estable. La varianza es un parámetro que no está definido para procesos estables con $\alpha < 2$.

- Tras la comparación tanto cualitativa como cuantitativa del ajuste proporcionado por la distribución α-estable con respecto a la distribución de Laplace Asimétrica estudiada en [Purdom & Holmes, 2005], se comprueba cómo la distribución de Laplace asimétrica no es capaz de ajustar siempre

de manera satisfactoria la distribución de la expresión de genes. El histograma de intensidades de la expresión genética presenta, normalmente, un comportamiento más suave alrededor del máximo que la distribución de Laplace.

- En [Khondoker *et al.*, 2006], la distribución de la expresión de genes se modela mediante una distribución de Cauchy. Nosotros, por otra parte, no asumimos que la distribución de la expresión de genes es Normal o Cauchy, pero ambas distribuciones son casos particulares de la distribución α-estable.

- El modelado de la distribución de la expresión de genes mediante la α-estable y los excelentes resultados obtenidos en el ajuste, son un primer acercamiento de esta distribución al estudio de los datos de micromatrices. Este estudio sirve, por ejemplo, de punto de partida para el diseño de un estadístico basado en las propiedades de la distribución α-estable que nos permite establecer un criterio sobre si un determinado gen está o no expresado. Los detalles sobre el diseño y funcionamiento de dicho estadístico se explican en el Capítulo 6.

0.2.2. Expresión genética mediante mezcla de distribuciones α-estables

Planteamiento del problema

En el Capítulo 5 se modela la distribución de la expresión de genes usando la distribución α-estable. Este enfoque, aunque puede ser útil en primera aproximación, no es realista debido al hecho de que los genes estudiados pertenecen a dos subpoblaciones distintas: genes expresados y no expresados. Recientemente, en la literatura, han aparecido diversos métodos que nos permiten determinar si un gen está o no expresado.

La estadística Bayesiana junto con los modelos de mezcla, nos permite la construcción de un estadístico para el estudio de la expresión genética. Para ello, se construye un modelo de mezcla con dos componentes. Uno de los componentes modela el conjunto de genes no expresados, mientras que el otro modela los que sí lo están.

Asumimos que los datos con los que trabajamos son el logaritmo en base-2 de intensidades R_{ij} y G_{ij} normalizadas mediante regresión local (LOESS) [Yang *et al.*, 2002a]. Siendo N el número de genes en cada matriz y n el número

de réplicas o matrices (es decir, el número de repeticiones idénticas del experimento). Los datos se describen, por lo tanto, como $M_{ij} = log\left(\frac{R_{ij}}{G_{ij}}\right)$, donde $i = 1...N$, $j = 1...n$.

La mayoría de los genes no están expresados, para dicho conjunto de genes, se considera $\mu_i = 0$. Sin embargo, los genes expresados se modelan mediante una distribución P centrada en cero. Por tanto, si la proporción de genes expresados es w, usamos un modelo de mezcla para expresar matemáticamente la distribución del parámetro μ_i:

$$\mu_i \sim wP + (1 - w)\delta(0) \tag{2}$$

Trabajos anteriores

Existen algunos trabajos en la literatura que plantean el diseño de un estadístico para discernir si un gen está expresado o no mediante el uso de los modelos de mezcla. Además, la estadística Bayesiana nos permite introducir, a través de distribuciones a priori, información teórica o experimental sobre el conjunto de datos de micromatrices antes de calcular el estadístico.

La estadística Bayesiana y el modelado mediante mezcla de distribuciones están adquiriendo gran popularidad en el contexto de el estudio de genes mediante micromatrices [Allison *et al.*, 2006]. Multitud de trabajos recientes en la literatura usan los modelos de mezcla en problemas de expresión genética [Newton *et al.*, 2004; Do *et al.*, 2005; Gottardo *et al.*, 2003].

De ellos, los que tienen una relación más estrecha con el método presentado en el Capítulo 6 de esta memoria son dos: [Lonnstedt & Speed, 2002] y [Bhowmick *et al.*, 2006].

En [Lonnstedt & Speed, 2002], los genes se modelan como pertenecientes a dos grupos distintos, genes expresados y no expresados. El modelo de mezcla requiere, por lo tanto, dos componentes. Para ello se elige una distribución de Dirac para la media μ_i de los genes no expresados y Gaussiana para los que sí lo están. Este trabajo supone independencia entre genes y, además, que las distintas repeticiones experimentales para cada gen siguen una distribución Gaussiana con media μ_i y varianza σ_i^2 con distribución Gamma Inversa, por lo que la distribución de la expresión de genes está modelada como una t-student. Este modelo, a pesar de que se remarca en varias ocasiones en [Lonnstedt & Speed, 2002] que supone que la distribución de la expresión de genes no expresados en Gaussiana y que la distribución de la expresión de genes presenta mayor grado de impulsividad, en realidad se trata de una mezcla escalada de Gaussianas con distribución Gamma Inversa para la varianza. El estadístico propuesto

se compara con 4 distintos métodos. La mejora obtenida es muy pequeña con respecto a otros estadísticos más sencillos.

En [Bhowmick *et al.*, 2006], el modelo es muy similar al presentado en [Lonnstedt & Speed, 2002], sólo que en vez de suponer para el modelo de mezcla de dos componentes la distribución Gaussiana, hace uso de la distribución de Laplace. En los resultados, compara para datos simulados el funcionamiento de ambos estadísticos entre sí y concluye que el funcionamiento cuando los datos tienen distribución dada por una mezcla de Dirac y Laplace, que es la que propone su propio modelo, no difiere del obtenido mediante el estadístico de [Lonnstedt & Speed, 2002]. Sin embargo, cuando los datos se simulan mediante el modelado propuesto por [Lonnstedt & Speed, 2002], este procedimiento es, efectivamente mejor.

Aportaciones en esta memoria

El estadístico diseñado está basado en la propiedad de mezcla escalada de Gaussianas y en la distribución α-estable.

En el Capítulo 6, asumimos que los valores M_{ij} son independientes. Además, suponemos que M_{ij} son variables aleatorias procedentes de una distribución Normal con media μ_i y varianza $\lambda_i\sigma^2$, donde λ_i tiene distribución α-estable positiva. En virtud de la propiedad del producto o representación mediante mezcla escalada de Gaussianas es un modelo α-estable.

Como ya hemos apuntado, los genes se consideran pertenecientes a uno de los dos grupos siguientes: expresados o no expresados. Sea w la probabilidad de que un gen esté expresado, modelamos el parámetro μ se modela como una variable aleatoria extraída de una distribución α-estable simétrica centrada en cero si el gen está efectivamente expresado, o como $\mu = 0$ en caso contrario.

$$p(\mu_i|\lambda_i, \sigma) = w f_{\alpha,0}(\sigma, 0) + (1 - w)\delta(0) \tag{3}$$

Matemáticamente, se trata de un modelo de mezcla de dos distribuciones: α-estable simétrica y Dirac.

A continuación se enumeran las principales aportaciones originales del Capítulo 6, donde se presenta el diseño de un nuevo estadístico basado en la la distribución α-estable para indicar si un determinado gen está, o no, expresado.

- El uso de la distribución α-estable para el modelado de la distribución de la expresión de genes está suficientemente motivado por el estudio detallado realizado en el Capítulo 5 de esta Memoria, donde se comprobó que la

distribución α-estable ajusta con gran exactitud la distribución de la expresión de genes, además de compartir con ella algunas de sus propiedades más importantes.

- El diseño de un estadístico mediante la suposición de que cada gen puede estar expresado o no, usando para ello un modelado matemático mediante mezcla de distribuciones, permite calcular la probabilidad de que un gen esté expresado sin la necesidad de calcular el valor P asociado a un resultado observado. Siendo el valor P la probabilidad de obtener un valor como el observado o más extremo si la hipótesis nula es cierta.

- El uso de la distribución α-estable como parte del modelo de mezcla nos permite simplificar notablemente el problema de cálculo del estadístico debido al uso de diversas propiedades de esta distribución. Así, la estimación de parámetros de la distribución puede realizarse mediante multitud de técnicas existentes en la literatura.

- Además, el modelo matemático se construye de manera relativamente simple, debido al uso de la distribución α-estable simétrica para modelar la distribución de la expresión de genes y a la representación mediante mezcla escalada de Gaussianas.

- En el diseño del estadístico hay que resolver varias integrales numéricamente. Una vez más, las propiedades de la distribución α-estable proporcionan un modo muy sencillo para aproximar las integrales por sumatorias sin más que extraer muestras con distribución α-estable.

- El uso de una distribución con gran peso en las colas como es la distribución α-estable, permite que las medidas tomadas para genes considerados por el modelo como no expresados tengan una mayor dispersión. Esto confiere al estadístico diseñado una ventaja sobre el uso de otros estadísticos basados en la distribución de Laplace y t-student en el caso en que los datos genéticos estudiados tengan una gran variabilidad entre las distintas repeticiones experimentales realizas en el laboratorio.

- Para mostrar el funcionamiento del estadístico α-estable, éste ha sido probado con datos simulados y comparado con un trabajo previo similar basado en la distribución t-student [Lonnstedt & Speed, 2002].

Parte I

Mezcla de distribuciones α-estables

CAPÍTULO

UNO

FUNDAMENTO TEÓRICO: DISTRIBUCIÓN α-ESTABLE, INFERENCIA BAYESIANA Y MODELOS DE MEZCLA

En este capítulo abordaremos las principales teorías matemáticas y conceptos sobre los que se sustenta esta Tesis Doctoral. Comenzaremos presentando la distribución α-estable en la Sección 1.1 y destacando su importancia en el modelado de fenómenos impulsivos. Además, detallaremos sus principales propiedades con especial énfasis a aquéllas que se usarán en los capítulos sucesivos de esta memoria. En la Sección 1.2, presentaremos el Teorema de Bayes y su aplicación a problemas de inferencia estadística y la importancia de los métodos Monte Carlo basados en cadenas de Markov para la resolución numérica de complejas integrales obtenidas bajo el paradigma Bayesiano. Por otro lado, en la Sección 1.3 introduciremos la formulación matemática de los modelos de mezcla y apuntaremos algunas de sus principales aplicaciones.

1.1. Distribución α-estable

Algunos fenómenos de la naturaleza no pueden ser descritos mediante la suposición Gaussiana ya que presentan un grado de impulsividad mayor que el que la distribución Normal es capaz de describir. Es decir, son posibles eventos o sucesos que, descritos mediante una distribución Gaussiana, serían considerados como muy poco probables. La distribución α-estable ha sido usada en

la literatura para describir ese tipo de fenómenos. La teoría de distribuciones estables fue desarrollada por primera vez en los años 20-30 del pasado siglo por Paul Lévy y Aleksander Khinchine [Samorodnitsky & Taqqu, 1994]. Desde entonces, esta distribución ha sido aplicada en diferentes áreas de conocimiento, tales como economía, física, hidrología, biología y procesado de señal [Adler *et al.*, 1998]. Una de las primeras aplicaciones de la distribución α-estable fue descubierta en el campo de la astrofísica. El astrónomo Danés Holtsmark en 1919 comprobó que las fluctuaciones aleatorias del campo gravitacional de las estrellas en el espacio tiene una distribución estable con exponente característico $\alpha = 1{,}5$. Sin embargo, no fue hasta los primeros trabajos de Mandelbrot en economía en la década de los 60 que no se popularizó la distribución α-estable. Mandelbrot propuso una revolucionaria teoría basada en dicha distribución para resolver el problema de la fluctuación de los precios. Más tarde se demostró que muchas otras variables en economía siguen una distribución α-estable.

Esta distribución también ha sido aplicada en procesado de señal y comunicaciones, véase [Nikias & Shao, 1995], por ejemplo, para el modelado de ruido impulsivo. Recientemente se ha demostrado que gran cantidad de procesos ruidosos en la naturaleza, atendiendo a las condiciones de generación y propagación de dicho ruido (ruido en líneas telefónicas, radares, series financieras, etc.), pueden modelarse satisfactoriamente usando una distribución α-estable.

En esta sección presentaremos la distribución α-estable y alguna de sus propiedades, remarcando que la distribución α-estable cumple el Teorema Central del Límite y la propiedad de estabilidad y que, además, contiene a la distribución Normal como caso particular de ésta. Por otro lado también mostraremos la principal dificultad de trabajar con este tipo de distribuciones, como es que carece, en general, de expresión analítica. Como consecuencia de este problema esta distribución no está muy extendida entre los investigadores que se dedican a métodos Bayesianos. Como veremos, esta dificultad puede ser superada haciendo uso de una propiedad de la distribución α-estable simétrica que permite escribir una distribución de este tipo como una Gaussiana condicionada a una variable aleatoria con distribución α-estable.

1.1.1. Definición

Introduciremos la distribución α-estable mediante 4 definiciones equivalentes.

Definición 1. *Una variable aleatoria X tiene distribución estable si, para cualquier constante positiva A y B, existe un número positivo C y un número real*

D tal que

$$AX_1 + BX_2 \stackrel{d}{=} CX + D \tag{1.1}$$

donde X_1 y X_2 son copias independientes de X y $\stackrel{d}{=}$ denota igualdad en distribución.

Por otro lado, una variable aleatoria tiene distribución estable simétrica si X y $-X$ tienen la misma distribución.

Definición 2. *Para cualquier variable aleatoria con distribución α-estable, hay un número $\alpha \in (0, 2]$ para el que la constante C en la expresión (1.1) cumple*

$$C^\alpha = A^\alpha + B^\alpha$$

El parámetro α se denomina índice de estabilidad o exponente característico. Una variable aleatoria X con índice α se denomina α-estable.

Ejemplo. *Si X es una variable aleatoria Gaussiana con media μ y varianza σ^2, ($X \sim N(\mu, \sigma^2)$), entonces, X es estable con exponente característico $\alpha = 2$, de modo que*

$$AX_1 + BX_2 \sim N((A + B)\mu, (A^2 + B^2)\sigma^2)$$

por ejemplo, la ecuación (1.1) se cumple para variables Gaussianas con $C = (A^2 + B^2)^{1/2}$ y $D = (A + B - C)\mu$.

El ejemplo anterior muestra un resultado de gran interés. La distribución Gaussiana es una distribución α-estable con exponente característico $\alpha = 2$.

Definición 3. *Una variable aleatoria X tiene distribución estable si para cada $n \geq 2$, existe un número positivo C_n y un número real D_n, tal que*

$$X_1 + X_2 + ... + X_n \stackrel{d}{=} C_n X + D_n \tag{1.2}$$

donde $X_1, X_2, ..., X_n$ representan copias independientes de X.

Definición. *Una variable aleatoria X tiene distribución estable si existe una secuencia de variables aleatorias $Y_1, Y_2, ...$ y una secuencia de números positivos d_n y números reales a_n, tal que*

$$\frac{Y_1 + Y_2 + ... + Y_n}{d_n} + a_n \stackrel{d}{\Rightarrow} X \tag{1.3}$$

donde $\stackrel{d}{\Rightarrow}$ representa convergencia en distribución.

Esta definición no es más que el Teorema del Límite Central Generalizado, el cual establece que la distribución de la suma de variables aleatorias tiende a una distribución estable cuando la cantidad de variables es muy grande.

Para la definición anterior, es suficiente que las variables que se suman sean independientes e idénticamente distribuidas. Si además, dichas variables tienen media y varianza finitas, obtenemos el Teorema del Límite Central y la variable X es, por lo tanto, una variable con distribución Gaussiana. Por lo tanto esta definición, equivalente a la definición anterior, presenta de nuevo a la distribución α-estable como una generalización de la distribución Gaussiana. Además, una de las principales justificaciones teóricas para usar la distribución Gaussiana (el Teorema del Límite Central) también se cumple, en su versión generalizada, para distribuciones α-estables.

Definición 4. *Una variable X tiene distribución α-estable si tiene la siguiente función característica:*

$$\varphi(\omega) = \begin{cases} \exp\{-|\gamma\omega|^{\alpha}\left[1 - isign(\omega)\beta\tan(\frac{\pi\alpha}{2})\right] + i\mu\omega\}, (\alpha \neq 1) \\ \exp\{-|\gamma\omega|\left[1 + isign(\omega)\beta\frac{2}{\pi}\log(|\omega|)\right] + i\mu\omega\}, (\alpha = 1) \end{cases} \tag{1.4}$$

donde $\alpha \in (0, 2]$ es el exponente característico. Este parámetro controla el grado de impulsividad de la variable aleatoria X. $\beta \in [-1, +1]$ es un parámetro que controla la simetría de la distribución. ($\beta = 0$, para la distribución α-estable simétrica, $\beta = 1$ y $\beta = -1$ para la familia de distribuciones α-estable positiva y negativa respectivamente). $\gamma > 0$ es un parámetro de escala, también denominado dispersión. μ es el parámetro posición.

En lo sucesivo, escribiremos la distribución α-estable como función de sus cuatro parámetros usando la siguiente notación:

$$f_{\alpha,\beta}(\cdot|\gamma, \mu) \tag{1.5}$$

Nótese que, si en la expresión para la función característica (1.4) hacemos $\alpha = 2$, el parámetro β pierde significado, ya que $\beta \tan \pi = 0$. En ese caso, la función característica queda como:

$$\varphi(\omega) = \exp\{-|\gamma\omega|^2 + i\mu\omega\} \tag{1.6}$$

La expresión anterior es la función característica de una variable aleatoria Gaussiana con media μ y varianza $\sigma^2 = 2\mu^2$. Por lo que, a partir de la definición anterior, también se demuestra que la distribución Normal es un caso particular de distribución α-estable.

1.1.2. Función densidad de probabilidad

A pesar de que las propiedades de la distribución α-estable apuntan a que su uso está justificado en la misma medida que el de las distribuciones Gaussianas y no sólo eso, sino que la distribución Gaussiana es un caso particular de la distribución estable y por lo tanto el rango de aplicación de las distribuciones α-estables es aún más amplio que el de la distribución Normal, el uso de ésta no está tan extendido como cabría esperar. Esto es debido, sobre todo, a que la función densidad de probabilidad α-estable existe y es continua, pero salvo unas cuantas excepciones, no puede expresarse de manera compacta. Dicho de otro modo, la integral respecto a ω de la función característica (1.4) sólo tiene solución analítica para los casos que se describen a continuación:

- Como apuntamos anteriormente, una distribución α-estable con parámetros $f_{2,0}(\cdot|\gamma,\mu)$ corresponde a una Gaussiana con media μ y varianza $2\gamma^2$.

$$f_{2,0}(x|\gamma,\mu) = N(x|\mu,2\gamma^2) = \frac{1}{2\gamma\sqrt{\pi}}\exp\{\frac{(x-\mu)^2}{4\gamma^2}\} \tag{1.7}$$

- $f_{1,0}(\cdot|\gamma,\mu)$ es una distribución de Cauchy con densidad

$$f_{1,0}(x|\gamma,\mu) = p_{Cauchy}(x|\mu,\gamma) = \frac{\gamma}{\pi((x-\mu)^2+\gamma^2)} \tag{1.8}$$

- $f_{1/2,1}(\gamma,\mu)$ es una distribución de Lévy con densidad

$$f_{1/2,1}(x|\gamma,\mu) = p_{Levy}(x|\mu,\gamma) = \Big(\frac{\gamma}{2\pi}\Big)^{1/2}\frac{1}{(x-\mu)^{3/2}}\exp\{-\frac{\gamma}{2(x-\mu)}\} \tag{1.9}$$

1.1.3. Propiedad del producto

El siguiente teorema se cumple para las distribuciones α-estables simétricas ($\beta = 0$) [Samorodnitsky & Taqqu, 1994]:

Teorema. *Sean X e $Y > 0$ dos variables aleatorias independientes con $X \sim f_{\alpha_1,0}(X|\gamma,0)$ y $Y \sim f_{\alpha_2,1}((\cos\frac{\pi\alpha_2}{2})^{\frac{1}{\alpha_2}},0)$. Entonces $Z = X\cdot Y^{1/\alpha_1}$ es α-estable con parámetros $Z \sim f_{\alpha_1\cdot\alpha_2,0}(\gamma,0)$.*

Si, en el teorema anterior, particularizamos al caso $\alpha_1 = 2$ (Gaussiana) y $\alpha_2 < 1$, puede usarse la propiedad del producto para escribir una expresión

compacta para las distribuciones estables simétricas, de modo que si v_i es una muestra i.i.d. extraída de una distribución simétrica α-estable con posición ($\mu = 0$) y dispersión γ:

$$\frac{v_i}{\gamma} \sim f_{\alpha,0}(1,0) \tag{1.10}$$

Aplicando la propiedad del producto de muestras aleatorias α-estables expresada en el anterior teorema, obtenemos que:

$$v_i \sim \mathcal{N}(0, \lambda_i \gamma^2) \tag{1.11}$$

$$\lambda_i \sim f_{\frac{\alpha}{2},1}(\lambda|2\left(\cos\frac{\pi\alpha_2}{2}\right)^{\frac{1}{\alpha_2}}, 0) \tag{1.12}$$

donde $\mathcal{N}(0, \lambda_i\gamma^2)$ representa la distribución Normal de media $\mu = 0$ y varianza $\sigma^2 = \lambda_i\gamma^2$. Este modelo equivalente es muy útil para realizar inferencia Bayesiana de parámetros, ya que, aunque la distribución α-estable no tiene expresión analítica, aplicando convenientemente la propiedad del producto v_i tiene distribución Gaussiana siempre que la variable λ_i tenga distribución estable con parámetro de asimetría $\beta = 1$.

La propiedad del producto para distribuciones α-estables es un caso particular de una propiedad estadística más general que relaciona la distribución Normal con varias familias de distribuciones más impulsivas que la Normal. Esta propiedad se conoce, en su caso más general como Mezcla Escalada de Gaussianas (*Scale Mixture of Normals*). Más información sobre dicha propiedad en [Fernandez & Steel, 2000].

1.1.4. Variable aleatoria α-estable

Obtener un vector de muestras aleatorias con distribución α-estable es posible gracias al método de Chambers-Mallows-Stuck [Chambers *et al.*, 1976]. Una variable aleatoria X con distribución $f_{\alpha,\beta}(X|1,0)$ puede generarse a partir de una transformación no lineal de dos variables aleatorias independientes, una uniforme V y otra exponencial W, usando el siguiente teorema:

Teorema. *Sea V una variable aleatoria uniforme en el intervalo $(-\frac{\pi}{2}, \frac{\pi}{2})$ y W una variable aleatoria exponencial con media 1. Si V y W son independientes, entonces*

$$X = S_{\alpha,\beta}\frac{\sin(\alpha(V + B_{\alpha,\beta}))}{(\cos V)^{1/\alpha}}\left(\frac{\cos((1-\alpha)V + B_{\alpha,\beta}\alpha)}{W}\right)^{(1-\alpha)/\alpha} \tag{1.13}$$

tiene distribución α-estable con $f_{\alpha,\beta}(X|1,0)$ donde

$$B_{\alpha,\beta} = \frac{\arctan(\beta \tan(\pi\alpha/2))}{\alpha} \tag{1.14}$$

$$S_{\alpha,\beta} = (1 + \beta^2 \tan^2 \frac{\pi\alpha}{2})^{\frac{1}{2\alpha}}. \tag{1.15}$$

Una vez obtenida la variable X, es sencillo generar una variable con distribución α-estable para cualquier valor de los parámetros α, β, σ y μ; ya que si $X \sim f_{\alpha,\beta}(X|1,0)$ entonces

$$f_{\alpha,\beta}(\gamma,\mu) \sim \gamma X + \mu \qquad \text{si } \alpha \neq 1 \tag{1.16}$$

y

$$f_{\alpha,\beta}(\gamma,\mu) \sim \gamma X + \frac{2}{\pi}\beta\gamma \ln\gamma + \mu \qquad \text{si } \alpha = 1 \tag{1.17}$$

1.1.5. Comportamiento asintótico

Si $\alpha < 2$, las probabilidades en las colas de la distribución $\{P < -\lambda\}$ y $\{P > \lambda\}$ cuando $\lambda \longrightarrow$ inf, decaen siguiendo una ley de potencias $\lambda^{-\alpha}$. Este comportamiento es conocido como ley-Pareto en la literatura. En concreto, si X es una variable aleatoria con distribución α-estable con exponente característico $\alpha < 2$, entonces [Samorodnitsky & Taqqu, 1994]:

$$\lim_{\lambda\to-\infty} \lambda^\alpha \{P < \lambda\} = C_\alpha \frac{1+\beta}{2}\gamma^\alpha \tag{1.18}$$

$$\lim_{\lambda\to-\infty} \lambda^\alpha \{P < -\lambda\} = C_\alpha \frac{1-\beta}{2}\gamma^\alpha \tag{1.19}$$

donde

$$C_\alpha = \frac{1-\alpha}{\Gamma(2-\alpha)\cos(\pi\alpha/2)} \text{ si } \alpha \neq 1 \tag{1.20}$$

$$C_\alpha = \frac{2}{\pi} \text{ si } \alpha = 1 \tag{1.21}$$

1.1.6. Gráficas

En esta sección, vamos a representar gráficamente algunas distribuciones α-estables para distintos valores de sus parámetros. Para ello hemos mantenido como distribución de referencia aquélla con parámetros $\alpha = 1{,}5$, $\beta = 0$, $\gamma = 1$ y $\mu = 0$. Para cuatro distintos escenarios, en los que mantenemos constante los valores de todos los parámetros excepto el de uno de ellos que hemos evaluado para tres valores diferentes. El resultado se muestra en la Figura 1.1. Donde, de manera gráfica, puede intuirse el porqué del nombre recibido por cada uno de los parámetros. El parámetro α regula el grado de impulsividad de la distribución, conforme menor es α, mayor es el grado de impulsividad de la distribución. β regula la asimetría y el signo de dicho parámetro, la orientación de la asimetría. γ es la dispersión y regula la concentración de la distribución alrededor de un valor determinado. Valores de γ más bajos se corresponden con una mayor concentración de la distribución estable. Por último, distintos valores de μ producen la misma pdf con posición desplazada en el eje x.

1.2. Inferencia Bayesiana

El teorema de Bayes, propuesto por Thomas Bayes (1702-1761) y publicado tras su muerte en 1763, permite obtener la distribución de probabilidad condicional de una variable aleatoria A dada B en términos de la distribución de probabilidad condicional de la variable B dada A y la distribución de probabilidad marginal de la variable A [Gelman *et al.*, 1995]. Este teorema relaciona las distribuciones marginales y condicionadas de dos eventos A y B del siguiente modo:

Teorema de Bayes.

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \tag{1.22}$$

Donde $p(A)$ es la distribución a priori o probabilidad marginal de A. Se denomina a priori ya que no lleva información sobre B. $p(A|B)$ es la probabilidad condicional de A, conocido B. Se denomina también probabilidad a posteriori. $p(B|A)$ es la probabilidad condicional de B dado A. $p(B)$ es la distribución a priori de B que en el Teorema de Bayes actúa como una constante.

El teorema de Bayes ha atraído la atención de gran cantidad de investigadores de distintas disciplinas en la última década. La teoría Bayesiana considera que las variables tanto conocidas como desconocidas, están modeladas mediante

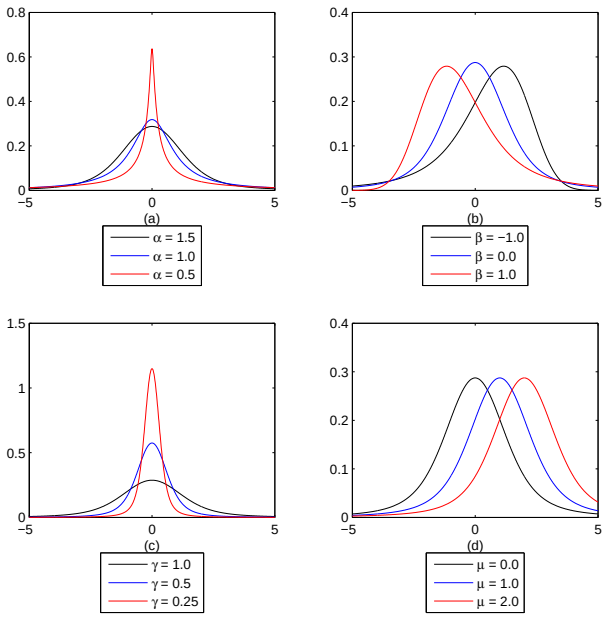

Figura 1.1: Densidad α-estable con parámetros de referencia $\alpha = 1{,}5$, $\beta = 0$, $\gamma = 1$ y $\mu = 0$. (a) Exponente característico α variable. (b) Parámetro de asimetría β variable. (c) Dispersión γ variable. (d) Posición μ variable.

distribuciones de probabilidad. Bajo este enfoque, el uso del Teorema de Bayes nos permite hacer inferencia de parámetros usando conjuntamente las variables cuyos valores queremos inferir y los datos. Por tanto, para aplicar el Teorema de Bayes al problema de inferencia de parámetros, podemos hacer una distinción de las variables en dos distintos tipos. Por un lado, las variables conocidas (los datos experimentales, la señal estudiada, etc.) y por otro lado, las variables desconocidas, es decir, aquéllas cuyos valores queremos inferir mediante la aplicación del Teorema de Bayes.

Ilustraremos mediante un sencillo ejemplo cómo realizar inferencia Bayesiana de parámetros.

Ejemplo. *Supongamos que estamos analizando una señal con $N = 1000$ muestras y_i i.i.d. con distribución Gaussiana con varianza σ_G y media μ_G desco-*

nocidas. Es sencillo realizar la estimación de la media y varianza de los datos y usando el teorema de Bayes sin más que considerar que, en dicha expresión A denota las variables conocidas, es decir, $A = y$, y B denota las variables desconocidas, $B = \{\mu_G, \sigma_G\}$. Sustituyendo en la expresión (1.22) *obtenemos que:*

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{1.23}$$

Por otro lado, $p(y)$ sólo actúa como constante de normalización, por lo que podemos reescribir (1.23) *como*

$$p(\theta|y) \propto p(y|\theta)p(\theta) \tag{1.24}$$

de donde $p(\theta) = p(\mu_G, \sigma_G)$ es la distribución a priori de las variables desconocidas y $p(y|\theta)$ es la función verosimilitud. En este caso en que suponemos un modelo Gaussiano para los datos la verosimilitud es $p(y|\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_G} \exp(\frac{(y_i-\mu_G)^2}{2\sigma_G^2})$

Una vez obtenida la distribución a posteriori de las variables θ es posible obtener valores esperados de cualquier función $h(\theta)$ como

$$E_Y[h(\theta)] = \int h(\theta)p(\theta|y)d\theta \tag{1.25}$$

Por ejemplo, el valor esperado para la media μ_G y la varianza σ_G en el ejemplo anterior es, respectivamente

$$E_Y[\mu_G] = \int \mu_G p(\theta|y)d\theta \tag{1.26}$$

$$E_Y[\sigma_G] = \int \sigma_G p(\theta|y)d\theta \tag{1.27}$$

El principal problema en la aplicación de los métodos Bayesianos a problemas de estimación, es que en la mayoría de los casos, las integrales involucradas no tienen solución analítica. Sin embargo, si somos capaces de extraer N_{iter} muestras con distribución $p(\theta|y)$

$$\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N_{iter})} \sim p(\theta|y) \tag{1.28}$$

entonces podemos estimar la integral en (1.25), mediante integración Monte Carlo, del siguiente modo

$$E_Y[h(X)] \approx \frac{1}{N} \sum_{i=1}^{N_{iter}} h(\theta^{(i)}) \tag{1.29}$$

Por este motivo, precisamos un método general de extracción de muestras con distribución cualquiera, para así estimar numéricamente las integrales involucradas. Los métodos Monte Carlo basados en cadenas de Markov (MCMC) tales como el algoritmo de Metropolis-Hasting o el algoritmo de Gibbs, son las herramientas mediante las cuales obtener muestras distribuidas con una distribución cualquiera.

1.2.1. Métodos Monte Carlo basados en cadenas de Markov

Los métodos Monte Carlo basados en cadenas de Markov nos permiten diseñar una cadena de Markov cuya distribución estacionaria sea una distribución $f(x)$ cualquiera. Para asegurar que la convergencia a la distribución estacionaria es única, es suficiente que la cadena de Markov cumpla las condiciones de *Irreducibilidad*, es decir, cada estado del sistema pueda ser alcanzado independientemente del estado inicial y la de *aperiodicidad*, la cadena no queda atrapada en un ciclo. Si se cumplen estas dos condiciones, el proceso es ergódico (para una información más detallada véase [Robert & Casella, 1999; Gilks *et al.*, 1996]). De este modo, una vez obtenidas N muestras $X^{(i=1:N_{iter})}$, es posible aproximar el valor de la integral

$$\int h(x)f(x)dx \tag{1.30}$$

por el obtenido mediante la siguiente sumatoria

$$\frac{1}{N_{iter}} \sum_{i=1}^{N_{iter}} h(X^{(i)}) \tag{1.31}$$

Definición. *Un método Monte Carlo basado en cadenas de Markov (MCMC por sus siglas en inglés) para la simulación de una distribución f cualquiera, es cualquier método que produce una cadena de Markov ergódica $X^{(i)}$ cuya distribución estacionaria es f.*

Algoritmo de Metropolis-Hastings

A continuación mostramos la implementación del algoritmo de Metropolis-Hastings [Hastings, 1970].

Donde la distribución $q(y|x^{(t)})$ recibe el nombre de distribución propuesta (*proposal distribution*) y $A(x^{(t)}, y)$ es la tasa de aceptación.

Algoritmo de Metropolis-Hastings En la iteración t

1. Generamos un nuevo valor y con distribución $y \sim q(y|x^{(t)})$.

2. Con probabilidad $A(x^{(t)}, y) = \min\{1, \frac{f(y)}{f(x^{(t)})} \frac{q(x^{(t)}|y)}{q(y|x^{(t)})}\}$
$x^{(t+1)} = y$ (aceptamos el valor propuesto)
y con probabilidad $1 - A(x^{(t)}, y)$

$x^{(t+1)} = x^{(t)}$ (rechazamos el valor propuesto)

3. $t \longrightarrow t + 1$

Algoritmo de Metropolis

Cronológicamente, el algoritmo de Metropolis [Metropolis *et al.*, 1953] es el precedente del algoritmo desarrollado por [Hastings, 1970] y denominado algoritmo de Metropolis-Hastings. La diferencia esencial entre ambos algoritmos es que en el algoritmo de metropolis la distribución propuesta es simétrica, por lo que $q(y|x^{(t)}) = q(x^{(t)}|y)$ y por lo tanto la tasa de aceptación se simplifica del siguiente modo: $A(x^{(t)}, y) = \min\{1, \frac{f(y)}{f(x^{(t)})}\}$

Algoritmo de Gibbs

Tanto el algoritmo de Metropolis como el de Metropolis-Hastings, son dos algoritmos MCMC que se pueden considerar genéricos, ya que no requieren introducir mucha información sobre el problema estudiado en particular. De hecho lo única información que usamos es la expresión analítica de la distribución f.

El muestreo de Gibbs es un caso especial de algoritmo de Metropolis-Hastings. Este algoritmo fue propuesto en [Geman & Geman, 1984]. El muestreo de Gibbs tiene gran interés cuando no se conoce la expresión analítica de la distribución conjunta de las variables de interés en nuestro problema de estimación Bayesiana, pero se conoce la distribución condicional para cada una de las variables. Este método Monte Carlo nos permite obtener muestras de la distribución de cada una de las variables, una a una, condicionada en los valores actuales del resto de variables. Las muestras aleatorias así obtenidas forman una cadena de Markov cuya distribución estacionaria es la distribución conjunta de las variables.

La clave del algoritmo de Gibbs es considerar como distribución propuesta la distribución condicional a posteriori de cada variable $q(y|x^{(t)}) = f(y|x^{(t)})$. Sustituyendo esta distribución en la expresión general de la tasa de aceptación del algoritmo de Metropolis-Hastings obtenemos $A(x^{(t)}, y) = \min\{1, \frac{f(y)}{f(x^{(t)})} \frac{q(x^{(t)}|y)}{q(y|x^{(t)})}\} = 1$, y por lo tanto, todos los valores propuestos son automáticamente aceptados. De este modo, se optimiza la convergencia del algoritmo.

Algoritmo de Gibbs
En la iteración t, dado $x^{(t)} = (x_1^{(t)}, ..., x_p^{(t)})$

1. $X_1^{(t+1)} \sim f_1(x_1|x_2^{(t)}, ..., x_p^{(t)})$

2. $X_2^{(t+1)} \sim f_2(x_2|x_1^{(t+1)}, x_3^{(t)}, ..., x_p^{(t)})$

$\vdots$

p. $X_p^{(t+1)} \sim f_p(x_p|x_1^{(t+1)}, ..., x_{p-1}^{(t+1)})$

Para más información sobre la implementación de este algoritmo, distintas variantes y sus propiedades, véase [Robert & Casella, 1999].

1.2.2. Método de Monte Carlo basado en cadenas de Markov con saltos reversibles

Para la estimación del número de componentes k de la mezcla es necesario el uso de métodos Monte Carlo de dimensión variable, ya que para diferentes valores de k, el número de parámetros desconocidos en el problema varía. Para un completo análisis del método de Monte Carlo basado en cadenas de Markov con saltos reversibles (*RJMCMC*) véase [Green, 1995].

Supongamos que proponemos un movimiento transdimensional, que denotaremos como m, de un estado con dimensión x a un nuevo estado de dimensión mayor x'. Este *salto* puede realizarse mediante la construcción de una biyección entre ambos espacios. Debido al hecho de que x y x' tienen distinta dimensión, existen $\dim(x') - \dim(x)$ grados de libertad para construir dicha biyección.

En [Green, 1995] se propone la introducción de $\dim(x') - \dim(x)$ variables aleatorias u. De este modo es posible saltar entre los espacios de diferente dimensión. En dicho trabajo, se extrae de una densidad $q(u)$, independiente de x,

un vector de variables aleatorias continuas u. Los nuevos valores x' se proponen usando una función invertible determinista $x'(x,u)$. Esta transformación de las variables $x \to x'$, se tiene en cuenta en la expresión para la tasa de aceptación por medio de la densidad $q(u)$ y el Jacobiano de la transformación. La probabilidad de aceptación A, para el método de Monte Carlo basado en cadenas de Markov con saltos reversibles, queda del siguiente modo:

$$A = \text{mín}\left\{1, \frac{p(x'|y)r_m(x')}{p(x|y)r_m(x)q(u)}\left|\frac{\partial x'}{\partial(x,u)}\right|\right\} \tag{1.32}$$

donde $r_m(x')$ es la probabilidad de elegir el movimiento de tipo m cuando el estado actual es x y $|\cdot|$ es el Jacobiano de la transformación.

En [Richardson & Green, 1997] se estudia la aplicación de *RJMCMC* a la mezcla de Gaussianas. Además, se estima el número de componentes Gaussianos satisfactoriamente. En dicho trabajo, se sugieren dos movimientos transdimensionales de nombre *birth-death* para componentes vacíos y *split-combine* para los no vacíos. Se considera a un componente j como vacío, si su correspondiente variable asignación es $z_j = 0$.

1.3. Modelos de mezcla

En la actualidad, es posible describir, estimar, predecir y hacer inferencia de parámetros en sistemas complejos gracias a, por un lado el desarrollo de nuevos métodos numéricos y el incremento de la potencia computacional y por otro, la aparición de nuevos métodos de modelado de sistemas complejos mediante mezcla de distribuciones.

El modelado mediante mezcla de distribuciones es una teoría muy útil para la descripción de sistemas multimodales. Este método incluye distintas variantes, como considerar la mezcla de distribuciones finitas o infinitas, de la misma distribución o mezcla de distintas distribuciones, etc. La potente descripción de los datos y señales conseguida con el modelado mediante mezclas ha hecho que este método sea aplicado en diferentes disciplinas, tales como en astronomía, ecología, bioinformática, ciencias de la computación, ecología, economía, ingeniería, robótica y bioestadística entre otras (véase [McLachlan & Peel, 2000] para un estudio detallado de dichos modelos).

1.3.1. Definición

El modelado mediante mezcla de distribuciones se describe, matemáticamente, mediante una suma de funciones densidad de probabilidad del siguiente modo:

$$p_Y(y) = \sum_{j=1}^{k} w_j p(y|\theta_j) \tag{1.33}$$

$$0 \leq w_j \leq 1(\forall j) \;\; ; \;\; \sum_{j=1}^{k} w_j = 1$$

donde w_j es la proporción del componente j en la mezcla y $p_Y(y)$ es la función densidad de probabilidad de la señal y de longitud N. $p(y|\theta_j)$ representa una distribución cualquiera con variable y y n parámetros que denotamos genéricamente como $\theta = [\theta_1, ..., \theta_k]$. Además, en lo sucesivo, por simplicidad en la notación, agruparemos las variables del siguiente modo: $w = [w_1, ..., w_k, ..., w_k]$, $y = [y_1, ..., y_i, ..., y_N]$ y $z = [z_1, ..., z_i, ..., z_N]$.

En la literatura, este modelo general ha sido estudiado para diversas distribuciones $p(y|\theta_j)$. En particular, aunque la mezcla de Gaussianas es el modelo más extendido, éste no es el único. Han sido realizados trabajos en los que la distribución $p(y|\theta_j)$ toma diversas formas: Gamma [M. Wiper & Ruggieri, 2001], Weibull [Tsionas, 2002], Poisson [Fernandez & Green, 2002] o t-student [McLachlan & Peel, 1998].

Para realizar inferencia de parámetros en este modelo es conveniente considerar que la señal observada y es un vector aleatorio extraído de cada una de las k subpoblaciones (distribuciones $p(y|\theta_j)$) etiquetadas como $j = 1, 2, ...k$. Por este motivo introducimos una nueva variable $z_i \in [1, 2, ..k]$ que asigne para cada una de las muestras i, a cuál de los componentes es más plausible que pertenezca. Condicionado a esta nueva variable z_i, denominada en lo sucesivo *variable asignación*, el modelo de mezcla puede reescribirse como:

$$y_i|z_i \sim p(y_i|\theta_{z_i}, w_{z_i}) \tag{1.34}$$

1.3.2. Enfoque Bayesiano

Los dos métodos más extendidos para la resolución de modelos de mezcla son el que se basa en el algoritmo de maximización de la expectación, EM (*expectation maximization algorithm* [Dempster *et al.*, 1977]) y el que resuelve la estimación de los parámetros θ, k, w en la expresión (1.33) mediante inferencia

Bayesiana. Este segundo enfoque se ha popularizado más en los últimos años por varios motivos. Por una parte el desarrollo de técnicas computacionales basadas en métodos Monte Carlo para resolver las integrales involucradas en el modelo Bayesiano, y por otro lado, porque a día de hoy sólo un planteamiento Bayesiano del problema permite realizar inferencia en el número de subpoblaciones k que componen la mezcla. Dotando a los modelos de mezcla de una flexibilidad adicional, puesto que pueden ser resueltos sin hacer suposiciones a priori sobre el número de componentes de dicha mezcla.

Modelo jerárquico

Bajo el paradigma Bayesiano, consideramos que las variables desconocidas k, w, θ son variables aleatorias extraídas de distribuciones a priori apropiadas. La distribución conjunta de todas las variables se puede escribir, en general, para el modelo de mezcla como

$$p(k, w, z, \theta, y) = p(y|\theta, z, w, k)p(\theta|z, w, k)p(z|w, k)p(w|k)p(k). \tag{1.35}$$

Es posible, teniendo en cuenta las dependencias entre los parámetros del modelo de mezcla $p(\theta|z, w, k) = p(\theta|k)$ y $p(y|\theta, z, w, k) = p(y|\theta, z)$, reescribir la expresión (1.35) del siguiente modo

$$p(k, w, z, \theta, y) = p(y|\theta, z)p(\theta|k)p(z|w, k)p(w|k)p(k). \tag{1.36}$$

Una vez planteada la distribución conjunta de los parámetros desconocidos y el vector observación en nuestro modelo, es conveniente, con el fin de introducir mayor flexibilidad en el modelo jerárquico, permitir que las distribuciones a priori de k,w y θ, dependan de unos hiperparámetros λ,δ y η respectivamente. A su vez, estos hiperparámetros son modelados independientemente mediante hiperdistribuciones a priori. Así, se reescribe la expresión para la distribución conjunta (1.36) como

$$p(\lambda, \delta, \eta, k, w, z, \theta, y) = p(y|\theta, z)p(\theta|k, \eta)p(z|w, k)p(w|k, \delta)p(k|\lambda)p(\eta)p(\delta)p(\lambda). \tag{1.37}$$

1.3.3. Mezcla de Gaussianas

El modelo de mezcla más ampliamente estudiado y que además, ha demostrado un buen funcionamiento en multitud de aplicaciones prácticas es el modelo de mezcla de Gaussianas (véase [McLachlan & Peel, 2000]). Para dicho modelo,

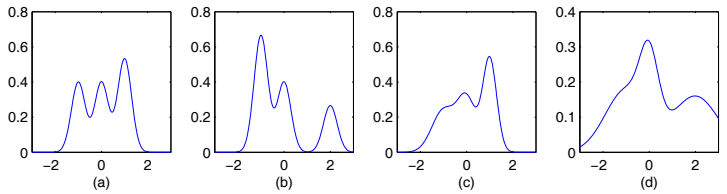

Figura 1.2: Mezcla de 3 distribuciones Gaussianas con parámetros $w = [w_1\ w_2\ w_3]$, $\mu = [\mu_1\ \mu_2\ \mu_3]$ y $\sigma = [\sigma_1\ \sigma_2\ \sigma_3]$ dados por: a) w = [0.3 0.3 0.4], μ = [-1 0 1] y σ = [0.3 0.3 0.3]. b) w = [0.5 0.3 0.2], μ = [-1 0 2] y σ = [0.3 0.3 0.3]. c) w = [0.3 0.3 0.4], μ = [-1 0 2] y σ = [0.9 0.4 1]. d) w = [0.4 0.2 0.4], μ = [-1 0 2] y σ = [0.9 0.4 0.3].

los parámetros desconocidos son los pesos w_j, las medias y las varianzas. Por lo tanto, $\theta_j = \{\mu_j, \sigma_j^2\}$ con $j = 1, 2, ..., k$. Así, el modelo de mezcla general expresado en (1.33) se reescribe como

$$p(y|\theta, z) = \sum_{j=1}^{k} w_j \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(y-\mu_j)^2}{2\sigma_j^2}\right) \tag{1.38}$$

La Figura 1.2 muestra distintas mezclas de distribuciones Gaussianas con diferentes parámetros. El modelado mediante mezcla de distribuciones trata de, a partir de un número de muestras u observaciones, hacer estimación de los parámetros de cada una de las distribuciones que componen la mezcla y el número de subpoblaciones que la componen.

Normalmente, se elige como distribución a priori de μ y σ en el modelo de mezclas Gaussiano a las distribuciones conjugadas para este modelo. Estas distribuciones conjugadas y el modelo de mezcla de distribuciones Normales Bayesiano, fue propuesto, para problemas de dimensión fija en [Diebolt & Robert, 1995].

Usaremos el método de Gibbs para obtener, en cada iteración, las estimaciones de cada parámetro. Por tanto, es necesario calcular las expresiones de las probabilidades a posteriori para cada uno de los parámetros.

Recordemos que el método de Gibbs es un método Monte Carlo basado en cadenas de Markov que propone, en cada iteración, como nuevos valores de los parámetros, variables aleatorias extraídas de la distribución a posteriori de las variables (véase la Sección 1.2.1). En las siguientes subsecciones, se detalla el cálculo de dichas distribuciones.

Actualización de los pesos (w)

En problemas de mezcla de distribuciones, la distribución a priori conjugada para los pesos w es la distribución Dirichlet simétrica $\mathcal{D}$:

$$\omega \sim \mathcal{D}(\zeta, ..., \zeta). \tag{1.39}$$

Por lo tanto, la distribución a posteriori para los pesos w es también Dirichlet, y usando el algoritmo de Gibbs, podemos obtener en cada iteración los nuevos valores de los pesos extrayendo muestras de dicha distribución con los siguientes parámetros:

$$\omega \mid ... \sim \mathcal{D}(\zeta + n_1, ..., \zeta + n_k) \tag{1.40}$$

donde n_j es el número de muestras del vector de observación y_i asignadas al componente j, es decir, el número de veces que el cálculo de la variable de asignación da como resultado $z_i = j$.

Actualización de la media (μ_j)

La distribución a priori conjugada para la media de la distribución Normal, es también una distribución Normal con media ξ y varianza κ^{-1}, donde ξ y κ son los hiperparámetros de la distribución a priori.

$$\mu_j \sim \mathcal{N}(\xi, \kappa^{-1}) \tag{1.41}$$

De este modo, la distribución a posteriori es también Gaussiana con los siguientes parámetros:

$$\mu_j \mid ... \sim \mathcal{N}\left(\frac{\sigma^{-2} \sum_{i=1:z_i=j}^{N} y_i + \kappa\xi}{(\sigma_j^{-2} n_j + \kappa)^{-1}}, (\sigma_j^{-2} n_j + \kappa)^{-1} \right) \tag{1.42}$$

Actualización de la varianza (σ_j)

Tal y como hemos hecho para la media, elegimos como distribución a priori la distribución conjugada para la varianza. En este caso es la distribución Gamma Inversa con hiperparámetros α_0,β_0.

$$\sigma_j^2 \sim \mathcal{IG}(\alpha_0, \beta_0). \tag{1.43}$$

Por lo tanto, en cada iteración, los nuevos valores de la varianza se obtienen extrayendo muestras de la distribución a posteriori que, puesto que la distribución a priori conjugada es $\mathcal{IG}$, ésta es también Gamma Inversa y tiene los siguientes parámetros:

$$\sigma_j^2 \mid ... \sim \mathcal{IG}(\alpha_0 + \frac{1}{2}n_j, \frac{1}{2}\sum_{i=1:z_i=j}^{N}(y_i - \mu_j)^2 + \beta_0) \tag{1.44}$$

donde n_j es el número de datos del vector de observación que pertenecen al componente j según lo calculado por la variable asignación z_i.

Actualización de la variable asignación (z_i)

La distribución conjugada a priori de la variable que asigna cada dato del vector observación con un componente j de la mezcla es:

$$p(z_i = j) = w_j. \tag{1.45}$$

La distribución a posteriori de dicha variable se obtiene fácilmente multiplicando la distribución a priori por la función verosimilitud (1.38), con lo que obtenemos:

$$p(z_i = j \mid ...) \propto \frac{w_j}{\sigma_j} \exp\left(-\frac{(y_i - \mu_j)^2}{2\sigma_j^2}\right). \tag{1.46}$$

El modelo hasta ahora descrito es invariante respecto a permutaciones del índice de los componentes $j = 1, 2, ..., k$. Para identificar unívocamente cada una de las subpoblaciones que componen la mezcla adoptamos el siguiente criterio, aunque hay que señalar que no es el único posible. Ordenamos cada uno de los componentes con el siguiente criterio $\mu_1 < ... < \mu_j < ... < \mu_k$. Es decir, ordenamos las medias estimadas por nuestro algoritmo en orden creciente. Existen otras alternativas a la ordenación creciente de los componentes [Celeux *et al.*, 2000; Stephens, 2000], aunque aumentan la carga computacional del algoritmo en gran medida y además, tal y como se verá en los capítulos posteriores, con la ordenación de los componentes según el orden creciente en sus medias obtenemos resultados más que satisfactorios.

CAPÍTULO

DOS

INFERENCIA BAYESIANA EN MEZCLA DE DISTRIBUCIONES α-ESTABLE SIMÉTRICAS

El modelo de mezcla de distribuciones simétricas estables para un vector observación y_i es el siguiente:

$$y_i \sim \sum_{j=1}^{k} w_j f_{\alpha_j,0}(\mu_j, \gamma_j) \tag{2.1}$$

donde k es el número de componentes de dicha mezcla, y w_j, μ_j y γ_j son los pesos, posición y dispersión de cada componente j, respectivamente. α_j es el exponente característico de la distribución estable que, al contrario del modelo propuesto en [Salas-Gonzalez *et al.*, 2006c], puede tomar distintos valores para cada una de las distribuciones que componen la mezcla. Usaremos, tal y como apuntamos en la Sección 1.3.3, para cada observación i, una variable $z_i \in [1, 2, ..., k]$ que asigne para cada una de las muestras i, a qué componente o subpoblación de la mezcla es más probable que pertenezcan. Condicionado a esta variable asignación, el modelo de mezcla puede reescribirse como:

$$y_i \sim f_{\alpha_{z_i},0}(\mu_{z_i}, \gamma_{z_i}) \tag{2.2}$$

Para cada observación y_i, la distribución α-estable simétrica puede expresarse, condicionada a la variable aleatoria λ_i, como una Gaussiana usando la propiedad del producto (véase la Sección 1.1.3). Por lo tanto, la verosimilitud

de este modelo puede escribirse:

$$p(\, y_i |\, \mu_j, \gamma_j^2, \lambda_i, z_i = j) = \frac{1}{\sqrt{2\pi\lambda_i}\gamma_j} \exp\left\{ -\frac{(y_i - \mu_j)^2}{2\lambda_i \gamma_j^2} \right\} \tag{2.3}$$

Como ya apuntábamos en la Sección 1.1, es interesante remarcar que la principal dificultad para realizar estimación Bayesiana con distribuciones α-estables es la no existencia de expresión analítica para su pdf.Este problema es superado para distribuciones α-estables simétricas mediante la aplicación de la propiedad del producto. Además, la propiedad del producto, no sólo permite expresar la pdf de la distribución estable simétrica de manera compacta, sino que además la convierte en una distribución Gaussiana condicionada a la variable aleatoria λ. Por lo tanto, es posible plantear el problema de estimación Bayesiana en mezcla de distribuciones simétricas α-estables, problema que a priori era analíticamente intratable, usando las mismas deseables propiedades de la distribución Normal [Salas-Gonzalez *et al.*, 2007c].

Este capítulo está estructurado del siguiente modo: en la Sección 2.1, desarrollamos el modelo jerárquico Bayesiano para el problema de mezcla de distribuciones simétricas α-estables considerado. En la Sección 2.2 mostramos las distribuciones a priori elegidas para cada uno de los parámetros del modelo y en 2.3 explicamos la actualización de cada una de las variables mediante métodos Monte Carlo. Finalmente, en las secciones 2.4 y 2.5, respectivamente, presentamos los resultados y extraemos las conclusiones.

2.1. Modelo Bayesiano

La estimación de parámetros en mdelos de mezcla ha sido estudiada, principalmente, bajo dos enfoques. Por una parte, mediante el uso del algoritmo EM, y por otra, haciendo uso de métodos Bayesianos. La ventaja principal de este último método respecto al primero es que permite hacer inferencia en el número de mezclas que componen el modelo, es decir, el número de componentes k es un parámetro desconocido y, por lo tanto, puede ser estimado. En general, los métodos Bayesianos se basan en la suposición de que las variables de interés en el problema analizado están modeladas mediante distribuciones de probabilidad. De este modo, es posible realizar inferencia de parámetros usando, conjuntamente, estas distribuciones y los datos experimentales u observaciones mediante el Teorema de Bayes.

Escribimos la expresión para la distribución a posteriori de las variables del modelo de mezcla simétrico α-estable considerado usando el Teorema de Bayes,

el cual establece que

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \tag{2.4}$$

donde $p(A)$ es la distribución a priori. $p(A|B)$ es la probabilidad de A, condicionada a B. En este caso, $p(B|A)$ es la verosimilitud de B dado A y $p(B)$ es la distribución a priori de B. $p(B)$ es una constante de normalización, por lo que el teorema de Bayes (2.4) puede escribirse como:

$$p(A|B) \propto p(B|A)p(A) \tag{2.5}$$

Es posible reescribir la expresión (2.5) para nuestro modelo de mezcla teniendo en cuenta que B es el vector de datos u observaciones disponibles (en este caso $B = \{y\}$) y A son las variables desconocidas en el modelo considerado (por lo tanto A son las variables cuyos valores queremos estimar).

$$A = \{k, w, z, \alpha, \gamma, \mu\} \tag{2.6}$$

El número de variables desconocidas es, por lo tanto $4k + 1$,

$$A = \{\alpha_j, \gamma_j, \mu_j, w_j, k\}. \tag{2.7}$$

El número de variables que queremos estimar, depende del número de componentes k. Este hecho plantea una dificultad añadida al problema de inferencia Bayesiana, ya que distintos valores de k producen un cambio en la dimensión del resto de variables desconocidas. Esto imposibilita el uso de métodos Monte Carlo clásicos para la obtención de dicha variable k, como el algoritmo de Metropolis-Hasting o Muestreo de Gibbs y obliga a usar nuevos y emergentes métodos de simulación Monte Carlo para problemas de dimensión variable, tal y como veremos más adelante.

El Teorema de Bayes nos permite escribir el modelo jerárquico para las variables de nuestro problema. La distribución conjunta es:

$$p(k, w, z, \alpha, \gamma, \mu, y) = p(y|k, w, z, \alpha, \gamma, \mu)p(k, w, z, \alpha, \gamma, \mu) \tag{2.8}$$

Con el fin de incrementar la flexibilidad en este modelo, permitimos a la distribución a priori $p(k, w, z, \alpha, \gamma, \mu)$ depender de una serie de nuevos parámetros que en el contexto de la teoría de Bayes se denominan hiperparámetros. Además, para simplificar la notación agruparemos las variables desconocidas de la distribución α-estable simétrica como $\theta = \{\alpha, \gamma, \mu\}$ y a los hiperparámetros de las variables agrupadas en θ como $\eta = \{a, b, \alpha_0, \beta_0, \xi, \kappa\}$. De este modo, la

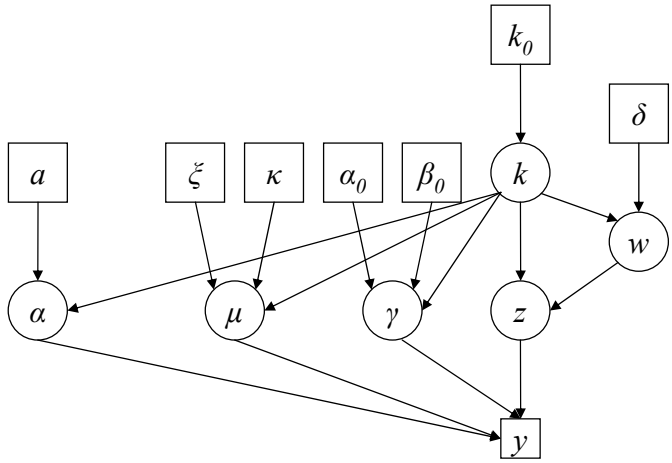

Figura 2.1: *Directed Acyclic Graph* (DAG) para el modelo Bayesiano representado por la ecuación (2.9)

expresión (2.8) puede expandirse teniendo en cuenta las dependencias condicionales para este modelo en particular que, en notación simplificada se escriben como sigue:

$$\begin{aligned} p(k,w,z,\theta,\eta,y) &= p(y|k,w,z,\theta,\eta)p(\theta|k,\eta)p(z|w,k) \\ &\times \; p(w|k,\zeta)p(k|k_0)p(k_0)p(\zeta)p(\eta) \end{aligned} \tag{2.9}$$

donde, por otra parte

$$p(\theta|k,\eta) = p(\mu|\xi,\kappa)p(\gamma|\alpha_0,\beta_0)p(\alpha|a) \tag{2.10}$$

Para una mayor claridad en la visualización de las dependencias entre variables, escribimos el gráfico DAG (*Directed Acyclic Graph*). En dicho gráfico, las variables en un rectángulo denotan parámetros fijos (hiperparámetros) o variables conocidas (datos experimentales u observaciones) mientras que las variables encerradas en un círculo representan las variables desconocidas cuyos valores queremos inferir. Las flechas entre variables muestran la dependencia entre cada una de ellas.

2.2. Distribuciones a priori

Una vez escrito el modelo Bayesiano correspondiente a la mezcla de distribuciones α-estables simétricas, es preciso elegir las distribuciones a priori para cada una de las variables.

2.2.1. Exponente característico (α)

Elegimos como la distribución a priori para el exponente α la distribución uniforme en el rango $0 < \alpha \leq 2$. Por lo tanto,

$$p(\alpha|a) = \frac{1}{a} = \frac{1}{2}; \quad \text{para} \quad 0 < \alpha \leq 2. \tag{2.11}$$

Esta elección se fundamenta en la sencillez de la distribución a priori. Los dos primeros trabajos en inferencia Bayesiana de parámetros en distribuciones α-estable [Buckle, 1995; Tsionas, 1999], también escogen la distribución uniforme entre 0 y 2 como distribución a priori para α. La distribución a priori podría elegirse de modo que incluyera la información conocida sobre el parámetro α en distribuciones α-estables. Como por ejemplo, que para valores de α diferentes pero cercanos a 2, dos distribuciones simétricas α-estables son muy parecidas entre sí. Por el contrario, para α diferentes pero no cercanos a dicho valor, pequeñas diferencias en el valor α pueden producir distribuciones α-estables con forma muy diferente, tal y como se muestra en la Figura 2.2, donde se representan gráficamente dos pares de distribuciones estables con una diferencia en el parámetro α de 0,3. En la figura, se observa que dicha diferencia ejerce una mayor influencia en la forma de la distribución cuando α es menor.

2.2.2. Dispersión (γ)

Elegimos como distribución a priori para el parámetro dispersión la distribución Gamma Inversa $\mathcal{IG}$. Para el modelo aquí considerado, del mismo modo que para el modelo de mezcla Gaussiano, esta distribución es la distribución a priori conjugada. Esta elección proporciona grandes ventajas a la hora de actualizar las variables, ya que la distribución a posteriori es también Gamma Inversa. Por lo tanto, la actualización de esta variable puede realizarse de manera sencilla mediante el algoritmo de Gibbs sin más que extraer muestras de una distribución Gamma Inversa. Esta distribución tiene 2 parámetros, α_0 y β_0. Su expresión analítica es:

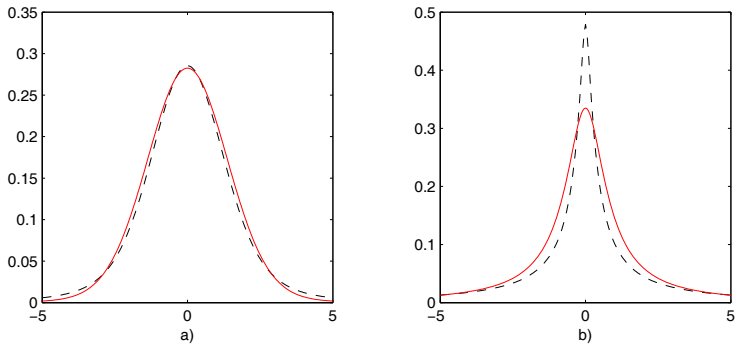

Figura 2.2: Distribución simétrica α-estable con (a) $\alpha = 1{,}6$ y $\alpha = 1{,}9$ (b) $\alpha = 0{,}6$ y $\alpha = 0{,}9$

Tabla 2.1: Dependencia de la media, moda y varianza de la distribución Gamma Inversa con los parámetros α_0 y β_0

Media	**Moda**	**Varianza**
$\frac{\beta_0}{\alpha_0-1}$	$\frac{\beta_0}{\alpha_0+1}$	$\frac{\beta_0^2}{(\alpha_0-1)^2(\alpha_0-2)}$
$\alpha > 1$		$\alpha > 2$

$$\mathcal{IG}(x|\alpha_0, \beta_0) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} x^{-\alpha_0-1} \exp\{-\frac{\beta_0}{x}\} \tag{2.12}$$

El dominio de esta distribución es $(0, \infty)$ y la forma de la distribución está controlada por los valores de sus dos parámetros α_0 y β_0. En la Tabla 2.1, se presenta la dependencia funcional de la media, moda y varianza de la distribución Gamma Inversa con los parámetros α_0 y β_0.

2.2.3. Posición (μ)

Para el parámetro desconocido μ, elegimos como distribución a priori la distribución conjugada, que para el modelo de mezcla de distribuciones α-estable simétricas considerado es la distribución Normal. Por lo tanto, considerando los hiperparámetros media (ξ) y varianza (κ^{-1}), la distribución a priori para la

variable posición es:

$$p(\mu|\xi,\kappa^{-1}) = \frac{\kappa^{1/2}}{\sqrt{2\pi}} \exp\{-\frac{(\mu-\xi)^2}{2\kappa^{-1}}\}. \tag{2.13}$$

2.2.4. Pesos (w)

La distribución a priori conjugada para los pesos en los modelos de mezcla es la distribución de Dirichlet simétrica [McLachlan & Peel, 2000].

$$\mathbf{w} \sim \mathcal{D}(\zeta, ..., \zeta) \tag{2.14}$$

la distribución de Dirichlet es una generalización para múltiples variables de la distribución Beta.

2.2.5. Número de componentes (k)

Elegimos como distribución a priori para el número de componentes de la mezcla k, la distribución discreta uniforme entre 1 y un valor k_0. Este valor de k_0 es el máximo valor que permitimos que tome nuestro algoritmo para el número de componentes k. De todos modos, tal y como veremos en los resultados de las simulaciones, elegimos k_0 suficientemente grande como para que en ninguna iteración el número de componentes k estimado sea mayor que k_0.

$$p(k|k_0) = \frac{1}{k_0} \tag{2.15}$$

Otra opción posible para la distribución a priori de la variable k es, si tenemos alguna información a priori sobre cuál podría ser, aproximadamente, el número de componentes en la mezcla, elegir como distribución a priori la distribución de Poisson con media en el valor más probable del número de componentes.

En la Tabla 2.2 se muestra un esquema de las variables, los hiperparámetros y las distribuciones a priori para el modelo de mezcla α-estable simétrico considerado.

Tabla 2.2: Sumario de las variables desconocidas en el modelo de mezcla α-estable simétrico, hiperparámetros y distribuciones a priori correspondientes.

Variable	**Nombre**	**Hiperparámetros**	**Distribución a priori**
α	exponente característico	a	$p(\alpha\|a) = \frac{1}{a}$
γ	dispersión	α_0,β_0	$\mathcal{IG}(\gamma\|\alpha_0, \beta_0)$
μ	posición	ξ,κ	$\mathcal{N}(\mu\|\xi, \kappa^{-}1)$
w	pesos	ζ	$\mathcal{D}(\zeta, ..., \zeta)$
k	número de componentes	k_0	$p(k\|k_0) = \frac{1}{k_0}$
z	asignación		$p(z_i = j) = w_j$

2.3. Actualización de las variables

En las siguientes subsecciones presentamos en detalle la resolución numérica mediante métodos Monte Carlo basados en cadenas de Markov del modelo de mezcla α-estable simétrico. El funcionamiento de este método, tal y como sucede normalmente con los algoritmos MCMC, no depende del orden de actualización de los parámetros, que podría incluso seguir un orden aleatorio en cada iteración. No obstante, nosotros seguiremos el siguiente esquema:

- Actualización de los pesos w_j usando muestreo de Gibbs.
- Actualización de la posición μ_j mediante muestreo de Gibbs.
- Actualización de la dispersión γ_j mediante muestreo de Gibbs.
- Actualización de la asignación de variables z_i mediante mustreo de Gibbs.
- Actualización de la variable aleatoria λ_i usando muestreo por rechazo.
- Actualización del número de componentes k usando métodos Monte Carlo basado en cadenas de Markov con saltos reversibles.
- Actualización del exponente característico α usando el algoritmo de Metropolis.

2.3.1. Actualización de los pesos (w_j)

La actualización de los pesos la realizamos mediante el algoritmo de Gibbs. Para ello calculamos la distribución a posteriori $p(w|\theta, y, k, z)$ y, para cada iteración, extraemos muestras de ella. La distribución a posteriori para los pesos w es también Dirichlet ya que hemos elegido como distribución a priori la distribución conjugada:

$$w \mid ... \sim \mathcal{D}(\zeta + n_1, \zeta + n_2, ..., \zeta + n_k) \tag{2.16}$$

donde n_j es el número de muestras que fueron asignadas al componente j mediante la actualización de la asignación de variables z_i.

2.3.2. Actualización del parámetro posición (μ_j)

El parámetro μ se actualiza mediante el muestreo de Gibbs. La distribución a posteriori, de la cual extraemos las muestras, es también Gaussiana con la siguiente media y varianza:

$$\mu_j \mid ... \sim \mathcal{N}\left(\frac{\frac{1}{\gamma^2}\sum_{i=1:z_i=j}^{N}\frac{y_i}{\lambda_i}+\kappa\xi}{\frac{1}{\gamma^2}\sum_{i=1:z_i=j}^{N}\frac{1}{\lambda_i}+\kappa}, \frac{1}{\frac{1}{\gamma^2}\sum_{i=1:z_i=j}^{N}\frac{1}{\lambda_i}+\kappa}\right) \tag{2.17}$$

2.3.3. Actualización de la dispersión (γ_j)

Los nuevos valores de la dispersión se obtienen extrayendo muestras de la distribución a posteriori que, al igual que la distribución a priori, es Gamma Inversa con los parámetros siguientes:

$$\gamma_j^2 \mid ... \sim \mathcal{IG}(\alpha_0 + \frac{n_j}{2}, \frac{1}{2}\sum_{i=1:z_i=j}^{N}\frac{(y_i-\mu_j)^2}{\lambda_i}+\beta_0) \tag{2.18}$$

2.3.4. Actualización de la asignación de índices (z_i)

Como ya apuntamos anteriormente, la distribución a priori de la variable z_i que asigna cada dato con un componente de la mezcla es $p(z_i = j) = w_j$. Asignamos a una muestra del vector de datos y_i la pertenencia a una subpoblación o distribución j de la mezcla con probabilidad proporcional a la distribución a posteriori de la variable z_i, $p(z_i = j \mid ...)$ evaluada en y_i. La distribución a posteriori de dicha variable se obtiene fácilmente multiplicando la distribución a priori por la función verosimilitud (2.3), con lo que obtenemos:

$$p(z_i = j \mid ...) \propto \frac{w_j}{\sqrt{\lambda_i}\gamma_j}\exp\left\{-\frac{(y_i-\mu_j)^2}{2\lambda_i\gamma_j^2}\right\}. \tag{2.19}$$

2.3.5. Muestreo del parámetro λ_i

La distribución a posteriori para la variable auxiliar λ_i es:

$$\begin{aligned} p(\lambda_i \mid \mu_j, \gamma_j^2, \lambda_i, z_i = j) \quad &\propto \tfrac{1}{\sqrt{\lambda_i}\gamma_j}\exp\left\{-\tfrac{(y_i-\mu_j)^2}{2\lambda_i\gamma_j^2}\right\} \\ &\times f_{\frac{\alpha}{2},1}(\lambda_i \mid \left(\cos\tfrac{\pi\alpha}{4}\right)^{\frac{2}{\alpha}}, 0) \end{aligned} \tag{2.20}$$

La extracción de muestras con distribución $p(\lambda_i \mid \mu_j, \gamma_j^2, \lambda_i, z_i = j)$, tal y como se comprueba en la expresión (2.20), involucra el producto de una función verosimilitud Normal y una distribución α-estable positiva. Esta última distribución no puede, en general, ser evaluada de manera exacta debido a que no existe una expresión analítica para la pdf estable. Sin embargo, el algoritmo de J. Chambers, C. Mallows and B. Stuck [Chambers *et al.*, 1976] puede usarse para extraer muestras de la distribución α-estable positiva. A su vez, las muestras de dicha distribución las usaremos como envolvente en el algoritmo de muestreo por rechazo o como distribución propuesta en el algoritmo de Metropolis-Hastings. Para ambos algoritmos, la probabilidad de aceptación conlleva sólo la evaluación de la verosimilitud y no de la distribución a priori. Por lo tanto, es posible obtener muestras de $p(\lambda_i \mid \mu_j, \gamma_j^2, \lambda_i, z_i = j)$ mediante el algoritmo de Metropolis-Hasting o usando el muestreo por rechazo.

En lo sucesivo, estudiaremos cómo realizar el muestreo de este parámetro usando el muestreo por rechazo y el algoritmo de Metropolis-Hasting. Para otros métodos de obtención de este parámetro, véase [Godsill & Kuruoglu, 1999].

Muestreo por rechazo

El valor máximo de la verosimilitud depende del valor del parámetro posición μ y está acotado del siguiente modo:

$$N(y_i|0, \lambda_i\gamma^2) \leq \frac{1}{\sqrt{2\pi y_i^2}} \exp\{-1/2\} \tag{2.21}$$

Conocida la cota máxima de la distribución, es sencillo diseñar el algoritmo Monte Carlo de muestreo por rechazo (*rejection sampling*):

Muestreo por rechazo.

1. Generamos un valor $\lambda_i \sim f_{\frac{\alpha}{2},1}$ usando el método de Chambers descrito en la Sección 1.1.4.

2. Generamos un número aleatorio uniforme con distribución $u \sim U(0, \frac{1}{\sqrt{2\pi y_i^2}} \exp\{-1/2\})$

3. Si $u < \mathcal{N}(y_i|0, \lambda_i\gamma^2)$ aceptamos el valor propuesto. En caso contrario volvemos otra vez a 1.

Mediante este algoritmo, en cada iteración obtenemos un nuevo valor de λ_i.

El problema es que este procedimiento puede llegar a ser lento si en la condición 3) no se acepta el valor de λ_i propuesto en un número razonable de intentos.

Algoritmo de Metropolis-Hasting

También es posible muestrear el parámetro λ usando el algoritmo de Metropolis-Hasting.

---

1. Proponemos un nuevo valor de λ'. $\lambda' \sim f_{\alpha/2,1}$

2. Aceptamos el valor propuesto con probabilidad $A = \min\{1, \frac{N(y_i|\mu_{z_i},\lambda'_i\gamma^2_{z_i})}{N(y_i|\mu_{z_i},\lambda_i\gamma^2_{z_i})}\}$

En este caso, el muestreo de λ_i se realiza de un modo sensiblemente más rápido que usando el muestreo por rechazo. Sin embargo, en cada iteración, no necesariamente se actualizan los valores de λ_i. Por lo tanto, es más conveniente usar el muestreo por rechazo, ya que con este método, sí que se actualizan todos los λ_i en cada iteración.

2.3.6. Actualización del exponente característico α

En cada iteración t, los nuevos valores de α se obtienen usando el algoritmo de Metropolis:

Elegimos como distribución propuesta, la distribución Gaussiana centrada en $\alpha^{(t)}$ y con varianza σ_α.

$$\alpha^{new} \sim (\alpha^{new}|\alpha^{(t)}) = \mathcal{N}\left(\alpha^{new}|\alpha^{(t)}, \sigma^2_\alpha\right) \tag{2.23}$$

2.3.7. Método Monte Carlo de dimensión variable para actualizar el número de componentes (k)

Al contrario que en el único trabajo previo de inferencia de parámetros en mezclas de distribuciones α-estables bajo el paradigma Bayesiano [Casarin, 2004], en nuestro modelo, el número de componentes en la mezcla k es un parámetro desconocido y por lo tanto debe ser estimado. Un cambio en el número de componentes de la mezcla k, conlleva un cambio de dimensión en los parámetros $\{w_j, \alpha_j, \mu_j, \gamma_j\}$. Por este motivo, no es posible usar técnicas Monte

---

Algoritmo de Metropolis

1. Proponemos un nuevo valor de α, que escribiremos como α^{new}, extrayéndolo de una distribución simétrica $q(\alpha^{new}|\alpha^{(t)})$

$$\alpha^{new} \sim q(\alpha^{new}|\alpha^{(t)})$$

2. Para el nuevo valor α^{new} propuesto, calculamos su vector λ correspondiente, que denotaremos como λ^{new} y también reasignamos los índices z_i, que denotaremos como z_i^{new}.

3. Aceptamos el valor propuesto α^{new} con probabilidad A_α, donde

$$A_\alpha = \text{mín}\left\{1, \frac{\prod_{i=1}^{N} p(\, y_i|\, \mu_{z_i}, \gamma_{z_i}^2, \lambda_i^{new}, z_i^{new})}{\prod_{i=1}^{N} p(\, y_i|\, \mu_{z_i}, \gamma_{z_i}^2, \lambda_i, z_i)}\right\} \tag{2.22}$$

4. Si el nuevo valor α^{new} no es aceptado, entonces

$$\alpha^{(t+1)} = \alpha^{(t)}$$

Carlo estándar tales como el muestreo de Gibbs o el algoritmo de Metropolis-Hasting y debemos hacer uso de técnicas Monte Carlo para dimensión variable, en particular el algoritmo RJMCMC (*reversible jump Markov chain Monte Carlo*) [Green, 1995].

De este modo, se incrementa notablemente la flexibilidad del algoritmo ya que es posible estimar el número de subpoblaciones k sin introducir ningún tipo de información sobre el número de componentes en la mezcla.

En [Richardson & Green, 1997] se aplica por vez primera el algoritmo RJMCMC al problema de modelos de mezcla, en concreto a mezcla de distribuciones Normales. En dicho trabajo, se sugieren dos movimientos entre espacios de diferente dimensión: *birth-death move* para componentes vacíos y *split-combine move* para componentes no vacíos. Donde un componente j se considera vacío cuando la asignación de índices correspondiente a dicho componente j es $z_j = 0$.

En esta memoria, se presenta una extensión del trabajo realizado en [Richardson & Green, 1997] al caso de mezcla de distribuciones α-estables simétricas. En un principio se implementó un movimiento transdimensional de tipo *birth-death move*, pero la tasa de aceptación para este tipo de movimiento fue muy baja y por lo tanto este movimiento se descartó debido a su baja eficiencia computacional. Sin embargo, el movimiento de tipo *split-combine* es suficiente para una correcta implementación del algoritmo, tal y como se mostrará en las simulaciones con datos sintéticos y reales.

A continuación detallamos ambos movimientos del algoritmo RJMCMC en el contexto de mezcla de distribuciones simétricas α-estables.

***Combine move*: combinación de dos componentes.**

Este movimiento conlleva un cambio en el número de componentes, que pasa de $k \longrightarrow k-1$. En cada iteración, con probabilidad $p = 0{,}5$ elegimos dos componentes consecutivos, es decir, con índices j y $j+1$ y parámetros $\{w_j, \alpha_j, \gamma_j, \mu_j\}$ y $\{w_{j+1}, \alpha_{j+1}, \gamma_{j+1}, \mu_{j+1}\}$ respectivamente y los combinamos en un nuevo componente con parámetros $\{w_{j*}, \alpha_{j*}, \gamma_{j*}, \mu_{j*}\}$.

Elegimos los nuevos valores de los pesos, posición y dispersión del componente combinado a partir de los dos componentes ya existentes del siguiente modo:

$$w_{j*} = w_{j1} + w_{j_2} \tag{2.24}$$

$$w_{j*}\mu_{j*} = w_{j1}\mu_{j1} + w_{j_2}\mu_{j_2} \tag{2.25}$$

$$w_{j*}(\mu_{j*}^2 + \gamma_{j*}^2) = w_{j1}(\mu_{j1}^2 + \gamma_{j1}^2) + w_{j2}(\mu_{j2}^2 + \gamma_{j2}^2) \tag{2.26}$$

Este movimiento es determinista, es decir, dados los valores actuales de los parámetros, los nuevos valores son automáticamente propuestos a partir de las expresiones (2.24), (2.25) y (2.26).

***Split move*: división de dos componentes.**

Aunque la combinación de componentes es determinista, el movimiento contrario, es decir, la división de un componente en dos, no lo es. De hecho, analizándolo detenidamente, comprobamos que en este movimiento proponemos 6 nuevos valores de los parámetros w_j, γ_j, μ_j y $w_{j+1}, \gamma_{j+1}, \mu_{j+1}$ a partir de 3 valores de parámetros existentes, por lo que, al diseñar la biyección entre el espacio de dimensión 3 con otro de dimensión 6 tenemos 3 grados de libertad. Por este motivo, al diseñar la biyección entre ambos espacios introducimos 3 variables aleatorias auxiliares. Una explicación general de este método puede encontrarse en la Sección 1.2.2.

Elegimos estas 3 nuevas variables auxiliares u_1, u_2 y u_3 aleatoriamente, extrayéndolas de sendas distribuciones Beta con parámetros

$$u_1 \sim Be(2,2) \tag{2.27}$$

$$u_2 \sim Be(2,2) \tag{2.28}$$

$$u_3 \sim Be(1,1) \tag{2.29}$$

En la Figura 2.3, se muestra la representación gráfica de las distribuciones $Be(1,1)$ y $Be(2,2)$.

El dominio de la distribución $Be(2,2)$ es $[0,1]$ y la evaluación de dicha distribución en 0 y 1 es $p(0) = 0$ y $p(1) = 0$ respectivamente. Además, el máximo de dicha función está en $p(0{,}5) = 1{,}5$. Por otra parte, la distribución Beta con parámetros $Be(1,1)$ es, simplemente, la distribución uniforme con dominio $[0,1]$. En la Figura 2.3 se representan las distribuciones $Be(1,1)$ y $Be(2,2)$ respectivamente. Por supuesto, la elección de la distribución Beta con los parámetros descritos para obtener las 3 variables necesarias para construir la biyección no es única. Sin embargo, esta elección proporciona una serie de ventajas que describimos a continuación: primero, los pesos propuestos w_{j1}, w_{j2} son siempre positivos y con valores comprendidos entre 0 y 1. Además, el valor propuesto para la dispersión es siempre positivo ($\gamma_j > 0$). Por supuesto, la elección de las distribuciones, $u_1 \sim Beta(2,2)$, $u_2 \sim Beta(2,2)$ y $u_3 \sim Beta(1,1)$ no es única, pero, tal y como se demostró en [Richardson & Green, 1997], ésta es una elección conveniente en el contexto de mezcla de distribuciones. Además, el correcto

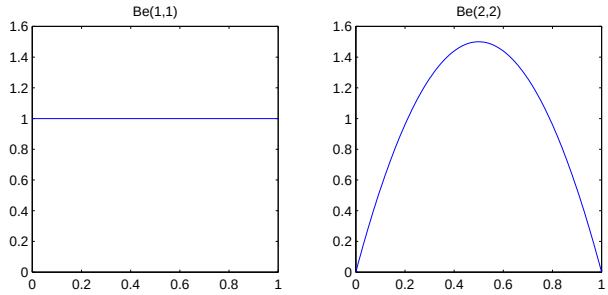

Figura 2.3: Distribuciones Beta con parámetros $Be(1,1)$ y $Be(2,2)$ usados en el movimiento de combinación del algoritmo RJMCMC.

funcionamiento del método no está basado en la elección de la distribución u, sino en la utilización de la expresión (1.32) para evaluar la tasa de aceptación.

Los nuevos valores propuestos para los pesos, posición y dispersión de los nuevos componentes de índices j_1 y j_2 obtenidos mediante división de un componente existente de índice j_* son:

$$w_{j1} = w_{j*}u_1 \tag{2.30}$$

$$w_{j2} = w_{j*}(1-u_1) \tag{2.31}$$

$$\mu_{j1} = \mu_{j*} - u_2\gamma_{j*}\sqrt{\frac{w_{j2}}{w_{j1}}} \tag{2.32}$$

$$\mu_{j2} = \mu_{j*} + u_2\gamma_{j*}\sqrt{\frac{w_{j1}}{w_{j2}}} \tag{2.33}$$

$$\gamma_{j1}^2 = u_3(1-u_2^2)\gamma_{j*}^2\frac{w_{j*}}{w_{j1}} \tag{2.34}$$

$$\gamma_{j2}^2 = (1-u_3)(1-u_2^2)\gamma_{j*}^2\frac{w_{j*}}{w_{j2}} \tag{2.35}$$

Los movimientos *split* y *combine* son reversibles, de ahí el nombre que recibe este método. Además, tras proponer los nuevos valores de las variables, y antes de aceptarlos según la expresión de la tasa de aceptación dada por la expresión (1.32), hay que comprobar que se cumple la condición $[\mu_1 < \mu_2 < ... < \mu_k]$, es decir, que los valores de las posiciones para cada uno de los componentes de

la mezcla están ordenados en orden creciente. Si no se cumple esta condición, el movimiento es automáticamente rechazado. Por otra parte, la obtención de unos nuevos valores para estos parámetros, provoca además un cambio en la asignación de variables z_i, la cual debe recalcularse mediante la expresión (2.19).

2.3.8. Tasa de aceptación

Los valores propuestos son aceptados con probabilidad A, donde este valor se calcula mediante la expresión propuesta en [Green, 1995] (véase la Sección 1.2.2). Remitimos al artículo de [Richardson & Green, 1997] para una explicación detallada de cómo hallar este valor para mezcla de Gaussianas y a [Salas-Gonzalez *et al.*, 2006a] para la deducción del índice de aceptación A para mezcla de α-estables donde los nuevos valores son propuestos con el movimiento reversible denominado división-combinación (*split-combine move*), tal y como realizamos en este trabajo. En [Salas-Gonzalez *et al.*, 2006c], fue propuesta una alternativa al método aquí expuesto, usando un movimiento transdimensional de tipo *birth-death*. Cabe resaltar que el método *split-combine* presentado en esta Memoria, es más eficiente, además de suficiente para una correcta estimación del número de componentes estables.

$$
\begin{aligned}
A &= \frac{\prod_{i:z_i=j_1}^{N} \frac{1}{\gamma_{j_1}} e^{-\frac{(y_i-\mu_{j_1})^2}{2\lambda_i\gamma_{j_1}^2}} \prod_{i:z_i=j_2}^{N} \frac{1}{\gamma_{j_2}} e^{-\frac{(y_i-\mu_{j_2})^2}{2\lambda_i\gamma_{j_2}^2}}}{\prod_{i:z_i=j_*}^{N} \frac{1}{\gamma_{j_*}} e^{-\frac{(y_i-\mu_{j_*})^2}{2\lambda_i\gamma_{j_*}^2}}} \\
&\times \frac{1}{a} \times (k+1) \times \frac{w_{j1}^{\zeta-1+n_1} w_{j2}^{\zeta-1+n_2}}{w_{j*}^{\zeta-1+n_1+n_2} B(\zeta, k\zeta)} \\
&\times \sqrt{\frac{\kappa}{2\pi}} e^{-0,5\kappa\left\{(\mu_{j1}-\xi)^2+(\mu_{j2}-\xi)^2-(\mu_{j*}-\xi)^2\right\}} \\
&\times \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \left(\frac{\gamma_{j1}^2\gamma_{j2}^2}{\gamma_{j*}^2}\right)^{-\alpha_0-1} e^{-\beta_0(\gamma_{j1}^{-2}+\gamma_{j2}^{-2}-\gamma_{j*}^{-2})} \\
&\times \frac{d_{k+1}}{b_k P_{alloc}} \times \left\{g_{2,2}(u_1) g_{2,2}(u_2) g_{1,1}(u_3)\right\}^{-1} \\
&\times \frac{w_{j*} \left|\mu_{j1}-\mu_{j2}\right| \gamma_{j1}^2\gamma_{j2}^2}{u_2(1-u_2^2)(1-u_3)\gamma_{j*}^2}
\end{aligned}
\tag{2.36}
$$

donde n_1 y n_2 son el número de muestras del vector y_i asignados a los componentes j_1 y j_2. $B(\cdot,\cdot)$ es la función Beta, P_{alloc} es la probabilidad de que la asignación de variables propuesta sea la elegida y b_k, $d_k = 1 - b_k$ son, respectivamente, las probabilidades de combinar dos componentes en uno sólo (*combine move*) o fraccionar un componente en dos (*split move*).

La primera línea de la expresión (2.36) es el cociente entre las verosimilitudes para ambos modelos. Los términos en la segunda línea son: $1/a$ es el cociente entre las distribuciones a priori para el exponente característico α. $(k+1)$ es un término que aparece como consecuencia de imponer orden creciente en el parámetro posición para cada uno de los componentes $[\mu_1 < \mu_2 < ... < \mu_k]$. El término $\frac{w_{j1}^{\zeta-1+n_1} w_{j2}^{\zeta-1+n_2}}{w_{j*}^{\zeta-1+n_1+n_2} B(\zeta,k\zeta)}$ es el cociente entre las distribuciones a priori para los pesos w y la variable asignación z. La tercera línea es el cociente entre las distribuciones a priori para el parámetro posición $\mathcal{N}(\mu|\xi,\kappa)/\mathcal{N}(\mu*|\xi,\kappa)$ y la cuarta para la dispersión $\mathcal{IG}(\gamma|\alpha_0,\beta_0)/\mathcal{IG}(\gamma*|\alpha_0,\beta_0)$. La quinta línea tiene dos términos, el primero de ellos, $\frac{d_{k+1}}{b_k P_{alloc}}$ es la probabilidad de que se produzca el movimiento y la asignación de variable actual y $\{g_{2,2}(u_1)g_{2,2}(u_2)g_{1,1}(u_3)\}^{-1}$ son las distribuciones Beta evaluadas en u_1, u_2 y u_3.

De este modo, en cada iteración, se proponen dos nuevos componentes con probabilidad b_k (*split move*) y son aceptados con probabilidad $\min\{1, A\}$. Si, por el contrario, proponemos un nuevo componente a partir de dos dados (*combine move*), éste es aceptado con probabilidad $\min\{1, A^{-1}\}$.

2.4. Resultados

2.4.1. Simulación 1: datos sintéticos

Vamos a aplicar la metodología propuesta para la estimación de parámetros en modelos de mezcla de distribuciones simétricas α-estables a $N = 2000$ muestras i.i.d. con la distribución siguiente:

$$p_Y(y) = 0{,}2 f_{1,4,0}(y|0{,}2,-2) + 0{,}3 f_{1,4,0}(y|0{,}5,0) + 0{,}5 f_{1,4,0}(y|0{,}6,3). \quad (2.37)$$

Los valores elegidos para los hiperparámetros de las distribuciones a priori y parámetros de la simulación son: $\alpha_0 = 1$, $\beta_0 = 1$, $\xi = 0$, $\kappa = 1/3^2$ y $\zeta = 1$. La probabilidad de elegir entre los movimientos *combine* o *split* es $b_k = d_k = 0{,}5$ y el número de iteraciones es $N_{iter} = 500$ con un periodo de calentamiento (*burn-in*) de $N_{burnin} = 100$. Inicializamos el número de componentes a $k = 5$. Por

Tabla 2.3: Simulación 1: Valores reales de los parámetros de la mezcla de tres distribuciones α-estables simétricas, valores estimados y desviación estándar.

Parámetro	Valores reales	Valores estimados	Desviación estándar
α_1	1.4	1.39	0.11
α_2	1.4	1.38	0.12
α_3	1.4	1.41	0.15
μ_1	-2	-1.983	0.018
μ_2	0	-0.00	0.04
μ_3	3	3.01	0.03
γ_1	0.2	0.233	0.014
γ_2	0.5	-0.51	0.04
γ_3	0.6	0.59	0.03
w_1	0.2	0.222	0.012
w_2	0.3	0.282	0.016
w_3	0.5	0.01	0.03

otra parte, el exponente característico α_j para cada componente fue inicializado a α = 1,2. El valor inicial de la dispersión de cada componente j se inicializó aleatoriamente, extrayendo muestras de la distribución $\gamma_j \sim \mathcal{IG}(\gamma_j|\alpha_0, \beta_0)$, mientras que la posición fue inicialmente considerada como μ_j = [-3 -1 1 3 5]. Los valores de los pesos fueron inicializados mediante la extracción de $k = 5$ variables aleatorias con distribución $w \sim \mathcal{D}(1, 1, 1, 1, 1)$.

Los resultados correspondientes a la estimación paramétrica del modelo expresado en la ecuación (2.37) obtenidos por nuestro algoritmo se muestran en la Tabla 2.3. Allí se observa que cada uno de los parámetros desconocidos es estimado con precisión. Además, tras poco más de 100 iteraciones obtenemos, por vez primera, el número verdadero de componentes en la mezcla $k = 3$.

Con el propósito de comparar los resultados obtenidos usando nuestro algoritmo con la metodología basada en mezcla de Gaussianas, el vector de $N = 2000$ datos con distribución dada por la expresión (2.37) ha sido estudiado bajo la suposición de que procede de una mezcla de distribuciones Normales. Para ello se ha usado el algoritmo propuesto en [Richardson & Green, 1997]. En la Fi-

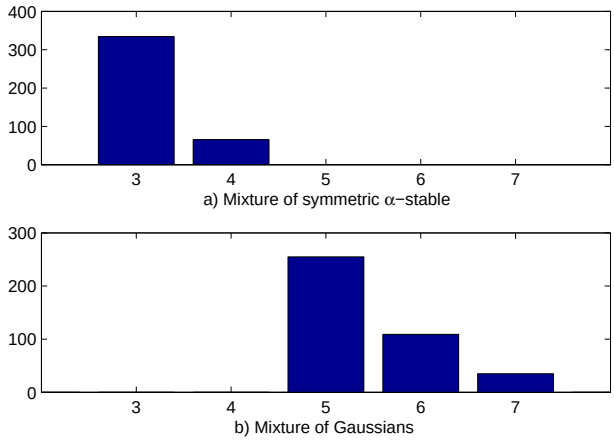

Figura 2.4: Histograma con la estimación del número de componentes k. Se comprueba cómo usando mezcla de distribuciones α-estables, el número real de componentes $k = 3$ es estimado correctamente, mientras que usando mezcla de Gaussianas, el número de componentes es sobreestimado.

gura 2.4 se muestra mediante sendos histogramas, el número de componentes estimado en cada iteración (tras el periodo *burn-in*) para el modelo de mezcla distribuciones α-estables simétricas aquí considerado y el modelo de mezcla de distribuciones Normales. En dicha gráfica se comprueba claramente cómo nuestro algoritmo es capaz de estimar el número verdadero de componentes de la mezcla, mientras que el modelo de mezclas de Gaussianas estima $k = 5$ como valor más probable de número de componentes. Por lo tanto, el algoritmo aquí propuesto presenta mejores resultados cuando el vector de datos es una mezcla de señales impulsivas. Por otra parte, cuando la señal es una mezcla de señales con distribución Normal, el resultado obtenido al aplicar nuestro algoritmo o el propuesto en [Richardson & Green, 1997] es el mismo, ya que, tal y como dijimos en la sección 2, la distribución Gaussiana es un caso particular de la distribución α-estable. Este caso lo analizaremos con más detalle en las siguientes simulaciones.

El modelado mediante mezcla de distribuciones α-estables simétricas es, por lo tanto, más robusto frente a datos alejados de la moda (los denominados *outliers*). Para mostrar esto, modelamos la secuencia de $N = 2000$ datos distri-

buidos como en la expresión (2.37) con una mezcla de tres componentes Normales. La densidad obtenida suponiendo mezcla de Gaussianas y el histograma discreto correspondiente a la secuencia de $N = 2000$ datos se muestra en la Figura 2.5. En esta figura, representamos la mezcla de distribuciones α-estables obtenida mediante los valores inscritos en la Tabla 2.3. Se puede observar con facilidad que la mezcla de Gaussianas no es capaz de explicar correctamente la distribución de los datos. Esto se debe, principalmente, al alto grado de impulsividad del vector observación y_i. Por ejemplo, el máximo y mínimo valor del vector de datos y es 26,40 y $-34{,}15$ respectivamente. Estos valores están muy alejados de la media y por tanto la suposición de que la señal es una mezcla de Gaussianas se aleja mucho de la realidad. De hecho, lo que obtenemos al tratar de explicar la distribución del vector observación mediante mezcla de Gaussianas es que uno de los componentes de la mezcla obtenido modela, más o menos acertadamente, al primer componente de la mezcla. El segundo y tercer componente de y_i se modelan mediante una sola Gaussiana y el tercer componente Gaussiano es una solución con peso w_j muy bajo (es decir, explica un número pequeño de datos) y varianza σ_j muy alta. Lo cual indica que este componente modela los datos muy alejados de la media (*outliers*). Más concretamente, para este tercer componente, obtuvimos los parámetros $\mu_3 = 2{,}80$, $\sigma_3 = 14{,}2$ y $w_3 = 0{,}01$.

2.4.2. Simulación 2: datos sintéticos

Tal y como se apuntó en la simulación anterior, el modelado mediante mezcla de Gaussianas es más robusto a señales impulsivas que la mezcla de Gaussianas. Además, una de las bazas de nuestro modelo es que es una generalización de la mezcla de Gaussianas. En esta simulación vamos a mostrar este hecho mediante un ilustrativo ejemplo. Para ello, simulamos $N = 1000$ muestras con distribución Normal y parámetros $w = [0{,}4, 0{,}3, 0{,}3]$, $\mu = [-3, 0, 2{,}5]$ y $\sigma = [0{,}8, 0{,}8, 0{,}4]$. Esto se corresponde con tres distribuciones α-estables simétricas con parámetros $\alpha = [2, 2, 2]$, dispersión $\gamma = \sigma/\sqrt{2} = [0{,}57, 0{,}57, 0{,}28]$ y parámetro posición $\mu = [-3, 0, 2{,}5]$. El vector observación de dimensión $N = 1000$ fue modelado mediante el algoritmo descrito en este capítulo. En la Tabla 2.4, están representados los valores de los parámetros de las tres distribuciones α-estables obtenidos junto a los valores verdaderos. Todos los parámetros fueron estimados correctamente. El número de iteraciones del algoritmo MCMC fue de 1000 iteraciones, con un periodo de entrenamiento de 500 iteraciones.

La media fue el estimador elegido para cada uno de los parámetros. En este caso y debido a que el dominio del exponente característico es $\alpha = (0, 2]$ es más

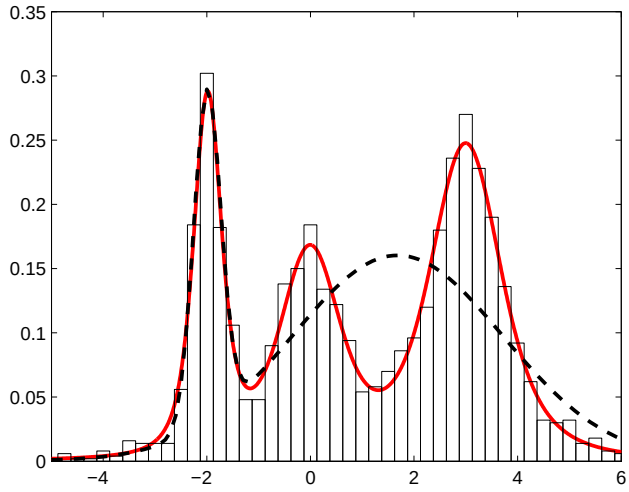

Figura 2.5: Histograma del vector de datos considerado. Línea continua: Mezcla de distribuciones simétricas α-estable. Línea continua: mezcla de 3 distribuciones Gaussianas.

conveniente usar la moda como estimador de este parámetro, en cuyo caso el valor estimado sería $\alpha = [2, 2, 2]$. Aún así, usando la media como estimador se obtienen también excelentes resultados.

En la Figura 2.6 están representados conjuntamente el histograma de los datos con distribución dada por mezcla de tres componentes Gaussianos y la densidad obtenida mediante los valores de los parámetros representados en la Tabla 2.4. El resultado obtenido ajusta perfectamente la densidad del vector de observación.

2.4.3. Simulación 3: datos reales

El método Bayesiano de estimación de parámetros en modelos de mezcla α-estable simétrica ha sido evaluado para tres conjuntos de datos reales de naturaleza muy diversa. El primero de ellos se trata de 245 medidas independientes de la actividad enzimática en la sangre, para un tipo de enzimas que se encuentra en el metabolismo de sustancias cancerígenas [Bechtel *et al.*, 1993; Richardson

Tabla 2.4: Simulación 2: Valores verdaderos, valores estimados y error.

Parámetro	Valor real	Valor estimado	Desviación estándar
α_1	2	1.96	0.05
α_2	2	1.93	0.07
α_3	2	1.95	0.05
μ_1	-3	-3.05	0.06
μ_2	0	0.06	0.07
μ_3	2.5	2.48	0.03
γ_1	0.57	0.56	0.03
γ_2	0.57	0.60	0.06
γ_3	0.28	0.29	0.02
w_1	0.4	0.42	0.02
w_2	0.3	0.29	0.02
w_3	0.3	0.29	0.02

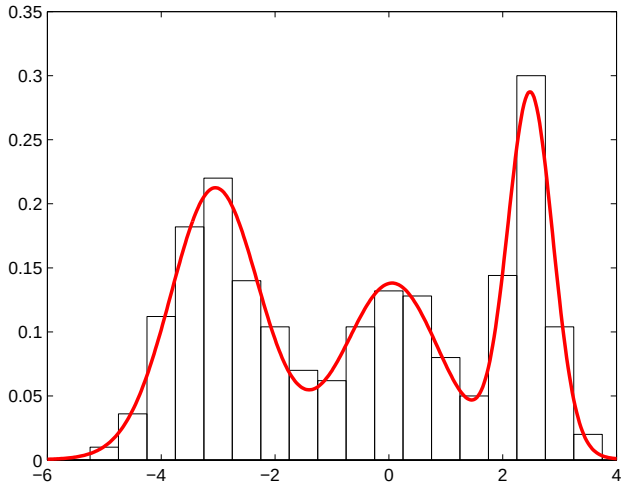

Figura 2.6: Histograma discreto con la mezcla de tres componentes Gaussianos de la simulación 2. *Línea continua:* Densidad α-estable simétrica calculada.

& Green, 1997]. El segundo, el índice de acidez en escala logarítmica medido en 155 lagos de los Estados Unidos [Crawford *et al.*, 1992; Crawford, 1994; Richardson & Green, 1997]. El tercer conjunto de datos consiste en la medida de la velocidad radial con respecto a la Vía Láctea de 82 galaxias distantes. Este último conjunto de datos ha sido analizado ampliamente en la literatura en trabajos realacionados con los modelos de mezcla [Escobar & West, 1995; Phillips & Smith, 1996; Stephens, 2000].

En la Tabla 2.5, se muestra el número de componentes y la estimación de parámetros obtenidos para estos tres conjuntos de datos. Por otro lado, en la Figura 2.7 están representados los histogramas de cada uno de los datos reales junto a la distribución mezcla α-estables simétricas obtenida.

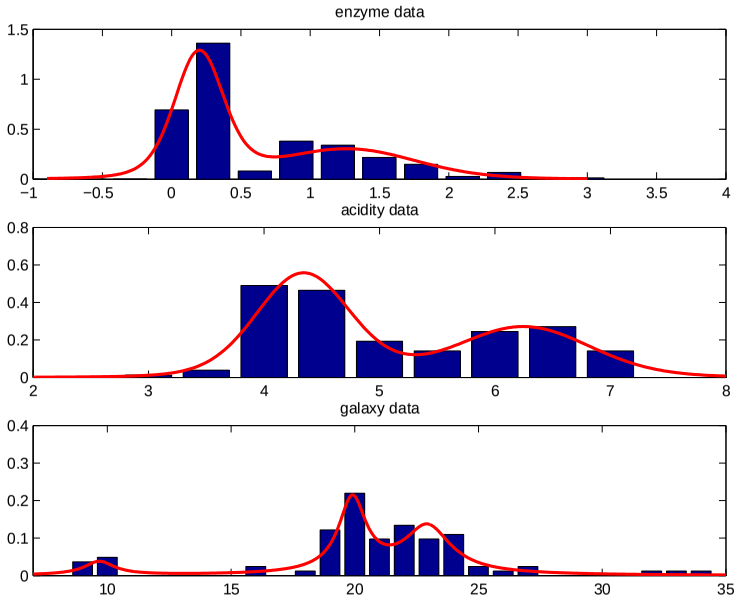

Figura 2.7: Histograma de cada uno de los conjuntos de datos reales analizados. *Línea continua:* Densidad α-estable simétrica calculada. 'Enzyme data': 2 componentes. 'Acidity data': 2 componentes. 'Galaxy data': 3 componentes.

2.5. Conclusiones

En este capítulo, se ha introducido un nuevo modelo de mezcla basado en distribuciones α-estables simétricas. Dicho modelo ha sido probado en mezcla sintéticas de señales impulsivas y se ha demostrado mediante simulación numérica que, para este tipo de señales, tanto el número verdadero de componentes en la mezcla como los distintos parámetros de las distribuciones que la componen, son estimados con precisión. Además, hemos comparado el trabajo aquí desarrollado con el modelo de mezcla de Gaussianas. El número de componentes fue sobreestimado cuando usamos el modelo de mezcla de Gaussianas, por lo que nuestro método se erige como una alternativa a dicho método cuando las señales que componen la mezcla son impulsivas. Otra de las ventajas de nuestro método es que el rango de aplicación no sólo se restringe a este tipo de señales. Puesto que la distribución Normal es un caso particular de distribución α-estable, el método aquí propuesto también puede usarse para los mismos casos que en los que la suposición de Gaussianidad era válida. Por último, en este capítulo, el modelo de mezcla de distribuciones α-estables simétricas ha sido probado en datos reales, en concreto en datos biológicos, geológicos y astrofísicos. La elección de datos de tres disciplinas tan dispares se ha hecho con el fin de mostrar el amplio rango de aplicación del algoritmo presentado en este capítulo.

Tabla 2.5: Valores estimados para los distintos conjuntos de datos reales analizados.

Parameter	Enzyme data (k=2)	Acidity data (k=2)	Galaxy data (k=3)
α_1	1.66	1.80	1.06
α_2	1.84	1.82	1.38
α_3	-	-	1.35
μ_1	0.20	4.34	9.71
μ_2	1.27	6.25	20.0
μ_3	-	-	23.0
γ_1	0.14	0.30	0.76
γ_2	0.37	0.42	0.74
γ_3	-	-	1.20
w_1	0.62	0.59	0.09
w_2	0.38	0.41	0.45
w_3	-	-	0.46

CAPÍTULO

TRES

INFERENCIA BAYESIANA EN MEZCLA DE DISTRIBUCIONES α-ESTABLES

En este capítulo generalizaremos el modelo desarrollado en el capítulo anterior a mezcla de distribuciones α-estables (simétricas o no) [Salas-Gonzalez *et al.*, 2008, 2006b,a; Kuruoglu *et al.*, 2006]. La inclusión del parámetro β en los cálculos añade un grado más de flexibilidad al modelo, que por otro lado sigue manteniendo sus deseables propiedades, entre ellas, la de ser una generalización del modelo de mezcla de Gaussianas. Aunque en este caso no existe una propiedad como la del Producto (Sección 1.1.3) que permita evaluar la función verosimilitud para este modelo, por lo que serán necesarias otras estrategias para resolver el modelo jerárquico Bayesiano. Afortunadamente, aunque no existe una fórmula analítica para la distribución α-estable, es posible su evaluación a partir de la integración numérica de su función característica.

Para la evaluación de la verosimilitud del modelo, usaremos la misma estrategia que en [Lombardi, 2007]. Además, compararemos extensivamente nuestro método con el propuesto en un trabajo no publicado [Casarin, 2004], ya que éste es el único trabajo previo en mezcla de distribuciones α-estables. Como veremos a lo largo del capítulo, el método que aquí se presenta es más eficiente computacionalmente, funciona mejor para datos cuya distribución consta de subpoblaciones muy mezcladas y además, nos permite calcular el número de componentes a través del método Monte Carlo basado en cadenas de Markov con saltos reversibles.

Este capítulo está estructurado del siguiente modo: en la Sección 3.1 se

presenta el modelo de mezcla α-estable. En la Sección 3.2, se muestran las distribuciones a priori para cada uno de los parámetros. En la Sección 3.3, se presenta el algoritmo Monte Carlo de cadenas de Markov. Finalmente, en las secciones 3.5 y 3.6, respectivamente, presentamos los resultados y extraemos las conclusiones.

3.1. Modelo Bayesiano

El modelo de mezcla de distribuciones α-estables, en su caso más general viene dado por:

$$p_Y(y) = \sum_{j=1}^{k} w_j f_{\alpha_j,\beta_j}(y|\gamma_j,\mu_j)$$

$$0 \leq w_j \leq 1(\forall j) \text{ y } \sum_{j=1}^{k} w_j = 1$$

donde w_j es el peso del componente j y $p_Y(y)$ es la función densidad de probabilidad del vector observación y. Consideramos en dicho modelo que los componentes del vector y han sido aleatoriamente extraídos de un número k de componentes, que denotamos como $j = 1, 2, ...k$. Introducimos una nueva variable $z_i \in [1, 2, ...k]$ denominada *asignación*. $z_i = j$ denota que la observación y_i pertenece al componente j de la mezcla.

Condicionado a los valores z_i, podemos considerar que las observaciones y_i han sido extraídas de los siguientes componentes individuales:

$$y_i|z_i \sim f_{\alpha_j,\beta_j}(y|\gamma_j,\mu_j) \qquad j = 1, 2, ..., k.$$

Este modelo es invariante bajo permutaciones de los índices j. Por lo tanto, es necesario establecer un criterio para que la variable asignación caracterice los distintos componentes de manera inequívoca. El criterio utilizado será la ordenación de los distintos componentes en orden creciente de los valores de la posición $\mu_1 < \mu_2 < ... < \mu_k$.

Particularizamos la ecuación del Teorema de Bayes (1.22) con la información proporcionada por nuestro modelo, considerando que B denota el vector observación y y A las variables desconocidas cuyos valores deseamos estimar.

$$A = \{k, w, z, \alpha, \beta, \gamma, \mu\} \tag{3.1}$$

Antes de presentar el método Monte Carlo basado en cadenas de Markov para inferir los valores de las variables desconocidas de nuestro modelo, es necesario construir el modelo jerárquico Bayesiano (véase la Sección 1.2). En este caso, el número de parámetros desconocidos es $5k + 1$ (α_j,β_j,γ_j,μ_j,w_j,k). La distribución conjunta de estas variables es:

$$\begin{aligned} p(k, w, z, \alpha, \beta, \gamma, \mu, y) &= p(y|k, w, z, \alpha, \beta, \gamma, \mu) \\ &\times \quad p(k, w, z, \alpha, \beta, \gamma, \mu) \end{aligned} \tag{3.2}$$

Para simplificar la notación, los parámetros de la distribución α-estable y los hiperparámetros de sus correspondientes distribuciones a priori los agruparemos como $\theta = \{\alpha, \beta, \gamma, \mu\}$ y $\eta = \{a, b, \alpha_0, \beta_0, \xi, \kappa\}$ respectivamente. Por lo tanto, la expresión (3.2), una vez tenida en cuenta las relaciones condicionales entre las distintas variables y los hiperparámetros, puede reescribirse como:

$$\begin{aligned} p(k, w, z, \theta, \eta, y) &= p(y|k, w, z, \theta, \eta)p(\theta|k, \eta)p(z|w, k) \\ &\times \quad p(w|k, \zeta)p(k|k_0)p(k_0)p(\zeta)p(\eta) \end{aligned} \tag{3.3}$$

donde k_0 y ζ son los hiperparámetros de k y w respectivamente.

La Figura 3.1 muestra el gráfico con las dependencias entre las distintas variables para el modelo representado por la ecuación (3.3). Como es habitual, los círculos denotan las variables desconocidas mientras que los rectángulos representan parámetros con valor fijo (hiperparámetros). y es el vector observación. La dirección de las flechas indica la dependencia condicional entre las distintas magnitudes.

3.2. Distribución a priori

A continuación, se presentan las distintas distribuciones a priori para este modelo. Hemos elegido las mismas que para las mezclas de distribuciones α-estables simétricas aunque en este caso no se trata de las distribuciones a priori conjugadas, ya que no podemos escribir la verosimilitud como una distribución Gaussiana, cosa que sí era posible para el caso simétrico.

3.2.1. Exponente característico (α)

Elegimos como la distribución a priori para el exponente α, a la distribución uniforme en el rango $0 < \alpha \leq 2$.

$$p(\alpha|a) = \frac{1}{a} = \frac{1}{2}; \quad \text{para} \quad 0 < \alpha \leq 2 \tag{3.4}$$

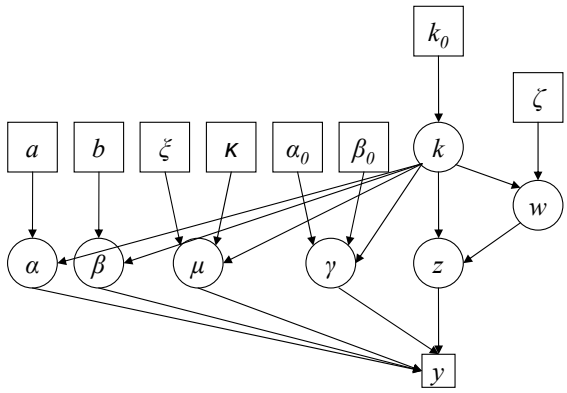

Figura 3.1: Dependencia entre las distintas variables y parámetros que componen el modelo jerárquico Bayesiano presentado en este capítulo. *Círculos:* variables. *Rectángulos:* hiperparámetros y vector observación . La dirección de las flechas representa la dependencia entre las variables y los hiperparámetros.

3.2.2. Parámetro de asimetría (β)

Como distribución a priori para este parámetro, también elegimos la distribución uniforme en el dominio de la variable β, es decir, en el rango $-1 < \beta < +1$.

$$p(\alpha|b) = \frac{1}{b} = \frac{1}{2}; \quad \text{para} \quad -1 < \beta \leq +1 \tag{3.5}$$

Esta elección es la misma que la realizada en [Lombardi, 2007; Buckle, 1995]. Tal y como se comentó en la Sección 2.2.1, podría ser tenido en cuenta el comportamiento de β para distintos valores de los parámetros α-estables, como por ejemplo, el hecho de que conforme α tiende a su valor extremo 2, el parámetro β tiene menos influencia en la asimetría de la distribución. Recordemos que cuando $\alpha = 2$, el parámetro β no está definido (véase la Sección 1.1.1). En la Figura 3.2, se muestran dos pares de distribuciones α-estable con $\gamma = 1$ y $\mu = 0$. La Figura 3.2a, presenta dos distribuciones con $\alpha = 1{,}9$, $\beta_1 = +1$ y $\beta_2 = -1$, mientras que en la Figura 3.2b, $\alpha = 1{,}2$, $\beta_1 = +1$ y $\beta_2 = -1$. Es inmediato

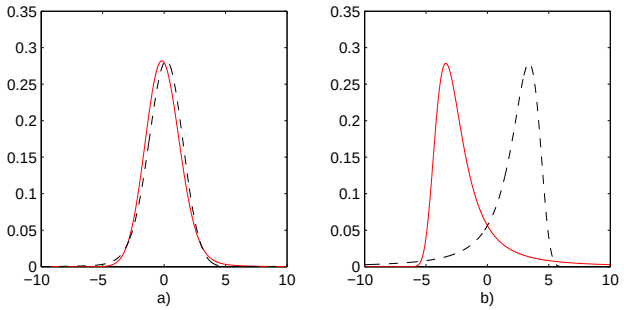

Figura 3.2: Distribución α-estable con $\mu = 0$ y $\sigma = 1$. *Línea continua:* $\beta = +1$. *Línea discontinua:* $\beta = -1$ (a) $\alpha = 1{,}9$ (b) $\alpha = 1{,}2$

comprobar que el parámetro β ejerce una influencia mayor en la forma de la distribución conforme el valor del exponente característico α es menor.

3.2.3. Dispersión (γ)

Elegimos como distribución a priori para el parámetro dispersión la distribución Gamma Inversa. En este caso, al contrario que en el modelo presentado en el anterior capítulo, no es la distribución a priori conjugada.

$$\mathcal{IG}(x|\alpha_0, \beta_0) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} x^{-\alpha_0 - 1} \exp\{-\frac{\beta_0}{x}\} \tag{3.6}$$

3.2.4. Posición (μ)

Para el parámetro desconocido μ, elegimos como distribución a priori la distribución Normal con hiperparámetros media (ξ) y varianza (κ^{-1}). La distribución a priori para la variable posición es, por lo tanto:

$$p(\mu|\xi, \kappa^{-1}) = \frac{\kappa^{1/2}}{\sqrt{2\pi}} \exp\{-\frac{(\mu - \xi)^2}{2\kappa^{-1}}\}. \tag{3.7}$$

3.2.5. Pesos (w)

La distribución a priori conjugada para los pesos en un modelo de mezcla es la distribución de Dirichlet simétrica, que es una generalización de la distribución

Beta para múltiples variables.

$$\mathbf{w} \sim \mathcal{D}(\zeta, ..., \zeta) \tag{3.8}$$

3.2.6. Asignación de índices (z)

La distribución a priori para la asignación de índices es

$$p(z_i = j) = w_j \tag{3.9}$$

por lo tanto, a la observación i asignada a la distribución $j = 1, 2, ...k$ (siendo k el número de componentes) le asignamos como distribución a priori el valor del peso estimado para dicho componente w_j.

3.2.7. Número de componentes (k)

Elegimos como distribución a priori para el número de componentes de la mezcla k, la distribución discreta uniforme entre 1 y un valor k_0. Donde k_0 es el máximo valor que permitimos que tome nuestro algoritmo para el número de componentes k. Tal y como hicimos en la Sección 2.15, elegimos k_0 suficientemente grande como para que en ninguna iteración el número de componentes estimados sea mayor que k_0.

$$p(k|k_0) = \frac{1}{k_0} \tag{3.10}$$

3.3. Actualización de las variables

Los métodos Bayesianos proporcionan el marco adecuado para hacer inferencia de parámetros, tal y como fue explicado en la Sección 1.2. El principal problema de esta técnica es que, a menudo, es necesario resolver complejas integrales multidimensionales sin solución analítica. Estas integrales involucradas en la inferencia Bayesiana de variables, pueden resolverse numéricamente usando Métodos Monte Carlo basados en cadenas de Markov [Robert & Casella, 1999; Gilks *et al.*, 1996].

El modelo de mezcla de distribuciones α-estables ha sido estudiado anteriormente en un trabajo no publicado [Casarin, 2004]. El método propuesto en este capítulo posee dos ventajas principales con respecto a aquél. El trabajo no publicado se basa en el algoritmo de Gibbs para distribuciones α-estables univariadas propuesto en [Buckle, 1995], el cual introduce una variable aleatoria

auxiliar de dimensión N igual a la dimensión del vector observación y que debe ser actualizada en cada iteración.

[Buckle, 1995] fue el primero en proponer un método de estimación Bayesiano de los parámetros de una distribución α-estable. Sin embargo, tal y como se analizó en [Lombardi, 2007], no es nada sencillo simular esta nueva variable auxiliar. Nuestro algoritmo, en cambio, no precisa de la introducción de ninguna variable auxiliar, por lo que la complejidad computacional es considerablemente menor (para una revisión más exhaustiva de [Buckle, 1995], véase [Lombardi, 2007]).

La segunda ventaja importante del método propuesto en este capítulo es que considera desconocido el número de componentes k de la mezcla. Esta variable está relacionada con la dimensión de los parámetros del modelo, por lo tanto, los métodos de Monte Carlo basados en cadenas de Markov usados en [Casarin, 2004], no son suficientes y es necesario usar un método Monte Carlo para dimensión variable como por ejemplo RJMCMC [Green, 1995]. Por lo tanto, el método que se presenta en este capítulo es capaz de determinar el número de componentes de la mezcla mediante una metodología Bayesiana.

Una vez escrito el modelo jerárquico Bayesiano, obtendremos muestras de cada parámetro siguiendo el siguiente esquema:

Actualización de las variables

1. Actualización de los pesos ($\mathbf{w}$) mediante el muestreo de Gibbs.

2. Actualización de los parámetros de las distribuciones α-estable $\theta = \{\alpha,\beta,\mu,\gamma\}$ mediante el algoritmo de Metropolis.

3. Actualización de la asignación de variables z.

4. Actualización del número de componentes k mediante el algoritmo RJMCMC.

3.3.1. Actualización de los pesos (w)

Combinando las ecuaciones (3.8) y (3.9), la distribución a posteriori para $\mathbf{w}$ es Dirichlet con parámetros $\zeta + n_j$. Por lo tanto, a cada iteración, los nuevos

valores de w se obtienen extrayendo muestras de la siguiente distribución:

$$\mathbf{w} \mid ... \sim D(\zeta + n_1, ..., \zeta + n_k) \tag{3.11}$$

donde n_j es el número de muestras asignadas al componente j, ($n_j = \sum_i \delta(z_i - j)$), siendo δ es la función delta de Dirac).

3.3.2. Actualización de α,β,μ,γ mediante el algoritmo de Metropolis

Estimamos el exponente característico α usando el algoritmo de Metropolis. Por lo tanto, para un componente j dado, los valores α_j lo obtendremos mediante el siguiente esquema:

1) En cada iteración t, proponemos un nuevo valor de α, que nombraremos α_j^{new}, extrayendo un valor aleatorio de una distribución $q(\cdot|\cdot)$

$$\alpha_j^{new} \sim q(\alpha_j^{new}|\alpha_j^{(t)})$$

2) Aceptamos el valor propuesto $\alpha_j^{(t+1)} = \alpha_j^{new}$, con probabilidad dada por mín $\{1, A_{\alpha_j}\}$, donde

$$\begin{aligned} A_{\alpha_j} &= \text{mín} \left\{ 1, \frac{\prod_{i:z_i=j}^{N} p(y_i|k, w_{z_i}, \alpha_j^{new}, \beta_{z_i}, \gamma_{z_i}, \mu_{z_i})}{\prod_{i:z_i=j}^{N} p(y_i|k, w_{z_i}, \alpha^{(t)}, \beta_{z_i}, \gamma_{z_i}, \mu_{z_i})} \right. \\ &\times \left. \frac{p(k, w_{z_i}, \alpha_j^{new}, \beta_{z_i}, \gamma_{z_i}, \mu_{z_i}) q(\alpha_j^{(t)}|\alpha_j^{new})}{p(k, w_{z_i}, \alpha_j^{(t)}, \beta_{z_i}, \gamma_{z_i}, \mu_{z_i}) q(\alpha_j^{new}|\alpha_j^{(t)})} \right\} \end{aligned} \tag{3.12}$$

si el nuevo valor α_j^{new} no es aceptado, entonces

$$\alpha_j^{(t+1)} = \alpha_j^{(t)}.$$

La expresión en (3.12) puede simplificarse para este modelo debido a que las distribuciones a priori son independientes entre sí, por lo tanto podemos escribir

$$\frac{p(k, w_{z_i}, \alpha_{new}, \beta, \gamma_{z_i}, \mu_{z_i})}{p(k, w_{z_i}, \alpha^{(t)}, \beta, \gamma_{z_i}, \mu_{z_i})} = \frac{p(\alpha_{new})}{p(\alpha^{(t)})}.$$

Por otro lado, si usamos una distribución simétrica $q(\alpha_{new}|\alpha^{(t)}) = q(\alpha^{(t)}|\alpha_{new})$ y tenemos en cuenta que la distribución a priori $p(\alpha)$ es la distribución uniforme de dominio $[0, 2]$ (véase la expresión (3.4)), la tasa de aceptación A (ec. (3.12)) se simplifica como:

$$A_{\alpha_j} = \text{mín}\left\{1, \frac{\prod_{i:z_i=j}^{N} p(y_i|k, w_{z_i}, \alpha_j^{new}, \beta_{z_i}, \gamma_{z_i}, \mu_{z_i})}{\prod_{i:z_i=j}^{N} p(y_i|k, w_{z_i}, \alpha_j^{(t)}, \beta_{z_i}, \gamma_{z_i}, \mu_{z_i})}\right\} \tag{3.13}$$

El mismo procedimiento detallado para el parámetro α lo usaremos para el resto de parámetros de la distribución α-estable. Para los parámetros de asimetría β_j, la dispersión γ_j y el parámetro posición μ_j proponemos nuevos valores $\beta_j^{new} \sim q(\beta_j^{new}|\beta_j^{(t)})$, $\gamma_j^{new} \sim q(\gamma_j^{new}|\gamma_j^{(t)})$ y $\mu_j^{new} \sim q(\mu_j^{new}|\mu_j^{(t)})$ respectivamente, y los aceptamos con probabilidad dada por las siguientes expresiones:

$$A_{\beta_j} = \text{mín}\left\{1, \frac{\prod_{i:z_i=j}^{N} p(y_i|k, w_{z_i}, \alpha_{z_i}, \beta_j^{new}, \gamma_{z_i}, \mu_{z_i})}{\prod_{i:z_i=j}^{N} p(y_i|k, w_{z_i}, \alpha_{z_i}, \beta_j^{(t)}, \gamma_{z_i}, \mu_{z_i})}\right\} \tag{3.14}$$

$$\begin{aligned} A_{\gamma_j} &= \text{mín}\left\{1, \frac{\prod_{i:z_i=j}^{N} p(y_i|k, w_{z_i}, \alpha_{z_i}, \beta_{z_i}, \gamma_j^{new}, \mu_{z_i})}{\prod_{i:z_i=j}^{N} p(y_i|k, w_{z_i}, \alpha_{z_i}, \beta_{z_i}, \gamma_j^{(t)}, \mu_{z_i})}\right. \\ &\times \left.\frac{IG(\gamma_j^{new}|\alpha_0, \beta_0)}{IG(\gamma_j^{(t)}|\alpha_0, \beta_0)}\right\} \end{aligned} \tag{3.15}$$

$$\begin{aligned} A_{\gamma_j} &= \text{mín}\left\{1, \frac{\prod_{i:z_i=j}^{N} p(y_i|k, w_{z_i}, \alpha_{z_i}, \beta_{z_i}, \gamma_{z_i}, \mu_j^{new})}{\prod_{i:z_i=j}^{N} p(y_i|k, w_{z_i}, \alpha_{z_i}, \beta_{z_i}, \gamma_{z_i}, \mu_j^{(t)})}\right. \\ &\times \left.\frac{N(\mu_j^{new}|\xi, \kappa^{-1})}{N(\mu_j^{(t)}|\xi, \kappa^{-1})}\right\} \end{aligned} \tag{3.16}$$

A pesar de que no existe una expresión analítica para la distribución α-estable, es posible evaluar la verosimilitud $p(y_i|k, w_j, \theta)$ en las ecuaciones (3.13)-(3.16) mediante la integración numérica de la función característica (ec. (1.4)). En la literatura, existen multitud de métodos de estimación de la verosimilitud [Menn & Rachev, 2006; Nikias & Shao, 1995; Bodenschatz & Nikias, 1999; Kuruoglu, 2001; Nolan, 1997].

Los valores propuestos $\theta_j^{new} = \{\alpha_j^{new}, \beta_j^{new}, \gamma_j^{new}, \mu_j^{new}\}$ son extraídos de distribuciones simétricas $q(\theta_j^{new}|\theta_j^{(t)}) = q(\theta_j^{(t)}|\theta_j^{new})$. En particular, en este trabajo, elegimos $q(\cdot|\cdot)$ como la distribución Normal de varianza σ_θ y centrada en el valor actual de cada variable:

$$\theta_j^{new} \sim \frac{1}{\sqrt{2\pi}\sigma_\theta} \exp\{-\frac{(\theta_j^{new} - \theta_j^{(t)})^2}{2\sigma_\theta^2}\}. \tag{3.17}$$

3.3.3. Actualización de la asignación de variables (z)

Como en el caso simétrico, es necesario calcular la variable asignación en cada iteración. Una observación y_i pertenece al componente j de parámetros $\theta_j = \{\alpha_j, \beta_j, \gamma_j, \mu_j\}$ con probabilidad

$$p(z_i = j|...) = w_j p(y_i|k, w_j, \alpha_j, \beta_j, \gamma_j, \mu_j) \tag{3.18}$$

3.3.4. Actualización del número de componentes usando RJMCMC

Al contrario que en [Casarin, 2004], en el algoritmo presentado en este capítulo, el número de componentes k de cada parámetro puede cambiar en cada iteración. En particular, al igual que en el Capítulo 2, usaremos el algoritmo RJMCMC propuesto en [Green, 1995] y aplicado para mezcla de distribuciones en [Richardson & Green, 1997].

Por lo tanto, la flexibilidad de nuestro algoritmo es superior a [Casarin, 2004] debido a la capacidad de estimar el número de componentes k. Aunque no mostramos aquí los resultados, el movimiento 'birth-death' sugerido en [Richardson & Green, 1997] también fue implementado en el contexto de mezcla de distribuciones α-estables en [Salas-Gonzalez *et al.*, 2006a]. El problema que presenta este movimiento transdimensional es que la tasa de aceptación es muy baja, por lo que finalmente, sólo consideraremos un salto de tipo *split-combine*.

Este movimiento o salto transdimensional debe ser reversible (véase la Sección 1.2.2). Además, la validez del algoritmo RJMCMC no viene dada por la

elección de los nuevos valores de las variables sino por el uso de la expresión (1.32) para calcular la tasa de aceptación de los nuevos valores propuestos.

Dichos valores propuestos son los mismos que en el caso de mezcla α-estable simétrica (sec. 2.3.7),

$$w_{j*} = w_{j1} + w_{j_2} \tag{3.19}$$

$$w_{j*}\mu_{j*} = w_{j1}\mu_{j1} + w_{j_2}\mu_{j_2} \tag{3.20}$$

$$w_{j*}(\mu_{j*}^2 + \gamma_{j*}^2) = w_{j1}(\mu_{j1}^2 + \gamma_{j1}^2) + w_{j2}(\mu_{j2}^2 + \gamma_{j2}^2) \tag{3.21}$$

donde dos componentes j_1 y j_2 con pesos, dispersión y posición $\{w_{j1}, \gamma_{j1}, \mu_{j1}\}$ y $\{w_{j2}, \gamma_{j2}, \mu_{j2}\}$ respectivamente se combinan en un nuevo componente, que señalamos como j^*, con parámetros $\{w_{j*}, \gamma_{j*}, \mu_{j*}\}$. En este movimiento, también cambia la asignación de variables, por lo que para cada observación con $z_i = j_1$ o $z_i = j_2$ hacemos $z_i = j_*$.

El movimiento de combinación es determinista, es decir, dado los valores de los componentes $\{w_{j1}, \gamma_{j1}, \mu_{j1}\}$ y $\{w_{j2}, \gamma_{j2}, \mu_{j2}\}$, los nuevos valores $\{w_{j*}, \gamma_{j*}, \mu_{j*}\}$ están inequívocamente determinados. Sin embargo, el movimiento complementario contrario *split-move* no lo es. De hecho, el cambio de dimensión provoca la existencia de 3 grados de libertad, por lo que es necesaria la introducción de 3 variables aleatorias continuas. Al igual que en [Richardson & Green, 1997], estas variables las extraemos de tres distribuciones Beta con los siguientes parámetros:

$$u_1 \sim Be(2, 2)$$

$$u_2 \sim Be(2, 2)$$

$$u_3 \sim Be(1, 1)$$

Los nuevos valores propuestos para los pesos, la posición y la dispersión de los nuevos componentes $(w_{j1}, \gamma_{j1}, \mu_{j1})$ y $(w_{j2}, \gamma_{j2}, \mu_{j2})$ obtenidos a partir de los valores $(w_{j*}, \gamma_{j*}, \mu_{j*})$ del componente j_* existente son:

$$w_{j1} = w_{j*}u_1 \tag{3.22}$$

$$w_{j2} = w_{j*}(1 - u_1) \tag{3.23}$$

$$\mu_{j1} = \mu_{j*} - u_2\gamma_{j*}\sqrt{\frac{w_{j2}}{w_{j1}}} \tag{3.24}$$

$$\mu_{j2} = \mu_{j*} + u_2\gamma_{j*}\sqrt{\frac{w_{j1}}{w_{j2}}} \tag{3.25}$$

$$\gamma_{j1}^2 = u_3(1-u_2^2)\gamma_{j*}^2 \frac{w_{j*}}{w_{j1}} \tag{3.26}$$

$$\gamma_{j2}^2 = (1-u_3)(1-u_2^2)\gamma_{j*}^2 \frac{w_{j*}}{w_{j2}} \tag{3.27}$$

Tras proponer los nuevos valores, comprobamos que se cumple la ordenación de componentes en orden creciente del parámetro posición $[\mu_1 < \mu_2 < ... < \mu_k]$. En caso negativo, los nuevos valores propuestos son rechazados. En caso afirmativo, recalculamos la asignación de variables. Los valores antes indicados como j_*, que pasarán a ser j_1 o j_2, se recalculan mediante la expresión (3.18).

El exponente característico α y el parámetro de asimetría β también podríamos incluirlos en el paso RJMCMC. Sin embargo, hemos comprobado que esto ralentiza la tasa de aceptación de los nuevos valores del algoritmo. Por este motivo, mantenemos en memoria los últimos valores obtenidos para α_{j*} y β_{j*} en la última iteración en al que se obtuvo un número de componentes de la mezcla k^* determinado. Cada vez que este valor k^* es propuesto de nuevo, establecemos como nuevos valores del exponente característico y parámetro de asimetría, los valores guardados en memoria α_{j*} y β_{j*}.

Reemplazando la información sobre el movimiento de tipo *split/combine* en las ecuaciones (3.19)-(3.27) junto con las expresiones para las distribuciones a priori en la ecuación (1.32), podemos escribir la siguiente expresión para la tasa de aceptación A:

$$\begin{aligned}
A &= \frac{p(y|k+1, w_{j1}, w_{j2}, z_{j1}, z_{j2}, \theta_{j1}, \theta_{j2})}{p(y|k, w_{j*}, z_{j*}, \theta_{j*})} \\
&\times \frac{1}{a} \times \frac{1}{b} \times (k+1) \times \frac{w_{j1}^{\zeta-1+n_1} w_{j2}^{\zeta-1+n_2}}{w_{j*}^{\zeta-1+n_1+n_2} B(\zeta, k\zeta)} \\
&\times \sqrt{\frac{\kappa}{2\pi}} e^{-0{,}5\kappa\left\{(\mu_{j1}-\xi)^2+(\mu_{j2}-\xi)^2-(\mu_{j*}-\xi)^2\right\}} \\
&\times \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \left(\frac{\gamma_{j1}^2\gamma_{j2}^2}{\gamma_{j*}^2}\right)^{-\alpha_0-1} e^{-\beta_0(\gamma_{j1}^{-2}+\gamma_{j2}^{-2}-\gamma_{j*}^{-2})} \\
&\times \frac{d_{k+1}}{b_k P_{alloc}} \times \{g_{2,2}(u_1) g_{2,2}(u_2) g_{1,1}(u_3)\}^{-1} \\
&\times \frac{w_{j*}\,|\mu_{j1}-\mu_{j2}|\,\gamma_{j1}^2\gamma_{j2}^2}{u_2(1-u_2^2)(1-u_3)\gamma_{j*}^2}
\end{aligned} \tag{3.28}$$

donde n_1 y n_2 son el número de muestras del vector observación y_i cuya variable asignación es j_1 y j_2. $B(\cdot,\cdot)$ es la función Beta, P_{alloc} es la probabilidad de obtener los valores de la asignación de variables actuales y b_k y $d_k = 1 - b_k$ son las probabilidades de elegir un movimiento de tipo *split-move* o *combine-move* respectivamente.

Resumiendo, en cada iteración proponemos dos nuevos componentes con probabilidad b_k y lo aceptamos con probabilidad mín $\{1, A\}$ (o combinamos dos de ellos en uno con probabilidad $d_k = 1 - b_k$ y tasa de aceptación mín $\left\{1, A^{-1}\right\}$).

Por último, remarcar que no está permitido para este algoritmo proponer la combinación de componentes cuando $k = 1$ o la creación de dos nuevos componentes a partir de otro cuando $k > k_0$.

La primera línea de la expresión (3.28) es el cociente entre verosimilitudes. La segunda representa el cociente entre distribuciones a priori de los parámetros α, β, w y z. El término $k+1$ aparece debido a la ordenación de componentes en orden creciente del parámetro posición $\mu_1 < \mu_2 < ... < \mu_k$. La tercera y cuarta línea son el cociente entre las distribuciones a priori de μ y γ. La quinta línea es el cociente entre las distribuciones a partir de las cuales se proponen los nuevos valores de los parámetros y la última línea es el Jacobiano de la transformación entre los espacios de diferente dimensión.

3.4. Comparación con otros trabajos en la literatura

El único trabajo previo de estimación paramétrica en mezcla de distribuciones α-estables es un trabajo no publicado [Casarin, 2004]. Nuestro algoritmo presenta algunas ventajas importantes con respecto a dicho trabajo:

- En [Casarin, 2004], la estimación paramétrica está basada en el algoritmo de Gibbs propuesto por [Buckle, 1995]. El algoritmo presentado en esta memoria, en cambio, realiza una evaluación numérica de la densidad de probabilidad α-estable basada en el método de [Mittnik *et al.*, 1999]. Ambos métodos son ampliamente comparados en [Lombardi, 2007] donde se muestra que el algoritmo de Gibbs es mucho más costoso computacionalmente. Además, en todas las simulaciones presentadas en [Buckle, 1995] se eligieron unos valores iniciales de los parámetros muy cercanos a los valores verdaderos.

- En [Casarin, 2004], el número de componentes de la mezcla k es un parámetro fijo que debe ser introducido a priori mientras que en nuestro algoritmo,

k se calcula mediante un algoritmo RJMCMC.

- En [Casarin, 2004], se introduce una variable auxiliar de dimensión igual al número de observaciones o tamaño del vector observación y. Esta variable auxiliar debe calcularse completamente en cada iteración mediante muestreo por rechazo. Por lo tanto, la eficiencia de este algoritmo decrece enormemente conforme aumenta el número de observaciones. Más concretamente, el número de variables desconocidas que debe actualizarse en nuestro método, incluyendo el número de componentes, es $5k+1$ mientras que en [Casarin, 2004] es $5k + N$, donde N es la dimensión del vector observación.

- Además en [Casarin, 2004], se prueba el algoritmo con unas simulaciones en las cuales los componentes están claramente muy separados unos de otros. En dicho trabajo, el vector observación está compuesto por 1000 muestras con la distribución siguiente: $p_Y(y) = 0{,}5f_{1,7,0,3}(y|1,1) + 0{,}5f_{1,3,0,5}(y|1,30)$.

3.5. Resultados

3.5.1. Simulación 1: datos sintéticos

El algoritmo de inferencia Bayesiana en mezcla de distribuciones α-estables lo aplicaremos a un vector observación y de dimensión $N = 1500$ con la siguiente distribución:

$$\begin{aligned} p_Y(y) &= 0{,}4f_{1,2,0,5}(y|1,-4{,}25) + 0{,}2f_{1,2,0}(y|0{,}5,0{,}3) \\ &+ 0{,}4f_{1,5,0,5}(y|0{,}3,3{,}25). \end{aligned} \tag{3.29}$$

Este vector es obtenido mediante el método descrito en la Sección 1.1.4 y propuesto en [Chambers *et al.*, 1976].

En esta simulación, los valores elegidos para los hiperparámetros son los siguientes: $\alpha_0 = \beta_0 = 1$ son los dos parámetros de la distribución Gamma Inversa. La media y varianza de la distribución Gaussiana a priori de $p(\mu|\xi,\kappa^{-1})$ es $\xi = 0{,}2$ y $\kappa^{-1} = 1/5$. Además $\zeta = 1$ y, puesto que la longitud del dominio de las variables α y β es 2, por lo tanto $a = b = 2$. El número de iteraciones del algoritmo MCMC y RJMCMC descrito en la Sección 3.3 es de $N_{iter} = 10000$ con un periodo de calentamiento (*burn-in*) de 1000 iteraciones. $k_0 = 10$, aunque el número de componentes nunca excedió $k = 6$ en esta simulación. Tal y como

se detalló en la Sección 3.3.4, usamos el algoritmo de Metropolis para estimar los parámetros $\theta = \{\alpha, \beta, \gamma, \mu\}$. La desviación estándar σ_θ de la distribución Normal $q(\theta_j^{new}|\theta_j^{(t)})$, para cada uno de los parámetros de la distribución α-estable es $\sigma_\alpha = 0{,}15$, $\sigma_\beta = 0{,}1$, $\sigma_\gamma = 0{,}1$ y $\sigma_\mu = 0{,}2$. Inicialmente, consideramos que el número de componentes de la mezcla es $k = 6$ y los valores iniciales de cada una de las variables: $w_j = 1/6$, $\alpha_j = 1{,}1$, $\beta_j = 0$, $\gamma_j = 1$ para cualquier componente j y $\mu =$ [-3 -1 1 2 3 5].

En la Tabla 3.1, mostramos el valor real de cada parámetro, el valor estimado y la desviación estándar para el vector observación con distribución dada por la expresión (3.29). La desviación obtenida para el parámetro de asimetría β_1 es mayor que para cualquier otro parámetro. Esto está justificado debido al hecho de que conforme se incrementa el valor de α, el factor de asimetría β se hace irrelevante. Por lo tanto, para valores de α cercanos a 2, valores de β muy diferentes entre sí no cambian demasiado la forma de la distribución estable (véase la Figura 3.2).

La Figura 3.3 muestra el histograma con el número de componentes estimados tras el periodo de calentamiento suponiendo un modelo de mezcla de distribuciones α-estables y una comparación con el número de componentes obtenido asumiendo mezcla de Gaussianas [Richardson & Green, 1997]. Usando el algoritmo propuesto en este capítulo, obtenemos el número de componentes verdadero. Sin embargo, el modelo de mezcla de Gaussianas sobreestima el verdadero número de componentes debido al carácter impulsivo del vector de datos.

En la Figura 3.4, está representado el número de componentes estimado en cada iteración. El número verdadero de componentes es alcanzado por primera vez en, aproximadamente, un centenar de iteraciones. El número de componentes inicial se estableció en $k = 6$ para ambos casos (α-estable y Gaussiano).

En la Figura 3.5, representamos el histograma del vector observación y junto a la densidad obtenida usando el algoritmo de mezcla de α-estables. En la misma figura, se representa la densidad obtenida mediante un modelado con mezcla de distribuciones Normales [Richardson & Green, 1997]. La densidad α-estable obtenida ajusta satisfactoriamente el histograma discreto de y. Por otra parte, también se muestra que la mezcla de 3 Gaussianas no es suficiente para ajustar los datos.

La distribución α-estable tiene 4 parámetros (véase la sec. 1.1.1) mientras que la distribución Normal tiene sólo 2. Por lo tanto, para establecer una comparación equitativa entre ambos modelos compararemos la mezcla de 3 distribuciones α-estable (que posee 5 parámetros por cada componente) con 5 componentes

Tabla 3.1: Simulación 1: datos sintéticos. Valor verdadero de los parámetros, valor estimado mediante el algoritmo de mezcla de α-estables y error.

Parámetro	**Valor verdadero**	**Valor estimado**	**Desviación típica**
α_1	1.20	1.27	0.09
β_1	0.50	0.65	0.08
γ_1	1.00	0.98	0.06
μ_1	-4.25	-4.3	0.6
w_1	0.40	0.40	0.02
α_2	1.20	1.30	0.17
β_2	0.00	0.04	0.3
γ_2	0.50	0.45	0.05
μ_2	0.30	0.4	0.3
w_2	0.20	0.198	0.018
α_3	1.50	1.37	0.12
β_3	0.50	0.34	0.20
γ_3	0.30	0.295	0.016
μ_3	3.25	3.24	0.06
w_3	0.40	0.398	0.018

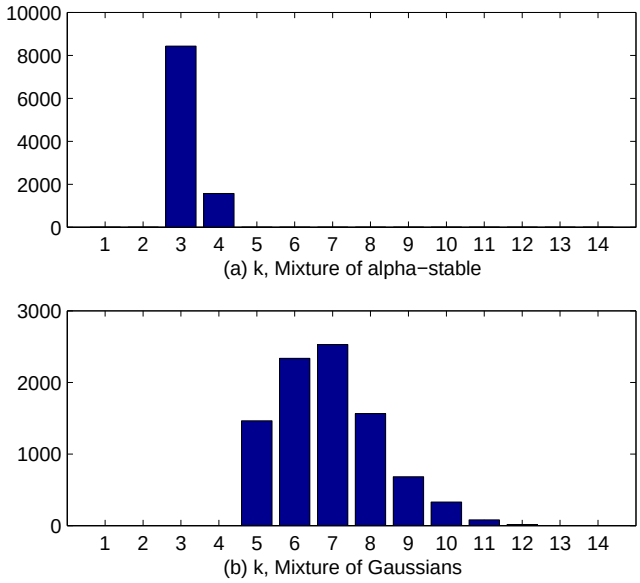

Figura 3.3: Histograma con el número de componentes estimado k. a) Mezcla de distribuciones α-estables. b) Mezcla de Gaussianas.

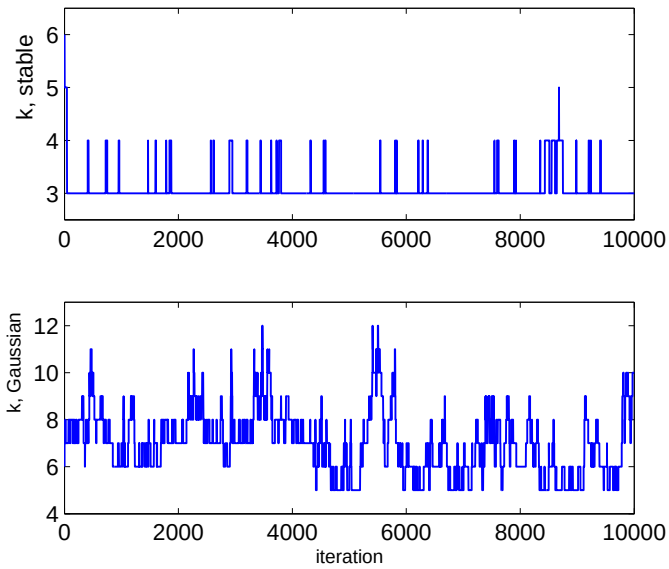

Figura 3.4: Evolución del número de componentes estimados en cada iteración. *Arriba.* Mezcla de distribuciones α-estables *: Mezcla de Gaussianas.*

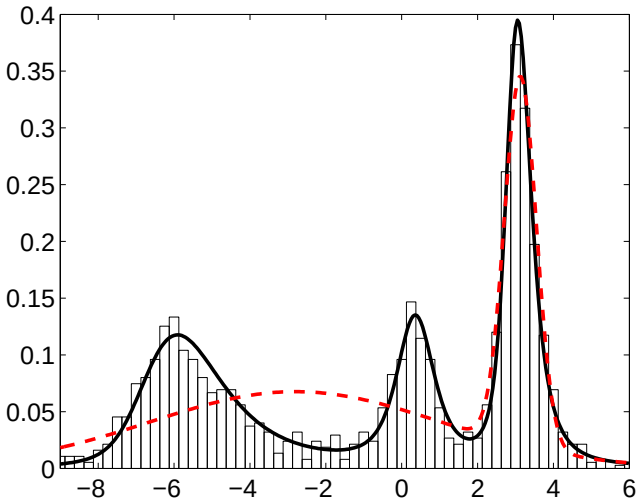

Figura 3.5: Histograma de las observaciones y con densidad α-estable. *Línea continua:* 3 componentes α-estables. *Línea discontinua:* 3 componentes con distribución Normal.

Tabla 3.2: Medidas de la distancia entre distribuciones. Comparación entre 3 distribuciones α-estables y 5 mezclas de Gaussianas.

Distancia	**Estable**	**Gaussiana**
Kullback-Leibler	0.0037	0.0141
Hellinger	0.0021	0.0060
χ^2	0.0080	0.0216

Gaussianos (ya que cada componente tiene 3 parámetros) De este modo, ambas mezclas de distribuciones tienen 15 parámetros.

En la Figura 3.6 se muestra conjuntamente el ajuste obtenido por la densidad Gaussiana con $k = 5$ y la distribución α-estable con $k = 3$ componentes. En esta figura, el funcionamiento de ambos métodos, al menos visualmente, no parece diferir demasiado debido a que la principal diferencia, para este ejemplo concreto, está en la parte de la distribución más alejada de los picos. Las colas de la distribución Normal tienen un comportamiento exponencial mientras que la distribución α-estable presenta comportamiento algebraico o Pareto (véase la sec. 1.1.5). Esta propiedad confiere a la distribución α-estable una mayor capacidad para modelar datos impulsivos comparado con la distribución Gaussiana. Con el fin de medir cuál de las dos distribuciones ajusta mejor los datos, hemos calculado la distancia de Kullback-Leibler, Hellinger y χ^2 [Borovkov, 1998]. Independientemente del método elegido, la distancia para el caso α-estable fue siempre menor tal y como se muestra en la Tabla 3.2, por lo tanto, la mezcla de 3 distribuciones α-estables ajusta mejor los datos analizados que la mezcla de 5 Gaussianas.

3.5.2. Simulación 2: datos sintéticos

En la simulación anterior, se calcularon varias distancias entre distribuciones y se comprobó que la mezcla de distribuciones α-estables ajustaba mejor al histograma del vector de datos y con distribución (3.29). No obstante, tal y como queda patente en la Figura 3.6, la principal diferencia entre el resultado obtenido usando mezcla de α-estables y mezcla de Gaussianas radicaba en el comportamiento asintótico en las colas de ambos modelos. Esta segunda simulación tiene como objetivo mostrar que, si el vector de observación tiene naturaleza aún más impulsiva que en el ejemplo anterior, el modelado median-

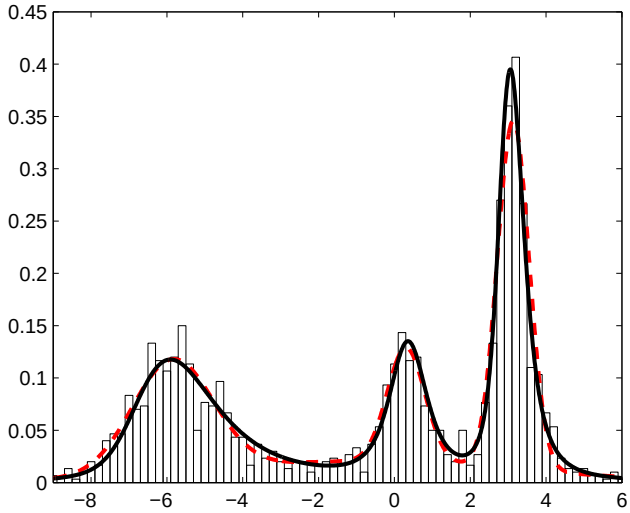

Figura 3.6: Simulación 1. Histograma de la distribución del vector observación y_i. *Línea continua:* mezcla de 3 distribuciones α-estables. *Línea discontinua:* mezcla de 5 componentes Normales.

te mezcla de Gaussianas no es válido y el modelo de mezcla α-estable se erige como una alternativa adecuada. Para ello simulamos un vector observación de longitud $N = 1500$ con un valor menor del exponente característico α que en la simulación 1, o lo que es lo mismo con una distribución más impulsiva (sec. 1.1.6).

$$\begin{aligned} p_Y(y) &= 0{,}4 f_{0{,}8,-0{,}5}(y|1,-1) + 0{,}3 f_{1,2,0}(y|0{,}3,0) \\ &+ 0{,}3 f_{0{,}8,-0{,}5}(y|0{,}6,4). \end{aligned} \tag{3.30}$$

Consideraremos en este caso los mismos valores para los hiperparámetros que en la simulación 1. La distribución del vector observación y fue ajustada usando un modelo de mezclas α-estable con 3 componentes y otro Gaussiano con 5, por el mismo motivo que el esgrimido en la Sección 3.5.1. En la Figura 3.7 se muestra la distribución obtenida con cada uno de los métodos junto con la distribución del vector observación y. En este caso, se comprueba claramente que la mezcla de Gaussianas no es capaz de modelar correctamente los datos. Esto prueba que, en caso de señales claramente impulsivas, la mezcla de distribuciones α-estables presenta una ventaja clara con respecto al modelo Gaussiano.

3.5.3. Simulación 3: datos reales

Debido a lo extendido que está el uso de la distribución α-estable en economía, el primer conjunto de datos reales en el que aplicaremos el modelo de mezcla α-estable consiste en datos diarios del indicador trimestral de los tipos de interés en eurodepósitos en Francia entre el 01/01/1988 y el 13/01/2003. Este conjunto de datos fue estudiado en [Casarin, 2004]. La Figura 3.8 muestra conjuntamente el histograma de los datos y la densidad mezcla de 2 distribuciones α-estables obtenida mediante el procedimiento descrito en este capítulo. En la Tabla 3.3, presentamos la comparación entre los valores obtenidos usando nuestro algoritmo y los obtenidos en [Casarin, 2004].

3.5.4. Simulación 4: datos reales

El segundo conjunto de datos reales en el que vamos a probar el funcionamiento de nuestro algoritmo es el mismo que ya se estudió en la Sección 2.4.3. Se trata de 245 valores de la actividad enzimática en la sangre [Bechtel *et al.*, 1993]. Existen claramente, según la figura, dos diferentes subpoblaciones de metabolizadores que podríamos considerar lentos y rápidos. La Figura 3.9 muestra conjuntamente el histograma discreto y el ajuste con mezcla de 2 distribuciones

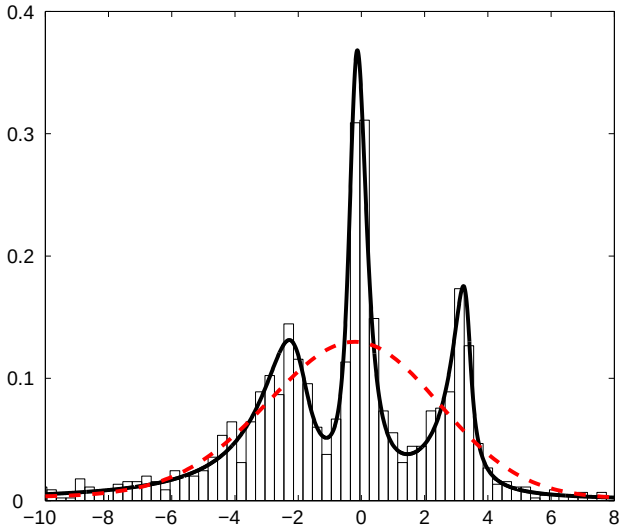

Figura 3.7: Simulación 2. Histograma del vector observación y_i con distribución dada por la ec. (3.30). *Línea continua:* Mezcla de 3 distribuciones α-estables. *Línea discontinua:* Mezcla de 5 distribuciones Normales.

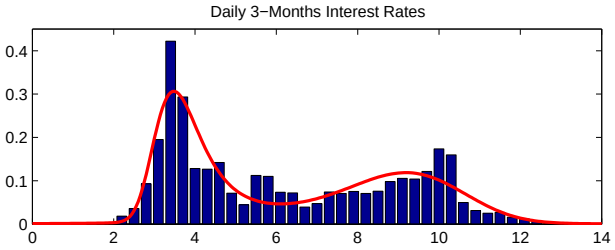

Figura 3.8: Histograma de los datos diarios del indicador trimestral de los tipos de interés en eurodepósitos en Francia entre 01/01/1988 y 13/01/2003. *Línea continua:* Mezcla de 2 distribuciones α-estable.

Tabla 3.3: Valores estimados mediante la distribución α-estable para los tipos de interés en Francia. *Estable(1):* algoritmo propuesto en esta memoria. *Estable(2):* algoritmo propuesto en [Casarin, 2004]

Parámetro	**Valor inicial**	**Estable(1)**	**Estable(2)**
α_1	1.5	1.3	1.2
α_2	1.5	1.7	1.2
β_1	0.01	0.97	0.02
β_2	0.01	-1.00	0.04
γ_1	1.5	0.493	0.307
γ_2	1.5	1.129	0.873
μ_1	4	4.443	3.012
μ_2	10	8.505	7.301
w_1	0.5	0.535	N/A
w_2	0.5	0.465	N/A

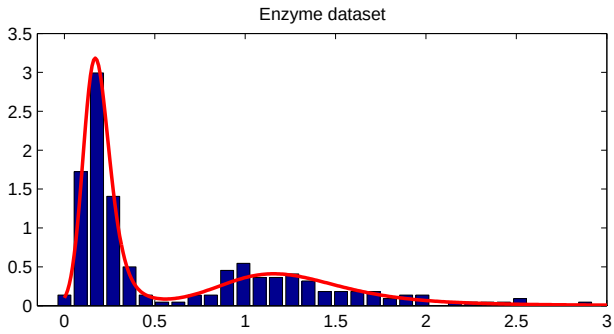

Figura 3.9: Actividad enzimática en la sangre para una enzima involucrada en el metabolismo de sustancias cancerígenas. *Línea continua:* mezcla de α-estables.

α-estables. En la Tabla 3.4 se presentan los valores obtenidos para los parámetros de la mezcla α-estable.

3.6. Conclusiones

En este capítulo hemos presentado el modelo de mezcla de distribuciones α-estables. Este modelo es una generalización del estudiado en el capítulo anterior. La estimación de los parámetros ha sido realizada mediante la construcción de un modelo jerárquico Bayesiano. Por otra parte, en dicho modelo, suponemos desconocido el número de componentes en la mezcla. Estimamos este parámetro mediante el método Monte Carlo de saltos reversibles (RJMCMC). La ausencia de una expresión analítica para la densidad de probabilidad de la distribución α-estable es solventada mediante la resolución numérica de la integral de la función característica α-estable.

El algoritmo propuesto ha sido probado tanto en datos sintéticos como reales. En todas las simulaciones realizadas, cada uno de los parámetros desconocidos del problema ha sido estimado correctamente. Además, en cada caso, el método propuesto ha obtenido el número verdadero de componentes. El algoritmo ha sido comparado con otros trabajos en la literatura. El modelo de mezcla α-estable presentado en este capítulo presenta muchas ventajas con respecto a el único trabajo previo de mezcla α-estable. Por ejemplo, es considerablemente más rápido, es más sencillo de implementar y además converge para datos reales donde los distintos componentes están muy mezclados entre sí. Para poner de

Tabla 3.4: Valores estimados de los parámetros de la mezcla de distribuciones α-estables para los 245 valores de la actividad enzimática en la sangre.

Parámetro	**Valor estimado**
α_1	1.6620
α_2	1.5545
β_1	0.8930
β_2	0.9064
γ_1	0.0552
γ_2	0.2589
μ_1	0.2047
μ_2	1.3911
w_1	0.6239
w_2	0.3761

manifiesto los diferentes campos donde puede encontrar aplicación el modelo de mezclas α-estable, los datos reales analizados proceden de dos disciplinas dispares, como son la biología y la economía.

En este capítulo, la mezcla de α-estables ha sido comparada con el modelo de mezclas de Gaussianas. Hemos probado que nuestro método es mucho más adecuado para el análisis de componentes impulsivos y asimétricos. Además el modelo presentado en este capítulo, es una generalización del modelo de mezclas Gaussianas, por lo que comparte muchas de las propiedades de aquél, con la flexibilidad añadida de que las mezclas estables nos permiten modelar componentes que presentan alto grado de impulsividad.

Parte II

Expresión genética y distribución α-estable

CAPÍTULO

CUATRO

FUNDAMENTO TEÓRICO: EXPRESIÓN GENÉTICA

4.1. Micromatrices y expresión genética

El estudio del ADN mediante micromatrices se basa en una colección de puntos microscópicos de ADN, cada uno de ellos representando un gen determinado, dispuestos en filas y columnas en una superficie sólida. Con dicho dispositivo es posible realizar medidas tanto cuantitativas como cualitativas de la expresión genética bajo distintas condiciones. Típicamente, un experimento de micromatrices, nos permite el estudio simultáneo de decenas de miles de genes.

4.1.1. Tecnología de micromatrices

El objetivo de la mayoría de experimentos de micromatrices es examinar la expresión genética mediante el análisis del nivel de expresión genética de miles de genes simultáneamente en lugar de realizar un análisis individual de cada gen. Normalmente existen decenas de miles de celdas en cada matriz. Cada una de estas celdas contiene una gran cantidad de moléculas de ADN idénticas (o fragmentos de moléculas idénticas), con una longitud que va de una veintena a varios centenares de nucleótidos.

En los estudios de la expresión genética mediante micromatrices, idealmente, cada una de estas moléculas identifica un gen o un exon del genoma. En la práctica, esto no siempre es así, debido a la existencia de familias de genes

similares en el genoma.

A continuación se resumen los pasos más importantes en el estudio de la expresión genética mediante tecnología de micromatrices. A grandes rasgos, estos pasos incluyen diseño experimental, extracción de datos, adquisición de imágenes y análisis de las mismas, preprocesado y normalización de datos; y por último, el uso de técnicas estadísticas para el estudio del conjunto de datos genéticos.

1. Se aisla ARN bien de diferentes tejidos o del mismo tejido pero distintos estados de desarrollo o de tejidos enfermos o de muestras sujetas a tratamientos o procedentes de diferentes condiciones experimentales.

2. El ARN es coloreado mediante dos tintas fluorescentes e hibridizado en la micromatriz. Esto permite la comparación de los niveles de expresión genética entre los pares de muestras.

3. La micromatriz es escaneada para obtener una imagen TIFF con el nivel de fluorescencia de cada celda de la matriz.

4. Estas imágenes se analizan y se mide la intensidad relativa para cada celda fluorescente de la matriz, descartándose aquellas celdas cuya imagen es defectuosa. Cada celda que proporciona una imagen de la intensidad relativa de calidad, se normaliza convenientemente. En la Figura 4.1, se muestra una imagen típica de una micromatriz tras el proceso de hibridación, con las distintas intensidades para cada celda.

5. Finalmente, los datos obtenidos ya están listos para realizar el análisis de la expresión genética.

En la Figura 4.2, se muestra un esquema de los pasos necesarios para obtener los datos genéticos de micromatrices.

4.1.2. Normalización

Los distintos tipos de métodos de normalización de datos de micromatrices existentes asumen lo siguiente:

- La mayoría de los genes estudiados no están diferentemente expresados. Es decir, el valor de M es cercano a 0.

- El número de genes regulados por incremento (*up-regulated*)y por decremento (*down-regulated*) es pequeño y aproximadamente el mismo.

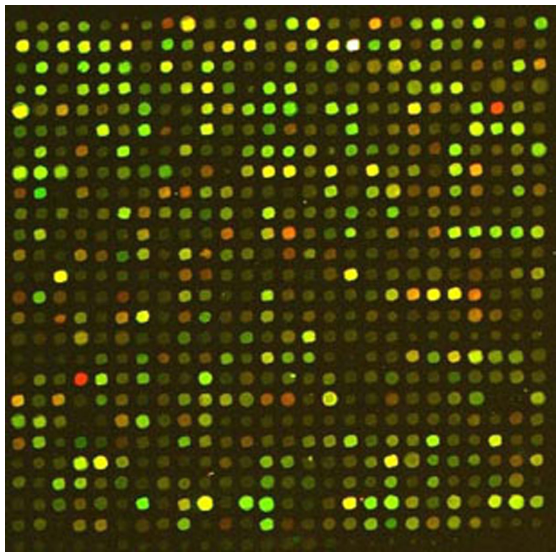

Figura 4.1: Imagen típica de una micromatriz escaneada tras el proceso de hibridación.

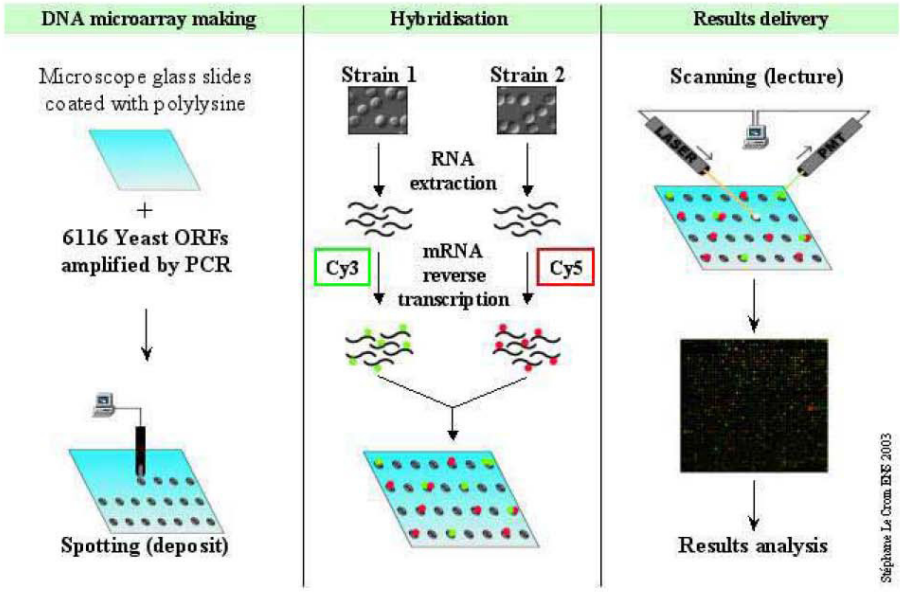

Figura 4.2: Esquema de los pasos necesarios para obtener los datos genéticos a partir de la tecnología de micromatrices.

- Los genes están expresados para un amplio rango de la intensidad total A.
- Los valores normalizados son representativos del genoma estudiado.

En esta Memoria, normalizamos los datos procedentes de micromatrices mediante regresión local, más concretamente usando el método *LOWESS*. Este método de normalización nos permite, por un lado, centrar la distribución de la expresión de genes en cero y por otro, eliminar, en los datos de micromatrices, la dependencia con la intensidad. Está ampliamente estudiado que los valores $log(R/G)$ tienen una dependencia con la intensidad de la fluorescencia medida. Este hecho hace que exista una desviación con respecto al cero para valores con baja intensidad. La regresión local *LOWESS* es capaz de eliminar dicha dependencia. Básicamente, este método consiste en la representación de las medidas del logaritmo de la intensidad relativa $lo7g(R/G)$ en función de la intensidad $log(R \cdot G)$. El algoritmo *LOWESS* detecta la desviación sistemática en la figura y la corrige mediante el cálculo de la recta que mejor ajusta los datos usando regresión local.

4.1.3. Aplicación de los modelos de mezcla en micromatrices

Sea N el número de genes en cada micromatriz, n el número de repeticiones para cada gen y $M_{ij} = log(R_{ij}/G_{ij})$,$i = 1 \ldots N$ y $j = 1 \ldots n$, el logaritmo de las intensidades relativas de las fluorescencias roja R y verde G. Modelamos M_{ij} como una variable aleatoria con distribución normal con media μ_i y varianza σ_i^2, de modo que:

$$M_{ij}|\mu_i, \sigma_i \sim N(M_{ij}|\mu_i, \sigma_i^2) \tag{4.1}$$

Los modelos de mezcla suponen dos tipos distintos de genes, expresados y no expresados. Cada uno de ellos se modela con una distribución diferente mediante un modelo de mezcla de dos distribuciones:

$$\mu_i \sim wP + (1 - w)\delta(0), \tag{4.2}$$

donde w es la proporción de genes expresados, $\delta(0)$ es la distribución de Dirac y P es una distribución centrada en cero y generalmente simétrica. En la literatura, se han propuesto distintas familias de distribuciones para P, entre ellas la distribución Gaussiana [Lonnstedt & Speed, 2002] y la de Laplace [Bhowmick *et al.*, 2006]. En el Capítulo 6 de esta Tesis Doctoral, proponemos el diseño

de un estadístico para establecer si un determinado gen está expresado usando un modelo de mezcla similar al de la ecuación (4.2), siendo la distribución P, una distribución α-estable simétrica. Además, en dicho capítulo, consideramos la distribución a priori para σ_i^2 como una distribución α-estable positiva. Por lo tanto, en virtud de la propiedad del producto presentada en la Sección 1.1.3 de este capítulo, la distribución de M_{ij} está modelada como una α-estable simétrica. Lo cual está en concordancia con los resultados experimentales que se presentarán con detalle en el Capítulo 5.

CAPÍTULO

CINCO

MODELADO DE MICROMATRICES MEDIANTE DISTRIBUCIONES α-ESTABLES

Una vez normalizados los datos de micromatrices, la distribución de la expresión de genes posee una forma similar. En general, presentan colas con peso superior a la distribución Gaussiana y cierto grado de asimetría. La distribución de la expresión de genes ha sido modelada en la literatura mediante distintas familias de distribuciones, como la distribución de Laplace Asimétrica [Hoyle *et al.*, 2002] o la distribución de Cauchy [Khondoker *et al.*, 2006]. Otra propiedad de esta distribución es que las colas tienen un comportamiento asintótico potencial y no exponencial. Además, la varianza de una matriz determinada crece con el número de genes. Todas estas características sugieren que la distribución α-estable es adecuada para modelar esta distribución.

En este capítulo, modelaremos la distribución de la expresión de genes de cuatro distintos conjuntos de datos reales usando la distribución α-estable [Salas-Gonzalez *et al.*, 2006d, 2007b; Kuruoglu *et al.*, 2007]. Además compararemos con la distribución de Laplace Asimétrica y Gaussiana, mediante la medida de la distancia entre distribuciones usando la distancia de Kullback-Leibler, χ^2 y Hellinger. Como veremos, la distribución α-estable modela mucho mejor los datos de micromatrices que el resto de distribuciones comparadas. Además, la distribución de la expresión de genes posee características empíricas similares a las de la distribución α-estable.

Este capítulo está estructurado del siguiente modo: en la Sección 5.1 se presentan algunos de los trabajos en la literatura que han modelado la distribución

de la expresión de genes con distintas distribuciones estadísticas. En la Sección 5.2, se explica la relación entre la distribución de la expresión de genes y la distribución α-estable. Además se presentan los resultados obtenidos tras modelar dicha distribución con la α-estable. En la Sección 5.3, se discuten los resultados y se comparan con otros trabajos previos en la literatura. Finalmente, en la Sección 5.4, se presentan las conclusiones de este capítulo.

5.1. Introducción

El estudio del ADN mediante el uso de micromatrices es una herramienta que permite el estudio simultáneo de miles de genes bajo diferentes condiciones. Estos experimentos comparan dos muestras diferentes de ADNc tintados con dos colores diferentes, normalmente rojo y verde. Tras el proceso de hibridación, mediante luz fluorescente se mide la intensidad de luz reflejada en cada celda de la micromatriz. Mayor intensidad se corresponde con mayor cantidad de ARN. Este método permite comparar una larga cantidad de información genética simultáneamente, con el fin de identificar qué genes están expresados bajo distintas condiciones.

Los datos genéticos procedentes de micromatrices no son independientes entre sí. Sin embargo, existen muchos trabajos en la literatura que suponen la independencia entre genes como hipótesis de trabajo [Lonnstedt & Speed, 2002; Gottardo *et al.*, 2003; Bhowmick *et al.*, 2006]. Además, los métodos basados en estadística Bayesiana, generalmente, asumen independencia entre genes. Dicha suposición, junto con la suposición Gaussiana se usa en multitud de trabajos como estrategia para obtener fórmulas analíticas. Aparte de la distribución Normal, la distribución de la expresión de genes ha sido modelada mediante diferentes distribuciones y métodos:

- En [Kuznetsov, 2001], se modela esta distribución usando diferentes clases de densidades asimétricas como la distribución de Poisson, exponencial, series logarítmicas y la distribución de Pareto. En dicho trabajo, muestra el resultado obtenido para la distribución Pareto, ya que es la que ofrece un mejor ajuste de la distribución de genes.

- En [Hoyle *et al.*, 2002], se analizan una gran cantidad de datos genéticos reales. La distribución de genes se aproxima mediante dos distribuciones: una Log-normal en el centro y una ley de potencias en las colas. Además, en dicho artículo, se resalta que la varianza de las intensidades muestra una correlación positiva con el número de genes. Es decir, la varianza de

los datos genéticos de micromatrices no se estabiliza conforme aumenta el número de datos estudiados.

- En [Purdom & Holmes, 2005], la distribución de la expresión genética se ajusta mediante la distribución de Laplace Asimétrica. La mejora de este método con respecto al uso de la distribución Gaussiana es notable, debido a que esta distribución presenta mayor peso en las colas y asimetría.
- En [Khondoker *et al.*, 2006], se presenta un modelo estadístico para estimar la expresión genética usando datos procedentes de múltiples escaneados mediante láser. En dicho trabajo, la distribución de la expresión de genes se modela mediante una distribución de Cauchy.

En este capítulo, propondremos el modelado de la distribución de genes usando la distribución α-estable. Además, el uso de la distribución α-estable tiene diversas ventajas teóricas comparado con las que presentan otras distribuciones.

5.2. Análisis de datos de micromatrices

Modelaremos la distribución de la expresión genética mediante la distribución α-estable para 4 distintos datos de micromatrices. El primer conjunto de datos (que nombraremos *self-self*) consiste en la auto hibridación de 19 células humanas cancerígenas, la referencia universal Stratagene ARN y ARN aislado de una muestra infectada [Yang *et al.*, 2002a]. El segundo conjunto de datos (*zebrafish*), consiste en dos muestras de experimentos para un total de 4 distintas repeticiones. Para cada una de estas hibridaciones, ADNc de *swirl* mutante fue tintado de color rojo y el tipo *wild-type mutant* mediante otro color distinto (verde). Este experimento fue realizado usando el pez cebra, que es un organismo que proporciona información valiosa sobre los vertebrados primitivos o poco desarrollados. El nombre *swirl* se corresponde con una mutación en el gen BMP2 que afecta el eje dorsal del organismo[1]. El tercer conjunto de datos estudiado (*lymphoma*) consiste en muestras de ADN de tumores procedente de pacientes con un tipo particular de linfomas [Alizadeh *et al.*, 2000]. El último de los conjuntos de datos que estudiaremos en este capítulo (*yeast*) consiste en un análisis de la variación y diferencia entre dos muestras, una de ellas sometida a estrés y otra de referencia, del organismo *Saccharomyces Cerevisiae* [Yvert *et al.*, 2003]. Estos cuatro conjuntos de datos reales fueron analizados en [Purdom & Holmes, 2005], por lo tanto es posible comparar los resultados obtenidos allí con los de

[1] Este conjunto de datos está disponible con el paquete marrayClasses del software R.

este capítulo. Antes de nada, normalizamos los datos usando el método de regresión local (LOWESS) [Cleveland & Delvin, 1988] (véase la Sección 4.1.2). La normalización nos permite eliminar la dependencia de los datos con la intensidad en la expresión del logaritmo de las intensidades relativas. El método de regresión local ha sido eficazmente usado en el contexto de datos genéticos procedentes de micromatrices [Yang *et al.*, 2002b]. Tras la normalización, cada distribución de la expresión de genes tiene una forma similar: mayor peso en las colas que la distribución Gaussiana y cierto grado de asimetría.

Existen diferentes métodos de estimación de parámetros de la distribución α-estable [Kuruoglu, 2001; Kogon & Williams, 1998]. Para cada micromatriz, estimamos los parámetros usando el método de maximización de la verosimilitud [Nolan, 2001]. La estimación de parámetros obtenida se muestra en la Figura 5.1. Se comprueba que la diferencia entre el parámetro posición μ y la dispersión γ para el conjunto de datos *self-self* es muy pequeña. Para este conjunto de datos, el mismo ARN es tintado separadamente de verde y rojo e hibridizado en la misma micromatriz, por lo que, en este caso, en ausencia de errores experimentales sistemáticos, la distribución de la expresión de genes debe ser simétrica. De hecho, obtenemos valores del factor de asimetría β muy cercanos a cero en casi todos los casos. Existen sólo tres en los que el parámetro de asimetría β no es cercano a cero: las matrices 8, 18 y 24, (en la Figura 5.1 se denotan con círculos en vez de puntos). Para estas tres matrices se obtienen los siguientes valores de los parámetros $\{\alpha_8 = 1{,}83, \beta_8 = -0{,}48\}$, $\{\alpha_{18} = 1{,}94, \beta_{18} = -0{,}65\}$ y $\{\alpha_{24} = 1{,}86, \beta_{24} = -0{,}77\}$, por lo que el exponente característico α es muy cercano a 2. Recordemos que una de las propiedades de la distribución α-estable es que conforme el parámetro α se aproxima a 2, la distribución se hace más simétrica y el parámetro de asimetría β afecta menos a la forma de la distribución (véase la sec. 3.2.2). Por lo tanto, en este caso, puede considerarse que los valores obtenidos para el parámetro β son consistentes con la simetría de la distribución.

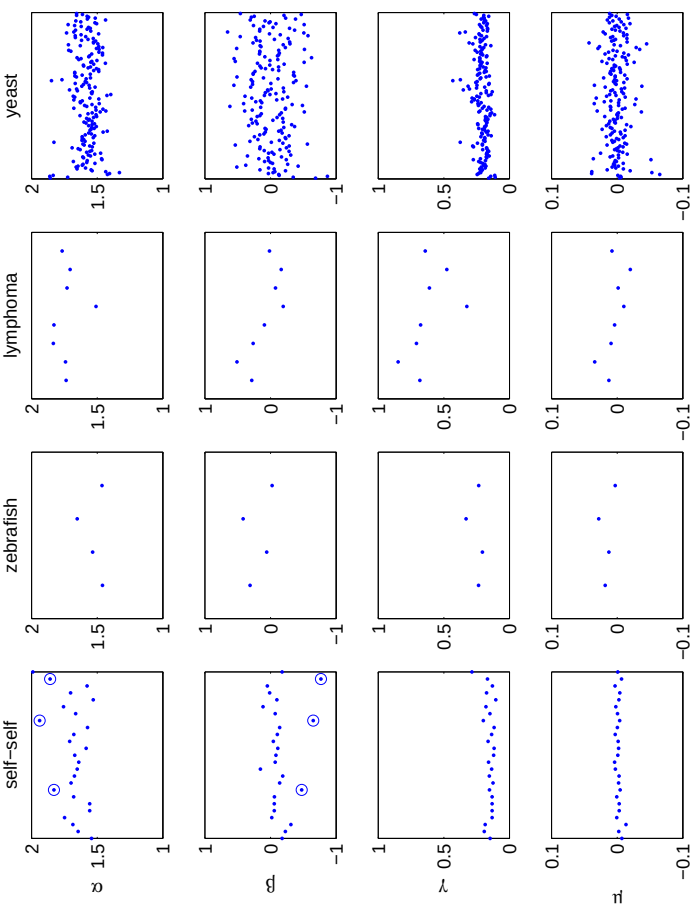

Figura 5.1: Valores de los parámetros obtenidos para cada uno de los cuatro conjuntos de datos. *Primera fila:* exponente característico α. *Segunda fila:* parámetro de asimetría β. *Tercera fila:* dispersión γ. *Cuarta fila:* posición μ.

La Figura 5.2 muestra la distribución de la expresión de genes para una micromatriz de cada uno de los conjuntos de datos considerados. A simple vista, se comprueba que la distribución α-estable ajusta el histograma con los datos mejor que la distribución de Laplace Asimétrica y la distribución Normal. También se comprueba que, la distribución de Laplace presenta un pico que no se observa en los datos de expresión genética. Por ejemplo, para los datos *self-self* la distribución de Laplace modela peor que para los datos *yeast*. La distribución α-estable, sin embargo, presenta un comportamiento más suave en su valor máximo, al igual que la que presenta en todos los casos el histograma de los datos estudiados. En la Figura 5.2, además, se comprueba que la distribución Gaussiana no es capaz de ajustar los datos, ya que el histograma presenta mayor peso en las colas que la distribución Normal.

Además de la comparación visual, calculamos la distancia de Kullback-Leibler, χ^2 y Hellinger [Borovkov, 1998] con el fin de comparar numéricamente cómo la distribución α-estable, Laplace Asimétrica y Normal ajustan los datos estudiados. La distancia χ^2 penaliza los posibles valores atípicos (*outliers*). Es decir, una pequeña cantidad de datos atípicos afecta más el valor obtenido para la distancia χ^2 que para la de *Hellinger* o *K-L*. De las tres, la distancia *K-L* es la más robusta frente a datos atípicos. Calculamos estas tres distancias para cada micromatriz. En la Tabla 5.2 se muestra la media y desviación típica para cada uno de los 4 conjuntos de datos estudiados. En todos los casos, la distribución α-estable presentó un mejor ajuste a los datos. Además, la desviación estándar es considerablemente menor para la distribución α-estable. Esto quiere decir que, tanto la distribución de Laplace Asimétrica como la Normal, al contrario que la α-estable, a veces ajusta la distribución de la expresión de genes de manera satisfactoria y otras lo hace francamente mal, dependiendo de la micromatriz considerada. Este hecho se ilustró mediante la representación gráfica de la Figura 5.2, donde se comprobó de manera visual, que el acentuado pico en la moda de la distribución de Laplace, propiciaba que, dependiendo de los datos considerados, éstos se ajusten o no con dicha distribución. Los valores más bajos para la distancia y la desviación típica obtenidos usando los métodos K-L, χ^2 y Hellinger en el caso de la distribución α-estable, muestran que dicha distribución ajusta mucho mejor los datos de micromatrices.

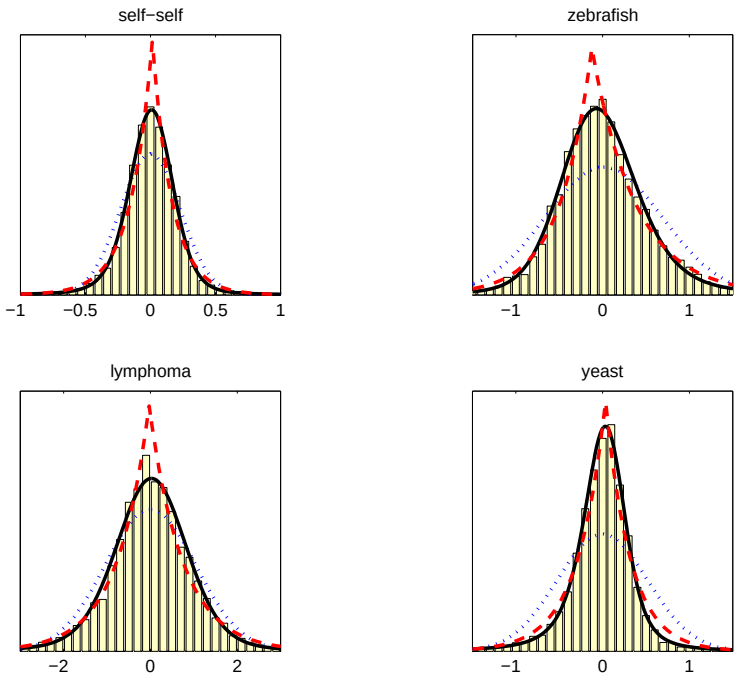

Figura 5.2: Histogramas con la distribución de los datos para un ejemplo de cada uno de los conjuntos de datos estudiados. Para los datos *self-self* representamos la matriz 9 (NT2.2(testis)). Para *zebrafish* y *lymphoma*, la matriz 2 y *DLCL-0024* respectivamente. Para los datos *yeast*, usamos la matriz de nombre *14-4-aCy3*. *Línea continua:* Distribución α-estable. *Línea discontinua:* Distribución de Laplace Asimétrica. *Línea punteada:* Distribución Gaussiana.

Tabla 5.1: Distancias de Kullback Leibler, χ^2 y Hellinger entre el histograma de los datos de expresión genética y los correspondientes ajustes mediante las distribuciones α-estable, Laplace Asimétrica y Gaussiana. Los valores son la media para cada uno de las micromatrices estudiadas. Entre paréntesis, el error calculado como la desviación estándar. En negrita los valores menores para la distancia y la desviación estándar.

	self-self	zebrafish	lymphoma	yeast
KL(Stable)	**0.013 (0.005)**	**0.0171 (0.0020)**	**0.021 (0.003)**	**0.018 (0.005)**
KL(ALaplace)	0.022 (0.019)	0.066 (0.021)	0.022 (0.003)	0.047 (0.017)
KL(Gauss)	0.10 (0.04)	0.35(0.10)	0.07 (0.03)	0.25 (0.09)
χ^2(Stable)	**0.015 (0.008)**	**0.015 (0.003)**	**0.022 (0.004)**	**0.016 (0.005)**
χ^2(ALaplace)	0.04 (0.06)	0.074 (0.019)	0.038 (0.008)	0.058 (07.022)
χ^2(Gauss)	0.12 (0.05)	0.6 (0.3)	0.09(0.05)	0.39 (0.21)
Hell.(Stable)	**0.0036 (0.0021)**	**0.0043 (0.0007)**	**0.0061 (0.0011)**	**0.0044 (0.0014)**
Hell.(ALaplace)	0.009 (0.009)	0.020 (0.005)	0.0087 (0.0015)	0.015 (0.005)
Hell.(Gauss)	0.030 (0.012)	0.11 (0.04)	0.025 (0.012)	0.07 (0.03)

5.3. Discusión

En la sección anterior, mostramos que la distribución α-estable puede ajustar de manera muy satisfactoria la distribución de la expresión de genes. En esta sección, comparamos el modelo estable propuesto con otros 4 distintos modelados de la distribución de la expresión de genes en la literatura:

- En [Kuznetsov, 2001], se modela la distribución de la expresión de genes mediante distintas familias de distribuciones asimétricas (aunque no usó la distribución α-estable). Finalmente, comprobó que la distribución Pareto era la que mejor se ajustaba a los datos de expresión genética. Para ello, tuvo que introducir además un parámetro posición adicional para generalizar la distribución Pareto. La distribución α-estable ya cuenta con dicho parámetro posición y proporciona un buen ajuste tanto en el centro de la distribución como en las colas. Además, la distribución α-estable también posee comportamiento asintótico de tipo Pareto (ley de potencias) en las colas cuando $\alpha < 2$. En concreto, sea X una variable aleatoria con distribución α-estable y exponente característico $\alpha < 2$, entonces [Samorodnitsky & Taqqu, 1994]:

$$\lim_{\lambda \to -\infty} \lambda^{\alpha} \{P < \lambda\} = C_{\alpha} \frac{1+\beta}{2} \gamma^{\alpha} \tag{5.1}$$

$$\lim_{\lambda \to -\infty} \lambda^{\alpha} \{P < -\lambda\} = C_{\alpha} \frac{1-\beta}{2} \gamma^{\alpha} \tag{5.2}$$

  donde

$$C_{\alpha} = \frac{1-\alpha}{\Gamma(2-\alpha)\cos(\pi\alpha/2)} \text{ si } \alpha \neq 1 \tag{5.3}$$

$$C_{\alpha} = \frac{2}{\pi} \text{ si } \alpha = 1 \tag{5.4}$$

  Por otra parte, Mandelbrot hizo hincapié en los primeros trabajos de aplicación de la distribución α-estable, en el hecho de que el uso de esta distribución para el estudio y descripción de datos biológicos era preferible al uso de distribuciones de tipo Zipf-Pareto debido a razones tanto teóricas como prácticas. (Véase [Zolotarev, 1986] para una explicación más detallada de dicha afirmación).

- En [Hoyle *et al.*, 2002], se analiza una gran cantidad de datos genéticos de micromatrices. La distribución de la expresión de genes se aproxima por

una Log-normal en el centro de la distribución. Las colas de la distribución, sin embargo, siguen una ley de potencias o ley Zipf, que es un caso especial de comportamiento asintótico de tipo Pareto [Newman, 2005]. Por lo tanto, en dicho trabajo, se usan dos distribuciones distintas para el modelado del histograma de los datos genéticos de micromatrices, Log-normal en el centro y ley de potencias en las colas. Dos explicaciones heurísticas para este distinto comportamiento se dan en [Hoyle *et al.*, 2002]. La distribución α-estable permite el modelado de la distribución de la expresión de genes de manera más compacta, mediante el uso de una sola distribución.

Además, en [Hoyle *et al.*, 2002], se apunta que la varianza σ^2 de las intensidades logarítmicas aumenta con el número de genes estudiado. Este resultado está en completo acuerdo con las propiedades de la distribución α-estable tal y como fue establecido en la Sección 1.1 La varianza es un parámetro que no está definido para procesos estables con $\alpha < 2$. La desviación estándar de una variable aleatoria α-estable aumenta conforme la longitud del vector considerado aumenta y no converge a un valor determinado, como sí lo hace en vectores de datos con distribución Gaussiana.

- En [Purdom & Holmes, 2005], la distribución de la expresión de genes se modela mediante la distribución de Laplace Asimétrica y se compara el ajuste con la distribución Normal. Por otro lado, en la sección anterior (sec. 5.2), se comparó el funcionamiento de la distribución α-estable y la de Laplace. Además, se mostró que la distribución de Laplace Asimétrica no siempre es capaz de ajustar convenientemente el histograma de intensidades de la expresión genética en datos de micromatrices. Dicho histograma, normalmente presenta un comportamiento más suave alrededor del máximo que la distribución de Laplace. Además, aunque la distribución de Laplace Asimétrica posee colas de mayor peso que la Normal, el comportamiento asintótico de dichas colas es exponencial, en lugar de potencial, y no sigue un comportamiento tipo ley de Pareto. Recientemente, esta suposición ha sido aceptada para la identificación de qué genes están realmente expresados y cuáles no [Bhowmick *et al.*, 2006]. Nuestra propuesta de modelar mediante la distribución α-estable la distribución de la expresión de genes es un primer paso para el desarrollo de estadísticos que permitan establecer si un gen está expresado o no. Éste será el objetivo del próximo capítulo de esta memoria.

- En [Khondoker *et al.*, 2006], la distribución de la expresión de genes se modela mediante una distribución de Cauchy. Dicha distribución se elige para

poder tener en cuenta los valores atípicos que la distribución Normal no puede modelar. Nosotros, por otra parte, no asumimos que la distribución de la expresión de genes es Normal o Cauchy, pero ambas distribuciones son casos particulares de la distribución α-estable. En concreto, cuando el exponente característico es $\alpha = 1$ y el parámetro de asimetría $\beta = 0$, la función característica es integrable y su solución es la distribución de Cauchy con parámetro posición μ y dispersión γ (ec. (1.8)). Si la distribución de la expresión de genes fuera realmente Cauchy, hubiéramos obtenido los valores $\alpha \approx 1$ y $\beta \approx 0$ para el exponente característico y el parámetro de asimetría respectivamente. Sin embargo, la Figura 5.1 muestra que los valores para el parámetro característico obtenidos típicamente están en el intervalo $\alpha \in [1{,}5 - 1{,}8]$.

Todas estas evidencias apoyan la hipótesis de que la distribución α-estable es apropiada para modelar la distribución de la expresión de genes. En particular, la distribución α-estable podría usarse en lugar de otra, por ejemplo la Normal, para el diseño de métodos que establecen si un gen está o no expresado. Además, la distribución estable posee propiedades que han sido ampliamente estudiadas en la literatura, como la representación mediante mezcla escalada de Normales, que nunca ha sido usada anteriormente en genética, y que será la base del diseño de un estadístico en el capítulo siguiente. Además, puesto que la Normal es un caso particular de la familia de distribuciones estables, modelar la distribución de la expresión de genes mediante una distribución α-estable, es una alternativa asimétrica y subgaussiana con la que obtendríamos el mismo resultado en la identificación de expresión genética que con la distribución Normal; siempre y cuando la distribución de la expresión de genes fuera realmente Normal. Sin embargo, tal y como se ha comprobado a lo largo de este capítulo, esta distribución está lejos de poder ser explicada mediante la distribución Gaussiana.

5.4. Conclusiones

En este capítulo, hemos presentado un nuevo modelo estadístico para la distribución de la expresión de genes usando la distribución α-estable. Este modelo proporciona la flexibilidad necesaria para trabajar con distribuciones con gran peso en las colas y asimétricas, característica que presenta el histograma de datos genéticos procedente de micromatrices. Además, la distribución α-estable comparte importantes propiedades con la distribución α-estable, tales como la ley Pareto en las colas de la distribución y la no convergencia de la desviación

estándar de los datos. Todas estas características hacen que la distribución α-estable sea adecuada para modelar los datos de micromatrices. Por último, el análisis estadístico de la distribución de los datos de expresión genética nos sugiere el uso de la distribución estable para el diseño de estadísticos con los cuales estimar si un conjunto de genes está o no expresado.

CAPÍTULO

SEIS

DISEÑO DE UN ESTADÍSTICO PARA LA EVALUACIÓN DE LA EXPRESIÓN GENÉTICA USANDO LA DISTRIBUCIÓN α-ESTABLE

El estudio de la expresión genética mediante micromatrices se ha consolidado en la actualidad como una importante herramienta que permite la comparación de los niveles de expresión genética de miles de genes simultáneamente. Estos experimentos nos permiten comparar dos muestras diferentes de ADNc obtenido bajo dos condiciones distintas. Tal y como se estudió en el Capítulo 5, algunas de las características de la distribución de la expresión de genes, por ejemplo, el comportamiento asintótico de tipo Pareto en las colas o la no convergencia de la varianza con el incremento del número de genes estudiados, sugieren el modelado de esta distribución empírica mediante la α-estable.

En este capítulo, presentamos un método para el análisis de micromatrices. Las caracterísicas del método propuesto lo hacen adecuado para los casos en que tengamos distintas réplicas de un mismo experimento de micromatrices, circunstancia que, por otro lado, es la más deseable [Allison *et al.*, 2006]. El método propuesto está basado en el análisis de la distribución de intensidades de las micromatrices como una mezcla de dos distribuciones: α-estable simétrica y Delta de Dirac. Este modelo nos permite el uso de propiedades de la distribución α-estable, tales como la representación mediante mezcla escalada de Gaussianas (véase la Sección 1.1.3) o la extracción de muestras aleatorias α-estables (Sección 1.1.4). El modelo α-estable para la distribución de la expresión de genes junto con la estadística Bayesiana, nos permite construir un estadísi-

tico que nos da información sobre la probabilidad de que un gen esté expresado bajo unas condiciones determinadas [Salas-Gonzalez *et al.*, 2007a]. Probaremos el funcionamiento del estadístico propuesto en datos tanto sintéticos como experimentales. Además lo compararemos con un algoritmo similar propuesto por [Lonnstedt & Speed, 2002], donde también hace uso de la Mezcla Escalada de Gaussianas, aunque allí, la distribución de la expresión de genes se modela mediante una distribución t-student.

Este capítulo está organizado del siguiente modo: en la Sección 6.1 se presentan algunos de los trabajos en la literatura que han modelado la distribución de la expresión de genes mediante mezcla de distribuciones. En la Sección 6.2, se desarrolla el estadístico basado en la representación mediante Mezcla Escalada de Gaussianas de la distribución α-estable simétrica. En la Sección 6.3 el estadístico es probado en datos sintéticos y reales de micromatrices. Finalmente, en la Sección 6.4 extraemos las conclusiones.

6.1. Introducción

El estudio de la expresión genética mediante micromatrices se basa en la comparación de dos diferentes muestras de ADN complementario teñidas de colores distintos (normalmente rojo y verde), a través de la medida de la intensidad lumínica de cada color tras la hibridación. Las micromatrices nos permiten la comparación de una gran cantidad de genes simultáneamente: típicamente miles de genes distintos con alguna repetición de los experimentos (normalmente menor a la decena). Debido a esta gran cantidad de datos y el número tan pequeño de repeticiones, el estudio de la expresión genética mediante micromatrices, es un problema estadístico bastante complejo.

Tras la normalización de los datos, la distribución de la expresión genética presenta, en general, mayor peso en las colas que la distribución Gaussiana. Esta distribución ha sido modelada mediante distintas familias de distribuciones: Cauchy [Khondoker *et al.*, 2006], distribución de Pareto [Kuznetsov, 2001], Laplace [Purdom & Holmes, 2005], t-student [Lonnstedt & Speed, 2002] o Log-Normal [Hoyle *et al.*, 2002]. En el Capítulo 5, estudiamos distintos conjuntos de datos reales y se comprobó que la distribución α-estable ajustaba con gran exactitud la distribución de la expresión de genes.

En este capítulo, diseñaremos un estadístico para la identificación de la expresión genética mediante el estudio de datos procedentes de micromatrices. Este estadístico está basado en el modelo de mezclas α-estable y la propiedad de Mezcla Escalada de Gaussianas. Estas dos propiedades, combinadas con la

estadística Bayesiana, nos permitirán obtener el estadístico deseado.

Introducimos un modelo Bayesiano de mezcla de α-estable, el cual es capaz de modelar genes como compuestos de dos poblaciones diferenciadas: genes expresados y no expresados. Estas dos condiciones conformarán los dos componentes de la mezcla. La estadística Bayesiana y el modelado mediante mezcla de distribuciones están adquiriendo gran popularidad en el contexto de las micromatrices [Lonnstedt & Speed, 2002; Bhowmick *et al.*, 2006; Newton *et al.*, 2004; Do *et al.*, 2005; Gottardo *et al.*, 2003].

6.2. Inferencia estadística

Asumimos que los datos con los que trabajamos son el logaritmo en base-2 de las intensidades del color rojo (R_{ij}) y verde (G_ij) normalizadas mediante regresión local (LOESS) [Yang *et al.*, 2002a]. Si N es el número de genes en cada matriz y n el número de réplicas o matrices (es decir, el número de veces que el experimento se ha repetido bajo las mismas condiciones), los datos son, por lo tanto, $M_{ij} = log\left(\frac{R_{ij}}{G_{ij}}\right)$, donde $i = 1...N$, $j = 1...n$. Nuestro objetivo es establecer cuáles son los genes que están verdaderamente expresados en el experimento.

En [Lonnstedt & Speed, 2002], se propone una mezcla de la distribución Normal y la de Dirac para la construcción de un estadístico B usando estadística Bayesiana. En dicho trabajo, se usa la distribución Gamma Inversa como distribución a priori para la varianza. En este capítulo, modelaremos las medidas del logaritmo entre el cociente de intensidades, asumiendo que dichos valores son independientes y suponiendo que M_{ij} son variables aleatorias procedentes de una distribución Normal con media μ_i y varianza $\lambda_i\sigma^2$.

$$M_{ij}|\mu_i, \lambda_i, \sigma \sim N(\mu_i, \lambda_i\sigma^2) \text{ para todo } i. \tag{6.1}$$

Aunque en realidad, los genes interaccionan unos con otros de modo desconocido, la independencia entre éstos es una simplificación que ha sido usada en multitud de trabajos [Bhowmick *et al.*, 2006; Gottardo *et al.*, 2003; Lonnstedt & Speed, 2002], y que nos permitirá la construcción del estadístico.

En este modelo, los genes se consideran pertenecientes a uno de los dos grupos siguientes: expresados o no expresados. Sea w la probabilidad de que un gen esté expresado, el parámetro μ se modela como una variable aleatoria extraída de una distribución α-estable simétrica si el gen está efectivamente

expresado o como $\mu = 0$ en caso contrario.

$$p(\mu_i|\lambda_i, \sigma) = w f_{\alpha,0}(\sigma, 0) + (1 - w)\delta(0) \tag{6.2}$$

Sea z_i una variable que puede tomar los valores 0 ó 1 y que indica si un gen está expresado ($z_i = 1$) o no ($z_i = 0$).

$$z_i = \begin{cases} 0 & \text{if } \mu_i = 0 \\ 1 & \text{if } \mu_i \sim f_{\alpha,0}(\sigma, 0) \end{cases}$$

El logaritmo del cociente entre las probabilidades de que un gen esté o no expresado, para un gen i, se puede escribir como

$$S_i = \log \frac{Pr(z_i = 1|M_{ij})}{Pr(z_i = 0|M_{ij})} \tag{6.3}$$

y, usando el teorema de Bayes (sec. 1.2) junto con la suposición de independencia entre genes:

$$S_i = \log \frac{w}{1 - w} \frac{Pr(M_i|z_i = 1)}{Pr(M_i|z_i = 0)} \tag{6.4}$$

donde M_i es el vector con las n repeticiones para el gen i. El objetivo es estimar las distribuciones de probabilidad $Pr(M_i|z_i = 1)$ y $Pr(M_i|z_i = 0)$ para, de este modo, calcular el estadístico S_i. Dicho estadístico establece la probabilidad de que un gen determinado i esté expresado.

La representación mediante Mezcla Escalada de distribuciones Normales nos permite escribir una distribución simétrica α-estable como una Gaussiana, condicionada a una variable aleatoria auxiliar α-estable positiva ($\beta = 1$), que denotaremos como λ. Por lo tanto, la distribución de μ_i condicionada en λ_i es una Gaussiana

$$\mu_i|\lambda_i, \sigma = \begin{cases} 0 & \text{si } z_i = 0 \\ N(0, \lambda_i \sigma^2) & \text{si } z_i = 1 \end{cases}$$

con la siguiente distribución a priori para la variable λ

$$p(\lambda_i) = f_{\frac{\alpha}{2},1}(2\{\cos\left(\frac{\pi\alpha}{4}\right)\}^{\frac{2}{\alpha}}, 0) \tag{6.5}$$

Teniendo en cuenta la información proporcionada por las expresión (6.2), la distribución de M_i condicionada por μ_i, λ_i y σ es

$$\begin{aligned} p(M_i|\mu_i, \lambda_i, \sigma) &= (2\pi\lambda_i)^{-\frac{n}{2}} \sigma^{-n} e^{-\frac{1}{2\lambda_i\sigma^2}\sum_j (M_{ij} - \mu_i)^2} \\ &= (2\pi\lambda_i)^{-\frac{n}{2}} \sigma^{-n} e^{-\frac{1}{2\lambda_i\sigma^2}\sum_j (M_{ij} - M_i)^2 + n(M_i - \mu_i)^2} \end{aligned} \tag{6.6}$$

por lo que, para el modelo estudiado, obtenemos:

$$\begin{aligned} p(M_i|z_i=1) &= \int\int p(M_i,\mu_i,\lambda_i)d\mu_i d\lambda_i \\ &= \int\int p(M_i|\mu_i,\lambda_i)p(\mu_i|\lambda_i,z_i=1)p(\lambda_i)d\mu_i d\lambda_i \end{aligned} \tag{6.7}$$

$$\begin{aligned} p(M_i|z_i=0) &= \int\int p(M_i|\mu_i,\lambda_i)p(\mu_i|\lambda_i,z_i=0)p(\lambda_i)d\mu_i d\lambda_i \\ &= \int\int p(M_i|\lambda_i)p(\mu_i|\lambda_i,z_i=0)p(\lambda_i)d\lambda_i \end{aligned} \tag{6.8}$$

Si sustituimos las expresiones (6.7) y (6.8) e identificamos la distribución a posteriori $p(\mu_i|\lambda_i,z_i=1)$ como el producto de una distribución Normal y una función exponencial

$$p(\mu_i|\lambda_i,z_i=1) = N(\mu_i|\frac{n}{n+1}M_i,\frac{\lambda_i\sigma^2}{n+1})\cdot e^{\frac{1}{2\lambda_i\sigma^2}\frac{n}{n+1}M_i^2} \tag{6.9}$$

obtenemos las siguientes integrales

$$\begin{aligned} p(M_i|z_i=1) &= \int (2\pi\lambda_i)^{-\frac{n}{2}}\sigma^{-n}e^{-\frac{1}{2\lambda_i\sigma^2}\sum_j(M_{ij}-M_i)^2} \\ &\times\ (n+1)^{1/2}e^{-\frac{1}{2\lambda_i\sigma^2}\frac{n}{n+1}M_i^2} \\ &\times\ f_{\frac{\alpha}{2},1}(2\{\cos\left(\frac{\pi\alpha}{4}\right)\}^{\frac{2}{\alpha}},0)d\lambda_i \end{aligned} \tag{6.10}$$

$$\begin{aligned} p(M_i|z_i=0) &= \int (2\pi\lambda_i)^{-\frac{n}{2}}\sigma^{-n}e^{-\frac{1}{2\lambda_i\sigma^2}\sum_j(M_{ij})^2} \\ &\times\ f_{\frac{\alpha}{2},1}(2\{\cos\left(\frac{\pi\alpha}{4}\right)\}^{\frac{2}{\alpha}},0)d\lambda_i \end{aligned} \tag{6.11}$$

$$\tag{6.12}$$

Debido a la falta de una expresión analítica para la distribución α-estable, las integrales (6.10) y (6.11) deben calcularse numéricamente. Existen multitud de técnicas numéricas para evaluar estas integrales. Nosotros las evaluaremos mediante técnicas Monte Carlo. Para ello, extraemos muestras de una distribución α-estable mediante el algoritmo de Chambers [Chambers *et al.*, 1976]

(véase la Sección 1.1.4). Si extraemos T muestras aleatorias $\left[\lambda_i^{(1)}...\lambda_i^{(t)}...\lambda_i^{(T)}\right]$ con distribución $p(\lambda_i)$ dada por la eq. (6.5), podemos estimar numéricamente mediante Monte Carlo, las integrales (6.10) y (6.11), aproximándolas mediante sumatorias:

$$\begin{aligned} p(M_i|z_i=1) &= \frac{1}{T}\sum_{t=1}^{T}(2\pi\lambda_i^{(t)})^{-\frac{n}{2}}\sigma^{-n}e^{-\frac{1}{2\lambda_i^{(t)}\sigma^2}\sum_j(M_{ij}-M_i)^2} \\ &\times \quad (n+1)^{1/2}e^{-\frac{1}{2\lambda_i^{(t)}\sigma^2}\frac{n}{n+1}M_i^2} \end{aligned} \tag{6.13}$$

$$p(M_i|z_i=0) = \frac{1}{T}\sum_{t=1}^{T}(2\pi\lambda_i^{(t)})^{-\frac{n}{2}}\sigma^{-n}e^{-\frac{1}{2\lambda_i^{(t)}\sigma^2}\sum_j(M_{ij})^2} \tag{6.14}$$

6.3. Resultados

6.3.1. Datos sintéticos

Para ilustrar el funcionamiento del estadístico S_i propuesto en este capítulo de la memoria, simularemos datos genéticos con $N = 10{,}000$ genes y $n = 4$ repeticiones. Los genes se modelaron siguiendo una distribución α-estable con los parámetros siguientes: $\alpha = 1{,}8$, $\beta = 0$, $\sigma = 0{,}1$ y $\mu = 0$. Dichos valores fueron elegidos, ya que son los valores típicos obtenidos en el análisis realizado en el Capítulo 5 de esta Tesis Doctoral.

Consideramos que una proporción $p = 0{,}01$ de los $N = 10{,}000$ genes están expresados. Este conjunto de genes los simularemos, con los mismos valores que para el caso de los genes no expresados, excepto que el parámetro posición es extraído de una distribución α-estable con dispersión $V\sigma$, siendo $V = 1{,}5$. Este parámetro V fue también introducido en [Lonnstedt & Speed, 2002; Bhowmick *et al.*, 2006]. La estimación de V es muy difícil de realizar, ya que tan sólo una pequeña proporción de genes están expresados, y por tanto afectados por el parámetro V. Por otra parte, no sabemos qué genes son los realmente expresados. Esta dificultad también fue apuntada en [Lonnstedt & Speed, 2002] para un modelo de mezcla Gaussiano y para mezcla de Laplace en [Bhowmick *et al.*, 2006], donde el parámetro V no es estimado correctamente. En las simulaciones, mostraremos que el funcionamiento del método presentado en este capítulo de la memoria no está afectado por el desconocimiento de este parámetro. Esto

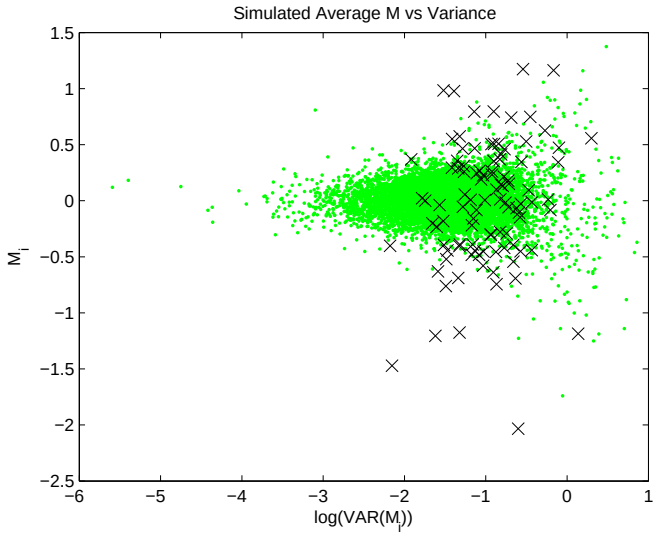

Figura 6.1: Representación de M_i frente al logaritmo de la varianza para uno de los conjuntos de datos simulados. *Cruces:* Genes realmente expresados.

es debido a dos causas: por una parte, la distribución α-estable es una distribución con un gran peso en las colas. Esto le confiere la propiedad de ser una distribución adecuada para el modelado de datos que contengan algunas muestras con valores atípico. Por otra parte, el número de genes expresado, como ya apuntamos anteriormente, es muy pequeño comparado con el total de los datos genéticos considerado.

Uno de los conjuntos de datos simulado se representa en la Figura 6.1, donde los genes realmente expresados están señalados mediante cruces. Éstos son los genes cuyo parámetro posición μ es distinto de cero. En dicha gráfica se observa cómo hay un número determinado de genes expresados cuya detección no es posible, independientemente del método utilizado.

La proporción p de genes expresados para cada conjunto de datos simulados es $p = 0{,}01$. El estadístico α-estable S_i y B_i fueron calculados 100 veces para 20 distintos valores de w, desde $w = 0{,}005$ hasta $w = 0{,}1$.

En la Figura 6.2, se muestra el histograma para los valores obtenidos en la estimación de los parámetros α-estables. Se comprueba en dicha figura, que los valores verdaderos para cada parámetro son estimados con exactitud. Además,

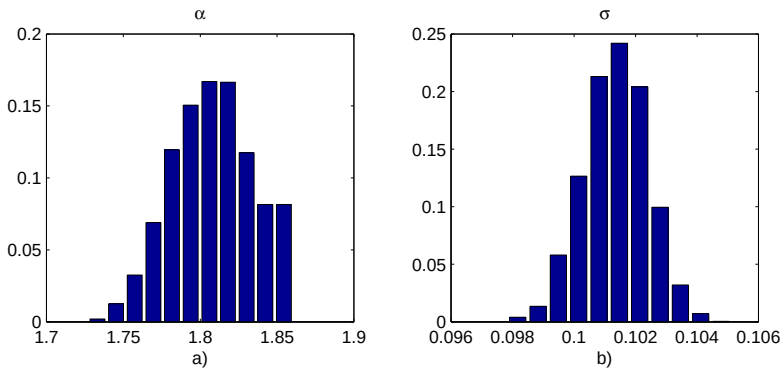

Figura 6.2: Histograma con los valores de los parámetros estimados para cada simulación. El valor verdadero de los parámetros es $\alpha = 1{,}8$ y $\sigma = 0{,}1$.

Tabla 6.1: Media de los valores estimados para los parámetros de la distribución α-estable simétrica. La proporción de genes expresados considerada es $p = 0{,}01$ y el número de repeticiones del experimento $n = 4$.

Parámetro	**Valor verdadero**	**Valor estimado**	**Error**
α	1.80	1.83	0.03
σ	0.100	0.102	0.001

en la Tabla 6.1 se presenta el valor medio de los parámetros obtenidos para todas las simulaciones realizadas. Se observa, de nuevo, cómo tanto el exponente característico α como la dispersión γ son estimados correctamente.

En la Figura 6.3 se representa el valor medio del número de falsos positivos y falsos negativos para los 20 distintos valores de w (desde $w = 0{,}005$ hasta $w = 0{,}1$) y 100 conjuntos de datos simulados para cada valor de w y para cada uno de los datos simulados considerados. En dicha figura, el valor obtenido para el estadístico α-estable S_i se representa junto al estadístico B_i basado en la distribución t-student propuesto en [Lonnstedt & Speed, 2002]. El estadístico S_i presenta un valor más bajo de falsos positivos y negativos para cualquier valor de w, por lo que el funcionamiento del estadístico S_i es considerablemente mejor que el de B_i para el conjunto de datos simulado.

Además, para cada valor de los pesos w (40 valores diferentes desde $w = 0$ hasta $w = 1$), se han simulado 100 conjunto de datos. También, en este caso, se

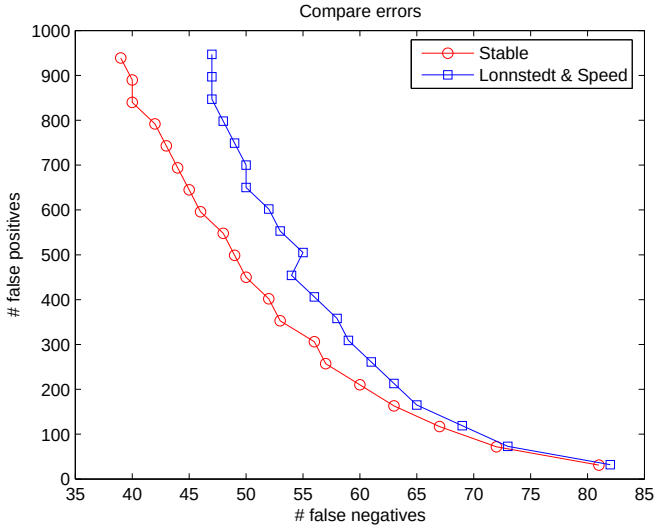

Figura 6.3: Representación gráfica del número de falsos positivos y falsos negativos para los estadísticos S_i y B_i. Los datos fueron simulados como una distribución α-estable con $\alpha = 1{,}8$, $\beta = 0$, $\sigma = 0{,}1$ y $\mu = 0$.

ha calculado el estadístico S y B.

La curva ROC (Receiver Operating Characteristic) está representada en la figura 6.4. En ella, se representa la fracción de verdaderos positivos frente a la de falsos positivos. En esta figura, el estadístico S presenta un mejor funcionamiento que B para el conjunto de datos de la simulación estudiada.

6.3.2. Datos experimentales

Probaremos el estadístico S_i para datos genéticos de micromatrices reales. En concreto, usaremos una comparación entre RNA de plantas de tipo *Arabidopsis*, infectadas con *rhizobacterium Pseudomonas thivervalensis (strain MLG45)*, y *axenic control plants* [Cartieaux *et al.*, 2003]. Este conjunto de datos consiste en el estudio de $N = 16{,}416$ genes y $n = 4$ repeticiones. Además, está disponible públicamente para su descarga en la web *Standford Microarray Database (SMD)* con número de experimento 27084, 27000, 26995 y 26718. Los datos fueron normalizados mediante regresión local (*LOESS*). Tras el proceso de nor-

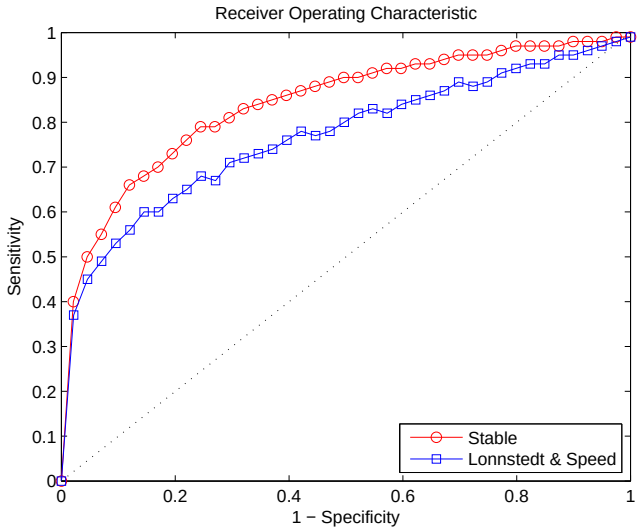

Figura 6.4: Curva ROC (Receiver Operatinng Characteristic) donde se muestran los valores medios para 100 realizaciones de los estadísticos S_i y B_i. Los datos fueron simulados como una distribución α-estable con $\alpha = 1{,}8$, $\beta = 0$, $\sigma = 0{,}1$ y $\mu = 0$.

malización, la distribución de genes fue modelada mediante una distribución α-estable simétrica. Los valores obtenidos para los parámetros fueron: $\alpha = 1{,}83$ y $\sigma = 0{,}15$. En la Figura 6.5, se representa la distribución de la expresión de genes y la distribución α-estable simétrica obtenida. En dicha gráfica, se comprueba que la densidad estable ajusta de manera exacta la distribución de la expresión de genes para el conjunto de datos reales considerado. Además, esta distribución empírica comparte importantes características con la distribución α-estable (véase la Sección 5.3)

En la Figura 6.6, representamos gráficamente el valor de M_i frente al logaritmo de la varianza. En dicha figura, los genes considerados expresados para los estadísticos S y B se denotan mediante cruces y círculos respectivamente. Existe una proporción de genes expresados según el estadístico B y no según S. Los genes en los que esto sucede, poseen, en general, un valor alto del estadístico M_i y la varianza entre las distintas repeticiones del experimento. Esto es así porque el estadístico S penaliza en mayor medida a los genes genes con

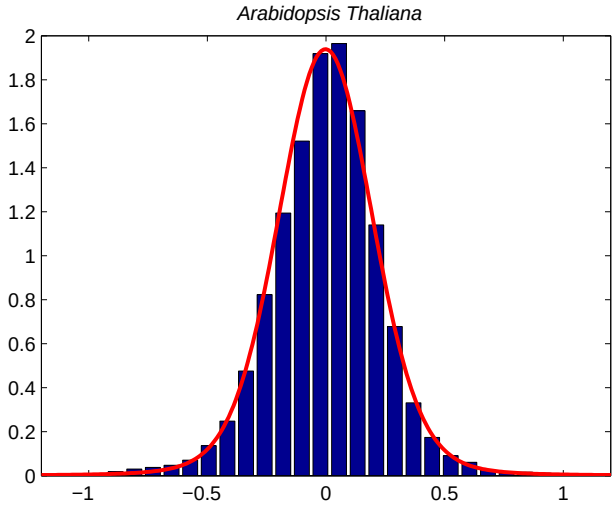

Figura 6.5: Histograma con la distribución de la expresión de genes y la distribución α-estable simétrica obtenida para el conjunto de datos *Arabidopsis Thaliana*.

valor alto de la varianza. Por otra parte, el modelo propuesto en este capítulo de la memoria asume una distribución con gran peso en las colas como la α-estable para la distribución de genes no expresados (eq. (6.14)), por lo tanto, este modelo permite a los genes no expresados poseer un valor alto de M_i si la varianza de los mismos también es alta. Esta característica hace al estadístico S adecuado para el análisis de datos de expresión genética de micromatrices cuyo error entre las distintas repeticiones del experimento sea grande.

La Figura 6.7 muestra la representación gráfica de tipo volcán en la cual se representa M_i frente a S_i. Las cruces denotan los genes expresados según el estadístico α-estable y, los círculos, los genes que el estadístico B ha considerado expresados. Además, esta figura presenta el mismo comportamiento que fue comentado para la figura anterior, es decir, hay una considerable cantidad de genes con alto valor de la variable M_i que, el estadístico S identifica como no expresados. La Figura 6.8 muestra el valor obtenido para el estadístico S_i frente a la varianza de las diferentes repeticiones de M_{ij}. En dicha figura, se muestra que algunos genes con un alto valor de la varianza son considerados expresados

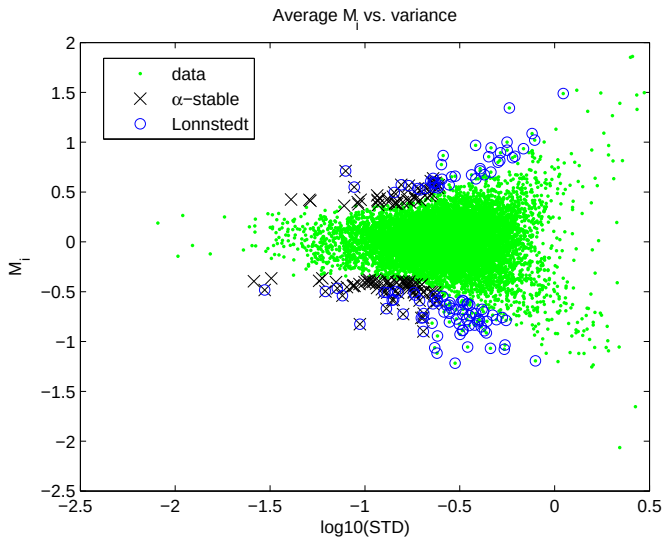

Figura 6.6: Representación gráfica de la media de M_i frente al logaritmo de al varianza. *Cruces:* Conjunto de genes expresados según el estadístico S. *Círculos:* Genes expresados según el estadístico B propuesto en [Lonnstedt & Speed, 2002].

cuando se usa el estadístico B y no para el estadístico α-estable S_i.

6.4. Conclusiones

En este capítulo, se ha propuesto un nuevo estadístico para identificar genes expresados. Este estadístico está diseñado para el estudio de experimentos de micromatrices con repeticiones y está basado en alguna de las propiedades de la distribución α-estable y los modelos de mezcla de distribuciones. En concreto, introducimos el modelo de mezcla α-estable y hacemos uso de la representación mediante mezcla escalada de Gaussianas para el cálculo Bayesiano del estadístico S_i. Este procedimiento nos permite calcular S_i usando distintas propiedades y métodos de las distribuciones α-estables, como la estimación de parámetros y la simulación de variables estables.

El estadístico propuesto fue probado con datos de micromatrices reales y simulados. Además, el funcionamiento de dicho estadístico se comparó con el

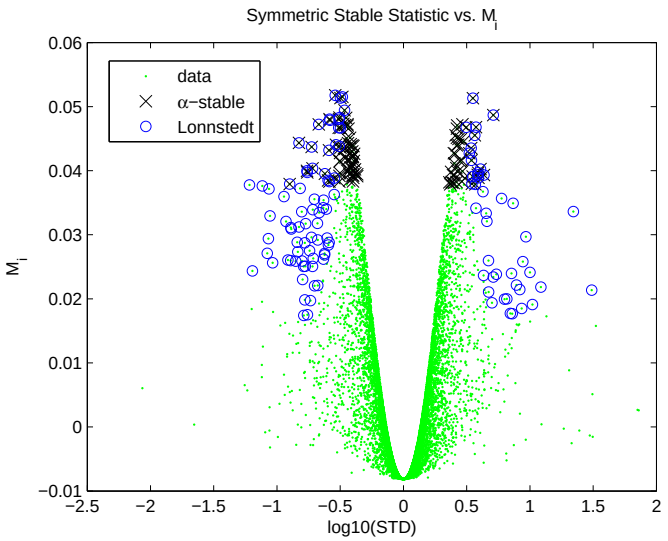

Figura 6.7: Representación gráfica del estadístico S_i frente a la media del nivel de expresión genética M_i. *Cruces:* Conjunto de genes expresados según el estadístico S_i. *Círculos:* Genes expresados según el estadístico B de [Lonnstedt & Speed, 2002].

estadístico B propuesto en [Lonnstedt & Speed, 2002] y basado en la distribución t-student. El estudio preliminar de la distribución de la expresión de genes realizado en el capítulo anterior sugería el uso de la distribución α-estable para el desarrollo de nuevos métodos de modelado de datos de micromatrices. En este capítulo se ha hecho uso de este hecho para la identificación de genes expresados. Los resultados obtenidos tienen validez general, pero en concreto, el uso de una distribución con gran peso en las colas como es la distribución α-estable, sugiere la utilización de este método para modelar datos de micromatrices con repeticiones cuando la varianza entre dichas repeticiones es alta.

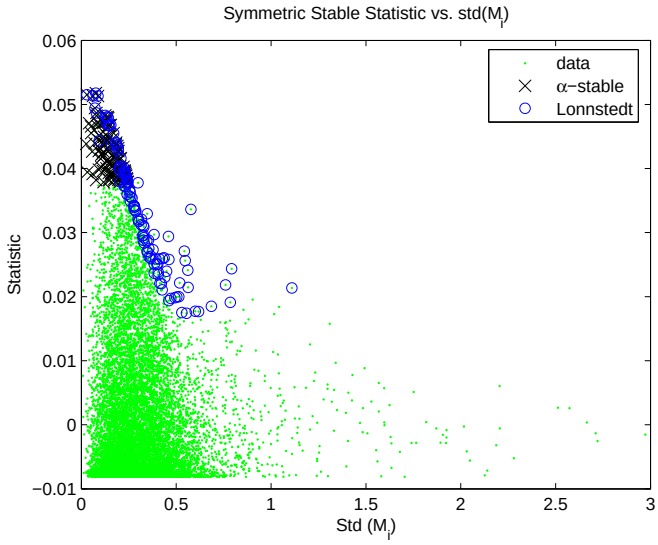

Figura 6.8: Representación gráfica del estadístico S_i frente a la desviación típica del nivel de expresión genética $std(M_i)$. *Cruces:* Conjunto de genes expresados según el estadístico S_i. *Círculos:* Genes expresados según el estadístico B de [Lonnstedt & Speed, 2002].

Parte III

Conclusiones Generales

CAPÍTULO
SIETE

CONCLUSIONES

7.1. Mezcla de distribuciones α-estable simétricas

Se ha introducido un nuevo modelo de mezcla basado en distribuciones α-estables simétricas. Dicho modelo ha sido probado en mezcla sintéticas de señales impulsivas y se ha demostrado mediante simulación numérica que, para este tipo de señales, tanto el número verdadero de componentes en la mezcla como los distintos parámetros de las distribuciones que la componen, son estimados con precisión. Además, hemos comparado el trabajo aquí desarrollado con el modelo de mezcla de Gaussianas. El número de componentes fue sobreestimado cuando usamos el modelo de mezcla de Gaussianas, por lo que nuestro método se erige como una alternativa a dicho método cuando las señales que componen la mezcla son impulsivas. Otra de las ventajas de nuestro método es que el rango de aplicación no sólo se restringe a este tipo de señales. Puesto que la distribución Normal es un caso particular de distribución α-estable, el método aquí propuesto también puede usarse para los mismos casos que en los que la suposición de Gaussianidad era válida. Por último, en el capítulo 2, el modelo de mezcla de distribuciones α-estables simétricas ha sido probado en datos reales, en concreto en datos biológicos, geológicos y astrofísicos. La elección de datos de tres disciplinas tan dispares se ha hecho con el fin de mostrar el amplio rango de aplicación del algoritmo presentado en este capítulo.

7.1.1. Resumen de las principales aportaciones

- Proponemos por vez primera en la literatura un modelo de mezcla de distribuciones α-estables simétricas, que entre otras características, es una generalización del ampliamente estudiado modelo de mezclas de Gaussianas.
- Usamos la representación mediante mezcla escalada de Gaussianas para obtener una expresión analítica de la función densidad de probabilidad α-estable simétrica. Esta representación, nunca había usada anteriormente en el contexto de la los modelos de mezcla.
- El uso de la representación mediante mezcla escalada de Gaussianas de la distribución simétrica α-estable, nos permite escribir la función densidad de probabilidad α-estable simétrica como una distribución Normal, condicionada a la variable auxiliar λ. Por tanto, aunque el algoritmo aquí propuesto resuelve el complejo modelo de mezclas α-estable simétricas, comparte la sencillez del modelo Bayesiano de mezcla de Gaussianas. Esto nos permite usar la distribución a priori conjugada y escribir una expresión analítica para las distribuciones a posteriori de algunos de los parámetros desconocidos del problema.
- Resolvemos la estimación de parámetros mediante un planteamiento estrictamente Bayesiano del problema. Usando métodos de muestreo Monte Carlo como el algoritmo de muestreo por rechazo, muestreo de Gibbs y Metropolis.
- En el contexto de mezcla de distribuciones α-estables, nunca con anterioridad se habían considerados métodos Monte Carlo de dimensión variable como el algoritmo Monte Carlo basado en cadenas de Markov con saltos reversibles (RJMCMC, por sus siglas en inglés). El cual nos permite calcular el número de componentes que componen la mezcla.
- Al ser este modelo una generalización del modelo de mezclas de Gaussianas, el rango de aplicación de éste se extiende a multitud de distintas disciplinas y materias. Por otro lado, presenta una ventaja bastante importante respecto al modelo Gaussiano, ya que la mezcla de distribuciones α-estables permite, además, modelar datos cuya distribución es una mezcla de componentes impulsivos.
- Este algoritmo es comparado con el modelo de mezcla Gaussiano. Se comprueba a partir del análisis de las simulaciones realizadas, que el modelo

α-estable simétrico permite modelar datos como mezcla de distribuciones impulsivas de manera más compacta que lo hace la distribución Gaussiana ya que precisa un menor número de componentes.

- Por otro lado, el modelo de mezcla α-estable simétrico demuestra funcionar muy bien y estimar correctamente todos los parámetros del modelo, incluso cuando los datos son mezcla de distribuciones Normales. No es posible decir lo mismo en el caso contrario, es decir, para un vector de datos mezcla de distribuciones α-estables simétricas.

- El amplio rango de aplicación y distintas posibilidades que posee este modelo es mostrado mediante tres simulaciones con datos reales de disciplinas dispares como la biología, astrofísica y geología.

7.2. Mezcla de distribuciones α-estable asimétricas

Hemos presentado el modelo de mezcla de distribuciones α-estables general. Este modelo es una generalización del estudiado en el capítulo 2. La estimación de los parámetros ha sido realizada mediante la construcción de un modelo jerárquico Bayesiano. Por otra parte, en dicho modelo, suponemos desconocido el número de componentes en la mezcla. Estimamos este parámetro mediante el método Monte Carlo de saltos reversibles (RJMCMC). La ausencia de una expresión analítica para la densidad de probabilidad de la distribución α-estable es solventada mediante la resolución numérica de la integral de la función característica α-estable.

El algoritmo propuesto ha sido probado tanto en datos sintéticos como reales. En todas las simulaciones realizadas, cada uno de los parámetros desconocidos del problema ha sido estimado correctamente. Además, en cada caso, el método propuesto ha obtenido el número verdadero de componentes. El algoritmo ha sido comparado con otros trabajos en la literatura. El modelo de mezcla α-estable presentado en este capítulo presenta muchas ventajas con respecto a el único trabajo previo de mezcla α-estable. Por ejemplo, es considerablemente más rápido, es más sencillo de implementar y además converge para datos reales donde los distintos componentes están muy mezclados entre sí. Para poner de manifiesto los diferentes campos donde puede encontrar aplicación el modelo de mezclas α-estable, los datos reales analizados proceden de dos disciplinas dispares, como son la biología y la economía.

En el capítulo 3, la mezcla de α-estables ha sido comparada con el modelo de mezclas de Gaussianas. Hemos probado que nuestro método es mucho más adecuado para el análisis de componentes impulsivos y asimétricos. Además el modelo presentado en este capítulo, es una generalización del modelo de mezclas Gaussianas, por lo que comparte muchas de las propiedades de aquél, con la flexibilidad añadida de que las mezclas estables nos permiten modelar componentes que presentan alto grado de impulsividad.

7.2.1. Resumen de las principales aportaciones

- Proponemos por primera vez en la literatura un análisis Bayesiano del modelo de mezcla α-estable que permite la estimación de todos los parámetros del problema de manera exacta.

- El modelo es una generalización de la mezcla de Gaussianas. En el caso en que los datos son mezcla de componentes impulsivos, la mezcla de Gaussianas no converge, mientras que nuestro modelo es robusto frente a señales impulsivas.

- Por otro lado, en caso de datos impulsivos, la mezcla de α-estables requiere un menor número de componentes para ajustar la distribución de los datos que en el caso Gaussiano.

- Al igual que en la aportación del Capítulo 2, en el modelo de mezcla α-estable más general, el número de componentes en la mezcla es calculado satisfactoriamente mediante técnicas Monte Carlo basadas en cadenas de Markov de dimensión variable, en concreto el algoritmo de saltos reversibles (RJMCMC, *Reversible jump Markov chain Monte Carlo*). Esta Tesis Doctoral presenta las dos únicas ocasiones en que este algoritmo se ha usado en el contexto de mezclas α-estables.

- El algoritmo ha sido ampliamente comparado con el método propuesto por [Casarin, 2004], las ventajas del método presentado en esta memoria son claras: menor complejidad computacional debido a la integración numérica de la función característica de la distribución α-estable. El algoritmo es mucho más robusto y la convergencia del mismo no depende de los valores iniciales de los parámetros.

- Del mismo modo que en el caso de la mezcla simétrica de distribuciones α-estable, hemos querido mostrar el amplio rango de aplicación de este

algoritmo mediante el estudio de datos reales de diversas disciplinas, como la economía y la biología.

7.3. Modelado de micromatrices usando la distribución α-estable

En el capítulo 5, hemos presentado un nuevo modelo estadístico para la distribución de la expresión de genes usando la distribución α-estable. Este modelo proporciona la flexibilidad necesaria para trabajar con distribuciones con gran peso en las colas y asimétricas, característica que presenta el histograma de datos genéticos procedente de micromatrices. Además, la distribución α-estable comparte importantes propiedades con la distribución α-estable, tales como la ley Pareto en las colas de la distribución y la no convergencia de la desviación estándar de los datos. Todas estas características hacen que la distribución α-estable sea adecuada para modelar los datos de micromatrices. Por último, el análisis estadístico de la distribución de los datos de expresión genética nos sugiere el uso de la distribución estable para el diseño de estadísticos con los cuales estimar si un conjunto de genes está o no expresado.

7.3.1. Resumen de las principales aportaciones

- En [Khondoker *et al.*, 2006], se modela la distribución de la expresión de genes mediante distintas familias de distribuciones asimétricas. Finalmente, la distribución que ofrece mejor ajuste fue la distribución Pareto aunque para ello tuvo que introducirse un parámetro posición adicional para generalizar dicha distribución. La distribución α-estable ya cuenta con dicho parámetro posición y proporciona un buen ajuste tanto en el centro de la distribución como en las colas. Además, la distribución α-estable también posee comportamiento asintótico de tipo Pareto (ley de potencias) en las colas cuando $\alpha < 2$.
- Mandelbrot hizo hincapié en los primeros trabajos de aplicación de la distribución α-estable, en el hecho de que el uso de dicha distribución para el estudio y descripción de datos biológicos era preferible al uso de distribuciones de tipo Zipf-Pareto debido a motivos tanto teóricos como prácticos.
- La distribución α-estable permite el modelado de la distribución de la expresión de genes de manera más compacta, mediante el uso de una sola

distribución. Al contrario de lo que sucede en [Hoyle *et al.*, 2002], donde la aproximación se realiza mediante una distribución Log-normal en el centro de la distribución y una ley de potencias o ley Zipf en las colas.

- Además, en [Hoyle *et al.*, 2002], se apunta que la varianza σ^2 de las intensidades logarítmicas aumenta conforme el número de genes estudiado aumenta. Este resultado está en completo acuerdo con las propiedades de la distribución α-estable. La varianza es un parámetro que no está definido para procesos estables con $\alpha < 2$.

- Tras la comparación tanto cualitativa como cuantitativa del ajuste proporcionado por la distribución α-estable con respecto a la distribución de Laplace Asimétrica estudiada en [Purdom & Holmes, 2005], se comprueba cómo la distribución de Laplace asimétrica no es capaz de ajustar siempre de manera satisfactoria la distribución de la expresión de genes. El histograma de intensidades de la expresión genética presenta, normalmente, un comportamiento más suave alrededor del máximo que la distribución de Laplace.

- En [Khondoker *et al.*, 2006], la distribución de la expresión de genes se modela mediante una distribución de Cauchy. Nosotros, por otra parte, no asumimos que la distribución de la expresión de genes es Normal o Cauchy, pero ambas distribuciones son casos particulares de la distribución α-estable.

- El modelado de la distribución de la expresión de genes mediante la α-estable y los excelentes resultados obtenidos en el ajuste, son un primer acercamiento de esta distribución al estudio de los datos de micromatrices. Este estudio sirve, por ejemplo, de punto de partida para el diseño de un estadístico basado en las propiedades de la distribución α-estable que nos permite establecer un criterio sobre si un determinado gen está o no expresado. Los detalles sobre el diseño y funcionamiento de dicho estadístico se explican en el Capítulo 6.

7.4. Expresión genética usando la distribución α-estable

En el capítulo 6, se ha propuesto un nuevo estadístico para identificar genes expresados. Este estadístico está diseñado para el estudio de experimentos de

micromatrices con repeticiones y está basado en alguna de las propiedades de la distribución α-estable y los modelos de mezcla de distribuciones. En concreto, introducimos el modelo de mezcla α-estable y hacemos uso de la representación mediante mezcla escalada de Gaussianas para el cálculo Bayesiano del estadístico S_i. Este procedimiento nos permite calcular S_i usando distintas propiedades y métodos de las distribuciones α-estables, como la estimación de parámetros y la simulación de variables estables.

El estadístico propuesto fue probado con datos de micromatrices reales y simulados. Además, el funcionamiento de dicho estadístico se comparó con el estadístico B propuesto en [Lonnstedt & Speed, 2002] y basado en la distribución t-student. El estudio preliminar de la distribución de la expresión de genes realizado en el capítulo 5 sugería el uso de la distribución α-estable para el desarrollo de nuevos métodos de modelado de datos de micromatrices. Se ha hecho uso de este hecho para la identificación de genes expresados. Los resultados obtenidos tienen validez general, pero en concreto, el uso de una distribución con gran peso en las colas como es la distribución α-estable, sugiere la utilización de este método para modelar datos de micromatrices con repeticiones cuando la varianza entre dichas repeticiones es alta.

7.4.1. Resumen de las principales aportaciones

- El uso de la distribución α-estable para el modelado de la distribución de la expresión de genes está suficientemente motivado por el estudio detallado realizado en el Capítulo 5 de esta Memoria, donde se comprobó que la distribución α-estable ajusta con gran exactitud la distribución de la expresión de genes, además de compartir con ella algunas de sus propiedades más importantes.
- El diseño de un estadístico mediante la suposición de que cada gen puede estar expresado o no, usando para ello un modelado matemático mediante mezcla de distribuciones, permite calcular la probabilidad de que un gen esté expresado sin la necesidad de calcular el valor P asociado a un resultado observado. Siendo el valor P la probabilidad de obtener un valor como el observado o más extremo si la hipótesis nula es cierta.
- El uso de la distribución α-estable como parte del modelo de mezcla nos permite simplificar notablemente el problema de cálculo del estadístico debido al uso de diversas propiedades de esta distribución. Así, la estimación de parámetros de la distribución puede realizarse mediante multitud de técnicas existentes en la literatura.

- Además, el modelo matemático se construye de manera relativamente simple, debido al uso de la distribución α-estable simétrica para modelar la distribución de la expresión de genes y a la representación mediante mezcla escalada de Gaussianas.
- En el diseño del estadístico hay que resolver varias integrales numéricamente. Una vez más, las propiedades de la distribución α-estable proporcionan un modo muy sencillo para aproximar las integrales por sumatorias sin más que extraer muestras con distribución α-estable.
- El uso de una distribución con gran peso en las colas como es la distribución α-estable, permite que las medidas tomadas para genes considerados por el modelo como no expresados tengan una mayor dispersión. Esto confiere al estadístico diseñado una ventaja sobre el uso de otros estadísticos basados en la distribución de Laplace y t-student en el caso en que los datos genéticos estudiados tengan una gran variabilidad entre las distintas repeticiones experimentales realizas en el laboratorio.
- Para mostrar el funcionamiento del estadístico α-estable, éste ha sido probado con datos simulados y comparado con un trabajo previo similar basado en la distribución t-student [Lonnstedt & Speed, 2002].

Part IV

Summary in English

CHAPTER

EIGHT

MIXTURE OF SYMMETRIC α-STABLE DISTRIBUTIONS

The stable distribution is a very useful tool to model impulsive data. In this chapter, a fully Bayesian mixture of symmetric stable distribution model is presented [Salas-Gonzalez *et al.*, 2007c, 2006c]. Despite the non existence of closed form for α-stable distributions, the use of the Product Property make it possible to infer on parameters using a straightforward Gibbs sampling. This model is compared to the mixture of Gaussians model. Our proposed methodology is proved to be more robust to outliers than the mixture of Gaussians therefore it is suitable to model mixture of impulsive data. Moreover, as Gaussian is a particular case of the α-stable distribution, the proposed model is a generalization of mixture of Gaussians. Mixture of symmetric α-stable is intensively tested in both, simulated and real data.

8.1. Introduction

Gaussian mixture modelling is nowadays a powerful approach that allows us to model data sampled from a population that is composed of distinct subpopulations. Mixture model have been successfully applied in model, in a parametric manner, data with non-standard distribution (see [McLachlan & Peel, 2000] for a review). Mixture of Gaussian distributions is the most widely studied mixture model due to the fact that the Gaussian distribution can be theoretically justified by the Central Limit Theorem and the Stability Property and, moreover,

it is analytically straightforward to work with.

The Gaussian distribution is a particular case of a more general family of distributions called α-stable laws. This distribution allow us to model data more impulsive than the Gaussian distribution. The α-stable distribution has many desirably properties which made it an alternative for modelling non-Gaussian signals in many fields as signal processing, electrical engineering, computer science, economics or physics [Nikias & Shao, 1995]. Nevertheless, this distribution does not have a general analytical expression for its probability density function.

In this chapter, we propose a mixture of symmetric α-stable distribution which is a generalization of the Gaussian mixture model. Furthermore, the symmetric α-stable distribution can be written as a Scale Mixture of Normals representation using the Product Property [Feller, 1966]. This allows us to write a symmetric α-stable distribution as a Gaussian conditionally on a positive α-stable random variable. Therefore, the non-existence of a closed form for the probability density function is surmounted and, as the symmetric α-stable distribution is written as a Gaussian, a straightforward Gibbs sampling can be used to estimate the unknown parameters.

Mixture of α-stable distributions were also studied in an unpublished technical report before [Casarin, 2004] in which it is claimed that its approach needs more evaluation, specially in the case of symmetric stable mixtures. In [Salas-Gonzalez *et al.*, 2006b, 2008], an alternative Bayesian α-stable mixture model is presented and compared extensively with the work in [Casarin, 2004]. The methodology used in [Salas-Gonzalez *et al.*, 2006b] was proved to be computationally more efficient, to work better for mixtures with closer modes, to be capable to estimate also the number of components in the mixture and to be easier to implement.

The mixture of symmetric α-stable proposed here has several advantages with respect to [Salas-Gonzalez *et al.*, 2006b; Casarin, 2004]. It is considerably easier to implement, the analytical expressions for the posterior distribution are very similar to that obtained for the Gaussian mixture model and new estimates can be easily sampled from the posterior distribution using Gibbs sampling, while in [Casarin, 2004; Salas-Gonzalez *et al.*, 2006c], the full pdf needs to be evaluated by numerical integration.

This chapter is organized as follows: in Section 8.2, the α-stable distribution and the main properties which will be used throughout this chapter are presented. In Section 8.3, the Bayesian symmetric α-stable mixture model is introduced. Section 8.4 provides the Markov chain Monte Carlo algorithm to infer on the Bayesian symmetric α-mixture model. Simulation results are provided in Section 8.5 and, finally, the conclusions are drawn in Section 8.6.

8.2. α-stable distribution

Due to the non existence of an analytical expression for the α-stable probability density function, this is usually expressed by means of the characteristic function of an α-stable distribution $f_{\alpha,\beta}(y|\gamma,\mu)$, which is given by:

$$\varphi(\omega) = \begin{cases} e^{-|\gamma\omega|^{\alpha}[1-i\text{sign}(\omega)\beta\tan(\frac{\pi\alpha}{2})]+i\mu\omega}, (\alpha \neq 1) \\ e^{-|\gamma\omega|[1+i\text{sign}(\omega)\frac{2}{\pi}\beta\log(|\omega|)]+i\mu\omega}, (\alpha = 1) \end{cases}$$

where the parameters of the stable distribution are: $\alpha \in (0, 2]$ is the characteristic exponent which sets the level of impulsiveness, $\beta \in [-1, +1]$ is the skewness parameter, ($\beta = 0$, for symmetric distributions and $\beta = \pm 1$ for the positive/negative stable family respectively), $\gamma > 0$ is the dispersion, a scale parameter, and μ is the location parameter.

8.2.1. Product property

The main drawback of the α-stable distribution is the non-existence of a closed expression for the probability density function. Nevertheless, this problem can be surmounted taking into account that the following property holds for symmetric α-stable distributions: [Samorodnitsky & Taqqu, 1994]

Let X and $Y > 0$ be independent random variables with $X \sim f_{\alpha_1,0}(\gamma, 0)$ and $Y \sim f_{\alpha_2,1}((\cos\frac{\pi\alpha_2}{2})^{\frac{1}{\alpha_2}}, 0)$. Then XY^{1/α_1} is stable with parameters $Z \sim f_{\alpha_1\cdot\alpha_2,0}(\gamma, 0)$.

We are interested in the case in which $\alpha_1 = 2$ (Gaussian case) and $\alpha_2 < 1$. For these parameter values, a scale mixtures of normals (SMiN) could be used for the symmetric stable law as follows (see [Godsill & Kuruoglu, 1999; Kuruoglu *et al.*, 1998, 1997] for more details). If y_i is a i.i.d. sample from a symmetric α-stable distribution with location parameter μ and scale parameter γ:

$$y_i \sim f_{\alpha,0}(\gamma, \mu) \tag{8.1}$$

The product property can be used to obtain the following equivalent representation:

$$y_i \sim \mathcal{N}(\mu, \lambda_i\gamma^2) \tag{8.2}$$

$$\lambda_i \sim f_{\frac{\alpha}{2},1}(2\left(\cos\frac{\pi\alpha}{4}\right)^{\frac{2}{\alpha}}, 0) \tag{8.3}$$

Where $\mathcal{N}(\mu, \lambda_i\gamma^2)$ is the Normal distribution with mean μ and variance $\lambda_i\gamma^2$. This equivalent model is very useful for Bayesian inference as conditionally on the auxiliary positive stable random variable λ_i, the symmetric α-stable variable y_i is Gaussian.

8.3. Symmetric α-stable mixture model

The symmetric α-stable mixture model is

$$p_Y(y) = \sum_{j=1}^{k} w_j f_{\alpha_j,0}(y|\gamma_j, \mu_j) \tag{8.4}$$

$$0 < w_j < 1 \text{ and } \sum_{j=1}^{k} w_j = 1 \tag{8.5}$$

For this model, a mixture of k distribution is considered. j is the index of every subpopulation, w_j is the weight of the distribution j. γ_j, μ_j and α_j are the dispersion, location parameter and characteristic exponent for the symmetric α-stable component j and y are the observations or data vector. We want to infer on the $4k+1$ unknown variables $\{w_j, \gamma_j, \mu_j, \alpha_j, k\}$ using the available data y. In order to accomplish this goal, it is useful to introduce a latent variable $z_i \in [1, 2, ..., k]$ named allocation variable which indicate that a given observation y_i belongs to the component z_i with parameter values $\{w_{z_i}, \gamma_{z_i}, \mu_{z_i}, \alpha_{z_i}\}$. Namely, $z_i = j$ denotes that the observation y_i has been drawn from the subpopulation j. Therefore, in that case:

$$y_i|z_i = j \sim f_{\alpha_{z_i},0}(y|\gamma_{z_i}, \mu_{z_i}) = f_{\alpha_j,0}(y|\gamma_j, \mu_j) \tag{8.6}$$

The Bayesian paradigm provide us a suitable methodology to infer on unknown quantities. Let $B = \{w_j, \gamma_j, \mu_j, \alpha_j, k\}$ be the unknown quantities and y the known data. The Bayes' Theorem allows us to build a hierarchical model in which the unknown quantities are estimated using the prior information and the available data via the Bayes' rule:

$$p(B|y) = \frac{p(y|B)p(B)}{p(y)} \tag{8.7}$$

where $p(B|y)$ denotes the posterior probability of B, given the data y, $p(y|B)$ is the likelihood of y given the parameters B and $p(B)$ is the prior distribution of the unknown variables. $p(y)$ is only a normalizing constant. Thus, the Bayes' Theorem can be rewritten as

$$p(B|y) \propto p(y|B)p(B) \tag{8.8}$$

Replacing the unknown quantities B and the latent variable z_i, in last expression we get

$$p(w_j, \gamma_j, \mu_j, \alpha_j, k, z|y) \propto p(y|w_j, \gamma_j, \mu_j, \alpha_j, k, z)p(w_j, \gamma_j, \mu_j, \alpha_j, k, z) \tag{8.9}$$

As usual in Bayesian methods, in order to introduce more flexibility in the model, we allow prior distribution to depend on hyperpriors with its corresponding hyperparameters. For the sake of clarity in the notation, we write the parameters of the symmetric α-stable as $\theta = \{\alpha, \gamma, \mu\}$ and its corresponding hyperparameters as $\eta = \{\alpha_0, \beta_0, \kappa, \xi, a\}$

$$p(k, w, z, \theta, \eta, y) = p(y|k, w, z, \theta, \eta)p(\theta|k, \eta)p(z|w, k)p(w|k, \zeta)p(k|k_0)p(k_0)p(\zeta)p(\eta) \tag{8.10}$$

Figure 8.1 shows the Direct Acyclic Graph for the proposed Bayesian symmetric α-stable mixture model.

8.3.1. Prior distributions

We choose the conjugate priors for location, dispersion and weights. As the likelihood for the SαS mixture model is Gaussian, the conjugate priors are the same that in the Gaussian mixture case. The conjugate prior for the location parameter is Gaussian:

$$p(\mu_j|\xi, \kappa^{-1}) = \mathcal{N}(\mu|\xi, \kappa) = \frac{1}{\sqrt{2\pi}\kappa^{-1}} \exp\{-\frac{(\mu - \xi)^2}{2\kappa^{-2}}\} \tag{8.11}$$

Inverse Gamma distribution is chosen for the dispersion parameter, γ_j:

$$p(\gamma_j^2|\alpha_0, \beta_0) = \mathcal{IG}(\alpha_0 + \frac{N}{2}, \frac{1}{2} \sum_{i=1:z_i=j}^{N} \frac{(y_i - \mu_j)^2}{\lambda_i} + \beta_0)) \tag{8.12}$$

and the conjugate prior on w is the Dirichlet distribution.

$$\mathbf{w} \sim D(\zeta, ..., \zeta) \tag{8.13}$$

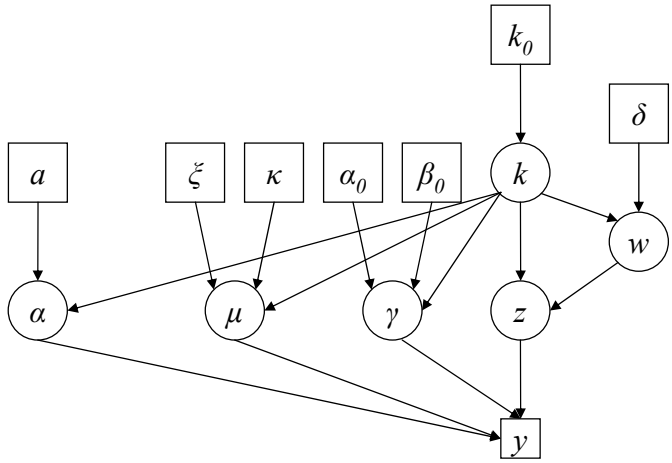

Figure 8.1: Directed Acyclic Graph (DAG) for the symmetric α-stable mixture model. *Circles* denote unknowns variables while *rectangles* represent fixed hyperparameters or vector observation and *arrows* denote the conditional dependence between variables.

For the α parameter, prior distribution is chosen to be the uniform distribution in its support $\alpha \in (0, 2]$.

$$p(\alpha_j|a) = \frac{1}{a} = \frac{1}{2}; \quad \text{for} \quad 0 < \alpha_j \leq 2 \tag{8.14}$$

8.4. Markov chain Monte Carlo implementation

For the Bayesian symmetric α-stable mixture model, the unknown parameters are estimated, at every iteration, using the following Markov chain Monte Carlo scheme:

1) Updating the weights w, μ, γ, using the Gibbs sampling.
2) Updating α using Metropolis sampling.
3) Updating the allocation of variables z.
4) Estimating the auxiliary variable λ.
5) Reversible jump Markov chain Monte Carlo (split/combine move) to estimate the number of components k.

8.4.1. Updating the weights (w) using Gibbs sampling

The full conditional for $\mathbf{w}$ is also a Dirichlet distribution, with parameters $\zeta + n_j$. Thus, at every iteration, every new estimate for the weights can be obtained by sampling from:

$$\mathbf{w} \mid ... \sim D(\zeta + n_1, ..., \zeta + n_k) \tag{8.15}$$

where n_j is the number of samples assigned to the component j, ($n_j = \sum_i \delta(z_i - j)$), and δ denotes the Dirac function).

8.4.2. Updating the location parameter μ using Gibbs sampling

We estimate the location parameter μ_i sampling from the posterior distribution using Gibbs sampling:

$$\mathcal{N}\left(\frac{\frac{1}{\gamma^2} \sum\limits_{i=1:z_i=j}^{N} \frac{y_i}{\lambda_i} + \kappa\xi}{\frac{1}{\gamma^2} \sum\limits_{i=1:z_i=j}^{N} \frac{1}{\lambda_i} + \kappa}, \frac{1}{\frac{1}{\gamma^2} \sum\limits_{i=1:z_i=j}^{N} \frac{1}{\lambda_i} + \kappa} \right) \tag{8.16}$$

8.4.3. Updating the dispersion γ using Gibbs sampling

The full conditional for γ^2 is an Inverse Gamma distribution $\mathcal{IG}$:

$$\mathcal{IG}(\alpha_0 + \frac{1}{2} n_j, \frac{1}{2} \sum_{i=1:z_i=j}^{N} (y_i - \mu_j)^2 + \beta_0) \tag{8.17}$$

8.4.4. Updating the characteristic exponent α using Metropolis-Hasting

For the parameter α, it is not possible to write the full conditional in a closed form. Hence, this parameter is estimated using the Metropolis-Hasting algorithm. For a given component j, samples α_j are obtained following the scheme:

1) At each iteration t we sample a candidate point for α_j (denoted as α_j^{new}) from a proposal distribution $q(\cdot|\cdot)$

$$\alpha_j^{new} \sim q(\alpha_j^{new} | \alpha_j^{(t)})$$

2) We accept the proposed value α_j^{new}, so we set $\alpha_j^{(t+1)} = \alpha_j^{new}$, with probability mín $\left\{1, A_{\alpha_j}\right\}$, where

$$A_{\alpha_j} = \text{mín}\left\{1, \frac{\prod_{i:z_i=j}^{N} p(y_i|k, w_{z_i}, \alpha_j^{new}, \beta_{z_i}, \gamma_{z_i}, \mu_{z_i})}{\prod_{i:z_i=j}^{N} p(y_i|k, w_{z_i}, \alpha^{(t)}, \gamma_{z_i}, \mu_{z_i})} \times \frac{p(k, w_{z_i}, \alpha_j^{new}, \gamma_{z_i}, \mu_{z_i}) q(\alpha_j^{(t)}|\alpha_j^{new})}{p(k, w_{z_i}, \alpha_j^{(t)}, \gamma_{z_i}, \mu_{z_i}) q(\alpha_j^{new}|\alpha_j^{(t)})}\right\} \tag{8.18}$$

if the new value is not accepted we set

$$\alpha_j^{(t+1)} = \alpha_j^{(t)}$$

It is possible to simplify Equation (8.18) for this model as priors are independent, therefore,

$$\frac{p(k, w_{z_i}, \alpha_{new}, \gamma_{z_i}, \mu_{z_i})}{p(k, w_{z_i}, \alpha^{(t)}, \gamma_{z_i}, \mu_{z_i})} = \frac{p(\alpha_{new})}{p(\alpha^{(t)})}.$$

Thus, using a symmetric proposal $q(\alpha_{new}|\alpha^{(t)}) = q(\alpha^{(t)}|\alpha_{new})$ and taking into account that the prior $p(\alpha)$ is chosen to be uniform on its support (see equation (8.14)), the acceptance rejection ratio A_{α_j} simplifies to:

$$A_{\alpha_j} = \text{mín}\left\{1, \frac{\prod_{i:z_i=j}^{N} p(y_i|k, w_{z_i}, \alpha_j^{new}, \gamma_{z_i}, \mu_{z_i})}{\prod_{i:z_i=j}^{N} p(y_i|k, w_{z_i}, \alpha_j^{(t)}, \gamma_{z_i}, \mu_{z_i})}\right\} \tag{8.19}$$

The full conditional of every unknown parameter is very similar to the Gaussian mixture model. Thus, it is straightforward to sample from the posterior using the Gibbs sampling. The posterior distribution for the parameters μ_i, γ_i and w are compared in the Table 8.1 where they are shown to be very similar.

8.4.5. Updating the allocation of variables z

The allocation of variables z is an index that indicates which subpopulation the data y_i is more likely to belong to. This step is accomplish by the full

conditional for allocation of variables ($p(z_i = j|...)$):

$$p(z_i = j|...) = p(y_i|k, w_j, \alpha_j, \gamma_j, \mu_j)p(z)$$

thus, an observation y_i is considered to be drawn from the α-stable component j with parameters $\theta_j = \{\alpha_j, \gamma_j, \mu_j\}$ with probability given by

$$p(z_i = j|...) = w_j p(y_i|k, w_j, \alpha_j, \gamma_j, \mu_j) = w_j \mathcal{N}(y_i|\mu_j, \lambda_i\gamma_j^2) \tag{8.20}$$

8.4.6. Estimating the auxiliary variable λ_i

The product property allow us to write a symmetric α-stable distribution as a Gaussian conditionally in a random positive stable variable λ_i. Different approaches can be used to obtain, at every iteration, samples from the posterior distribution $p(\lambda)$.

$$p(\lambda_i|y, \mu, \gamma, z_i = j) \propto N(y_i|\mu_j, \lambda_i\gamma_j^2) f_{\frac{\alpha}{2},1}(2\left(\cos\frac{\pi\alpha}{4}\right)^{\frac{2}{\alpha}}, 0) \tag{8.21}$$

In [Godsill & Kuruoglu, 1999; Tsionas, 1999], Metropolis-Hasting algorithm is proposed to draw samples from (8.21). The chosen proposal distribution is $\lambda_i^* \sim f_{\frac{\alpha}{2},1}(2\left(\cos\frac{\pi\alpha}{4}\right)^{\frac{2}{\alpha}}, 0)$, so the acceptance rejection ratio $A_{\lambda i}$ can be calculated by:

$$A_{\lambda i} = \text{mín}\left(1, \frac{N(y_i|\mu_j, \lambda_i^*\gamma_j^2)}{N(y_i|\mu_j, \lambda_i\gamma_j^2)}\right) \tag{8.22}$$

where λ_i^* is the proposed new value for the auxiliary variable. λ_i^* can be easily drawn using the Chambers-Mallows-Stuck algorithm [Chambers *et al.*, 1976].

The main drawback of the Metropolis-Hasting algorithm is that the proposed new values are not always accepted in every iteration. Due to that, instead of using the Metropolis-Hasting algorithm, we calculate the auxiliary variable using the rejection sampling. The rejection sampling allows us to accept a new value of λ in every iteration. Note that the maximum value of the likelihood is bounded by a function which depends on $|y_i - \mu_{z_i}|$:

$$N(y_i|\mu_i, \lambda_i\gamma^2) \leq \frac{1}{\sqrt{2\pi}|y_i - \mu_{z_i}|}\exp\left(-\frac{1}{2}\right) \tag{8.23}$$

Therefore, the rejection sampling can be used to draw samples from λ_i.

1. We draw samples from the positive stable distribution with parameters

$\lambda_i^* \sim f_{\frac{\alpha}{2},1}(2\left(\cos\frac{\pi\alpha}{4}\right)^{\frac{2}{\alpha}}, 0)$.
2. We draw samples from the following uniform distribution
$u_i \sim U\left(0, \frac{1}{\sqrt{2\pi}|y_i-\mu_{z_i}|}\exp(-1/2)\right)$.
3. If $u > N\left(u_i|\mu_i, \lambda_i\gamma^2\right)$ goto 1.

Table 8.1: Comparison between the full conditional of every unknown parameter for Gaussian and symmetric α-stable mixture model.

parameter	full conditional Gaussian mixture model	full conditional symmetric α-stable
γ_j^2	$\mathcal{IG}(\alpha_0 + \frac{1}{2}n_j, \frac{1}{2}\sum_{i=1:z_i=j}^{N}(y_i-\mu_j)^2 + \beta_0)$	$\mathcal{IG}(\alpha_0 + \frac{1}{2}n_j, \frac{1}{2}\sum_{i=1:z_i=j}^{N}\frac{(y_i-\mu_j)^2}{\lambda_i} + \beta_0)$
μ_j	$\mathcal{N}\left(\frac{\frac{1}{\sigma^2}\sum_{i=1:z_i=j}^{N} y_i + \kappa\xi}{\frac{1}{\sigma_j^2}n_j + \kappa}, \frac{1}{\sigma_j^{-2}n_j + \kappa}\right)$	$\mathcal{N}\left(\frac{\frac{1}{\gamma^2}\sum_{i=1:z_i=j}^{N}\frac{y_i}{\lambda_i} + \kappa\xi}{\frac{1}{\gamma^2}\sum_{i=1:z_i=j}^{N}\frac{1}{\lambda_i} + \kappa}, \frac{1}{\frac{1}{\gamma^2}\sum_{i=1:z_i=j}^{N}\frac{1}{\lambda_i} + \kappa}\right)$
w_j	$\mathcal{D}(\zeta + n_1, ..., \zeta + n_k)$	$\mathcal{D}(\zeta + n_1, ..., \zeta + n_k)$

8.4.7. Updating the number of components k using RJMC-MC

One of the main advantage of our algorithm is that it is possible to estimate the number of components in the mixture using variable dimension Markov chain Monte Carlo methods. This algorithm jumps between parameters subspaces of different dimension and accept or reject the proposed values of the parameters and number of components using the expression for the acceptance-rejection ratio given by the reversible jump Markov chain Monte Carlo technique. See [Green, 1995] for more details.

In [Green, 1995], it is explained how to jump between spaces with different dimension attaining detailed balance. If a trans-dimensional move denoted by m is proposed, from a given state x to a new state x'. The reversible jump Markov chain Monte Carlo propose to build a bijection between both spaces with different dimension. This can be accomplish introducing $\dim(x') - \dim(x)$ random variables u. A vector of continuous random variables u is drawn from a density $q(u)$, independent of x, and the new values x' are proposed using an invertible deterministic function $x'(x, u)$. This transformation in the variables $x \rightarrow x'$, is taking into account in the expression of the acceptance ratio by means of the density $q(u)$ and the Jacobian of the transformation. Thus, the acceptance probability A, is

$$A = \text{mín}\left\{1, \frac{p(x'|y)r_m(x')}{p(x|y)r_m(x)q(u)}\left|\frac{\partial x'}{\partial(x,u)}\right|\right\} \tag{8.24}$$

where $r_m(x')$ is the probability of choosing move type m when the actual state is x and $|\cdot|$ is the Jacobian of the transformation.

In [Richardson & Green, 1997], an application of RJMCMC to the estimation of the number of components in a mixture of Gaussian model is proposed. We estimate the number of components following a similar approach.

In [Richardson & Green, 1997], two trans-dimensional moves: birth-death move for empty components and split-combine move for non-empty components are suggested, where a component j is said to be empty when its corresponding allocation of variable is $z_j = 0$.

We extend that work to mixture of α-stable densities. The birth-death move suggested in [Richardson & Green, 1997] was implemented, but the acceptance rate for this move was found to be very low. For this reason, we only consider the split-combine move which was found to be enough for our purposes.

For the split-combine-move, the reversible jump mechanism is needed. Two moves in tandem need to be designed as they form a reversible pair.

The new parameters setting are the same as in [Richardson & Green, 1997]:

$$w_{j*} = w_{j1} + w_{j_2} \tag{8.25}$$

$$w_{j*}\mu_{j*} = w_{j1}\mu_{j1} + w_{j_2}\mu_{j_2} \tag{8.26}$$

$$w_{j*}(\mu_{j*}^2 + \gamma_{j*}^2) = w_{j1}(\mu_{j1}^2 + \gamma_{j1}^2) + w_{j2}(\mu_{j2}^2 + \gamma_{j2}^2) \tag{8.27}$$

where two components j_1 and j_2 with weights, dispersion and location parameters $(w_{j1}, \gamma_{j1}, \mu_{j1})$ and $(w_{j2}, \gamma_{j2}, \mu_{j2})$ respectively, are combined in a new component, denoted as j^*, with parameters $(w_{j*}, \gamma_{j*}, \mu_{j*})$. The allocation of variables changes in this move. Thus, for every data y_i which $z_i = j_1$ or $z_i = j_2$ we set $z_i = j_*$.

Although the combine move is deterministic, the reverse split move is not. There are 3 degrees of freedom, due to the change of dimensionality so three continuous random variables u must be introduced at this point. As in [Richardson & Green, 1997], Beta distributions $Be(\cdot, \cdot)$ are used with the following parameters:

$$u_1 \sim Be(2, 2)$$

$$u_2 \sim Be(2, 2)$$

$$u_3 \sim Be(1, 1)$$

The proposed new values for weights, location and dispersion parameters of the new components j_1 and j_2, split from a given existing component j_*, are

$$w_{j1} = w_{j*}u_1 \tag{8.28}$$

$$w_{j2} = w_{j*}(1 - u_1) \tag{8.29}$$

$$\mu_{j1} = \mu_{j*} - u_2\gamma_{j*}\sqrt{\frac{w_{j2}}{w_{j1}}} \tag{8.30}$$

$$\mu_{j2} = \mu_{j*} + u_2\gamma_{j*}\sqrt{\frac{w_{j1}}{w_{j2}}} \tag{8.31}$$

$$\gamma_{j1}^2 = u_3(1 - u_2^2)\gamma_{j*}^2\frac{w_{j*}}{w_{j1}} \tag{8.32}$$

$$\gamma_{j2}^2 = (1 - u_3)(1 - u_2^2)\gamma_{j*}^2\frac{w_{j*}}{w_{j2}} \tag{8.33}$$

After proposing these new values, we need to test if the condition $[\mu_1 < \mu_2 < ... < \mu_k]$ holds. If not, the move is rejected. The new values for the allocation of

variables must be calculated after this move by assigning to the values labeled as j_* the new allocation, either j_1 or j_2, using the expression (8.20).

$$\begin{aligned} A &= \frac{\prod_{i:z_i=j_1}^{N} \frac{1}{\gamma_{j_1}} e^{-\frac{(y_i-\mu_{j_1})^2}{2\lambda_i\gamma_{j_1}^2}} \prod_{i:z_i=j_2}^{N} \frac{1}{\gamma_{j_2}} e^{-\frac{(y_i-\mu_{j_2})^2}{2\lambda_i\gamma_{j_2}^2}}}{\prod_{i:z_i=j_*}^{N} \frac{1}{\gamma_{j_*}} e^{-\frac{(y_i-\mu_{j_*})^2}{2\lambda_i\gamma_{j_*}^2}}} \\ &\times \frac{1}{a} \times (k+1) \times \frac{w_{j1}^{\zeta-1+n_1} w_{j2}^{\zeta-1+n_2}}{w_{j*}^{\zeta-1+n_1+n_2} B(\zeta, k\zeta)} \\ &\times \sqrt{\frac{\kappa}{2\pi}} e^{-0,5\kappa\left\{(\mu_{j1}-\xi)^2+(\mu_{j2}-\xi)^2-(\mu_{j*}-\xi)^2\right\}} \\ &\times \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \left(\frac{\gamma_{j1}^2\gamma_{j2}^2}{\gamma_{j*}^2}\right)^{-\alpha_0-1} e^{-\beta_0(\gamma_{j1}^{-2}+\gamma_{j2}^{-2}-\gamma_{j*}^{-2})} \\ &\times \frac{d_{k+1}}{b_k P_{alloc}} \times \left\{g_{2,2}(u_1) g_{2,2}(u_2) g_{1,1}(u_3)\right\}^{-1} \\ &\times \frac{w_{j*} \left|\mu_{j1}-\mu_{j2}\right| \gamma_{j1}^2\gamma_{j2}^2}{u_2(1-u_2^2)(1-u_3)\gamma_{j*}^2} \end{aligned} \tag{8.34}$$

where n_1 and n_2 are the number of samples from y_i assigned to the components j_1 and j_2. $B(\cdot,\cdot)$ is the Beta function, P_{alloc} is the probability that the current allocation is chosen and b_k and $d_k = 1 - b_k$ are the probabilities of choosing between split and combine moves respectively. Thus, at every iteration, two-split new components are proposed with probability b_k (otherwise one-combined component with probability $d_k = 1 - b_k$ is proposed) and it is accepted with probability mín $\{1, A\}$. If one-combined new component is proposed, this is accepted with probability mín $\left\{1, A^{-1}\right\}$. Lastly, we remark that it is not allowed to propose a combine move when $k = 1$ or a split move when k is greater than a given integer k_0.

The first line in expression (8.34) is the likelihood ratio, the second one is the ratio between priors for α, β, w and z. The term $k+1$ in this line is obtained due to the restriction to the set $\mu_1 < \mu_2 < ... < \mu_k$. The third and forth line are the ratio between priors for the location parameter μ and dispersion γ. The fifth line is the proposal ratio and the last one is the Jacobian of the transformation.

8.5. Simulation results

8.5.1. Synthetic data

- Simulation 1.

The proposed methodology is tested to $N = 2000$ samples with the following distribution:

$$p_Y(y) = 0{,}2f_{1,4,0}(y|0{,}2,-2) + 0{,}3f_{1,4,0}(y|0{,}5,0) + 0{,}5f_{1,4,0}(y|0{,}6,3). \quad (8.35)$$

The settings for hyperparameters of prior distributions are: $\alpha_0 = 1$, $\beta_0 = 1$, $\xi = 0$, $\kappa = 1/3^2$ and $\delta = 1$. The probability of choosing the split or combine move was set to $b_k = d_k = 0{,}5$ and the number of iterations is set to 500 with a burn-in period of 100 iterations. The number of components is initialized to $k = 5$. For $p(k)$, a discrete uniform distribution between 1 and 10 is chosen.

After a few iterations, the true number of components $k = 3$ is obtained and every parameter in the simulation is estimated very accurately. Table 8.2 shows the true values for every parameter and the estimated values using the proposed methodology.

Figure 8.2 shows two histograms with the number of components k estimated for the symmetric α-stable mixture and Gaussian mixture model. It is easily seen that the true number of components $k = 3$ is obtained for the mixture of symmetric α-stable case whereas for mixture of Gaussians, the number of component is overestimated.

- Simulation 2

Mixture of SαS is more robust to outliers than Gaussian mixture model. In order to show that, we try to model the $N = 2000$ samples distributed as (8.35) with a mixture of three (fixed) Normal components. The predicted density and the discrete histogram for the data are plotted together in Figure 8.3. In this figure, the predicted symmetric α-stable with parameters given in Table 8.2. The mixture of Gaussian approach with 3 components fail in order to model the data. This is due to the fact that the maximum and minimum value for data vector y are 26,40 and $-34{,}15$ respectively. These values are very far from the mean hence they cannot be explain under a Gaussian framework. A component in the mixture with very high variance and very low weight is obtained to model the outliers.

- Simulation 3

Table 8.2: True value and estimated value for every unknown parameter.

Parameter	True value	Estimated value	Standard deviation
α_1	1.4	1.39	0.11
α_2	1.4	1.38	0.12
α_3	1.4	1.41	0.15
μ_1	-2	-1.983	0.018
μ_2	0	-0.00	0.04
μ_3	3	3.01	0.03
γ_1	0.2	0.233	0.014
γ_2	0.5	-0.51	0.04
γ_3	0.6	0.59	0.03
w_1	0.2	0.222	0.012
w_2	0.3	0.282	0.016
w_3	0.5	0.01	0.03

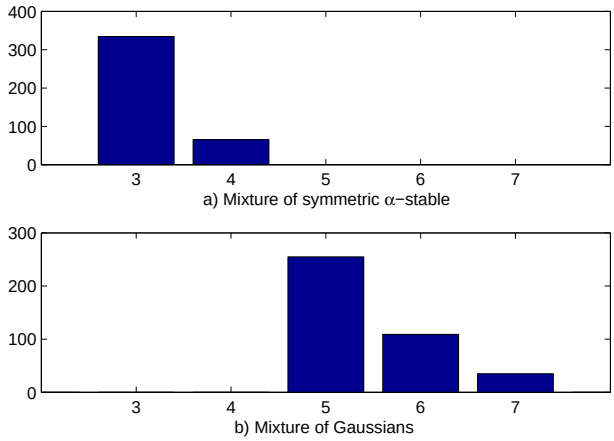

Figure 8.2: Histograms of the number of components estimated after the burn-in period. *Top figure* show the number of components estimated considering mixture of symmetric α-stable model. *Bottom figure* show the results assuming mixture of Gaussians.

As it was extensively explain in this chapter, one of the main advantages of this model is that the Gaussian mixture model is a particular case of the symmetric α-stable mixtures. N=1000 samples with distribution given by a mixture of three Normals with parameters $w = [0{,}4, 0{,}3, 0{,}3]$, $\mu = [-3, 0, 2{,}5]$, $\sigma = [0{,}8, 0{,}8, 0{,}4]$ are simulated. This corresponds to a mixture of symmetric α-stable mixture with $\alpha = [2, 2, 2]$ and dispersion $\gamma = \sigma/\sqrt{2} = [0{,}57, 0{,}57, 0{,}28]$. The data is fitted using the mixture of symmetric α-stable approach. In Table 8.3, the estimated symmetric α-stable parameters and the true values are given. It is shown that all the parameters are estimated accurately. The number of iterations of the MCMC was set to 1000 with a burn-in period of 500 iterations. The minimum mean square error estimator was used to obtain the estimated values. In that case, it would be more convenient to use the mode as the estimator of the characteristic exponent α due to the fact that the domain of the characteristic exponent is $\alpha = (0, 2]$. Using the mode, the estimated value for alpha would be $\alpha = [2, 2, 2]$.

The discrete histogram of the mixture of Gaussian data and the predicted density are depicted in Figure 8.4. The predicted density fits very accurately

Table 8.3: Simulation 3: True value and estimated value for every unknown parameter.

Parameter	True value	Estimated value	Std deviation
α_1	2	1.96	0.05
α_2	2	1.93	0.07
α_3	2	1.95	0.05
μ_1	-3	-3.05	0.06
μ_2	0	0.06	0.07
μ_3	2.5	2.48	0.03
γ_1	0.57	0.56	0.03
γ_2	0.57	0.60	0.06
γ_3	0.28	0.29	0.02
w_1	0.4	0.42	0.02
w_2	0.3	0.29	0.02
w_3	0.3	0.29	0.02

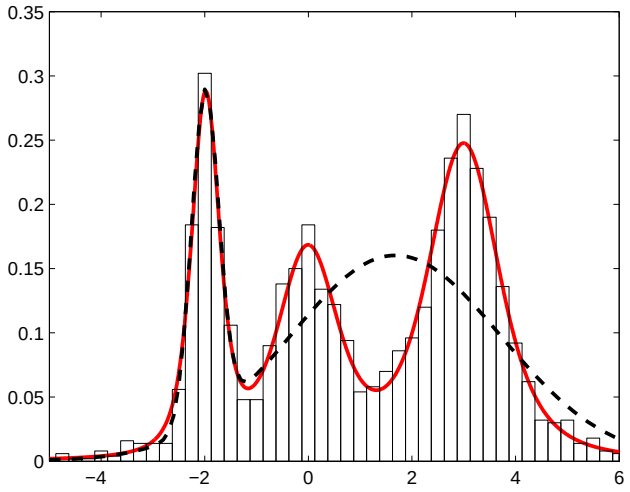

Figure 8.3: Discrete histogram of the mixture of three components in equation (8.35). *Continuous line:* Predicted SαS density with parameters given in Table 8.2. *Dashed line:* predicted density considering mixture of three Gaussian components.

the data.

8.5.2. Real data

The proposed methodology is tested in three different data sets. These data sets have been analyzed before using mixture models. The 'Enzyme data' concerns the distribution of enzymatic activity in the blood, for an enzyme involved in the metabolism of carcinogenic substances, among a group of 245 unrelated individuals [Bechtel *et al.*, 1993; Richardson & Green, 1997]. The 'Acidity data' data set involves a log scale acidity index measured in a sample of 155 lakes in the Northeastern of USA [Crawford *et al.*, 1992; Crawford, 1994; Richardson & Green, 1997]. The 'Galaxy data' consists of the measure of the velocities of 82 distant galaxies, diverging from the Milky Way. This data set has been analyzed widely in the mixture model literature [Escobar & West, 1995; Phillips & Smith, 1996; Stephens, 2000; Richardson & Green, 1997].

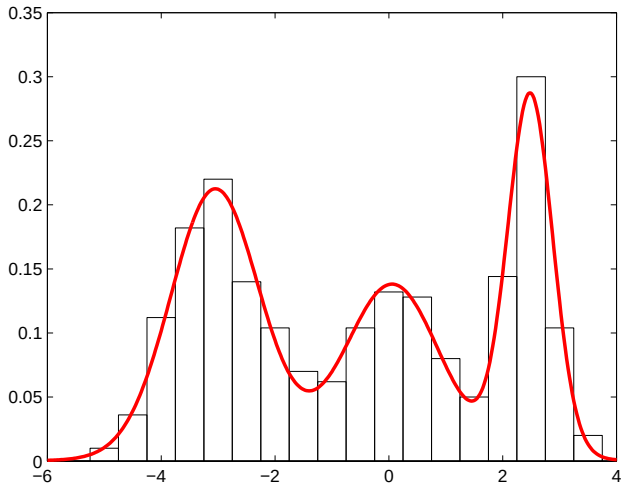

Figure 8.4: Discrete histogram of the mixture of three Gaussian components. *Continuous line:* Predicted SαS density with parameters given in Table 8.3

.

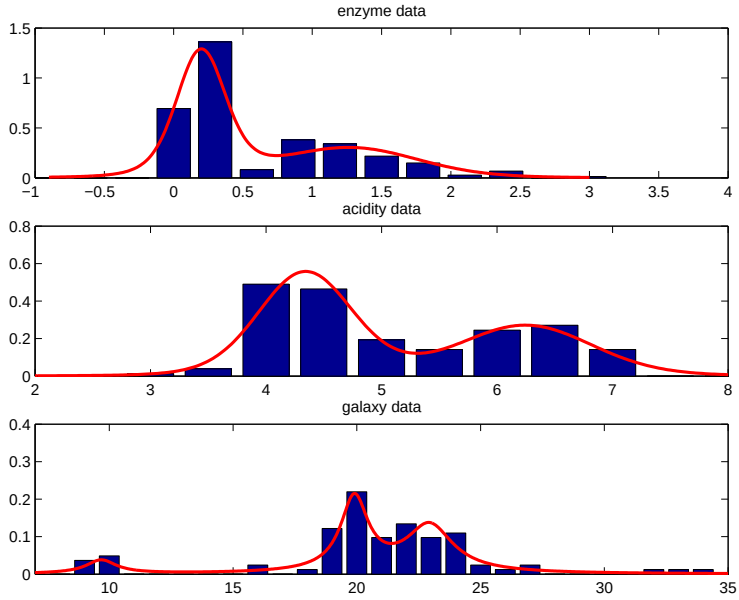

Figure 8.5: Discrete histogram for every real data set and predictive symmetric α-stable density. 'Enzyme data': 2 components. 'Acidity data': 2 components. 'Galaxy data': 3 components.

The number of components and the parameter estimates for these three different datasets are shown in Table 8.4. The histograms for every data set and the predictive symmetric α-stable mixture model are depicted in Figure 8.5.

8.6. Conclusion

This chapter presents a new mixture model based on mixture of symmetric α-stable distribution. Despite the non-existence of closed form for the α-stable probability density function, Bayesian inference is possible as the density of a symmetric α-stable distribution can be written as a Gaussian, conditionally on a positive α-stable random variable. The main advantage of this model is that as the Gaussian distribution is a particular case of the symmetric α-stable, mixture of symmetric α-stable distribution is a generalization of the well-studied

Table 8.4: Estimated values for three different real datasets.

Parameter	Enzyme data (k=2)	Acidity data (k=2)	Galaxy data (k=3)
α_1	1.66	1.80	1.06
α_2	1.84	1.82	1.38
α_3	-	-	1.35
μ_1	0.20	4.34	9.71
μ_2	1.27	6.25	20.0
μ_3	-	-	23.0
γ_1	0.14	0.30	0.76
γ_2	0.37	0.42	0.74
γ_3	-	-	1.20
w_1	0.62	0.59	0.09
w_2	0.38	0.41	0.45
w_3	-	-	0.46

Gaussian mixture model.

The proposed methodology was tested and every unknown parameter was proved to be estimated very accurately. The MCMC realization for the full Bayesian model considered was compared with the fully Bayesian Gaussian mixture model and both were shown to be very similar except for the auxiliary variable. Moreover the symmetric α-stable mixture model was shown to be more robust to outliers than the Gaussian mixture model. The mixture of α-stable model was also tested in three different real data sets.

CHAPTER

NINE

MIXTURE OF SKEWED α-STABLE DISTRIBUTIONS

Over the last decades, the α-stable distribution has proved to be a very efficient model for impulsive data. In this chapter, we propose an extension of stable distributions, namely mixture of α-stable distributions to model multimodal, skewed and impulsive data. A fully Bayesian framework is presented for the estimation of the stable density parameters and the mixture parameters. As opposed to most previous work on mixture models, the model order is assumed unknown and is estimated using reversible jump Markov chain Monte Carlo. It is important to note that the Gaussian mixture model is a special case of the presented model which provides additional flexibility to model skewed and impulsive phenomena. The algorithm is tested using synthetic and real data, accurately estimating α-stable parameters, mixture coefficients and the number of components in the mixture.

9.1. Introduction

Mixture distributions and α-stable distributions have been two important statistical model families due to the large amount of potential applications in various areas (see [McLachlan & Peel, 2000] and [Samorodnitsky & Taqqu, 1994] respectively and references therein). Mixture models allow us to describe, estimate and infer on complex multimodal data, considering them as sampled from different subpopulations. In particular, mixture of Gaussian distributions have

found wide applications ranging from image processing to radar signal processing, thanks to possessing the advantages of both being multimodal and the subpopulations belonging to Gaussian distributions. Other than the Gauss mixtures, mixtures of alternative distributions such as mixtures of Gamma distributions [M. Wiper & Ruggieri, 2001], Weibull distributions [Tsionas, 2002], Poisson [Fernandez & Green, 2002] or t-student [McLachlan & Peel, 1998] among others have been studied in the literature.

There are many approaches for making inference on parameters in mixture models, but two of them are the most common: Expectation-Maximization (EM) algorithm and the Bayesian techniques. The Bayesian methods are getting more and more popular due to their flexibility and potential to include prior information in the estimation process while EM based methods suffer from local optimality.

α-stable distributions have been proved to be a successful alternative for modeling non-Gaussian data. Historically, the use of Gaussian distribution has been justified theoretically by the Central Limit Theorem. However, in electrical engineering, computer science, economics, physics and astronomy, among other disciplines, some signals present impulsiveness (see [Nikias & Shao, 1995] for a review) and asymmetry [Kuruoglu & Zerubia, 2003; Herranz *et al.*, 2004]. For such data, the Gaussian assumption does not lead to satisfactory modeling results.

In this chapter, we are interested in inference on impulsive, asymmetric and multimodal signals using α-stable distributions under a Bayesian approach [Salas-Gonzalez *et al.*, 2008, 2006b,a; Kuruoglu *et al.*, 2006]. Due to the lack of an analytical expression for the probability density function for α-stable signals, few works use Bayesian inference and a Monte Carlo approach to infer on α-stable parameters and, furthermore, most work considers only unimodal α-stable models. In this context, Buckle in [Buckle, 1995], exploited a particular mathematical representation involving the stable density, that allow to use the Gibbs sampler to make inference on parameters. Tsionas in [Tsionas, 1999] developed a Gibbs and Metropolis sampler in models with symmetric α-stable disturbances using the Scale Mixture of Normals property. In that work, as in [Buckle, 1995], an additional random variable is introduced and the location and dispersion parameters are estimated using a straightforward Gibbs sampling, since the full conditional for these two parameters are Gaussian and Inverse Gamma respectively. More recently, Lombardi in [Lombardi, 2007] introduced a random walk MCMC approach for Bayesian inference in stable distributions using a numerical approximation of the likelihood function. Work on mixtures of α-stable distributions in the literature is very limited generally referring

to only special members of the stable family: for example, Swami in [Swami, 1999], studied mixtures of Cauchy and Gaussian distributions using the EM algorithm to capture heavy-tails in signals with α-stable disturbances; Ilow and Hatzinakos in [Ilow & Hatzinakos, 1998] employed Cauchy-Gauss mixtures in the detection problem. Using Buckle's work [Buckle, 1995], to estimate distribution parameters, Casarin [Casarin, 2004] studied an α-stable mixture model with fixed number of components, introducing a random auxiliary vector, with dimension equal to the length of the observation, at every iteration. Monno et al. [Monno *et al.*, 2004] applied this technique in volatility modeling in economics.

In this chapter, a Bayesian mixture model of α-stable distributions is proposed to make inference on impulsive and multimodal signals. For distribution parameter estimation, we use the same strategy as in [Lombardi, 2007]. However, our proposed α-stable mixture model is more flexible than [Casarin, 2004] since the number of components in the mixture is assumed unknown a priori and is estimated using a numerical Bayesian sampling technique namely reversible jump Markov chain Monte Carlo[Green, 1995] (RJMCMC). Furthermore, we use a numerical approximation of the stable distribution, and thus an auxiliary random vector is not required, reducing the computational load and increasing the efficiency of the proposed approach.

In a previous work, we have considered symmetric α-stable mixtures where the components were assumed to possess equal characteristic exponents [Salas-Gonzalez *et al.*, 2006c]. There, an auxiliary parameter had been introduced to sample indirectly using a Scale Mixture of Normals representation. In this chapter, we present an unified Bayesian analysis which allows us to make inference on parameters for general α-stable mixtures allowing skewed components which can possibly have diverse shape parameters. We accomplish this goal using a numerical calculation of the α-stable distribution and a full Bayesian methodology via the reversible jump Markov chain Monte Carlo (RJMCMC) technique to estimate blindly the number of mixtures.

This chapter is organized as follows: In Section 9.2, the α-stable distribution is presented. In Section 9.3, a Bayesian hierarchical model for mixture of α-stable is introduced. In Section 9.4, the MCMC and RJMCMC methodology adopted to solve this model is described. Simulation results are shown in Section 9.5 and, finally, conclusions are drawn in Section 9.6.

9.2. α-stable distributions

The characteristic function of an α-stable distribution $f_{\alpha,\beta}(y|\gamma,\mu)$ is given by:

$$\varphi(\omega) = \begin{cases} e^{-|\gamma\omega|^{\alpha}[1-i\text{sign}(\omega)\beta\tan(\frac{\pi\alpha}{2})]+i\mu\omega}, (\alpha \neq 1) \\ e^{-|\gamma\omega|[1+i\text{sign}(\omega)\frac{2}{\pi}\beta\log(|\omega|)]+i\mu\omega}, (\alpha = 1) \end{cases}$$

where the parameters of the stable distribution are: $\alpha \in (0, 2]$ is the characteristic exponent which sets the level of impulsiveness, $\beta \in [-1, +1]$ is the skewness parameter, ($\beta = 0$, for symmetric distributions and $\beta = \pm 1$ for the positive/negative stable family respectively), $\gamma > 0$ is the scale parameter, also called dispersion, and μ is the location parameter.

The α-stable density function is the inverse Fourier transform of the characteristic function, thus it can be obtained evaluating the following integral:

$$f_{\alpha,\beta}(y|\gamma,\mu) = \mathcal{F}_w^{-1}[\varphi(\omega)](y) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\omega y}\varphi(\omega)d\omega \tag{9.1}$$

The expression (9.1) does not have analytical solution other than a few particular cases: When $\alpha = 2$ we get the Gaussian case and then $\gamma = \sigma/\sqrt{2}$, where σ is the standard deviation of the Gaussian distribution. Furthermore, for $\alpha = 1$ and $\beta = 0$ the distribution reduces to a Cauchy distribution and for $\alpha = 1/2$ and $\beta = 1$ to a Levy distribution. The non-existence of an analytical expression for the α-stable distribution is the reason why Bayesian and maximum-likelihood methods have not been extensively exploited in the literature. Nolan in [Nolan, 1997] proposes a numerical procedure to calculate the stable density. He uses a special integral representation of the α-stable density proposed by Zolotarev in [Zolotarev, 1986]. Based on this representation integral formulas for the density distribution are derived. The integrand in these formulas is a continuous, bounded, non-oscillating function and the interval of integration is bounded. For this kind of functions it is straightforward to use adaptive quadrature, a well-known numerical integration technique. We use the method proposed by Nolan to evaluate the α-stable distribution numerically.

Summarizing, some properties of the α-stable distributions are (see [Samorodnitsky & Taqqu, 1994] and references therein):

- An α-stable distribution is completely described by only four parameters.
- Stable distributions can fit asymmetry and heavy tailed data better than a Normal.

- The Gaussian distribution is a particular case of α-stable distributions.
- α-stable distributions satisfy the Generalized Central Limit Theorem which states that the sum of a number of random variables with infinite variance will tend to an α-stable distribution as the number of variables grows.
- α-stable distributions have the Stability Property: the output of a linear system in response to α-stable inputs is α-stable distributed.
- α-stable distributions have been widely studied in the literature and their properties are well known [Samorodnitsky & Taqqu, 1994].

9.3. Bayesian stable mixture model

The mixture of alpha-stables density $f_{\alpha,\beta}(y|\gamma,\mu)$ is given by:

$$p_Y(y) = \sum_{j=1}^{k} w_j f_{\alpha_j,\beta_j}(y|\gamma_j,\mu_j)$$

$$0 \leq w_j \leq 1(\forall j) \text{ and } \sum_{j=1}^{k} w_j = 1$$

where w_j is the mixture proportion or weight for component j and $p_Y(y)$ is the probability density function (pdf) of the data vector y. It is convenient to consider a mixture model as a missing data problem in the estimation of its parameters. Hence, we assume that the data vector y has been randomly drawn from k subpopulations (labeled as $j = 1, 2, ...k$). We introduce a new variable $z_i \in [1, 2, ...k]$ named allocation variable; $z_i = j$ denotes that observation y_i belongs to the subpopulation (or component) j of the mixture. The z_i are supposed to be drawn from distributions

$$p(z_i = j) = w_j \qquad \text{for } j = 1, 2, ..., k. \tag{9.2}$$

Conditional on the values z_i, the observations are considered to be drawn from their individual subpopulations, i. e.

$$y_i|z \sim f(\cdot|\alpha_j,\beta_j,\gamma_j,\mu_j) \qquad j = 1, 2, ..., k.$$

It is important to note that this model is invariant under permutations of the label j. In order to avoid this lack of identifiability, a criterion for unique

labeling is needed. We choose an increasing ordering in the location parameter $\mu_1 < \mu_2 < ... < \mu_k$.

Mixture models have been widely studied under many approaches since the early work of Pearson at the end of the 19th century [Pearson, 1894]. In the past, the EM algorithm was used to estimate the mixture parameters despite its drawback of local convergence [Dempster *et al.*, 1997]. On the other hand, the Bayesian inference framework allows us to build a hierarchical model in which the unknown quantities are estimated via the prior information and the available data using the Bayes' rule:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \tag{9.3}$$

where $p(A)$ is the prior probability. $p(A|B)$ denotes the posterior probability of A, given B. $p(B|A)$ is the likelihood of B given A and $p(B)$, the prior probability of B. Equation (9.3) becomes:

$$p(A|B) \propto p(B|A)p(A) \tag{9.4}$$

It is possible to rewrite the equation (9.4) for our mixture model by considering that B is the available data (or vector observation y) and A are the unknown variables

$$A = \{k, w, z, \alpha, \beta, \gamma, \mu\} \tag{9.5}$$

9.3.1. Priors

The following priors are chosen for this model: a Normal distribution with mean ξ and variance κ^{-1} for location (μ) which can be written as:

$$p(\mu) = N(\mu|\xi, \kappa^{-1}).$$

For the dispersion (γ), an inverse gamma distribution with hyperparameters α_0 and β_0 is chosen

$$p(\gamma) = IG(\gamma|\alpha_0, \beta_0).$$

These priors are conjugate priors in Bayesian inference for Gaussian models for the mean and the variance respectively [Richardson & Green, 1997].

The priors for the exponent (α) and the skewness parameter (β) are chosen to be the Uniform distribution on their supports as in [Lombardi, 2007; Buckle, 1995].

$$p(\alpha_j|a) = \frac{1}{a} = \frac{1}{2}; \quad \text{for} \quad 0 < \alpha_j \leq 2 \tag{9.6}$$

$$p(\beta_j|b) = \frac{1}{b} = \frac{1}{2}; \quad \text{for} \quad -1 \leq \beta_j \leq 1$$

Although it could be useful to use the information we know about properties of α-stable parameters (e. g. that as α tends toward 2, β has less influence on asymmetry), priors for α,β,γ,μ were chosen to be independent, as in previous works on Bayesian estimation for α-stable distribution [Lombardi, 2007; Buckle, 1995].

In accordance with other works in the literature on mixing problems [McLachlan & Peel, 2000], the prior on the weights $\mathbf{w} = [w_1, w_2, ..., w_k]$ is taken as symmetric Dirichlet D, which is the conjugate prior for this model:

$$\mathbf{w} \sim D(\zeta, ..., \zeta). \tag{9.7}$$

$p(k|k_0)$ is chosen to be a discrete uniform distribution between 1 and an integer k_0.

9.3.2. Hierarchical model

Thus, we construct a Bayesian hierarchical model in order to make inference on the parameters for this mixture model in which $5k+1$ variables are unknown (α_j,β_j,γ_j,μ_j,w_j,k). The joint distribution of all the variables is:

$$\begin{aligned} p(k, w, z, \alpha, \beta, \gamma, \mu, y) &= p(y|k, w, z, \alpha, \beta, \gamma, \mu) \\ &\times \quad p(k, w, z, \alpha, \beta, \gamma, \mu) \end{aligned} \tag{9.8}$$

It is convenient to increase flexibility adding an extra layer and allowing priors in this model to depend on hyperpriors. For the sake of simplicity, parameters of the α-stable distribution and hyperpriors for these variables are grouped together $\theta = \{\alpha, \beta, \gamma, \mu\}$ and $\eta = \{a, b, \alpha_0, \beta_0, \xi, \kappa\}$ respectively. Therefore, the expression in equation (9.8) can be expanded by taking into account the conditional dependences for this model, in particular:

$$\begin{aligned} p(k, w, z, \theta, \eta, y) &= p(y|k, w, z, \theta, \eta) p(\theta|k, \eta) p(z|w, k) \\ &\times \quad p(w|k, \zeta) p(k|k_0) p(k_0) p(\zeta) p(\eta) \end{aligned}$$

where k_0 and ζ are the hyperparameters for k and w.

For the sake of clarity, Figure 9.1 shows the Direct Acyclic Graph for this Bayesian mixture model. As usual, circles denote unknown variables while rectangles represent fixed (hyperparameters), y is the vector observation and arrows denote conditional dependence.

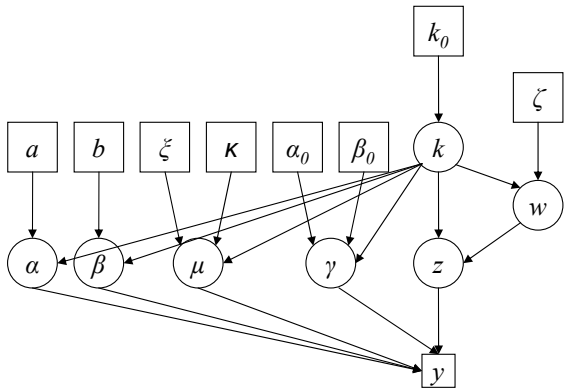

Figure 9.1: Directed Acyclic Graph (DAG) for the α-stable mixture model. $\alpha, \beta, \gamma, \mu$ are the distribution parameters. w are the mixture weights, k is the number of components, z is the allocation variable and $a, b, \xi, \kappa, \alpha_0, \beta_0$ are the hyperparameters.

9.4. MCMC and RJMCMC implementation

Bayesian methods provide the adequate framework to infer on parameters as it was explained in Section 9.3. The main problem of this theoretical framework is that we need to solve multidimensional integrals which often do not have an analytical solution. The posterior expectation of a function of the unknown quantities A, denoted as $f(A)$ is:

$$E\left[f(A)|B\right] = \frac{\int f(A)p(B|A)p(A)dA}{\int p(B|A)p(A)dA}$$

The numerical solution of this kind of integrals involved in the Bayesian estimation problem can be calculated using numerical Monte Carlo methods based on Markov chains [Robert & Casella, 1999].

Mixtures of α-stable models under a Bayesian approach was addressed in an unpublished technical report [Casarin, 2004]. Two main advantages of our

methodology with respect to this work can be pointed out. First, in [Casarin, 2004] the Gibbs sampler for univariate α-stable distribution proposed by Buckle [Buckle, 1995] is used. The stable density is represented in integral form introducing an auxiliary vector λ with dimension equal to the length of the vector observation. Thus, a bivariate density function $f_{\alpha,\beta}(y, \lambda|\gamma, \mu)$ is obtained. And the univariate stable density $f_{\alpha,\beta}(y|\gamma, \mu)$ is calculated integrating numerically respect to λ (via Gibbs sampling)

$$f_{\alpha,\beta}(y|\gamma, \mu) = \int f_{\alpha,\beta}(y, \lambda|\gamma, \mu)d\lambda \tag{9.9}$$

Buckle's work is the first which considers the estimation of parameters of α-stable distributions under a Bayesian approach. However, as it was pointed in [Lombardi, 2007], it is not easy to draw samples from the auxiliary variable λ. Furthermore, the auxiliary vector λ has the same dimension as the vector observation y and must be calculated at every iteration. Rejection sampling is used to obtain λ. Hence, this is a very slow and time consuming procedure. On the contrary to [Buckle, 1995], in the work presented in this chapter, we do not need to introduce any auxiliary variable, hence the computational complexity is considerably lower, as it was analyzed in [Lombardi, 2007].

Another advantage of our approach with respect to [Casarin, 2004] is that we consider unknown number of components in the mixture. The number of components k is related to the dimension of the α-stable parameters of every component in the mixture. Therefore, standard Monte Carlo methods based on Markov chains, which were used in [Casarin, 2004], are not sufficient to infer on this parameter. Trans-dimensional Markov chain Monte Carlo methodology must be used. Specifically, we use reversible jump Markov chain Monte Carlo in order to accomplish this goal [Green, 1995]. Thus, our algorithm is capable of estimating blindly the number of components (k) in the mixture using a fully Bayesian methodology.

Once our model is written in fully Bayesian form, samples for every parameter are obtained, at every iteration, following the scheme:

1) Update the weights ($\mathbf{w}$) using the Gibbs sampling.

2) Update $\theta = \{\alpha,\beta,\mu,\gamma\}$ using Metropolis sampling.

3) Update the allocation of variables z.

4) Reversible jump MCMC (split/combine move) to estimate the number of components k.

9.4.1. Updating the weights (w)

The full conditional distribution for $\mathbf{w}$ is straightforward to calculate. Combining equations (9.7) and (9.2), the full conditional for $\mathbf{w}$ is also a Dirichlet distribution, with parameters $\zeta + n_j$. Thus, at every iteration, every new estimate for the weights can be obtained drawing from the distribution:

$$\mathbf{w} \mid ... \sim D(\zeta + n_1, ..., \zeta + n_k) \tag{9.10}$$

where n_j is the number of samples assigned to the component j, ($n_j = \sum_i \delta(z_i - j)$), where δ denotes the Dirac function).

9.4.2. Updating α-stable parameters using MCMC (α,β,μ,γ)

Parameter α is estimated using the Metropolis-Hasting algorithm. For a given component j, samples α_j are obtained following the scheme:

1) At each iteration t we sample a candidate point for α_j (denoted as α_j^{new}) from a proposal distribution $q(\cdot|\cdot)$

$$\alpha_j^{new} \sim q(\alpha_j^{new}|\alpha_j^{(t)})$$

2) We accept the proposed value α_j^{new}, so we set $\alpha_j^{(t+1)} = \alpha_j^{new}$, with probability mín $\left\{1, A_{\alpha_j}\right\}$, where

$$\begin{aligned} A_{\alpha_j} &= \text{mín} \left\{ 1, \frac{\prod_{i:z_i=j}^{N} p(y_i|k, w_{z_i}, \alpha_j^{new}, \beta_{z_i}, \gamma_{z_i}, \mu_{z_i})}{\prod_{i:z_i=j}^{N} p(y_i|k, w_{z_i}, \alpha^{(t)}, \beta_{z_i}, \gamma_{z_i}, \mu_{z_i})} \right. \\ &\times \left. \frac{p(k, w_{z_i}, \alpha_j^{new}, \beta_{z_i}, \gamma_{z_i}, \mu_{z_i}) q(\alpha_j^{(t)}|\alpha_j^{new})}{p(k, w_{z_i}, \alpha_j^{(t)}, \beta_{z_i}, \gamma_{z_i}, \mu_{z_i}) q(\alpha_j^{new}|\alpha_j^{(t)})} \right\} \end{aligned} \tag{9.11}$$

if the new value is not accepted we set

$$\alpha_j^{(t+1)} = \alpha_j^{(t)}$$

Equation (9.11) can be simplified for this model due to the fact that priors are independent, therefore,

$$\frac{p(k, w_{z_i}, \alpha_{new}, \beta, \gamma_{z_i}, \mu_{z_i})}{p(k, w_{z_i}, \alpha^{(t)}, \beta, \gamma_{z_i}, \mu_{z_i})} = \frac{p(\alpha_{new})}{p(\alpha^{(t)})}.$$

Thus, using a symmetric proposal $q(\alpha_{new}|\alpha^{(t)}) = q(\alpha^{(t)}|\alpha_{new})$ and taking into account that the prior $p(\alpha)$ is chosen to be uniform on its support (see equation (9.6)), equation (9.11) simplifies to:

$$A_{\alpha_j} = \text{mín}\left\{1, \frac{\prod_{i:z_i=j}^{N} p(y_i|k, w_{z_i}, \alpha_j^{new}, \beta_{z_i}, \gamma_{z_i}, \mu_{z_i})}{\prod_{i:z_i=j}^{N} p(y_i|k, w_{z_i}, \alpha_j^{(t)}, \beta_{z_i}, \gamma_{z_i}, \mu_{z_i})}\right\} \tag{9.12}$$

The same strategy can be used for skewness, location and dispersion. For β_j,γ_j and μ_j we propose new values $\beta_j^{new} \sim q(\beta_j^{new}|\beta_j^{(t)})$, $\gamma_j^{new} \sim q(\gamma_j^{new}|\gamma_j^{(t)})$ and $\mu_j^{new} \sim q(\mu_j^{new}|\mu_j^{(t)})$ respectively, and they are accepted with probability given by the following expressions:

$$A_{\beta_j} = \text{mín}\left\{1, \frac{\prod_{i:z_i=j}^{N} p(y_i|k, w_{z_i}, \alpha_{z_i}, \beta_j^{new}, \gamma_{z_i}, \mu_{z_i})}{\prod_{i:z_i=j}^{N} p(y_i|k, w_{z_i}, \alpha_{z_i}, \beta_j^{(t)}, \gamma_{z_i}, \mu_{z_i})}\right\} \tag{9.13}$$

$$\begin{aligned} A_{\gamma_j} &= \text{mín}\left\{1, \frac{\prod_{i:z_i=j}^{N} p(y_i|k, w_{z_i}, \alpha_{z_i}, \beta_{z_i}, \gamma_j^{new}, \mu_{z_i})}{\prod_{i:z_i=j}^{N} p(y_i|k, w_{z_i}, \alpha_{z_i}, \beta_{z_i}, \gamma_j^{(t)}, \mu_{z_i})}\right. \\ &\times \left.\frac{IG(\gamma_j^{new}|\alpha_0, \beta_0)}{IG(\gamma_j^{(t)}|\alpha_0, \beta_0)}\right\} \end{aligned} \tag{9.14}$$

$$\begin{aligned} A_{\gamma_j} &= \text{mín}\left\{1, \frac{\prod_{i:z_i=j}^{N} p(y_i|k, w_{z_i}, \alpha_{z_i}, \beta_{z_i}, \gamma_{z_i}, \mu_j^{new})}{\prod_{i:z_i=j}^{N} p(y_i|k, w_{z_i}, \alpha_{z_i}, \beta_{z_i}, \gamma_{z_i}, \mu_j^{(t)})}\right. \\ &\times \left.\frac{N(\mu_j^{new}|\xi, \kappa^{-1})}{N(\mu_j^{(t)}|\xi, \kappa^{-1})}\right\} \end{aligned} \tag{9.15}$$

Despite the non-existence of an analytical expression for the α-stable distribution, it is possible to evaluate the likelihood $p(y_i|k, w_j, \theta)$ in equations (9.12)-(9.15) numerically using existing techniques as it was stated in Section 9.2.

In addition, candidate values $\theta_j^{new} = \{\alpha_j^{new}, \beta_j^{new}, \gamma_j^{new}, \mu_j^{new}\}$ are sampled from a symmetrical distribution $q(\theta_j^{new}|\theta_j^{(t)}) = q(\theta_j^{(t)}|\theta_j^{new})$. In particular, a Normal distribution centered in the previous value for this variable and with variance σ_θ is chosen:

$$\theta_j^{new} \sim \frac{1}{\sqrt{2\pi}\sigma_\theta} \exp\{-\frac{(\theta_j^{new} - \theta_j^{(t)})^2}{2\sigma_\theta^2}\}. \tag{9.16}$$

9.4.3. Updating the allocation (z)

In a mixture model considered as a missing data problem, it is very important to estimate, at every iteration, which subpopulation the data y_i is more likely to belong to. This step is done using the full conditional for allocation of variables (z):

$$p(z_i = j|...) = p(y_i|k, w_j, \alpha_j, \beta_j, \gamma_j, \mu_j)p(z)$$

thus, an observation y_i is considered to be drawn from the α-stable component j with parameters $\theta_j = \{\alpha_j, \beta_j, \gamma_j, \mu_j\}$ with probability

$$p(z_i = j|...) = w_j p(y_i|k, w_j, \alpha_j, \beta_j, \gamma_j, \mu_j) \tag{9.17}$$

9.4.4. Reversible jump move for the number of components (k)

Unlike previous work which consider mixtures of α-stable distributions in the literature [Casarin, 2004], in our model, the dimension k of every parameter can change at every iteration. Our algorithm jumps between parameters subspaces of different dimension using reversible jump Markov chain Monte Carlo technique [Green, 1995]. Therefore, the flexibility of our algorithm is increased as the number of components k is estimated blindly. A general RJMCMC scheme is explained in the following (see [Green, 1995] for more details).

Suppose a general move denoted by m is proposed, from a state x to a new state x' with higher dimension. This can be accomplished by building a bijection between both spaces. Due to the fact that x and x' have different dimension, there are $\dim(x') - \dim(x)$ degrees of freedom in order to build the bijection.

P. J. Green [Green, 1995] realized that introducing $\dim(x') - \dim(x)$ random variables u, it was possible to jump between spaces with different dimension attaining detailed balance. In that work, a vector of continuous random variables u is drawn from a density $q(u)$, independent of x, and the new values x' are

proposed using an invertible deterministic function $x'(x,u)$. This transformation in the variables $x \to x'$, is taken into account in the expression of the acceptance ratio by means of the density $q(u)$ and the Jacobian of the transformation. Thus, the acceptance probability, here denoted by A is

$$A = \min\left\{1, \frac{p(x'|y)r_m(x')}{p(x|y)r_m(x)q(u)}\left|\frac{\partial x'}{\partial(x,u)}\right|\right\} \tag{9.18}$$

where $r_m(x')$ is the probability of choosing move type m when the actual state is x. In the above equation, $|\cdot|$ denotes the Jacobian of the transformation.

Richardson and Green studied the application of RJMCMC to mixture of Gaussians in [Richardson & Green, 1997] where they estimated the number of Gaussian mixtures successfully using a fully Bayesian approach. They suggested two trans-dimensional moves: birth-death move for empty components and split-combine move for non-empty components, where a component j is said to be empty when its corresponding allocation of variable is $z_j = 0$.

We extend that work to mixture of α-stable densities. The birth-death move suggested in [Richardson & Green, 1997] was implemented but the acceptance rate for this move was found to be very low, for this reason, we only consider the split-combine move. For this split-combine-move, the reversible jump mechanism is needed. Two moves in tandem need to be designed as they form a reversible pair. We remark that the validity of the algorithm is not compromised by the choice of the proposals, since detailed balance is conformed using (9.18).

The new parameters setting is as:

$$w_{j*} = w_{j1} + w_{j_2} \tag{9.19}$$

$$w_{j*}\mu_{j*} = w_{j1}\mu_{j1} + w_{j_2}\mu_{j_2} \tag{9.20}$$

$$w_{j*}(\mu_{j*}^2 + \gamma_{j*}^2) = w_{j1}(\mu_{j1}^2 + \gamma_{j1}^2) + w_{j2}(\mu_{j2}^2 + \gamma_{j2}^2) \tag{9.21}$$

where two components j_1 and j_2 with weights, dispersion and location parameters $(w_{j1}, \gamma_{j1}, \mu_{j1})$ and $(w_{j2}, \gamma_{j2}, \mu_{j2})$ respectively are combined in a new component, denoted as j^*, with parameters $(w_{j*}, \gamma_{j*}, \mu_{j*})$. In this move, the allocation of variables has changed. For every data which $z_i = j_1$ or $z_i = j_2$ we set $z_i = j_*$.

Although the combine move is deterministic, the reverse split move is not. There are 3 degrees of freedom, due to the change of dimensionality so three continuous random variables must be introduced at this point. As in [Richardson & Green, 1997], Beta distributions $Be(\cdot,\cdot)$ are used with the following parameters:

$$u_1 \sim Be(2,2)$$

$$u_2 \sim Be(2,2)$$

$$u_3 \sim Be(1,1)$$

The proposed new values for weights, location and dispersion parameters of the new components j_1 and j_2, split from a given existing component j_*, are

$$w_{j1} = w_{j*}u_1 \tag{9.22}$$

$$w_{j2} = w_{j*}(1-u_1) \tag{9.23}$$

$$\mu_{j1} = \mu_{j*} - u_2\gamma_{j*}\sqrt{\frac{w_{j2}}{w_{j1}}} \tag{9.24}$$

$$\mu_{j2} = \mu_{j*} + u_2\gamma_{j*}\sqrt{\frac{w_{j1}}{w_{j2}}} \tag{9.25}$$

$$\gamma_{j1}^2 = u_3(1-u_2^2)\gamma_{j*}^2\frac{w_{j*}}{w_{j1}} \tag{9.26}$$

$$\gamma_{j2}^2 = (1-u_3)(1-u_2^2)\gamma_{j*}^2\frac{w_{j*}}{w_{j2}} \tag{9.27}$$

After proposing these new values, we need to test whether the condition $[\mu_1 < \mu_2 < ... < \mu_k]$ holds. If not, the move is rejected. Allocation of variables is done for the new values by assigning to the values labeled as j_* the new allocation, either j_1 or j_2, using the expression (9.17).

Up to now, we have not considered how to assign new values α_{j_1} and α_{j_2} from α_* or β_{j_1},β_{j_2} from β_*. This is a very important issue due to the fact that these parameters are concerned with the shape of the α-stable distribution. Different values for α and β produce very different forms of the α-stable distribution. This fact adds an extra level of difficulty to the estimation problem. It is clear that there is not a general law for assigning these values and various approaches may be used.

We keep in memory the last values α_j and β_j for every iteration of our algorithm to a given value k for the number of components in the mixture. Every time that this given value k is proposed again in a split/combine move, we set the previous values for α_j and β_j as new values.

Replacing the information about the split/combine move in equations (9.19)-(9.27) together with the priors in equation (9.18) allows us to write the following

expression for the acceptance/rejection ratio A.

$$\begin{aligned} A &= \frac{p(y|k+1, w_{j1}, w_{j2}, z_{j1}, z_{j2}, \theta_{j1}, \theta_{j2})}{p(y|k, w_{j*}, z_{j*}, \theta_{j*})} \\ &\times \frac{1}{a} \times \frac{1}{b} \times (k+1) \times \frac{w_{j1}^{\zeta-1+n_1} w_{j2}^{\zeta-1+n_2}}{w_{j*}^{\zeta-1+n_1+n_2} B(\delta, k\delta)} \\ &\times \sqrt{\frac{\kappa}{2\pi}} e^{-0,5\kappa\left\{(\mu_{j1}-\xi)^2+(\mu_{j2}-\xi)^2-(\mu_{j*}-\xi)^2\right\}} \\ &\times \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \left(\frac{\gamma_{j1}^2 \gamma_{j2}^2}{\gamma_{j*}^2}\right)^{-\alpha_0-1} e^{-\beta_0(\gamma_{j1}^{-2}+\gamma_{j2}^{-2}-\gamma_{j*}^{-2})} \\ &\times \frac{d_{k+1}}{b_k P_{alloc}} \times \left\{g_{2,2}(u_1) g_{2,2}(u_2) g_{1,1}(u_3)\right\}^{-1} \\ &\times \frac{w_{j*} \left|\mu_{j1} - \mu_{j2}\right| \gamma_{j1}^2 \gamma_{j2}^2}{u_2(1-u_2^2)(1-u_3)\gamma_{j*}^2} \end{aligned} \tag{9.28}$$

where n_1 and n_2 are the number of samples from y_i assigned to the components j_1 and j_2. $B(\cdot, \cdot)$ is the Beta function, P_{alloc} is the probability that the current allocation is chosen and b_k and $d_k = 1 - b_k$ are the probabilities of choosing between split and combine moves respectively. Thus, at every iteration, two-split new components are proposed with probability b_k (otherwise one-combined component with probability $d_k = 1 - b_k$ is proposed) and it is accepted with probability mín $\{1, A\}$. If one-combined new component is proposed, this is accepted with probability mín $\left\{1, A^{-1}\right\}$. Lastly, we remark that it is not allowed to propose a combine move when $k = 1$ or a split move when k is greater than a given integer k_0.

The first line in expression (9.28) is the likelihood ratio, the second one is the ratio between priors for α, β, w and z. The term $k+1$ in this line is obtained due to the restriction to the set $\mu_1 < \mu_2 < ... < \mu_k$. The third and forth line are the ratio between priors for the location parameter μ and dispersion γ. The fifth line is the proposal ratio and the last one is the Jacobian of the transformation.

9.5. Simulation results

9.5.1. Synthetic data

We test the proposed methodology on the following α-stable mixture model:

$$\begin{aligned} p_Y(y) &= 0{,}4 f_{1,2,0,5}(y|1,-4{,}25) + 0{,}2 f_{1,2,0}(y|0{,}5,0{,}3) \\ &+ 0{,}4 f_{1,5,0,5}(y|0{,}3,3{,}25). \end{aligned} \tag{9.29}$$

Random samples of size $N = 1500$ and distribution provided in equation (9.29) are generated using the algorithm proposed by Chambers et al. [Chambers *et al.*, 1976]. We consider the following settings for the hyperparameters: $\alpha_0 = \beta_0 = 1$ are the parameters of the Inverse Gamma distribution. The hyperparameters for the mean and the variance of the Gaussian distribution are $\xi = 0{,}2$ and $\kappa^{-1} = 1/5$. The hyperparameter of the Dirichlet prior for the weights w is set to $\zeta = 1$ and the size of the support of parameters α and β is 2, hence $a = b = 2$. The MCMC and RJMCMC described in Section 9.4 were run for 10000 iterations and a burn-in period of 1000 iterations was considered. k_0 is chosen to be equal to 10, although the number of components never exceeded $k = 6$. As was stated in Section 9.4.4, a Metropolis algorithm is used to estimate the parameters $\theta = \{\alpha, \beta, \gamma, \mu\}$ and a Gaussian is chosen as the symmetric proposal distribution. The standard deviation σ_θ of the Normal distribution, for every α-stable parameter, is set to $\sigma_\alpha = 0{,}15$, $\sigma_\beta = 0{,}1$, $\sigma_\gamma = 0{,}1$ and $\sigma_\mu = 0{,}2$. Initially, we consider that the number of components is 6 and the initial values for mixture model parameters are: $w_j = 1/6$, $\alpha_j = 1{,}1$, $\beta_j = 0$, $\sigma_j = 1$ for every j and $\mu =$ [-3 -1 1 2 3 5].

In Table 9.1, we display the true value, the estimated value and the root mean standard (RMS) deviation of the posterior distribution of the α-stable parameters for the studied mixture model. The standard deviation obtained for β_1 is greater than for any other parameter. This is justified by the fact that for α-stable distribution, as α increases, the skewness parameter β becomes irrelevant. Thus, for values of α close to 2, different values of β do not change very much the shape of the distribution.

Figure 9.2 shows a histogram with the number of components estimated after burn-in period for the generated data with distribution (9.29) using mixture of α-stable and a comparison with the number of components estimated assuming mixture of Gaussian [Richardson & Green, 1997]. The true number of components ($k = 3$) is obtained most of the times for our algorithm. Nevertheless, mixture of Gaussian model overestimated the true number of components

Table 9.1: Simulation results

Parameter	**True value**	**Estimated value**	**Standard deviation**
α_1	1.20	1.27	0.09
β_1	0.50	0.65	0.08
γ_1	1.00	0.98	0.06
μ_1	-4.25	-4.3	0.6
w_1	0.40	0.40	0.02
α_2	1.20	1.30	0.17
β_2	0.00	0.04	0.3
γ_2	0.50	0.45	0.05
μ_2	0.30	0.4	0.3
w_2	0.20	0.198	0.018
α_3	1.50	1.37	0.12
β_3	0.50	0.34	0.20
γ_3	0.30	0.295	0.016
μ_3	3.25	3.24	0.06
w_3	0.40	0.398	0.018

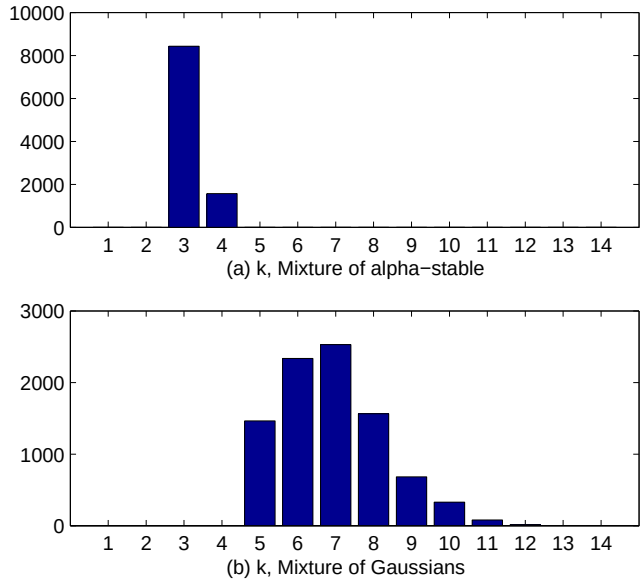

Figure 9.2: Histogram of the number of components estimated in every iteration. 10000 iterations are considered. a) Mixture of α-stable: the true number of components, $k = 3$, is obtained most of times. b) Mixture of Gaussians.

in order to model the impulsiveness and never obtained the true number of components.

In Figure 9.3, the number of components estimated for every iteration is depicted. The true number of components was reached very quickly. The initial value for k was set to 6 (in the mixture of α-stable and Gaussian case). Our algorithm obtained the true value $k = 3$ the first time after less than 100 iterations.

In Figure 9.4, we plot the discrete histogram of the data sequence y together with the predicted multimodal density obtained using mixture of α-stables. In the same figure, the predicted density assuming mixture of 3 Normal distributions is plotted as well for comparison. The predicted multimodal α-stable density fits the discrete histogram very well. Nevertheless, mixture of 3 Gaussians is not capable of fitting the data. As it was pointed before, the mixture of

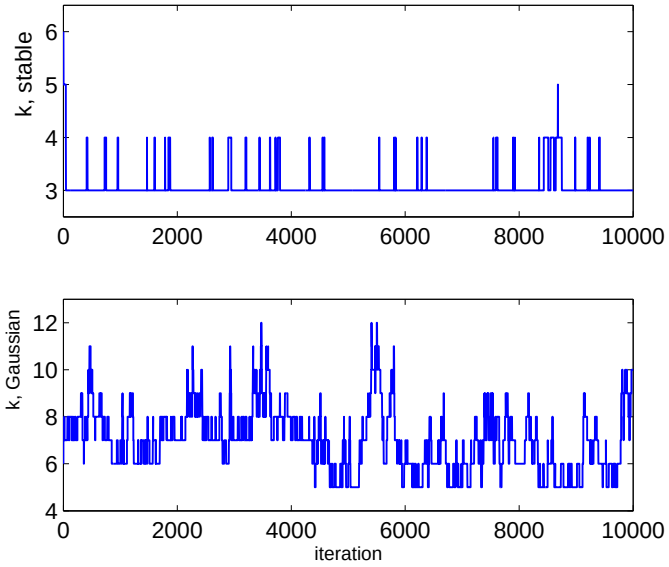

Figure 9.3: Evolution of the number of components estimated at every iteration. Top: Mixture of α-stable: the true number of components $k = 3$ is reached at first time in less than 100 iterations. Bottom: Mixture of Gaussians.

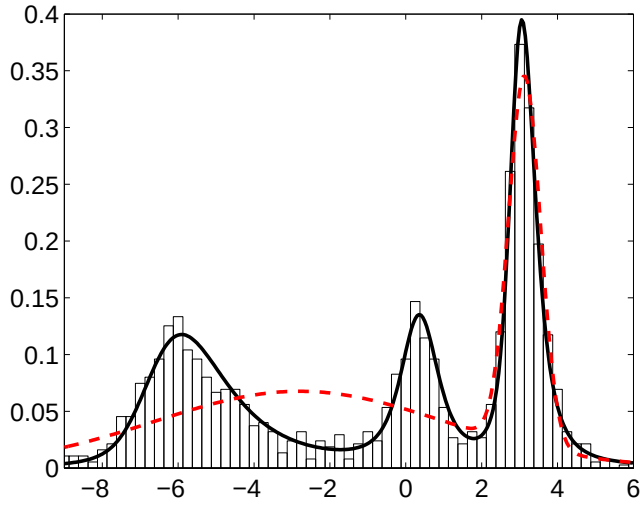

Figure 9.4: Histogram for observations of α-stable mixtures y_i. Solid line: predicted mixture of 3 α-stable components. Dashed line: Mixture of Normals with 3 components.

Gaussians model overestimates the number of components when the data has a heavy-tailed distribution.

The α-stable distribution has four parameters while the Gaussian distribution only two. Due to this fact, in order to perform a comparison between these two different models, it is necessary to compare scenarios in which the same number of unknown parameters are involved. Every α-stable component has 5 parameters while a Gaussian component has only 3, where the corresponding unknown weight w_j was also considered. Therefore, the number of parameters in a mixture of 5 Gaussian components and 3 α-stable are both the same.

The predicted $k = 5$ Gaussian mixture density, the $k = 3$ α-stable mixture and the discrete histogram are plotted jointly in Figure 9.5. In this figure, the performance of mixture of Stable and Gaussian looks very similar but the main difference between both approaches is in the outliers. The tails of the Gaussian distribution are exponential, not algebraic, and the α-stable exhibits Paretian behaviour in the tails. This makes mixture of α-stables more suitable to model

Table 9.2: Measures of probability distance. Comparison between 3 stable mixtures and 5 Gaussian mixtures.

Distance	**Stable**	**Gaussian**
Kullback-Leibler	0.0037	0.0141
Hellinger	0.0021	0.0060
χ^2	0.0080	0.0216

mixture of impulsive data than the Gaussian distribution. The Kullback-Leibler, the Hellinger and the χ^2 distance [Borovkov, 1998] were calculated to measure which of the two mixtures models fits the data better. The calculated distances are shown in Table 9.2. The measured distance is lower for the α-stable case, therefore, the mixture of α-stables fits the analyzed data better.

In the previous simulation, various distance measure methods were calculated and it was shown that the α-stable mixture model was more suitable to fit the impulsive data considered. Nevertheless, as it was pointed out in Figure 9.5 the main difference between this approach and mixture of Gaussians was in the outliers and not in the bulk of the data. A second simulation in which Stable Mixture model performs very accurately and mixture of Gaussians fails also in the bulk will be considered. In this case a mixture of α-stable distribution with lower value of the parameter α than in the previous simulation will be considered. This case corresponds to a more impulsive data.

Random samples of size $N = 1500$ and distribution

$$\begin{aligned} p_Y(y) &= 0{,}4 f_{0{,}8,-0{,}5}(y|1,-1) + 0{,}3 f_{1,2,0}(y|0{,}3,0) \\ &+ 0{,}3 f_{0{,}8,-0{,}5}(y|0{,}6,4). \end{aligned} \tag{9.30}$$

are generated. The settings for the hyperparameters are considered the same as in the previous case. This data was fitted using both, a mixture of α-stable with 3 components and a mixture of 5 Gaussian distributions. The predicted density and discrete histogram of the data are plotted joined in Figure 9.6. In this case, mixture of Gaussians is not able to work due to the outliers. The modes are very near and Gauss mixture cannot fit that resolution while alpha stable mixture can. In this scenario, the proposed algorithm presents a clear advantage over the mixture of Normal model.

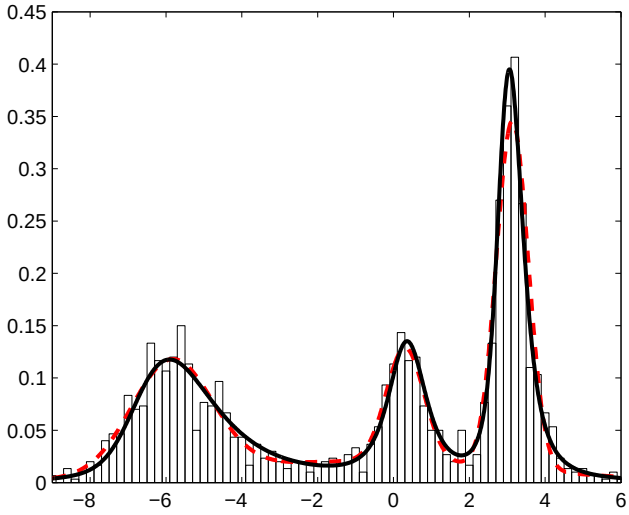

Figure 9.5: Simulation 1. Histogram for observations of α-stable mixtures y_i. Solid line: predicted mixture of α-stable density. Dashed line: Mixture of Normals with 5 components.

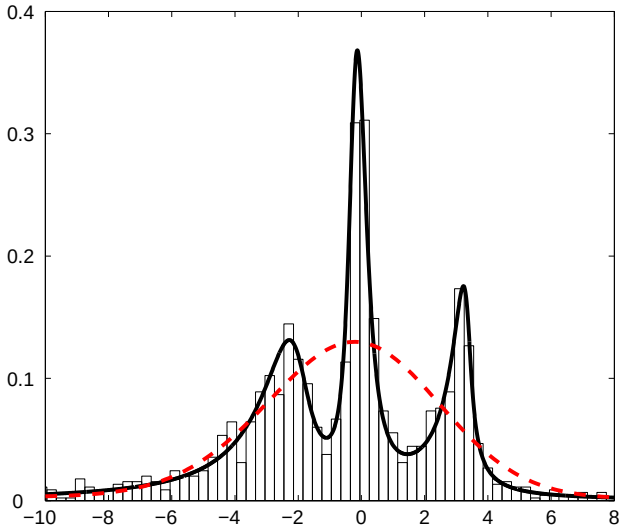

Figure 9.6: Simulation 2. Histogram for observations of α-stable mixtures y_i. Solid line: predicted mixture of α-stable density. Dashed line: Mixture of Normals with 5 components.

9.5.2. Comparison with previous work

Our proposed methodology presents several advantages over the unpublished technical report [Casarin, 2004]:

- In [Casarin, 2004], every parameter is estimated using the Gibbs sampler proposed in [Buckle, 1995]. On the contrary, we evaluate numerically the stable density. Both approaches are compared in [Lombardi, 2007] and the Gibbs sampler is proved to take twice the time as the numerical evaluation of the likelihood. Furthermore, the parameters in the simulations presented in [Buckle, 1995] are initialized to values very near the true values.

- In [Casarin, 2004], the number of components in the mixture is assumed to be known. Our proposed methodology is more flexible and allows us to estimate the number of subpopulations in the mixture.

- In [Casarin, 2004], an auxiliary variable with dimension equal to the number of observations is introduced. This auxiliary variable is updated at every iteration using rejection sampling. Therefore, the efficiency of this algorithm decreases with the number of observations. Namely, the number of unknown quantities in our approach, considering unknown number of components, is $5k + 1$ while in [Casarin, 2004] is $5k + N$, where N is the dimension of the observation vector.

- In [Casarin, 2004], the algorithm is tested on a mixture which has components with modes well away from each other. A synthetic dataset of 1000 observations is generated from the following stable mixture:

$$0{,}5 f_{1,7,0,3}(y|1,1) + 0{,}5 f_{1,3,0,5}(y|1,30) \tag{9.31}$$

9.5.3. Real data

The proposed algorithm is also tested on two different datasets. The first dataset contains daily 3-months interest rates on Euro-Deposits in France between 01/01/1988 and 13/01/2003. This dataset was also studied in [Casarin, 2004]. Figure 9.7 shows jointly the histogram and the predicted mixture of two α-stable densities. The Table 9.3, presents a comparison between the estimated values obtained for a mixture of two components using our algorithm and the values obtained in [Casarin, 2004].

The second dataset consists on 245 values of enzymatic activity in the blood, for an enzyme involved in the metabolism of carcinogenic substances [Bechtel

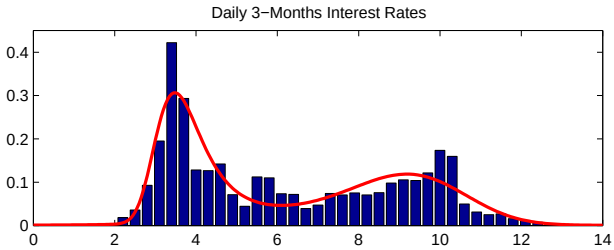

Figure 9.7: Histogram for daily 3-Months Interest Rates on Euro-Deposits in France between 01/01/1988 and 13/01/2003. Solid line: predicted mixture of α-stable density.

Table 9.3: Comparison between estimated values for every parameter of the mixture of α-stable for the daily interest rate dataset. Estimate$_1$ denotes the proposed algorithm. Estimate$_2$ the values obtained in [Casarin, 2004]

Parameter	**Starting Value**	**proposed method**	**method of [Casarin, 2004]**
α_1	1.5	1.3	1.2
α_2	1.5	1.7	1.2
β_1	0.01	0.97	0.02
β_2	0.01	-1.00	0.04
γ_1	1.5	0.493	0.307
γ_2	1.5	1.129	0.873
μ_1	4	4.443	3.012
μ_2	10	8.505	7.301
w_1	0.5	0.535	N/A
w_2	0.5	0.465	N/A

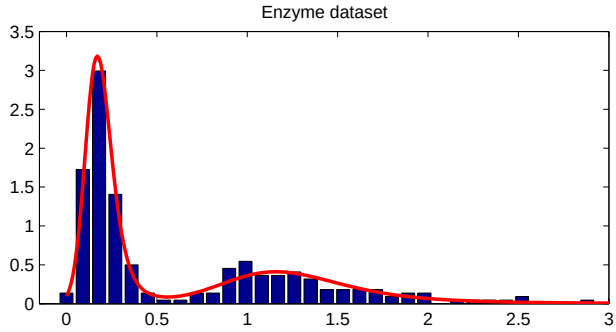

Figure 9.8: Enzymatic activity in the blood for an enzyme involved in the metabolism of carcinogenic substances. Solid line: predicted mixture of α-stable density.

et al., 1993]. There are clearly two different subpopulations of slow and fast metabolisers for this population. Figure 9.8 shows jointly the discrete histogram and the predicted mixture of two α-stable densities. In this case, the data was fitted by two skewed components with parameters $\alpha_1 = 1{,}6620$, $\beta_1 = 0{,}8930$, $\gamma_1 = 0{,}0552$, $\mu_1 = 0{,}2047$, $w_1 = 0{,}6239$, $\alpha_2 = 1{,}5545$, $\beta_2 = 0{,}9064$, $\gamma_2 = 0{,}2589$, $\mu_2 = 1{,}3911$ and $w_2 = 0{,}3761$.

9.6. Conclusion

The alpha-stable mixture distribution is presented in this chapter. The estimation problem is studied and a fully Bayesian approach methodology is proposed to infer on parameters for this mixture model. Furthermore, the number of components in the mixture is assumed to be unknown and the reversible jump Markov chain Monte Carlo method is used to infer on it. The lack of an analytical expression for the probability density function of the α-stable distribution is overcome via a numerical approximation of the stable pdf.

The methodology was tested on synthetic data and every parameter was estimated very accurately. The proposed method was shown to obtain the true number of mixtures and every parameter precisely and very quickly. The convergence of the proposed algorithm was reached after only one hundred iterations. Furthermore, the algorithm was tested on real data and compared to previous works in the literature. Our method presented several advantages, it was con-

siderably faster and easier to implement and, furthermore, was able to work in more difficult scenarios than the previous works in mixture of α-stable distributions.

The proposed methodology was also compared with the Gaussian mixture model. Our method has shown to be more successful in modeling impulsive and skewed data. Besides, the proposed method allows to model not only impulsive data but also non-symmetric data. In this chapter, we have accomplished a mixture model which is a generalization of the Gaussian mixture model and hence satisfies many desiring properties as the Gaussian case, with the added flexibility of being able to model impulsive and skewed data with much less components than required by the Gaussian case.

CHAPTER

TEN

MODELLING MICROARRAY GENE EXPRESSION USING α-STABLE DISTRIBUTIONS

After normalization, the distribution of gene expressions for very different organisms have a similar shape, they usually exhibit heavier tails than a Gaussian distribution and a certain degree of asymmetry. Therefore, this distribution has been modelled in the literature using different parametric families of distributions, such the Asymmetric Laplace or the Cauchy distribution. Moreover, it is known that the tails of spot intensity distributions are described by a power law and the variance of a given array increases as the number of gene considered increases. These features of the distribution of gene expression strongly suggest that the α-stable distribution is suitable to model it.

In this chapter, we model the error distribution for gene expression data using the α-stable distribution [Salas-Gonzalez *et al.*, 2006d, 2007b; Kuruoglu *et al.*, 2007]. This distribution is tested successfully for four different datasets. The Kullback-Leibler, χ^2 and Hellinger tests are performed to compare how α-stable, Asymmetric Laplace and Gaussian fit the spot intensity distribution. The α-stable is proved to perform much better for every array in every dataset considered.

10.1. Introduction

Dna microarray has been established as a powerful tool to study the RNA expression levels of thousands of genes simultaneously under different condi-

tions. Namely, these experiments compare two different samples of cDNA dyed with different colours (red and green) by the mean of the fluorescence intensity measured in the microarray after hybridization. This methodology allows us to compare a large amount of information simultaneously in order to identify and quantify the genes which are differentially expressed.

It is well known that the independence assumption between genes is not true, but many of the works in identification of differential expression in microarray are based in this assumption [Lonnstedt & Speed, 2002; Gottardo *et al.*, 2003; Bhowmick *et al.*, 2006]. Moreover, independency is assumed in the majority of the approaches based in Bayesian statistical methods. Most of these works are based on the assumption of independency between genes and Gaussian distribution as a device to obtain an analytic formula. But the distribution of gene expression, also known as the error distribution for gene expression data, has been also modelled under different approaches:

- [Kuznetsov, 2001] models this distribution using different classes of skewed probability functions such Poisson, exponential, logarithmic series and Pareto-like distribution. He shows the results only for the Pareto-like distribution as it is claimed that this distribution fits the discrete gene expression distribution better than the other distributions.

- In [Hoyle *et al.*, 2002], a wide range of datasets are analyzed empirically and the error distribution is approximated by two distributions: a log-normal in the bulk of microarray spot intensities and a power law in the tails. Furthermore, in this article it is pointed out that the variance of log spot intensity shows a positive correlation with the number of genes considered. Namely, the variance increases as the length of the arrays increases.

- In [Purdom & Holmes, 2005], the gene expression distribution is fitted using the Asymmetric Laplace distribution. The improvement upon the Gaussian distribution is notable, as the Asymmetric Laplace presents asymmetry and heavy tails. One justification for the use of this distribution is based on the fact that it can be represented as the log-ratio of two independent random variables with Pareto distribution.

- In [Khondoker *et al.*, 2006], a statistical model for estimating gene expression using data from multiple laser scans is presented. They also point out that the distribution of gene expression exhibits heavy tails and a Cauchy distribution is adopted to model it.

In this chapter, we propose to model the gene expression distribution with α-stable distributions. We demonstrate that this distribution is able to fit the error distribution for gene expression very accurately. Furthermore, the α-stable distribution has many advantages when compared to other existing approaches in the literature, as will be emphasized in the chapter.

This chapter is organized as follows: In Section 10.2, the α-stable distribution and its main properties are presented. In Section 10.3, we model the arrays from four different datasets with an α-stable distribution. In Section 10.4, the motivation and comparison of the proposed methodology to other existing approaches in the literature is discussed. Lastly, in Section 8.6, we summarize the conclusions.

10.2. An overview of the α-stable distribution

The α-stable distribution is a family of distributions that presents heavy tails and is also capable of exhibiting a certain degree of asymmetry. This distribution has been used in the literature successfully to model skewed and impulsive phenomena. Furthermore, the α-stable distribution is a generalisation of the Gaussian distribution and allows us to describe impulsive processes by means of a small number of parameters.

This distribution has been widely studied in the literature and its properties are very well understood. It satisfies the Generalized Central Limit theorem which states that the limit distribution of infinitely many i.i.d. random variables, possibly with infinite variance distribution, is a stable distribution [Feller, 1966]. The α-stable distribution also satisfies the stability property which states that any combination of random variates with α-stable distribution is also α-stable. More information about the main properties of this distribution can be found in [Samorodnitsky & Taqqu, 1994].

The α-stable distribution has four parameters, the shape parameter $\alpha \in (0, 2]$ is the characteristic exponent which sets the level of impulsiveness. $\beta \in [-1, +1]$ is a skewness parameter, ($\beta = 0$, for symmetric distributions and $\beta = \pm 1$ for the positive/negative stable family respectively). $\gamma > 0$ is the dispersion, a scale parameter and $\mu \in [-\infty, +\infty]$ is a shift parameter called location parameter.

There is not a general closed expression for the α-stable probability density function (pdf) so it is usually defined by its characteristic function which is given by:

$$\varphi(\omega) = \begin{cases} e^{-|\gamma\omega|^{\alpha}[1-i\mathrm{sign}(\omega)\beta\tan(\frac{\pi\alpha}{2})]+i\mu\omega}, (\alpha \neq 1) \\ e^{-|\gamma\omega|[1+i\frac{2}{\pi}\mathrm{sign}(\omega)\beta\log(|\omega|)]+i\mu\omega}, (\alpha = 1) \end{cases} \tag{10.1}$$

Only for three particular cases it is possible to write the α-stable pdf. A distribution with characteristic exponent $\alpha = 2$ corresponds to a Gaussian distribution with $\gamma = \sigma/\sqrt{2}$ where σ is the standard deviation. The $\alpha = 1$ and $\beta = 0$ case corresponds to a Cauchy distribution and for $\alpha = 1/2$ and $\beta = 1$ to a Pearson distribution. Thus, the α-stable distribution can be seen as a generalization of the Normal distribution and some features of linear system theory developed for Gaussian distribution can be extended directly to the α-stable distribution.

The α-stable density, except for the three particular cases mentioned above, must be calculated numerically. Moreover, it exhibits heavier tails than a Gaussian distribution. In other words, it is more likely to obtain samples far from the mean for i.i.d. distributed as an α-stable distribution with characteristic exponent $\alpha < 2$ than for the Gaussian case. This impulsive behaviour is a very well known feature of the distribution of gene expressions.

When $\alpha < 2$, the tails probability $\{P < -\lambda\}$ and $\{P > \lambda\}$ as $\lambda \to \infty$, behave like the power law $\lambda^{-\alpha}$. This is also a known property of the distribution of gene expressions: the tails of the error distribution for gene expression data is also well described by a power law (Paretian tail behaviour).

Let X be a vector with α-stable distribution and $0 < \alpha < 2$. Then,

$$E|X|^p < \infty \text{ for any } 0 < p < \alpha, \tag{10.2}$$

$$E|X|^p = \infty \text{ for any } p \geq \alpha. \tag{10.3}$$

Thus, α-stable random variables with $\alpha < 2$ have an infinite second moment and therefore the second order statistics has no meaning for these variables. The standard deviation for a given random variable with α-stable distribution does not converge to a meaningful value and an increase in the standard deviation is observed as the length of the α-stable random vector increases.

In order to show the behaviour of the α-stable pdf, the stable density for varying α with $\beta = 0$ and varying β with $\alpha = 1{,}5$ are plotted in Figure 10.1. On one hand, in Figure 10.1a shows how the α parameter governs the degree of impulsiveness. Lower values of this parameter means heavier tails and higher peak of the α-stable distribution. On the other hand, as it is plotted in Figure 10.1b, an α-stable distribution with $\beta = 0$ is symmetric and as the skewness parameter β goes to ± 1, the distribution exhibits a higher degree of asymmetry.

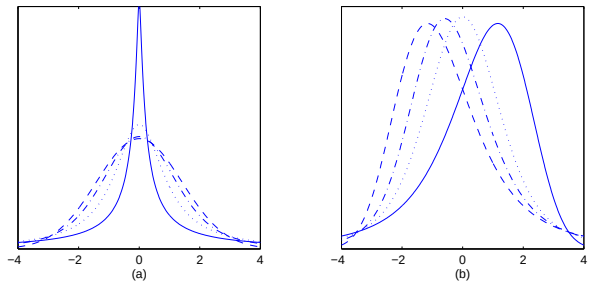

Figure 10.1: Density plot of α-stable distribution with location parameter $\mu = 0$ and $\gamma = 1$. (a) $\alpha = 1{,}5$. Solid line: $\beta = -1$. Dotted line: $\beta = 0$. Dash-dotted line: $\beta = 0{,}5$. Dashed line: $\beta = 1$. (b) $\beta = 0$. Solid line: $\alpha = 0{,}5$. Dotted line: $\alpha = 1$. Dash-dotted line: $\alpha = 1{,}5$. Dashed line: $\alpha = 2$.

10.3. Microarray data analysis

We model the distribution of gene expressions using the α-stable distribution for 4 different microarray datasets. The first dataset (labelled as 'self-self') consists of self-self hybridization of 19 different human cancer cell lines, the Stratagene universal reference RNA and RNA isolated from a tumor specimen [Yang *et al.*, 2002a]. The second dataset ('zebrafish') is two sets of dye-swap experiments for a total of four replicate hybridizations. For each of these hybridizations, target cDNA from the swirl mutant was labeled using one of the Cy3 or Cy5 dyes and the target cDNA wild-type mutant was labeled using the other dye. This experiment was carried out using zebrafish as a model organism to study early development in vertebrates. Swirl is a point mutant in the BMP2 gene that affects the dorsal/ventral body axis[1]. The third dataset ('lymphoma') consists of tumor samples from diffuse large B-cell lymphoma patients [Alizadeh *et al.*, 2000]. The last dataset ('yeast') is an analysis of regulatory variation in a cross between laboratory and wild strains of Saccharomyces Cerevisiae [Yvert *et al.*, 2003]. These four datasets were chosen because they were also analyzed in [Purdom & Holmes, 2005], therefore it is possible to compare their results with our proposed methodology. Every dataset was normalized using locally weighted linear regression (LOWESS) [Cleveland & Delvin, 1988]. This method is capable of removing intensity dependence in $\log_2(R_i/G_i)$ values and has been success-

[1]This data is available as a dataset with the R package marrayClasses.

fully applied to microarray data [Yang *et al.*, 2002b]. After normalization, each distribution of the gene expression has a similar shape: it exhibits heavier tails compared to Gaussian distribution and a certain degree of asymmetry.

There are different approaches to estimate the α-stable parameters [Kuruoglu, 2001; Kogon & Williams, 1998]. For every array in the dataset, we estimate them using the maximum likelihood approach [Nolan, 2001]. The parameter estimates are shown in Figure 10.2. It is seen that the difference between the location (μ) and dispersion (γ) parameters estimated for the 'self-self' data is very low. For this dataset, the same RNA sample is labelled separately with green and red fluorescent dyes and hybridized to the same microarray, therefore, the gene expression distribution is expected to be symmetric. We obtain values of the skewness parameter β very close to zero in almost every case. There are only three cases in which β parameters are not near zero. They are the arrays 8, 18 and 24, (note that they are plotted with big circles in the figure). These three values are $\{\alpha_8 = 1{,}83, \beta_8 = -0{,}48\}$, $\{\alpha_{18} = 1{,}94, \beta_{18} = -0{,}65\}$ and $\{\alpha_{24} = 1{,}86, \beta_{24} = -0{,}77\}$, so the α parameter for every one of them is very near 2. A well known property of the α-stable distribution is that as the exponent α tends to the limiting value 2, more symmetric is the α-stable distribution and the β parameter less affect on the shape. Therefore, these values of β are consistent with the expected symmetry of the distribution.

Figure 10.3 shows the distribution of the gene expression for an example array of every dataset. It is seen that α-stable distribution fits the discrete distribution of the gene expression very accurately and better than the Asymmetric Laplace and Gaussian. It is also seen that, despite the heavy tails and skewness of the Asymmetric Laplace distribution, this distribution has a very thin peak which is not always shown in gene expression data. See how in the figure, the 'self-self' array considered is fitted worse than the 'yeast' array using the Asymmetric Laplace distribution. α-stable distribution, however, presents a smoother behaviour in the peak which allows us to fit the data better. It is also seen that the Gaussian distribution is not able to fit the gene expression data as the discrete histograms present heavier tails than the Normal distribution.

In order to compare numerically how α-stable, Asymmetric Laplace and the Normal distribution fit the gene expression distributions, we calculated the Kullback-Leibler, χ^2 and Hellinger distance [Borovkov, 1998]. The χ^2 distance penalizes possible outliers in the fitting. Namely, a small amount of samples affects more the measured χ^2 distance than for the Hellinger and K-L distance. The former is the most robust to outliers among the three distances considered. These tests were applied to each array and better performance for the α-stable distribution was obtained for all of them. Table 10.3 shows the corresponding

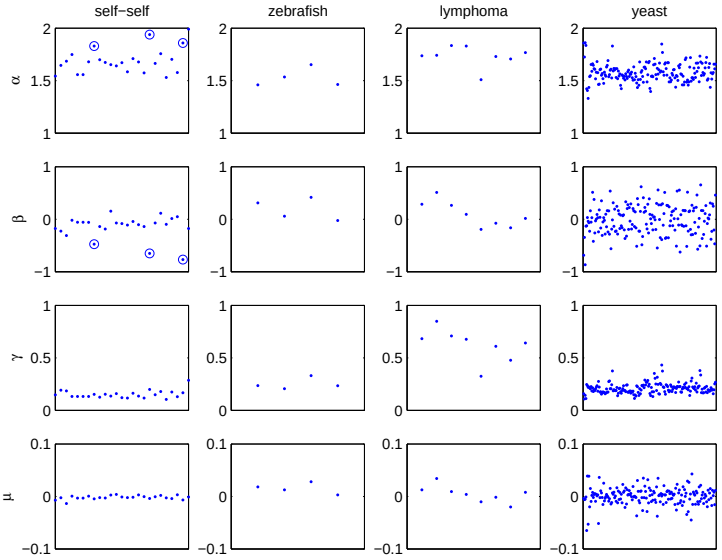

Figure 10.2: Estimated parameters for every dataset. First row: characteristic exponent α. Second row: skewness parameter β. Third row: dispersion γ. Fourth row: location parameter μ.

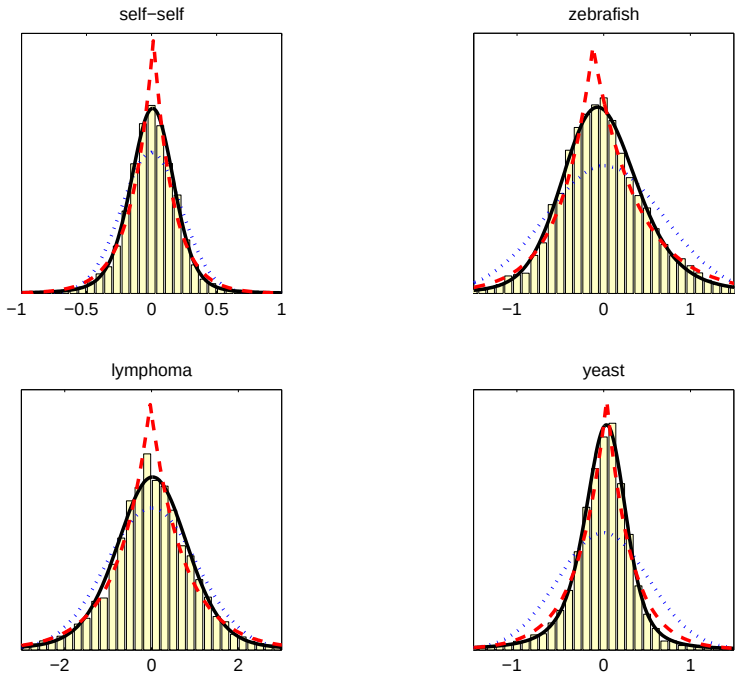

Figure 10.3: Discrete gene expression histogram and predicted density for one array of each dataset. From the 'self-self' dataset we choose the array 9 (NT2.2(testis)). From the 'zebrafish' and 'lymphoma' dataset, the 2 array and DLCL-0024 are chosen respectively. From 'yeast' dataset, we used 14-4-aCy3. Solid line: α-stable distribution. Dashed line: Asymmetric Laplace distribution. Dotted line: Gaussian distribution.

mean and standard deviation of these tests for every dataset. It is shown that the α-stable distribution fits the discrete gene expression distribution much better than the Asymmetric Laplace and Gaussian. Furthermore, the standard deviation is considerably lower for the α-stable case. This means that the Asymmetric Laplace and Gaussian, contrary to α-stable, fit the gene expression distribution accurately or poorly depending on the array. This fact was remarked on the last paragraph and it was illustrated in Figure 10.3. The lower mean and standard deviation obtained for the K-L, χ^2 and Hellinger distance for the α-stable distribution, shows that this distribution fits accurately for each array.

Table 10.1: Kullback Leibler, χ^2 and Hellinger distance between the discrete gene expression distribution and the predicted stable, Asymmetric Laplace and Gaussian density for each dataset. The number denotes the mean of the distance calculated for each dataset. In brackets, the error (standard deviation). In bold the lowest distance and standard deviation.

	self-self	zebrafish	lymphoma	yeast
KL(Stable)	**0.013 (0.005)**	**0.0171 (0.0020)**	**0.021 (0.003)**	**0.018 (0.005)**
KL(ALaplace)	0.022 (0.019)	0.066 (0.021)	0.022 (0.003)	0.047 (0.017)
KL(Gauss)	0.10 (0.04)	0.35(0.10)	0.07 (0.03)	0.25 (0.09)
χ^2(Stable)	**0.015 (0.008)**	**0.015 (0.003)**	**0.022 (0.004)**	**0.016 (0.005)**
χ^2(ALaplace)	0.04 (0.06)	0.074 (0.019)	0.038 (0.008)	0.058 (0.022)
χ^2(Gauss)	0.12 (0.05)	0.6 (0.3)	0.09(0.05)	0.39 (0.21)
Hell.(Stable)	**0.0036 (0.0021)**	**0.0043 (0.0007)**	**0.0061 (0.0011)**	**0.0044 (0.0014)**
Hell.(ALaplace)	0.009 (0.009)	0.020 (0.005)	0.0087 (0.0015)	0.015 (0.005)
Hell.(Gauss)	0.030 (0.012)	0.11 (0.04)	0.025 (0.012)	0.07 (0.03)

10.4. Discussion

In the previous section, the α-stable distribution has been proved to model very accurately the distribution of gene expression for different datasets. In this section, the α-stable methodology is compared to the four other approaches in the literature which were presented in the introduction.

- [Kuznetsov, 2001] noted that the gene expression distribution follows a Pareto-like distribution. He modelled the gene expression distribution using several classes of skewed probability functions and obtained a better performance using Pareto-like. He introduced an artificial location parameter to generalize the Pareto distribution. α-stable distribution already accounts for this parameter and provides a good fit in both the main lobe and the tails of the distribution. He demonstrated that the empirical histograms of gene expression levels are well fitted by a power law distribution. The α-stable distribution also has a Paretian tail behaviour when $\alpha < 2$. Specifically, if X is a random variable with α-stable distribution with $\alpha < 2$, then [Samorodnitsky & Taqqu, 1994]:

$$\lim_{\lambda \to -\infty} \lambda^{\alpha} \{P < \lambda\} = C_{\alpha} \frac{1+\beta}{2} \gamma^{\alpha} \tag{10.4}$$

$$\lim_{\lambda \to -\infty} \lambda^{\alpha} \{P < -\lambda\} = C_{\alpha} \frac{1-\beta}{2} \gamma^{\alpha} \tag{10.5}$$

  where

$$C_{\alpha} = \frac{1-\alpha}{\Gamma(2-\alpha)\cos(\pi\alpha/2)} \text{ if } \alpha \neq 1 \tag{10.6}$$

$$C_{\alpha} = \frac{2}{\pi} \text{ if } \alpha = 1 \tag{10.7}$$

  Furthermore, Mandelbrot remarked the fact that the use of the Stable distribution for describing empirical principles was preferable to the use of Zipf-Pareto distribution for both, theoretical and practical reasons. (See [Zolotarev, 1986] for a deeper explanation regarding Stable laws in biology).

- In [Hoyle *et al.*, 2002], a wide range of datasets are analyzed. The error distribution is approximated by a log-normal in the bulk of microarray spot intensities which is claimed to be a good approximation for the distribution of most of the spot intensity values. It is also pointed out that

the tails of the distribution show good agreement with Zipf's law, a special case of Pareto behaviour (or power law) [Newman, 2005]. Therefore, two different distributions are used to model the distribution of gene expression, log-normal in the bulk and power law in the tails. Two possible, and heuristic, explanations for this different behaviour are given in [Hoyle *et al.*, 2002]. Contrary to that, the α-stable distribution allow us to model the gene expression distribution with only one distribution, modelling very accurately both the center and the tails.

Furthermore, in [Hoyle *et al.*, 2002], it is pointed out that the variance σ^2 of log spot intensity increases as the number of genes considered increases. This is in complete agreement with the properties of the α-stable distribution as it was stated in Section 10.2. The variance is not defined for stable processes with $\alpha < 2$, so the second order statistics can not help us to gain an insight into stable random variables. Due to this fact, the standard deviation of a random variable with α-stable distribution increases as the length of the random vector increases and it does not converge to a given value.

- In [Purdom & Holmes, 2005], an Asymmetric Laplace distribution is used to fit the gene expression distribution and its performance is compared to the Gaussian distribution. In the previous section, our methodology was compared experimentally to the Asymmetric Laplace distribution and it was shown that this distribution does not always fit properly the gene expression distribution. The histogram of gene expression levels often presents a smoother behaviour in the maximum. Furthermore, although the Asymmetric Laplace distribution presents heavier tails than the Gaussian distribution, the tails of this distribution are exponential, not algebraic, and do not exhibit Paretian behaviour. A Laplace and Assymmetric Laplace distribution for identification of differential expression in microarray experiments have been assumed recently in [Bhowmick *et al.*, 2006]. We believe that an α-stable assumption for the gene expression distribution could help to build new statistical methods in order to asses whether a gene is differentially expressed or not.

- In [Khondoker *et al.*, 2006], the distribution of gene expression is modelled using a Cauchy distribution as a part of a statistical model for estimating gene expression using data from multiple laser scans. The Cauchy distribution is chosen rather than assuming a Normal distribution in order to take into account the outliers. We do not assume neither Gaussian nor

Cauchy but both are particular cases of the α-stable family. Specifically, for $\alpha = 1$, $\beta = 0$, the inverse Fourier transform of the characteristic function in equation (10.1) has analytical solution. In that particular case, the pdf of the α-stable is

$$\frac{\gamma}{\pi((x-\mu)^2+\gamma^2)} \tag{10.8}$$

which corresponds to a Cauchy distribution with location parameter μ and dispersion parameter γ. If the distribution of gene expression were Cauchy, we would have obtained $\alpha \approx 1$ and $\beta \approx 0$ most of the times for the characteristic exponent and the skewness parameter respectively and the Figure 10.2 showed that the values obtained in the estimation of the shape parameter α are typically in the interval $[1{,}5 - 1{,}8]$.

All these evidences suggest us to use α-stable distribution to model the distribution of the gene expression. In particular, the stable distribution could be used instead of the Gaussian simplification in order to asses whether differential expression has ocurred. Furthermore, the stable distribution has other well known properties, as the Scale Mixture of Normals representation, which has never been used before in microarray data. As the Gaussian distribution is a particular case of the more flexible α-stable distribution family, the α-stable assumption is a long-tailed and skewed alternative that allow us to obtain the same results in identification of differential expression as them based on the Normal distribution [Lonnstedt & Speed, 2002] if the distribution of the gene expression were Gaussian. But the distribution of the gene expression is found empirically not to be Gaussian. We believe that the α-stable approach can help to solve new and more challenging identification of differential expression in microarray problems.

10.5. Conclusion

In this chapter we have presented a new statistical model for the distribution of gene expressions. The model provides the flexibility for modelling impulsiveness and skeweness required for gene expression data. We stress the fact that it is not an ad-hoc model but has strong theoretical justifications such as the generalised central limit theorem. It confirms with earlier observations made by other researchers such as Paretian tails and non-converging standard deviation. Both impulsiveness and skewness are parametrised in a parsimonious way using

the α-stable distribution. A rich variety of techniques exist in the literature for parameter estimation. We believe that the statistical model presented in this chapter will be very useful in estimation and detection problems involving gene expression array data.

CHAPTER

ELEVEN

ASSESSING DIFFERENTIAL GENE EXPRESSION USING THE α-STABLE DISTRIBUTION

DNA microarray has been established as an important tool to study the expression of thousands of genes simultaneously. These experiments allow us to compare two different samples of cDNA obtained under different conditions. In this chapter, we present a novel method for the analysis of replicated microarray experiments based on the modelling of the gene expression distribution as a mixture of α-stable distribution [Salas-Gonzalez *et al.*, 2007a]. Some features of the distribution of genes expression such the Pareto tails and that the variance of a given arrays increases as the number of genes studied increases, suggest to model the gene expression distribution using the α-stable density. The proposed methodology uses very well known properties of the α-stable distribution, as the Scale Mixture of Normals representation. A Bayesian log-posterior odds is calculated which allows us to decide whether a gene is differently expressed or not. The proposed methodology is illustrated using simulated and experimental data and the results are compared with other existing statistical approaches. The proposed heavy-tail model improve the performance of other distributions and it is a convenient model to work with microarray gene data, specially if the dataset contains outliers or it presents high variance between replicates.

11.1. Introduction

Dna microarray has been established as a powerful tool to study the RNA expression levels of thousands of genes simultaneously under different conditions. These experiments compare two different samples of cDNA coloured with different dyes (red and green) measuring the fluorescence intensity after hybridization. This method allows us to compare a large amount of information simultaneously in order to identify and quantify the genes which are differentially expressed. Typically, microarray experiments consist in intensity measures of thousands of genes with a few, if any, replicates for each gene.

After normalization, the gene expression distribution presents, in general, heavier tails than a Gaussian distribution. This distribution has been modelled using several densities: Cauchy [Khondoker *et al.*, 2006], Pareto distribution [Kuznetsov, 2001], Laplace [Purdom & Holmes, 2005], t-student [Lonnstedt & Speed, 2002] or Log-normal [Hoyle *et al.*, 2002]. Recently, we studied several different microarray datasets in [Salas-Gonzalez *et al.*, 2006b] and the α-stable distribution was seen to fit better the gene expression distribution.

In this chapter, we derive a novel statistics that can be used to identify differential expression in microarray experiments. This statistic is based on an α-stable mixture model and the Scale Mixture of Normals property. The very well knowns properties of the α-stable distribution allow us to calculate the parameters of the model. The posterior probability is used to build the statistic and infer in the Bayesian mixture model.

We introduce a Bayesian α-stable mixture model which model genes as composed of two different populations: differentially expressed genes and not differently expressed. The Bayesian mixture model is a very popular method in the gene expression literature [Lonnstedt & Speed, 2002; Bhowmick *et al.*, 2006; Newton *et al.*, 2004; Do *et al.*, 2005; Gottardo *et al.*, 2003].

The article is organized as follows: the main properties of the α-stable distribution are presented in Section 11.2. In Section 11.3, a novel statistic based in the Scale Mixture of Normals Property is presented. The statistic is tested in both, synthetic and real microarray data in Section 11.5. Lastly, the conclusions are drawn in 11.6.

11.2. α-stable distribution

11.2.1. Properties

The α-stable distribution is a family of distributions which presents heavy tails and a certain degree of asymmetry. Its properties are very well understood and it has been used to model impulsive phenomena in many different fields such biology, electrical engineering, computer science, economics, physics and astronomy (see [Zolotarev, 1986; Nikias & Shao, 1995]). Moreover, the α-stable distribution satisfies the Stability Property and the Generalized Central Limit Theorem. In a certain manner, this distribution can be considered a generalisation of the Gaussian distribution which allows us to describe impulsive and skewed processes using only four parameters. More information about the α-stable properties can be found in [Samorodnitsky & Taqqu, 1994].

The α-stable distribution has four parameters, the shape parameter $\alpha \in (0, 2]$ is the characteristic exponent which sets the level of impulsiveness. $\beta \in [-1, +1]$ is a skewness parameter, ($\beta = 0$, for symmetric distributions and $\beta = \pm 1$ for the positive/negative stable family respectively). $\gamma > 0$ is the dispersion, a scale parameter, and $\mu \in [-\infty, +\infty]$ is the location parameter.

There is not a general closed expression for the α-stable probability density function (pdf) and the α-stable pdf can only be written for the following three particular cases: a distribution with a characteristic exponent $\alpha = 2$ corresponds to a Gaussian distribution with $\gamma = \frac{\sigma}{\sqrt{2}}$ where σ^2 is the variance. The $\alpha = 1$ and $\beta = 0$ case corresponds to a Cauchy distribution and for $\alpha = 1/2$ and $\beta = 1$ to a Pearson distribution. Due to this lack of an analytical expression, this distribution is usually defined by its characteristic function which is given by:

$$\varphi(\omega) = \begin{cases} e^{-|\gamma\omega|^{\alpha}[1 - i\,\mathrm{sign}(\omega)\beta\tan(\frac{\pi\alpha}{2})] + i\mu\omega}, (\alpha \neq 1) \\ e^{-|\gamma\omega|[1 + i\frac{2}{\pi}\mathrm{sign}(\omega)\beta\log(|\omega|)] + i\mu\omega}, (\alpha = 1) \end{cases} \tag{11.1}$$

Other important properties of the α-stable related to the gene expression distribution is that α-stable exhibits heavier tails than a Gaussian distribution when $\alpha < 2$. Moreover, when $\alpha < 2$, the tail probability behaves like a power law (Paretian tail behaviour). Lastly, the standard deviation for a given random vector with α-stable distribution does not converge and increases as the length of the α-stable vector considered increases. These three properties were also observed for the microarray gene expression distribution (see the Section 11.2.5).

11.2.2. Scale Mixture of Normals property

Let X and $Y > 0$ be independent random variables with $X \sim f_{\alpha_1,0}(\sigma, 0)$ and $Y \sim f_{\alpha_2,1}((\cos\frac{\pi\alpha_2}{2})^{\frac{1}{\alpha_2}}, 0)$. Then XY^{1/α_1} is stable with parameters $Z \sim f_{\alpha_1\cdot\alpha_2,0}(\sigma, 0)$.

We are interested in the case in which $\alpha_1 = 2$ (Gaussian case) and $\alpha_2 < 1$. For these parameter values, a scale mixtures of normals (SMiN) could be used for the symmetric stable law as follows (see [Godsill & Kuruoglu, 1999] for more details). If y_i is a i.i.d. sample from a symmetric α-stable distribution with location parameter μ and scale parameter σ:

$$y_i \sim f_{\alpha,0}(\sigma, \mu) \tag{11.2}$$

The product property can be used to obtain the following equivalent representation:

$$y_i \sim N(\mu, \lambda_i\sigma^2) \tag{11.3}$$

$$\lambda_i \sim f_{\frac{\alpha}{2},1}(2\left(\cos\frac{\pi\alpha}{4}\right)^{\frac{2}{\alpha}}, 0) \tag{11.4}$$

Where $N(\mu, \lambda_i\sigma^2)$ is the Normal distribution with mean μ and variance $\lambda_i\sigma^2$. This equivalent model is very useful for Bayesian inference as conditionally on the auxiliary positive stable random variable λ_i, the symmetric α-stable variable y_i is Gaussian. Furthermore, it allows us to write a symmetric α-stable distribution as a Scale Mixture of Normals:

$$p(x) = \int_0^\infty N(x|0, \lambda\sigma^2)p(\lambda)d\lambda \tag{11.5}$$

When $p(\lambda) = f_{\alpha/2,1}(\lambda|2\left(\cos\frac{\pi\alpha}{4}\right)^{\frac{2}{\alpha}}, 0)$ then $p(x)$ is symmetric α-stable with distribution $f_{\alpha,0}(x|\sigma, 0)$. Equation (11.5) represents not only a symmetric α-stable but also a large class of distributions which includes the t-Student, Laplace or exponential power law among others [Fernandez & Green, 2002]. The distribution $p(x)$ depends on the mixing distribution $p(\lambda)$. In particular, when this distribution is Inverted Gamma $p(\lambda) = IG(\lambda|\alpha, \beta)$, the distribution $p(x)$ is t-student as it was obtained for the B statistics proposed in [Lonnstedt & Speed, 2002].

11.2.3. Stable random number generation

It is straightforward to draw samples from stable distributions using the method proposed by Chambers, Mallows and Stuck [Chambers *et al.*, 1976]. A

$f_{\alpha,\beta}(1,0)$ random variable X can be generated via a non-linear transformation of an uniformly distributed variate V and an exponential variate W using the following theorem:

Let V be an uniform random variable in the interval $(-\frac{\pi}{2}, \frac{\pi}{2})$ and W be an independent exponential random variable with mean 1. If V and W are independent, then

$$X = S_{\alpha,\beta}\frac{\sin(\alpha(V+B_{\alpha,\beta}))}{(\cos V)^{1/\alpha}}\left(\frac{\cos((1-\alpha)V+B_{\alpha,\beta}\alpha)}{W}\right)^{(1-\alpha)/\alpha} \tag{11.6}$$

is α-stable distributed with $f_{\alpha,\beta}(1,0)$ where

$$B_{\alpha,\beta} = \frac{\arctan(\beta\tan(\pi\alpha/2))}{\alpha} \tag{11.7}$$

$$S_{\alpha,\beta} = (1+\beta^2\tan^2\frac{\pi\alpha}{2})^{1/(2\alpha)}. \tag{11.8}$$

Once the standard stable random variable X is simulated, it is straightforward to generate a stable random variable for any value of the parameters α, β, σ and μ using the following property:

If $X \sim f_{\alpha,\beta}(1,0)$ then

$$Y = \begin{cases} \sigma X + \mu & \text{if } \alpha \neq 1 \\ \sigma X + \frac{2}{\pi}\beta\sigma\log\sigma + \mu & \text{if } \alpha = 1 \end{cases} \tag{11.9}$$

is $f_{\alpha,\beta}(\sigma,\mu)$.

11.2.4. Parameter estimation

One of the main advantages of the α-stable distribution is that its properties has been widely studied in the literature. Therefore, there are many differents parameter estimation techniques. We are interested in estimation of symmetric α-stable distributions with $\mu = 0$. For that case, there are several estimation methods: the fractional moment estimation (see [Nikias & Shao, 1995]), a maximum likelihood alternative for the symmetric case is given in [Bodenschatz & Nikias, 1999]. In [Fan, 2006], an unbiased estimator for the stable index α is developed with the structure of U-statistics. We made use of the log-absolute moment method proposed in [Nikias & Shao, 1995] but any other method could be used in order to estimate the values of the characteristic exponent α and the dispersion σ.

11.2.5. α-stable and microarray gene expression

The α-stable distribution was seen to fit very accurately the distribution of gene expression for different datasets in [Salas-Gonzalez *et al.*, 2006b]. Furthermore, this distribution is able to suit some earlier observations made by other researches:

- In [Kuznetsov, 2001], the gene expression distribution was modelled using several distributions and the better performance was obtained using a Pareto-like distribution. The α-stable distribution also has Paretian tail behaviour when $\alpha < 2$ (see [Samorodnitsky & Taqqu, 1994]).
- In [Hoyle *et al.*, 2002], the gene expression distribution was fitted by two different distributions: a Log-Normal in the bulk and a power law or Zipf's law (a special case of Pareto behaviour) in the tails. The α-stable pdf allows us to model the gene expression distribution using only one distribution. Moreover, it was noticed that the variance of log spot intensities increases as the number of genes increases. This is also an important property of the α-stable distribution. The variance is not defined for stable processes with $\alpha < 2$. Therefore, the variance of random vector with α-stable distribution increases as the length of the random vector increases and does not converge to a given value.
- In [Purdom & Holmes, 2005], the gene expression distribution was modelled using an Asymmetric Laplace distribution. This approach was extensively compared in [Salas-Gonzalez *et al.*, 2006b], where it was shown that the α-stable density fitted much better the gene expression distribution than the Asymmetric Laplace. Furthermore, the Asymmetric Laplace distribution was presented as a long-tailed alternative to the Gaussian distribution but the tails of this distribution are also exponential (as in the Gaussian case) and do not exhibit a Paretian behaviour. More recently, in [Bhowmick *et al.*, 2006], a Laplace mixture model was introduced in order to identity differential expression in microarray experiments. We believe that an α-stable model can help to build novel and robust statistics to state whether a gene is deferentially expressed or not.
- In [Khondoker *et al.*, 2006], the distribution of gene expression is modelled using a Cauchy distribution. In that work it is explained than this long-tailed distribution is chosen rather than a Gaussian distribution to take into account the outliers. We propose to model the error distribution for gene expression using the α-stable distribution. We do not assume neither

Gaussian nor Cauchy, but both are particular cases of the α-stable distribution as it was explained in Section 11.2.1. Therefore, α-stable distribution is a very flexible density which allow us to model the gene expression distribution in a parsimonious way using only one well known density.

11.3. Statistical inference

In this section, we introduce a novel statistic based on the Scale Mixture of Normals and the heavy-tailed α-stable distribution. Let N be the number of genes on each array, n the number of replicates (arrays), $i = 1...N$ and $j = 1...n$. We assume that the data M_{ij} are base 2 logarithms of red dye intensity (denoted as R_{ij}) and green (G_{ij}) conveniently normalized using LOESS normalization [Yang *et al.*, 2002a]. Therefore,

$$M_{ij} = log\left(\frac{R_{ij}}{G_{ij}}\right). \tag{11.10}$$

Our goal is to state which genes are differently expressed. In [Lonnstedt & Speed, 2002], a mixture of Normal and a Dirac distribution is proposed in order to build a statistic B which is a Bayes log posterior odds. In that work, an Inverse Gamma distribution is chosen as the prior distribution for the variance.

Contrary to that, in this work, we model the normalized relative expression measures independently on each gene, assuming that M_{ij} are random variables from a Normal distribution with mean μ_i and variance $\lambda_i\sigma^2$. Although genes interact with each other, the independence between genes is a useful simplification that allows us to build a statistic to state whether a gene is differentially expressed or not. This simplification is also considered in other related works [Bhowmick *et al.*, 2006; Gottardo *et al.*, 2003; Lonnstedt & Speed, 2002]

$$M_{ij}|\mu_i, \lambda_i, \sigma \sim N(\mu_i, \lambda_i\sigma^2) \text{ for all } i. \tag{11.11}$$

Assuming that genes are either differentially expressed or not, let w denote the probability that a given gene is differently expressed. The parameter μ can be regarded as drawn from a symmetric α-stable distribution if the gene is expressed or $\mu_i = 0$, modelled as a Dirac delta $\delta(0)$, if it is not differently expressed.

$$p(\mu_i|\lambda_i, \sigma) = wf_{\alpha,0}(\sigma, 0) + (1 - w)\delta(0). \tag{11.12}$$

Let z_i indicate if a given gene is differentially expressed ($z_i = 1$) or not ($z_i = 0$):

$$z_i = \begin{cases} 0 & \text{if } \mu_i = 0, \\ 1 & \text{if } \mu_i \sim f_{\alpha,0}(\sigma, 0). \end{cases}$$

The log posterior ratio for a given gene i can be calculated as

$$S_i = \log \frac{Pr(z_i = 1|M_{ij})}{Pr(z_i = 0|M_{ij})}, \tag{11.13}$$

and following the Bayes' Theorem and assuming independence between genes

$$S_i = \log \frac{w}{1-w} \frac{Pr(M_i|z_i = 1)}{Pr(M_i|z_i = 0)}, \tag{11.14}$$

where M_i is the vector of the n replicates for gene i. Our goal is to calculate the posterior probabilities $Pr(M_i|z_i = 1)$ and $Pr(M_i|z_i = 0)$ in order to compute the log posterior odds S_i which, for a given gene i, computes the probability of being differently expressed.

We use the Scale Mixture of Normals property to write a symmetric α-stable distribution as a Gaussian, conditional on an auxiliary positive α-stable random variable λ. Therefore, the distribution of μ_i given λ_i is Gaussian

$$\mu_i|\lambda_i, \sigma = \begin{cases} 0 & \text{if } z_i = 0, \\ N(0, \lambda_i\sigma^2) & \text{if } z_i = 1, \end{cases}$$

with the following prior distribution for λ:

$$p(\lambda_i) = f_{\frac{\alpha}{2},1}(2\{\cos\left(\frac{\pi\alpha}{4}\right)\}^{\frac{2}{\alpha}}, 0). \tag{11.15}$$

Considering $M_{i.}$ as the average of M_{ij} for $j = 1 \ldots n$ for a given gene i. The distribution of M_i conditional on μ_i, λ_i and σ is

$$\begin{aligned} p(M_i|\mu_i, \lambda_i, \sigma) &= (2\pi\lambda_i)^{-\frac{n}{2}} \sigma^{-n} e^{-\frac{1}{2\lambda_i\sigma^2}\sum_j (M_{ij}-\mu_i)^2} \\ &= (2\pi\lambda_i)^{-\frac{n}{2}} \sigma^{-n} e^{-\frac{1}{2\lambda_i\sigma^2}(\sum_j (M_{ij}-M_{i.})^2 + n(M_{i.}-\mu_i)^2)}, \end{aligned} \tag{11.16}$$

and, for the proposed model:

$$\begin{aligned} p(M_i|z_i = 1) &= \int\int p(M_i, \mu_i, \lambda_i) d\mu_i d\lambda_i \\ &= \int\int p(M_i|\mu_i, \lambda_i) p(\mu_i|\lambda_i, z_i = 1) p(\lambda_i) d\mu_i d\lambda_i \end{aligned} \tag{11.17}$$

and

$$\begin{aligned} p(M_i|z_i=0) &= \int\int p(M_i|\mu_i,\lambda_i)p(\mu_i|\lambda_i,z_i=0)p(\lambda_i)d\mu_i d\lambda_i \\ &= \int p(M_i|\lambda_i)p(\lambda_i)d\lambda_i. \end{aligned} \tag{11.18}$$

Substituting in the expressions (11.17) and (11.18) the distributions for our model and considering the following equality

$$N(\mu_i|0,\lambda_i\sigma^2)\cdot e^{-\frac{n(M_{i.}-\mu_i)^2}{2\lambda_i\sigma^2}} = N(\mu_i|\frac{n}{n+1}M_{i.},\frac{\lambda_i\sigma^2}{n+1})\cdot(n+1)^{-1/2}\cdot e^{-\frac{1}{2\lambda_i\sigma^2}\frac{n}{n+1}M_{i.}^2}, \tag{11.19}$$

it is possible to integrate out the distribution $N(\mu_i|\frac{n}{n+1}M_{i.},\frac{\lambda_i\sigma^2}{n+1})$ and to obtain the following integrals:

$$\begin{aligned} p(M_i|z_i=1) &= \int (2\pi\lambda_i)^{-\frac{n}{2}}\sigma^{-n}e^{-\frac{1}{2\lambda_i\sigma^2}\sum_j(M_{ij}-M_{i.})^2} \\ &\times \ (n+1)^{-1/2}e^{-\frac{1}{2\lambda_i\sigma^2}\frac{n}{n+1}M_{i.}^2} \\ &\times \ f_{\frac{\alpha}{2},1}(2\{\cos\left(\frac{\pi\alpha}{4}\right)\}^{\frac{2}{\alpha}},0)d\lambda_i \end{aligned} \tag{11.20}$$

and

$$\begin{aligned} p(M_i|z_i=0) &= \int (2\pi\lambda_i)^{-\frac{n}{2}}\sigma^{-n}e^{-\frac{1}{2\lambda_i\sigma^2}\sum_j(M_{ij})^2} \\ &\times \ f_{\frac{\alpha}{2},1}(2\{\cos\left(\frac{\pi\alpha}{4}\right)\}^{\frac{2}{\alpha}},0)d\lambda_i. \end{aligned} \tag{11.21}$$

Due to the non existence of an analytical expression for the α-stable pdf, the integrals (11.20) and (11.21) need to be calculated numerically. Some different approaches could be used in order to accomplish this goal. We made use of the fact that drawing samples from an α-stable distribution can be easily accomplished using the Chambers' algorithm (see the section 11.2.3). If we have T random samples $\left[\lambda_i^{(1)}...\lambda_i^{(t)}...\lambda_i^{(T)}\right]$ with distribution $p(\lambda_i)$ given by eq. (11.15), a Monte Carlo empirical estimate of the integrals (11.20) and (11.21) is

$$\begin{aligned} p(M_i|z_i=1) &= \frac{1}{T}\sum_{t=1}^{T}(2\pi\lambda_i^{(t)})^{-\frac{n}{2}}\sigma^{-n}e^{-\frac{1}{2\lambda_i^{(t)}\sigma^2}\sum_j(M_{ij}-M_{i.})^2} \\ &\times \ (n+1)^{-1/2}e^{-\frac{1}{2\lambda_i^{(t)}\sigma^2}\frac{n}{n+1}M_{i.}^2} \end{aligned} \tag{11.22}$$

and

$$p(M_i|z_i=0)=\frac{1}{T}\sum_{t=1}^{T}(2\pi\lambda_i^{(t)})^{-\frac{n}{2}}\sigma^{-n}e^{-\frac{1}{2\lambda_i^{(t)}\sigma^2}\sum_j(M_{ij})^2}. \tag{11.23}$$

11.3.1. Extension to asymmetric gene expression

In the last section, the expressed genes were modelled using a symmetric α-stable distribution, therefore, it was possible to write an analytical expression for the α-stable distribution using the Scale Mixture of Normals property. This led us to a symmetric modelling of the gene expression distribution. However, in some cases, this distribution can be asymmetric due to a global under or over-expression. In this section, three different skewed alternatives using the α-stable are suggested to deal with asymmetric gene expression distributions.

First, the α-stable density could also be used to model the degree of asymmetry, including the skewness parameter β. In that case, the distribution of μ_i for the expressed genes is expressed as:

$$p(\mu_i|\alpha,\beta,\sigma,z_i=1)=f_{\alpha,\beta}(\mu_i|\sigma,0), \tag{11.24}$$

and β could be estimated from the distribution of microarray gene data using existing estimation techniques in the literature [Kuruoglu, 2001; Kogon & Williams, 1998]. In Figure 11.1a), 11.1b) and 11.1c), three different α-stable densities with varying β are depicted. The main difficulty using the asymmetric version of the α-stable distribution is the lack of an analytical expression for this density. Therefore $f_{\alpha,\beta}(\mu_i|\sigma,0)$ should be evaluated using numerical techniques [Nolan, 1997]. Furthermore, in that case the equality given in equation (11.19) does not hold and therefore, $d\mu_i$ can not be integrated out analytically and the integral (11.17) should be evaluated numerically.

The second asymmetric alternative is to model the expressed genes using a mixture of two (or more) symmetric α-stable distributions (see [Salas-Gonzalez *et al.*, 2006c] for more details about parameter estimation in this model):

$$p(\mu_i|\alpha_1,\alpha_2,\sigma_1,\sigma_2,\mu_1,\mu_2)=w_1f_{\alpha_1,0}(\mu_i|\sigma_1,\mu_1)+w_2f_{\alpha_2,0}(\mu_i|\sigma_2,\mu_2). \tag{11.25}$$

This approach allows us to use the Scale Mixture of Normals property to evaluate the α-stable density introducing an auxiliary variable λ with distribution given in equation (11.15):

$$p(\mu_i\mid ...)=w_1N(\mu_i\mid\mu_1,\lambda_{1i}\sigma_1^2)+w_2N(\mu_i\mid\mu_2,\lambda_{2i}\sigma_2^2). \tag{11.26}$$

The proposed mixture model can be simplified by considering the same values for some parameters in both components, $\sigma_1 = \sigma_2$ and $\alpha_1 = \alpha_2$. Furthermore, the latter equality leads to

$$p(\lambda_1) = p(\lambda_2). \tag{11.27}$$

Figure 11.1d) and 11.1e) show two asymmetric distributions modelled as a mixture of two symmetric α-stables densities with different parameter values, the dashed lines are the symmetric components and the solid line denotes the mixture.

The third approach to model asymmetry in the gene expression distribution using α-stable densities is the simplest and most straightforward to apply. Furthermore it shares some desirably properties with the symmetric α-stable model proposed in this chapter. Introducing a location parameter μ_1 different to zero for the distribution of the expressed genes,

$$p(\mu_i \mid \lambda_i, \sigma) = w f_{\alpha,0}(\sigma, \mu_1), \tag{11.28}$$

will led to a global asymmetric gene expression distribution:

$$p(\mu_i \mid \lambda_i, \sigma, \mu_1) = w f_{\alpha,0}(\sigma, \mu_1) + (1 - w)\delta(0). \tag{11.29}$$

Where μ_1 can be estimated using the Bayesian α-stable mixture model approach presented in [Salas-Gonzalez *et al.*, 2006c]. In Figure 11.1f), a symmetric α-stable distribution centered in $\mu_1 = 0{,}15$ is plotted. Obviously, in this case it is also possible to use the Scale Mixture of Normals property introducing an auxiliary variable λ_i with distribution given in equation (11.15), therefore, μ_i is modelled as:

$$\mu_i \mid \lambda_i, \sigma = \begin{cases} 0 & \text{if } z_i = 0, \\ N(\mu_1, \lambda_i \sigma^2) & \text{if } z_i = 1, \end{cases}$$

and substituting the last expression in equation (11.17) and following an analogous procedure to that used in the previous section for the symmetric case, it is possible to integrate out $d\mu_i$ and to evaluate numerically $p(M_i|z_i = 1)$ using the following expression:

$$\begin{aligned} p(M_i|z_i = 1) &= \frac{1}{T}\sum_{t=1}^{T}(2\pi\lambda_i^{(t)})^{-\frac{n}{2}}\sigma^{-n}e^{-\frac{1}{2\lambda_i^{(t)}\sigma^2}\sum_j (M_{ij}-M_{i.})^2} \\ &\times \quad (n+1)^{-1/2}e^{-\frac{1}{2\lambda_i^{(t)}\sigma^2}\frac{n}{n+1}(M_{i.}-\mu_1)^2} \end{aligned} \tag{11.30}$$

See that equations (11.22) and (11.30) are equivalent when μ_1 is zero.

11.4. Material and methods

11.4.1. Simulated data

In order to illustrate the performance of the statistic S, we simulated a dataset containing $N = 10,000$ genes and $n = 4$ replicates . The non-expressed genes were simulated following an α-stable distribution with parameters: $\alpha = 1{,}8$, $\beta = 0$, $\sigma = 0{,}1$ and $\mu = 0$. Thus, the samples are simulated from the following mathematical model (equations (11.11) and (11.15)):

$$M_{ij}|\mu_i, \lambda_i, \sigma \sim N(0, \lambda_i 0{,}1^2) \text{ for } i = 1 : N. \tag{11.31}$$

$$p(\lambda_i) = f_{\frac{1{,}8}{2},1}(2\{\cos\left(\frac{1{,}8\pi}{4}\right)\}^{\frac{2}{1{,}8}}, 0). \tag{11.32}$$

They were typical values obtained in the analysis of the four gene expression datasets studied in [Salas-Gonzalez *et al.*, 2006b]. These four datasets were: first, self-self hybridization of 19 different human cancer cell lines [Yang *et al.*, 2002a]. The second dataset consisted of gene expression data from Swirl zebrafish cDNA microarray experiment.The third dataset consisted of tumor samples from diffuse large B-cell lymphoma patients [Alizadeh *et al.*, 2000] and the last dataset was an analysis of regulatory variation in a cross between laboratory and wild strains of Saccharomyces Cerevisae [Yvert *et al.*, 2003].

A proportion $p = 0{,}01$ of the $N = 10,000$ genes were considered to be differently expressed. For that set of genes, the values of the α-stable parameters were chosen the same as the non-expressed but the location parameter μ was simulated as an α-stable (condition $z_i = 1$ in Equation (11.15)) with dispersion parameter set to $V\sigma$ with $V = 1{,}5$, where V represents a type of generalized signal to noise ratio. The parameter V is also introduced in [Lonnstedt & Speed, 2002; Bhowmick *et al.*, 2006]. The estimation of this parameter is very difficult due to the fact that only a very small proportion of genes are expressed and we do not know which ones. This difficulty was also pointed out in [Lonnstedt & Speed, 2002] for a Gaussian mixture model and for a Laplace mixture model in [Bhowmick *et al.*, 2006] where the parameter V is not estimated correctly. It will be shown in the simulations that the performance of our algorithm is not affected by the ignorance of this parameter. This is due to two different facts: on the one hand, the α-stable distribution is a heavy-tailed distribution, therefore it is a proper distribution to accommodate outliers in the data and, on the other hand, the number of genes differently expressed is usually a very small proportion of the whole data.

One of the simulated datasets is plotted in Figure 11.2, where the average M-values versus the logarithm of the variance is plotted. The expressed genes are denoted with crosses. These are the genes with expectation different to zero. It is easily seen that many of the true influenced genes have a negligible value of the average, therefore it is not possible to detect them using any method.

11.4.2. Experimental data

We test the proposed statistic using real data. Namely, a comparison between RNA transcript levels of Arabidopsis plants, infected by the rhizobacterium Pseudomonas thivervalensis (strain MLG45), and axenic control plants [Cartieaux *et al.*, 2003]. This dataset contains 16,416 spots and $n = 4$ replicates and it is available from the Standford Microarray Database (SMD) with SMD experiment ID numbers 27084, 27000, 26995 and 26718. The data was normalized using LOESS. After normalization, the gene expression distribution was modelled as a symmetric α-stable and we obtained the following values of the parameters: $\alpha = 1{,}83$ and $\sigma = 0{,}15$. In Figure 11.3, the gene expression distribution and the α-stable density are plotted. It is easily seen that the Stable density fits very accurately the gene expression data. Furthermore, the gene expression data shares some of its features with the α-stable distribution (see the section 11.2.5).

11.5. Results

11.5.1. Simulated data

For each different values of the cutoff w considered (20 different values from $w = 0{,}005$ to $w = 0{,}1$), 100 different datasets were simulated, with α-stable parameters given in Section 11.4.1. The Stable statistic and B were calculated for each dataset.

In the Figure 11.4, the histograms of the estimated values for each α-stable parameter are depicted. Thus, the variance of the estimated values between different simulated datasets can be compared. It is easily seen that the true values are estimated very accurately. Moreover, Table 11.1 gives the minimum mean square error estimator for each parameter. It is shown that our method estimates the parameters of the model very accurately.

The average number of false positives and false negative genes for the 20 different cutoffs w are plotted in Figure 11.5 for each synthetic dataset. This

Table 11.1: Estimated values of the parameters of the symmetric α-stable mixture model. The proportion of differentially expressed genes was set to $p = 0{,}01$ and the number of replicates $n = 4$.

Parameter	**true value**	**estimated value**	**estimated variance**
α	1.80	1.83	0.03
σ	0.100	0.102	0.001

plot allows us to compare type I and type II errors without the using p-values. For each of the 20 values of the cutoff, 100 different datasets were simulated. The numbers of false-negatives and false-positives genes are averaged over the 100 datasets. In this figure, the statistic based on the α-stable distribution is compared with B, the statistic based on the scale t-statistics proposed in [Lonnstedt & Speed, 2002]. The Stable statistic exhibits lower values of false positives and false negatives than B for each value of the cutoff.

For each different values of the cutoff w considered (40 different values from $w = 0$ to $w = 1$), 100 different datasets were simulated. The Stable statistic S and B were calculated for each dataset.

The Receiver Operating Characteristic curve for the 40 different cutoffs w is plotted in Figure 11.6 for each synthetic dataset. The fraction of true positives and false-positives genes is averaged over the 100 datasets. In this figure, the statistic based on the α-stable distribution is compared with B, the statistic based on the scale t-statistics proposed in [Lonnstedt & Speed, 2002]. The Stable statistic exhibits higher values of true positives and true negatives than B for each value of the cutoff.

11.5.2. Experimental data

In Figure 11.7, $M_{i.}$ (the average M_i for $j = 1 \dots n$) versus the log-variance is plotted. Moreover, in that figure, the location of the differently expressed genes for the S and B statistic are depicted with crosses and circles respectively. It can be seen that, roughly, a proportion of the genes which are considered differently expressed by the B statistic, has greater value of $M_{i.}$ and the variance than for the S statistic case. This is due to the fact that the S statistic penalizes more the genes with high variance. Furthermore, our model assumes a heavy-tail model for the distribution of the non-expressed genes (eq. (11.23)), therefore this model allows non-expressed genes to have a high value of M_i if the variance

is high. This feature makes the statistic S suitable for gene expression data with high errors between the different replicates.

Figure 11.8 shows the volcano plot in which the fold-change and the statistical significance are depicted. The Stable log posterior odds S_i is plotted on the ordinate and the corresponding $M_{i\cdot}$ values on the abscissa. The upper corners of the figure represent genes with statistical significance and large fold changes. Crosses denote the expressed genes according to the Stable statistics and circles the expressed genes according to the B statistic. This figure displays the same behaviour that was commented for the previous figure, that is, there are a considerably amount of genes with high $|M_{i\cdot}|$ that the Stable statistic S considers as not expressed. The Figure 11.9 shows the statistics S_i versus the variance of the different replicates of M_{ij}. It is seen that some spots with high variance are declared as differently expressed for the B statistic and not for the Stable one.

11.6. Conclusion

In this chapter, a new statistic to identify expressed genes in replicated microarray data is proposed. This statistic is based on the properties of the α-stable distribution. An α-stable mixture model is introduced and the Scale Mixture of Normals property is used to calculate the Bayes log posterior odds. This procedure allows us to calculate the proposed statistic very easily using some known properties of the α-stable density. The proposed statistic was tested in both, synthetic and real data and its performance was compared to the B statistic based on the t-student. On the contrary to that approach, the empirical Bayes method proposed in this chapter is based on the modelling of the distribution of gene expression as a heavy-tailed distribution. This choice is suggested by a preliminary study of the gene error distribution and make this approach suitable to model replicated microarray data when the variance between replicates is very high.

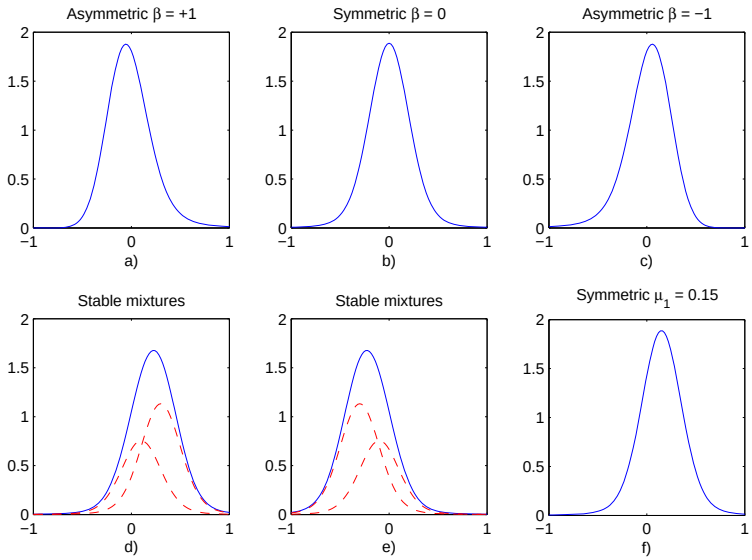

Figure 11.1: a) α-stable distribution with $\alpha = 1{,}8$, $\mu = 0$, $\sigma = 0{,}15$ and skewness parameter $\beta = +1$ b) α-stable distribution with $\alpha = 1{,}8$, $\mu = 0$, $\sigma = 0{,}15$ and skewness parameter $\beta = 0$ c) α-stable distribution with $\alpha = 1{,}8$, $\mu = 0$, $\sigma = 0{,}15$ and skewness parameter $\beta = -1$. d) Mixture of two symmetric α-stable distributions with parameters $w_1 = 0{,}6$,$\alpha_1 = 1{,}8$,$\sigma_1 = 0{,}15$ and $\mu_1 = 0{,}3$. And for the second component: $w_2 = 0{,}4$,$\alpha_2 = 1{,}8$,$\sigma_2 = 0{,}15$ and $\mu_2 = 0{,}1$. e) Mixture of two symmetric α-stable distributions with parameters $w_1 = 0{,}6$,$\alpha_1 = 1{,}8$,$\sigma_1 = 0{,}15$ and $\mu_1 = -0{,}3$. And for the second component: $w_2 = 0{,}4$,$\alpha_2 = 1{,}8$,$\sigma_2 = 0{,}15$ and $\mu_2 = -0{,}1$. f) Symmetric α-stable distribution with parameters $\alpha = 1{,}8$, $\sigma = 0{,}15$ and $\mu_1 = 0{,}15$.

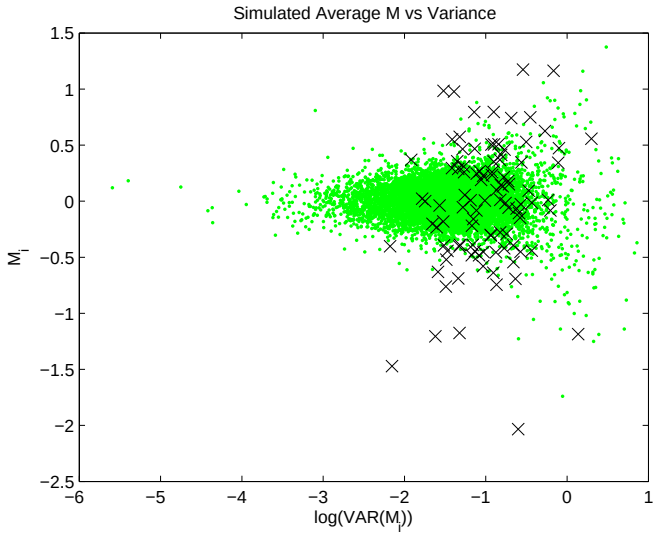

Figure 11.2: $M_{i\cdot}$ vs log-variance for one of the simulated datasets. *Crosses:* True expressed genes.

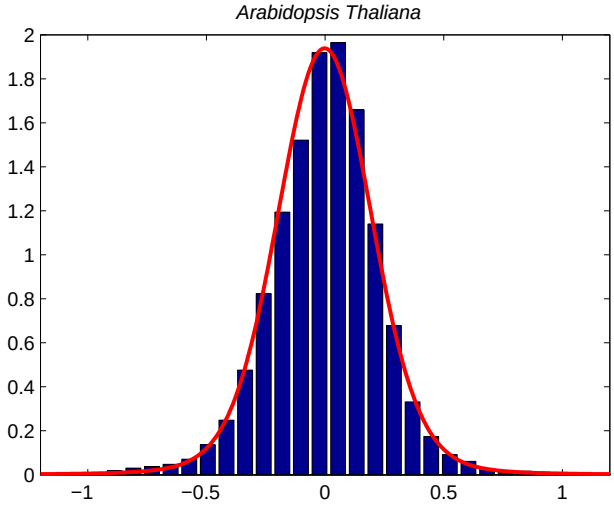

Figure 11.3: Discrete gene expression histogram and predicted symmetric α-stable density for the *Arabidopsis Thaliana* dataset.

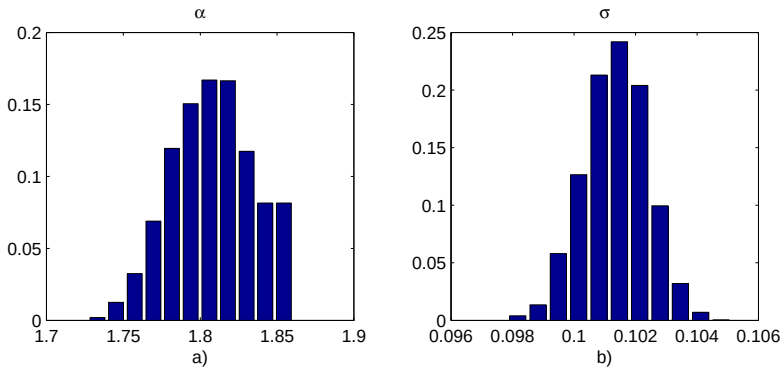

Figure 11.4: Histogram of the parameter estimates. The true values are $\alpha = 1{,}8$ and $\sigma = 0{,}1$.

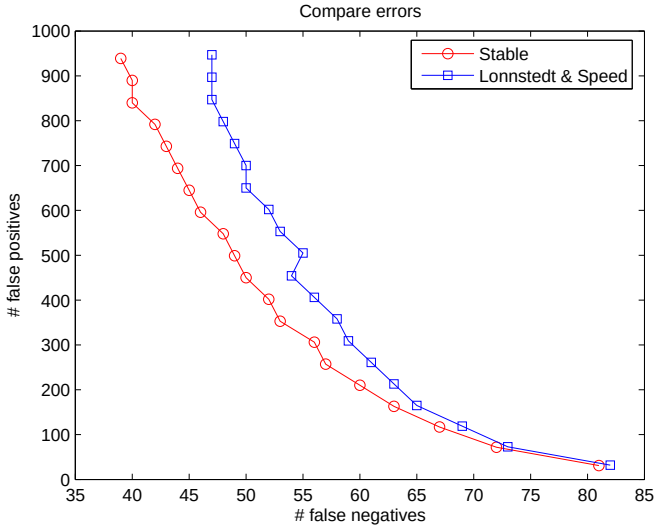

Figure 11.5: Type I vs. type II error for S_i and B_i statistics computed on the simulated α-stable data with $\alpha = 1{,}8$, $\beta = 0$, $\sigma = 0{,}1$ and $\mu = 0$.

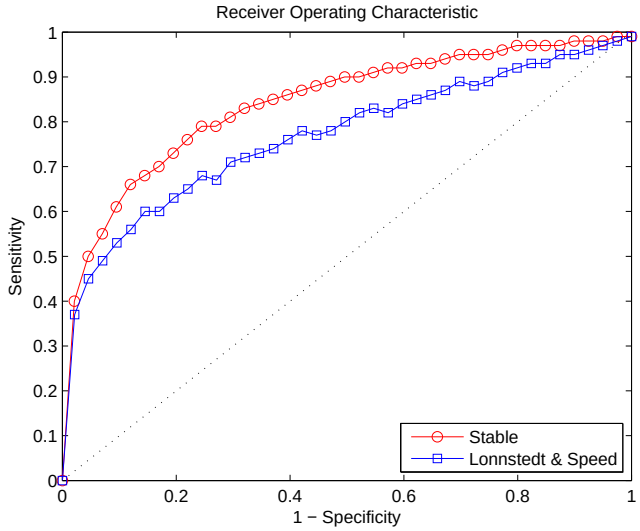

Figure 11.6: Receiver Operatinng Characteristic curve for S_i and B_i statistics computed on the simulated α-stable data with $\alpha = 1{,}8$, $\beta = 0$, $\sigma = 0{,}1$ and $\mu = 0$.

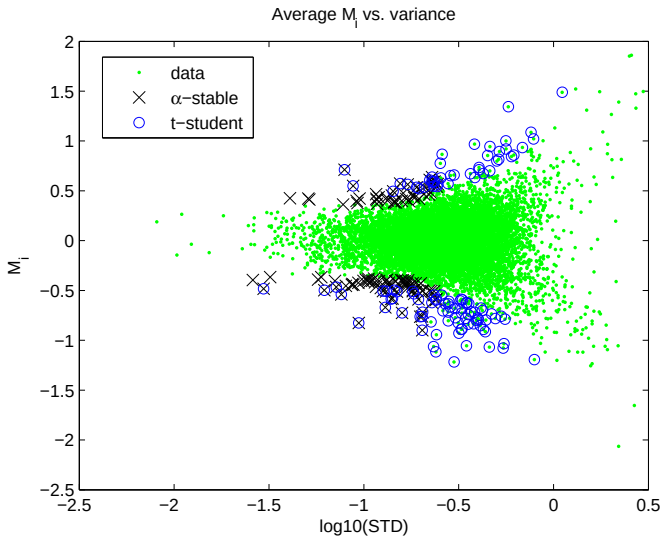

Figure 11.7: Average M_i versus the log-variance. *Crosses:* Set of genes which are differently expressed for the Stable statistic. *Circles:* Genes differently expressed for the statistic B based on t-student [Lonnstedt & Speed, 2002].

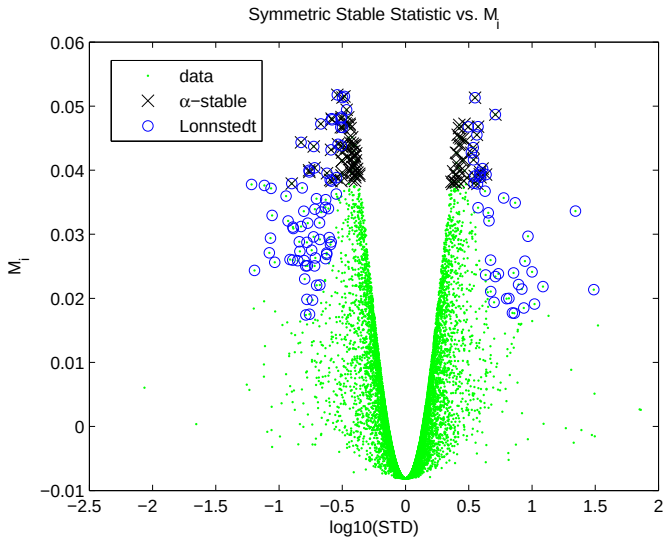

Figure 11.8: Log posterior odds S_i vs. the average expression level M_i.. *Crosses:* Set of genes which are differently expressed for the Stable statistic. *Circles:* Genes differently expressed for the statistic B based on t-student.

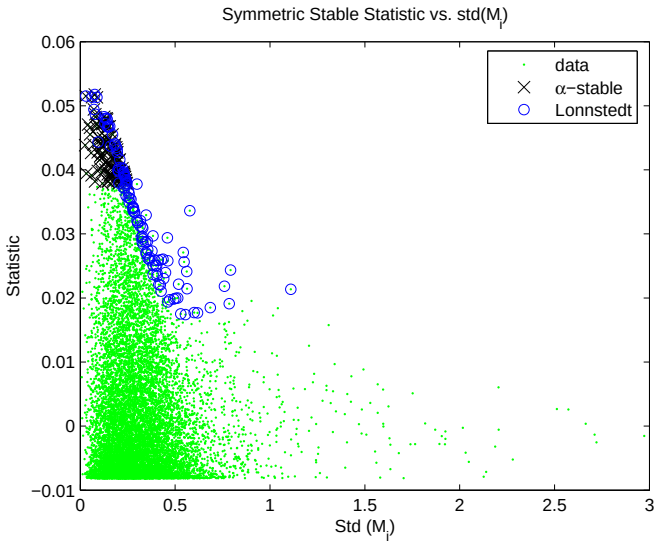

Figure 11.9: Log posterior odds S_i vs. the standard deviation of the expression level $std(M_i)$. *Crosses:* Set of genes which are differently expressed for the Stable statistic. *Circles:* Genes differently expressed for the statistic B based on the t-student distribution.

BIBLIOGRAFÍA

Adler, Robert J., Feldman, Raisa E., & Taqqu, Murad S. (Eds.). 1998. *A practical guide to heavy tails: statistical techniques and applications.* Boston: Birkhausser.

Alizadeh, Ash A., Eisen, Michael B., Davis, R. Eric, Ma, Chi, Lossos, Izidore S., Rosenwald, Andreas, Boldrick, Jennifer C., Sabet, Hajeer, Tran, Truc, Yu, Xin, Powell, John I., Yang, Liming, Marti, Gerald E., Moore, Troy, James Hudson, Jr, Lu, Lisheng, Lewis, David B., Tibshirani, Robert, Sherlock, Gavin, Chan, Wing C., Greiner, Timothy C., Weisenburger, Dennis D., Armitage, James O., Warnke, Roger, Levy, Ronald, Wilson, Wyndham, Grever, Michael R., Byrd, John C., Botstein, David, Brown, Patrick O., , & Staudt, Louis M. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Allison, David B., Cui, Xiangqin, Page, Grier P., & Sabripour, Mahyar. 2006. Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, **7**(January), 55–65.

Bechtel, Y. C., Bonaiti-Pellie, C., Poisson, N., Magnette, J., & Bechtel, P. R. 1993. A population and family study of N-acetyltransferase using caffeine urinary metabolities. *Clinical pharmacology and therapeutics*, **54**, 134–141.

Bhowmick, Debjani, Davison, A. C., Goldstein, Darlene R., & Ruffieux, Yann. 2006. A Laplace mixture model for identification of differential expression in microarray experiments. *Biostatistics*, **7**(4), 630–641.

Bodenschatz, John S., & Nikias, Chrysostomos L. 1999. Maximum-Likelihood Symmetric alpha-Stable Parameter Estimation. *IEEE Transactions on Signal Processing*, **47**(5), 1382–1384.

Borovkov, A. A. 1998. *Mathematical statistics.* Amsterdam: Gordon and Breach science.

Buckle, D. J. 1995. Bayesian inference for stable distribution. *Journal of American Statistical Association*, **90**, 605–613.

Cartieaux, Fabienne, Thibaud, Marie-Christine, Zimmerli, Laurent, Lessard, Philippe, Sarrobert, Catherine, David, Pascale, Gerbaud, Alain, Robaglia, Christophe, Somerville, Shauna, & Nussaume, Laurent. 2003. Transcriptome analysis of Arabidopsis colonized by a plant-growth promoting rhizobacterium reveals a general effect on disease resistance. *The plant journal*, **36**(2), 177–188.

Casarin, R. 2004. *Bayesian inference for mixture of stable distributions.* Tech. rept. Working paper n. 0428, CEREMADE. University Paris IX.

Celeux, G., Hurn, M., & Robert, C.P. 2000. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, **95**, 957–970.

Chambers, J., Mallows, C., & Stuck, B. 1976. A method for simulating stable random variables. *Journal of the American Statistical Association*, **71**, 340–344.

Cleveland, W. S., & Delvin, S. J. 1988. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of American Statistical Association*, **83**(403), 596–610.

Crawford, S. L. 1994. An application of the Laplace method to finite mixture distributions. *Journal of the American Statistical Association.*, **89**, 259–267.

Crawford, S. L., DeGroot, M. H., Kadane, J. B., & Small, M. J. 1992. Modeling lake chemistry distributions: approximate Bayesian methods for estimating a finite mixture model. *Technometrics.*, **34**, 441–453.

Dempster, A. P., Laird, N. M., & Rubin, D. B. 1997. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1), 1–38.

Dempster, Arthur, Laird, Nan, & Rubin, Donald. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**(1), 1–38.

Diebolt, J., & Robert, C. 1995. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B*, **56**, 577–588.

Do, Kim-Anh, Muller, Peter, & Tang, Feng. 2005. A bayesian mixture model for differential gene expression. *Biostatistics*, **54**(3), 627–644.

Escobar, M. D., & West, M. 1995. Bayesian density estimation and inference using mixture. *Journal of the American Statistical Association.*, **90**, 577–588.

Fan, Zhaozhi. 2006. Parameter estimation of Stable distributions. *Communications in statistics - Theory and methods*, **35**, 245–255.

Feller, W. 1966. *An introduction to probability theory and its applications.* Vol. II. New York: Wiley & Sons.

Fernandez, C., & Green, P. J. 2002. Modelling spatially correlated data via mixtures: a bayesian approach. *Journal of the royal statistical society: series B (Statistical methodology)*, **64**(4), 805–826.

Fernandez, Carmen, & Steel, Mark F. J. 2000. Bayesian regression analysis with scale mixture of Normals. *Econometric Theory*, **16**, 80–101.

Gelman, Andrew, Carlin, John B., Stern, Hal S., & Rubin, Donald B. 1995. *Bayesian Data Analysis.* London: Chapman & Hall/CRC.

Geman, S., & Geman, D. 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (eds.). 1996. *Markov chain Monte Carlo in Practice.* London: Chapman & Hall.

Godsill, S., & Kuruoglu, E. E. 1999. Bayesian inference for time series with heavy-tailed symmetric alpha stable noise processes. *In: Proc. Applications of heavy tailed distributions in economics, engineering and statistics, June 1999. Washington DC, USA.*

Gottardo, Raphael, Panucci, James A., Kuske, Cheryl R., & Brettin, Thomas. 2003. Statistical analysis of microarray data: a Bayesian approach. *Biostatistics*, **4**(4), 597–620.

Green, P. J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**(4), 711–732.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their application. *Biometrika*, **57**, 97–109.

Herranz, D., Kuruoglu, E. E., & Toffolatti, L. 2004. Using alpha-stable distributions to model point source distributions in CMB sky maps. *Astronomy and astrophysics*, **424**(3), 1081–1096.

Hoyle, D. C., Rattray, M., Jupp, R., & Brass, A. 2002. Making sense of microarray data distributions. *Bioinformatics*, **18**(4), 576–584.

Ilow, J., & Hatzinakos, D. 1998. Applications of the empirical characteristic function to estimation and detection problems. *Signal processing*, **65**(2), 199–219.

Khondoker, Mizanur R., Glasbey, Chris A., & Worton, Bruce J. 2006. Statistical estimation of gene expression using multiple laser scans of microarrays. *Bioinformatics*, **22**(2), 215–219.

Kogon, S. M., & Williams, D. B. 1998. *Characteristic function based estimation of stable parameters.* Boston, MA: R. Feldman, and M. Taqqu (Eds.), A Practical Guide to Heavy Tailed Data.

Kuruoglu, E. E. 2001. Density Parameter Estimation of Skewed alpha-Stable Distributions. *IEEE Transactions on Signal Processing*, **49**(10), 2192–2201.

Kuruoglu, E. E., & Zerubia, J. 2003. Skewed alpha-stable distributions for modelling textures. *Pattern recognition letters*, **24**(1-3), 339–348.

Kuruoglu, E. E., Molina, C., Godsill, S. J., & Fitzgerald, W. J. 1997 (August). A new analytic representation for the alpha-Stable probability density function. *In: The Fifth World Meeting of the International Society for Bayesian Analysis (ISBA), Istanbul. Turkey.*

Kuruoglu, E. E., Molina, C., & Fitzgerald, W. J. 1998 (September). Approximation of alpha stable probability densities using finite mixtures of Gaussian. *In: Proceedings of the European Signal Processing Conference. Rhodes, Greece.*

Kuruoglu, E. E., Salas-Gonzalez, D., & Ruiz, D. P. 2006 (June). Bayesian Inference on Mixtures of Stable Densities. *In: ISBA Eighth World Meeting on Bayesian Statistics. Valencia, Spain.*

Kuruoglu, E. E., Salas-Gonzalez, D., & Ruiz, D. P. 2007 (June). Microarray gene expression and Stable laws. *In: IEEE 15th Signal Processing and Communications Applications Conference, SIU2007. Eskisehir, Turkey.*

Kuznetsov, Vladimir A. 2001. Distribution Associated with Stochastic Processes of Gene Expression in a Single Eukaryotic Cell. *EURASIP Journal on applied signal processing*, **4**, 285–296.

Lombardi, M. J. 2007. Bayesian inference for α-stable distributions: A random walk MCMC approach. *Computational Statistics and Data Analysis*, **51**(5), 2688–2700.

Lonnstedt, Ingrid, & Speed, Terry. 2002. Replicated microarray data. *Statistica Sinica*, **12**, 31–46.

M. Wiper, D. R. Insua, & Ruggieri, F. 2001. Mixture of Gamma distributions with applications. *Journal of computational and graphical statistics*, **10**(3), 440–454.

McLachlan, G., & Peel, D. 2000. *Finite mixture models.* Wiley series in probability and statistics.

McLachlan, Geoffrey J., & Peel, David. 1998. Robust Cluster Analysis via Mixtures of Multivariate t-Distributions. *Pages 658–666 of: SSPR '98/SPR '98: Proceedings of the Joint IAPR International Workshops on Advances in Pattern Recognition.* London, UK: Springer-Verlag.

Menn, C, & Rachev, S. T. 2006. Calibrated FFT-based density approximations for stable distributions. *Computational Statistics & Data Analysis*, **50**(8), 1891–1904.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. H., & Teller, E. 1953. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092.

Mittnik, S., Rachev, T., Doganoglu, D., & Chenyao, D. 1999. Maximum likelihood estimation of stable Paretian models. *Mathematical and Computer modelling*, **829**(10-12), 275–293.

Monno, L., Petrella, L., & Tancredi, A. 2004. Bayesian modelling volatility with mixture of alpha-stable distributions. *In: Proceedings of the 19th International workshop on statistical modelling, Florence (Italy), 4-8 July, 2004.*

Newman, M. E. J. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, **46**, 323.

Newton, Michael A., Noueiry, Amine, Sarkar, Deepayan, & Ahlquist, Paul. 2004. Detecting differentially gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, **5**(5), 155–176.

Nikias, C. L., & Shao, M. 1995. *Signal processing with alpha-stable distributions and applications.* New York, USA: Wiley-Interscience.

Nolan, J. P. 1997. Numerical calculation of stable densities and distribution functions. *Commun. Statist.-Stochastic Models*, 759–774.

Nolan, J. P. 2001. Maximum likelihood estimation of stable parameters. *in Levy Processes (O. Barndorff-Nielsen, T. Mikosch, and S. Resnick, eds.)*, 379–400.

Pearson, K. 1894. Contribution to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London A*, **185**, 71–110.

Phillips, D. B., & Smith, A. F. M. 1996. *Bayesian model comparison via jump diffusions, chapter 13 (pp. 215-239).* Vol. 90. London: W. R. Gilks, S. Richardson and D. J. Spiegelhalter and D. J. Spiegelgalter, eds. Chapman and Hall.

Purdom, Elizabeth, & Holmes, Susan. 2005. Error distribution for gene expression data. *Statistical applications in genetics and molecular biology*, **4**(1).

Richardson, S., & Green, P. J. 1997. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, B*, 731–792.

Robert, C., & Casella, G. 1999. *Monte Carlo Statistical Methods.* New York: Springer-Verlag.

Salas-Gonzalez, D., Kuruoglu, E. E., & Ruiz, D. P. 2006a (September). Bayesian estimation of mixtures of skewed alpha stable distributions with an unknown number of components. *In: European Signal Processing Conference. Firenze, Italy.*

Salas-Gonzalez, D., Kuruoglu, E. E., & Ruiz, D. P. 2006b. *Bayesian inference on mixture of alpha-stable distributions.* Tech. rept. ISTI-2006-PP-02. ISTI-CNR Pisa.

Salas-Gonzalez, D., Kuruoglu, E. E., & Ruiz, D. P. 2006c (May). Estimation of mixtures of symmetric alpha stable distributions with an unknown number of components. *In: IEEE International Conference on Acoustics, Speech and Signal Processing. Toulouse, France.*

Salas-Gonzalez, D., Kuruoglu, E. E., & Ruiz, D. P. 2006d. *Modelling microarray gene expression using alpha-stable distributions.* Tech. rept. ISTI-2006-PP-01. ISTI-CNR Pisa.

Salas-Gonzalez, D., Kuruoglu, E. E., & Ruiz, D. P. 2007a. A heavy-tailed empirical Bayes method for replicated microarray data. *Computational Statistics and Data Analysis (Submitted to).*

Salas-Gonzalez, D., Kuruoglu, E. E., & Ruiz, D. P. 2007b (January). Modelling the distribution of gene expressions using stable distributions. *In: 5th Asia-Pacific Bioinformatics Conference, APBC2007. Hong Kong.*

Salas-Gonzalez, D., Kuruoglu, E. E., & Ruiz, D. P. 2007c. Modelling with mixture of symmetric stable distributions using Gibbs sampling. *Computational Statistics and Data Analysis (Submitted to).*

Salas-Gonzalez, D., Kuruoglu, E. E., & Ruiz, D. P. 2008. Finite mixtures of Stable distributions. *Digital Signal Processing.*

Samorodnitsky, G., & Taqqu, M.S. 1994. *Stable Non-Gaussian Random Process: Stochastic Models with Infinite Variance.* New York: Chapman-Hall.

Stephens, M. 2000. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society. Series B, statistical methodology.*, **62**, 795–809.

Swami, A. 1999. Non-Gaussian mixture models for detection and estimation in heavy-tailed noise. *In: IEEE Internatoinal Conference on Acoustics, Speech, and Signal Processing. June 2000. Istanbul, Turkey.*

Tsionas, E. 2002. Bayesian analysis of finite mixtures of Weibull distributions. *Communications in Statistics, Part A - Thery and methods*, **31**(1), 37–48.

Tsionas, E. G. 1999. Monte Carlo inference in econometric models with symmetric stable disturbances. *Journal of econometrics*, 365–401.

Yang, Ivana, Chen, Emily, Hasseman, Jeremy, Liang, Wei, Frank, Bryan, Wang, Shuibang, Sharov, Vasily, Saeed, Alexander, White, Joseph, Li, Jerry, Lee, Norman, Yeatman, Timothy, & Quackenbush, John. 2002a. Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biology*, **3**(11), research0062.1–research0062.12.

Yang, Yee Hwa, Dudoit, Sandrine, Luu, Percy, Lin, David M., Peng, Vivian, Ngai, John, & Speed, Terence P. 2002b. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucl. Acids Res.*, **30**(4), e15–.

Yvert, Gael, Brem, Rachel B., Whittle, Jacqueline, Akey, Joshua M., Foss, Eric, Smith, Erin N., Mackelprang, Rachel, & Kruglyak, Leonid. 2003. Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nature Genetics*, **35**(1), 57–64.

Zolotarev, V. M. 1986. *One dimensional Stable Distributions*. Providence: Translation on Mathematical Monographs 65. American Math. Soc.