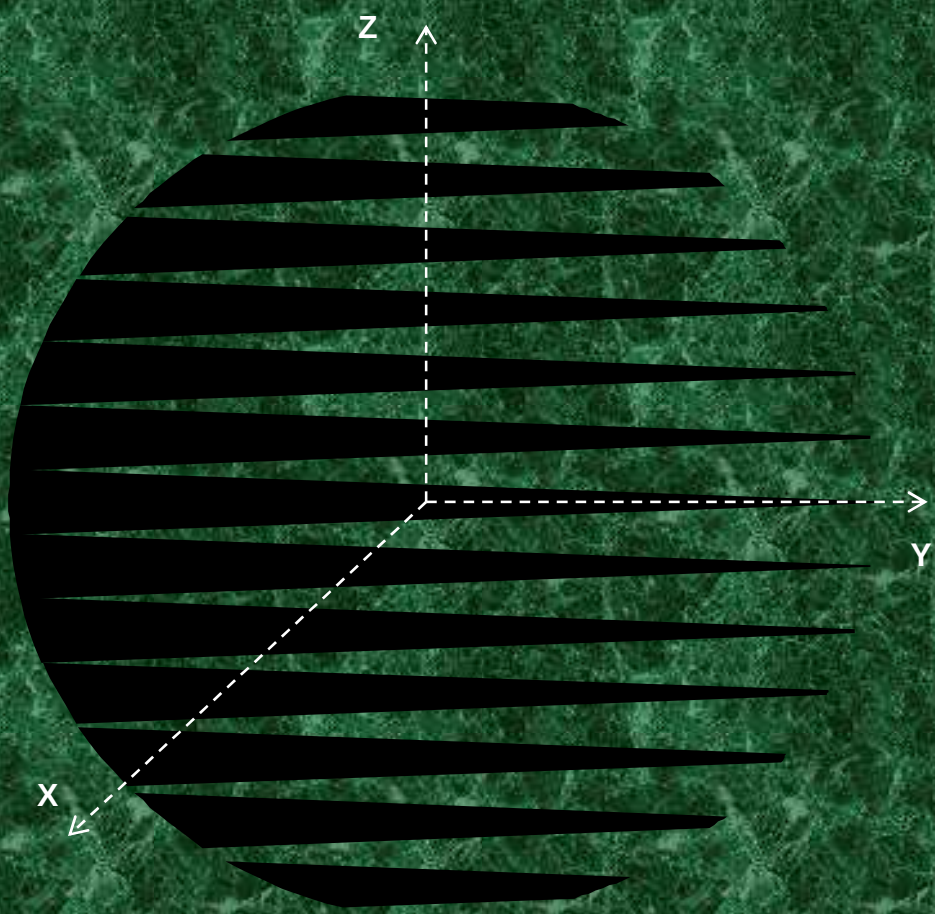
ISSN Nº 1390 - 3802

matemática

UNA PUBLICACIÓN DEL ICM - ESPOL

Volumen 10 Número 1 Octubre 2012



Escuela Superior Politécnica del Litoral - ESPOL
Instituto de Ciencias Matemáticas - ICM

INSTITUTO DE CIENCIAS MATEMÁTICAS

El Instituto de Ciencias Matemáticas (ICM) es una unidad académica de la ESPOL. Desde el inicio la función del ICM ha sido la docencia en Matemáticas, Ciencias Gráficas e Informática, para la formación de profesionales en ingeniería, tecnología y ciencias, habiendo tenido a su cargo en los albores de la ESPOL, el dictado de 10 materias. Con el transcurso del tiempo y acorde con la era de la información, el ICM creó en mayo de 1995 la carrera de "Ingeniería en Estadística Informática", como alternativa en ingeniería en información y servicios. Posteriormente, con el fin de garantizar la eficiencia en el control y gestión empresarial con profesionales capacitados y de excelencia se creó la carrera de "Auditoría y Control de Gestión" en mayo de 2000. También el Instituto ha incursionado en una de las más importantes ramas de la matemática aplicada que tiene grandes aplicaciones en el mundo moderno, esto es la Investigación de Operaciones, la Teoría de Optimización, y particularmente las aplicaciones logísticas, a través del ofrecimiento de programas de pre-grado y post-grado en estas áreas. Así es como desde el año 2005 se viene ofreciendo la maestría en Control de Operaciones y Gestión Logística y desde el año 2006 la carrera de Ingeniería en Logística y Transporte.

El ICM también cuenta con el CENTRO DE INVESTIGACIONES ESTADÍSTICAS, a través del cual, se realizan: estudios de predicción, estudios actuariales, estudios de mercado, diseños de experimentos, planificación y dirección de censos, análisis financieros, bases de datos estadísticos, formulación de proyectos, ingeniería de la calidad, etc.

Entre otras actividades que desarrolla el ICM anualmente están: las JORNADAS EN ESTADÍSTICA E INFORMÁTICA que actualmente está en su decimoséptima versión, el CONCURSO INTERCOLEGIAL DE MATEMÁTICAS que se viene realizando en forma continúa desde 1988.

.

Más información: www.icm.espol.edu.ec o escribirnos al e-mail: icm@espol.edu.ec, palvarez@espol.edu.ec, erivaden@espol.edu.ec, 30 ½ vía Perimetral: Edificios 25 – B Planta alta (Área de Institutos) Telfs.: (593-4) 2269525 – 2269526, fax: (593–4) 853138.
Guayaquil – Ecuador

matemática

UNA PUBLICACIÓN DEL ICM – ESPOL

Volumen 10 **Número 1** **Octubre 2012**

matemática es una publicación del Instituto de Ciencias Matemáticas de la Escuela Superior Politécnica del Litoral, y pretende constituirse en un órgano de difusión científico – tecnológico, con el fin de incentivar y motivar el desarrollo y avance de la matemática y sus aplicaciones.

matemática publica artículos teóricos y de tipo experimental tales como ensayos, resúmenes de tesis de grado y trabajos de investigación relacionados con la aplicación de la matemática en los diferentes ámbitos de la realidad.

CONTENIDO

EDITORIAL

Uno de los desafíos fundamentales en Ingeniería es la elaboración de modelos matemáticos robustos que describan el comportamiento de los fenómenos observados en cierta aplicación, sea esta de naturaleza física, química, económica, etc. El objetivo del Modelaje Matemático es precisamente identificar las variables y parámetros relevantes de un fenómeno así como también las leyes y principios que gobiernan su comportamiento para después integrar estos elementos en un modelo; comúnmente ecuaciones diferenciales ordinarias/parciales, ecuaciones en diferencias, etc. Esta tarea constituye un arte por sí misma ya que se requiere encontrar un adecuado equilibrio para que el modelo sea de utilidad. El mismo debe tomar en cuenta los aspectos relevantes del fenómeno en cuestión, lo cual frecuentemente incrementa la complejidad del modelo, pero este debe ser al mismo tiempo lo suficientemente sencillo de tal forma que sea posible su estudio y uso para predecir el comportamiento del fenómeno de interés.

En la revista ***Matemática*** ponemos a disposición de la comunidad politécnica diversos artículos de investigación orientados al estudio y elaboración de modelos matemáticos aplicados a una gran variedad de problemas de Ingeniería. Para este efecto el consejo editorial enfoca su esfuerzo en garantizar que los temas tratados estén al nivel del estado del arte del área así como también asegurar un buen balance entre las publicaciones teóricas y aquellas orientadas a las aplicaciones.

matemática: Una publicación del ICM – ESPOL
2012, Vol. 10, No. 1

DESARROLLO DE UNA APLICACIÓN PARA CALENDARIZAR EL CAMPEONATO ECUATORIANO DE FÚTBOL PROFESIONAL POR MEDIO DE UNA APROXIMACIÓN HEURÍSTICA UTILIZANDO PROGRAMACIÓN ENTERA

Cabezas Xavier[1], Morales Jorge[2]

Resumen. *En este trabajo se presenta una aplicación de las técnicas de Investigación de Operaciones, el cual consiste en el desarrollo de una aplicación para calendarizar el Campeonato Ecuatoriano de Fútbol profesional por medio de una aproximación heurística utilizando Programación Entera, este documento presenta mucho interés personal en los Scheduling Problems, debido a que involucra un procedimiento heurístico con la programación lineal entera (ILP). Se consideran restricciones basadas a las características actuales del campeonato con unas variantes, como la asignación de dos equipos o más equipos a un canal TV en una determinada fecha. Este procedimiento está dividido en tres fases: búsqueda de conjuntos de esquemas factibles, búsqueda de calendarios factibles y emparejamiento de equipos a esquemas, el cual se lo implementa en GAMS® [3] como motor de optimización y Wolfram Mathematica ® [4] para generar conjuntos de entrada, obteniendo buenos resultados en poco tiempo, con la posibilidad de generar distintos calendarios factibles para que puedan ser alternativas para los distintos equipos y la Federación Ecuatoriana de Fútbol.*

Palabras Claves: Calendarización, Heurísticas, Wolfram Mathematica, Gams, Quiebres, Esquemas.

Abstract. *This paper presents an application of the techniques of Operations Research, which is the development of an application to schedule the Ecuadorian Professional Football Championship by a heuristic approach using Integer Programming, this document presents much interest in the scheduling Problems, because it involves a heuristic procedure with Integer Linear Programming (ILP). Constraints are considered based on the current characteristics of the championship with some variants, such as assigning two or more teams to a channel at a certain date. This procedure is divided into three phases: searching for feasible pattern sets, searching for feasible schedules and matching teams to patterns, which is implemented in GAMS® [3] as motor optimization and Wolfram Mathematica® [4] to generate sets of input, obtaining good results in short time, with the ability to generate feasible schedules, these may be alternatives for the different team and the FEF.*

Keywords: Scheduling, Heuristics, Wolfram Mathematica, Gams, Breaks, Pattern

Recibido: Mayo, 2012
Aceptado: Junio, 2012

1. INTRODUCCIÓN

1.1. PROBLEMAS COMBINATORIOS

Un problema de optimización combinatoria es especificado por un conjunto P de las instancias que están asociadas a la solución del problema, las cuales son de tipología de minimización o maximización. Por problema se entiende, una pregunta general a ser contestada, teniendo muchas variables y parámetros con valores no especificados. La instancia del problema, a su vez se entiende como un caso particular del problema, con valores específicos para todos los parámetros y variables. Es importante recalcar que las instancias no son dadas explícitamente, sino más bien, se expresan en un conjunto de valores relacionados directamente con los parámetros que han sido usados para modelar el problema. En algunos casos, además de encontrar una solución óptima, debemos hallar una solución que satisfaga un conjunto de restricciones dentro del espacio de solución, es decir, esta podría no ser válida, aunque sea la mínima o máxima globalmente.

1.2. METODOLOGÍA DE LAS SOLUCIONES PARA PROBLEMAS COMBINATORIOS

Definición de Heurística y Metaheurística.
En un problema de optimización, se califica de heurístico a un procedimiento para el que se tiene un alto grado de confianza en que encuentran soluciones de alta calidad con un costo computacional razonable, aunque no se garantice su optimalidad, e incluso, en algunos casos, no se llegue a establecer lo cerca que se está de dicha situación. En optimización matemática, se usa el término heurístico en contraposición a exacto, que se aplica los procedimientos a los que se les exige que la solución aportada sea óptima y factible.

---

[1] *Cabezas Xavier, M.Sc., Profesor de la Escuela Superior Politécnica del Litoral (ESPOL); (e_mail: xcabezas@espol.edu.ec).*
[2] *Morales Jorge, Ing. en Logística y Transporte (ESPOL); (e_mail: jorlumors@espol.edu.ec).*
[3] *Sistema General de Modelaje Algebraico GAMS, diseñado específicamente para modelar problemas de optimización.*
[4] *Wolfram Matemática, desarrollado por Wolfram Research Inc, herramienta especializada en análisis numérico.*

Además, es usual aplicar este término, cuando utilizando el conocimiento que se tiene del problema, se realizan modificaciones en el procedimiento de solución del problema, que aunque no afecta a la complejidad del mismo, mejoran el rendimiento en su comportamiento práctico. Las metaheurísticas son estrategias inteligentes para diseñar o mejorar procedimientos heurísticos muy generales con un alto rendimiento, estas se refieren al diseño de los tipos fundamentales de procedimientos heurísticos de solución de un problema de optimización.

2. ESTADO DEL ARTE

2.1. ENFOQUES DE ALGUNOS MÉTODOS DE SOLUCIONES PARA CALENDARIZACIÓN

Los problemas de Calendarización en general han sido aplicados en diferentes casos, como: educacional, deportes, tareas de actividades, entre otras. Este trabajo trata de resolver la calendarización para deportes las cuales tienen características propias incluso diferentes entre las otorgadas por la Federación Ecuatoriana de Fútbol (FEF). El esfuerzo (humano y computacional) de quien o quienes planifican los calendarios de juegos, ha hecho que el problema de calendarización para este caso particular sea considerado uno de los más importantes en la optimización combinatoria. En muchos de los casos, las soluciones encontradas de forma manual suelen dejar muchas de las restricciones del problema sin cumplir, y el descontento de los actores que de una u otra forma están relacionados con la construcción del calendario final, se hace notar de diferentes maneras. Un camino para resolver el problema es la Programación Lineal Entera (ILP) [3], que considera un conjunto de ecuaciones que representan las restricciones naturales de la elaboración de horarios, además de una función objetivo (FO). Formulaciones de programación lineal del problema y sus restricciones, se pueden encontrar a lo largo de la literatura, sin embargo muchas de ellas no contienen algunas restricciones de fuerte interés y que se esperan se cumplan en la planificación. Muchas veces, esto es debido a que algunas restricciones suelen ser difíciles de formular y limitan los problemas a instancias que no son aplicables a algunos de los casos reales.

Aquí es donde se consideran alternativas de solución Heurísticas o Metaheurísticas como Algoritmos Genéticos [7], Búsqueda Tabú [6], Recocido Simulado, entre otros, que encuentran soluciones aproximadas, pero que con un buen diseño pueden estar muy cerca del óptimo.

2.2. ENTIDADES QUE PARTICIPAN EN LA CALENDARIZACIÓN DE DEPORTES TTP

De acuerdo a la literatura de documentos desarrollados por otros autores, se ha encontrado una serie de presentaciones para la formulación de este problema, lo cual es debido a que no es posible escribir una que contenga todos los casos posibles que pueden presentarse en la vida real, cada federación de distintos países poseen características diferentes que hacen que el problema se vuelva muy particular. Con el fin de tener una descripción general de este problema se ha tomado como referencia diferentes propuestas que contienen las restricciones más populares (duras y suaves) del problema. Primero se comienza con definir las entidades que participan en el problema:

- **Esquema (Pattern).-** Cadena de símbolos que pertenecen al conjunto H, A indicando una sucesión de partidos de Local (H, por Home) y Visitante (A, por Away) de un equipo en el torneo.
- **Quiebre (Break).-** Un Quiebre ocurre cuando dos partidos consecutivos se juegan o bien de local o bien de visitante.
- **Conjunto de esquemas (Pattern Set).-** Conjunto de esquemas diferentes con cardinalidad igual al número de equipos del torneo.
- **Televisoras.-** Son las distintas empresas televisivas que poseen los derechos de transmisión de los diferentes equipos.
- **Esquemas Complementarios.-** Son los esquemas correspondientes a los partidos de vuelta por cada equipo.
- **Semanas.-** Es el número total de encuentros realizados por cada equipo.
- **Emparejamiento.-** Es la asignación de los equipos a los esquemas factibles.
- **Round Robin.-** Es el tipo de torneo cuando todos los equipos jugaban todos contra todos una sola vez por cada etapa.

3. DESCRIPCIÓN Y DEFINICIÓN DEL PROBLEMA

3.1. SISTEMA DEL CAMPEONATO ECUATORIANO DE FÚTBOL 2011

El Campeonato Ecuatoriano de Fútbol de 2011, se jugó con la misma modalidad que el año 2010, de acuerdo a lo que decidieron los dirigentes de la Federación Ecuatoriana de Fútbol. El Campeonato

Ecuatoriano de Fútbol de 2011, según lo establecido en el Reglamento 2011 otorgado por la Federación Ecuatoriana de Fútbol, éste se jugó con 12 equipos que se disputaron el título en tres etapas. De las cuales dos etapas fue de todos contra todos y la tercera etapa fue para definir al campeón y a los mejores equipos de la Tabla Acumulada, que no eran los finalistas, los cuales pelearon por el tercer lugar que otorgó un cupo a la Primera Fase (repesca) de la Copa Libertadores. Una de las tareas de la FEF es la programación de los partidos de cada torneo de la Primera y Segunda División en general. Hace algunos años, este era uno de los más altos problemas que tenía de FEF, debido a que por lo general cada año cambiaban la modalidad (características) del campeonato. Una de las modalidades que tenía inicialmente la FEF era de elegir aleatoriamente unos números los cuales eran asignados a los equipos participantes en el torneo, de esta forma generaban el calendario utilizado en los torneos. Esta modalidad les permitía realizar una programación que cumplía los requerimientos básicos del torneo, pero no permitía considerar una serie de otros criterios, lo que conllevaba a múltiples quejas de parte de los equipos. En base a este tipo de problemas que se presenta en la elaboración del calendario en este tipo de torneos, ya sean estos de Fútbol, Basket, etc., es creado este documento, debido a que se presenta un mecanismo para la elaboración de los mismo (calendarios de juegos), en el cual nos basamos a las características básicas de un torneo todos contra todos mediante técnicas de Investigación de Operaciones (IO). Adicionalmente se agrega restricciones basadas en los derechos de televisivos de las TVs, para de una u otra forma definir los recursos de estas (TVs); ésta es una de las características más importantes que presentamos en el desarrollo del calendario, la cual además incorporó requerimientos de la FEF y de los clubes. Las condiciones consideradas integraron múltiples aspectos tales como partidos considerados clásicos (equipos de la misma ciudad), partidos de mucho interés para el público (aficionados), entre otros.

3.1.1. EQUIPOS PARTICIPANTES

TABLA I

Desarrollo de una aplicación para calendarizar el campeonato Ecuatoriano de fútbol profesional por medio de una aproximación Heurística utilizando programación entera

Equipos Participantes

Equipos Participantes	
Equipo	**Ciudad**
Barcelona	Guayaquil
Emelec	Guayaquil
Deportivo Quito	Quito
El Nacional	Quito
Espoli	Quito (Santo Domingo)
Liga de Quito	Quito
Deportivo Cuenca	Cuenca
Imbabura S.C.	Ibarra
Independiente del Valle	Sangolquí
Liga de Loja	Loja
Manta F.C.	Manta
Olmedo	Riobamba

TABLA II

Desarrollo de una aplicación para calendarizar el campeonato Ecuatoriano de fútbol profesional por medio de una aproximación Heurística utilizando programación entera

Equipos por Provincia

Equipos por Provincia	
Provincia	Equipos
Pichincha	El Nacional, Independiente del Valle, Liga de Quito, Espoli y Deportivo Quito
Guayas	Barcelona y Emelec
Azuay	Deportivo Cuenca
Chimborazo	Olmedo
Imbabura	Imbabura S.C.
Loja	Liga de Loja
Manabí	Manta F.C.

3.2. DEFINICIÓN DEL PROBLEMA

Cuando el problema de calendarización de un torneo todos contra todos (con un número par de equipos) no tiene restricciones, su construcción puede ser fácilmente obtenido en tiempo lineal respecto al número de partidos y tiene una interpretación teórica sobre grafos, porque corresponde al problema de encontrar la 1-factorización de un grafo completo no dirigido con un número par de vértices.

El problema de calendarización de deportes consiste básicamente en:

Dados un conjunto de equipos, un conjunto de televisoras, un conjunto de esquemas; en asignar los equipos a las respectivas televisoras que poseen sus derechos de transmisión con un

determinado esquema específico disminuyendo la cantidad de quiebres, todo esto están sujetos a un conjunto de restricciones que son consideradas suaves y duras. La mayor dificultad de este problema consiste en encontrar soluciones que involucren un emparejamiento de los equipos con las televisoras y su respectivo esquema para una determinada semana. Debido a estas características este problema estaría dentro de la categoría de los problemas de optimización combinatoria. Por este motivo no siempre es posible dar una solución que sea del todo aceptable para todas las partes involucradas.

3.2.1. RESTRICCIONES DEL PROBLEMA

Dado un conjunto N de n equipos provenientes de diferentes ciudades, crearemos un calendario de juegos para la primera y segunda etapa, con la modalidad de todos contra todos (DOUBLE ROUND ROBIN TOURNAMENT), cada equipo jugará dos veces contra un equipo, uno de local y el otro de visitante.

Las restricciones que asumiremos para este problema son las siguientes:

- Esta es basada a la posición que los equipos del campeonato obtuvieron en la temporada inmediata anterior. Los mejores equipos de la seria A del año anterior los consideraremos cabezas de serie para el siguiente año y para estos equipos se sugiere que no jueguen entre ellos las primeras y últimas γ semanas (asumiremos $\gamma = 2$).

TABLA III

Desarrollo de una aplicación para calendarizar el campeonato Ecuatoriano de fútbol profesional por medio de una aproximación Heurística utilizando programación entera

Equipos Cabezas de Serie

Equipos cabeza de serie
Emelec
Liga de Quito
Barcelona
Deportivo Quito

- Una restricción típica que consideraremos es relacionado a la sucesión de partidos de local y visitante, es decir no se permitiría más de dos partidos consecutivos de local (visitante) para cada equipo.
- Como este es un caso particular de las aplicaciones de TTP para deportes, debemos tener particularidades en su formulación. En lo presente, tenemos equipos que pertenecen a las mismas ciudades como equipos representantes, es decir que será necesario que los equipos de la misma ciudad deberán tener programaciones de juegos local-visitante complementarios, por ejemplo Barcelona y Emelec son únicos representantes de Guayaquil, lo que implica que si Barcelona juega de local en una determinada fecha, Emelec jugaría de visitante en la misma jornada.
- De la misma forma que para los equipos cabezas de serie, los equipos entre equipos de una misma ciudad no se les permitiría jugar ente ellos en las primeras y últimas *k* semanas, asumiremos $k = 2$. Tabla IV

TABLA IV

Desarrollo de una aplicación para calendarizar el campeonato Ecuatoriano de fútbol profesional por medio de una aproximación Heurística utilizando programación entera

Equipos de una misma ciudad

Equipos de una misma ciudad
Barcelona / Emelec
Liga de Quito / Deportivo Quito / El Nacional

- Como suele suceder en todos los países, existen equipos que son únicos representantes de una ciudad, es comúnmente preferible que estos equipos no jueguen un partido de local cuando exista en la ciudad algún evento festivo obligatorio, como éstas de fundación, entre otras. Esta restricción se considera suave.
- Finalmente desde hace algunos años las compañías de TV poseen derechos de transmisión de los partidos del Campeonato Nacional Copa Credife. Los respectivos equipos son asignados a estas empresas como: Ecuavisa, Teleamazonas y GamaTv; lo que significa que la asignación de un equipo j a la televisora p implica que p puede transmitir en vivo los partidos de *j*.

TABLA 5

Desarrollo de una aplicación para calendarizar el campeonato Ecuatoriano de fútbol profesional por medio de una aproximación Heurística utilizando programación entera

Televisoras con los derechos de transmisión

Equipos asignados a cada Televisora		
TV1	**TV2**	**TV3**
Emelec	Liga de Quito	Barcelona S.C.
El Nacional	Dep. Quito	Independiente
Deportivo Cuenca	Liga de Loja	
Olmedo	Manta F.C.	
	Espoli	
	Imbabura	

En este trabajo se impone una restricción de balance en la asignación de equipos a las televisoras, donde los partidos de local en todas las semanas son proporcionalmente divididos. Esto es dividido a que al menos los 2/3 de los partidos se juegan al mismo tiempo. Es decir, un buen calendario balanceado debería minimizar los requerimientos de personal y equipo disponible en paralelo.

4. MODELO MATEMÁTICO DEL PROBLEMA

El modelo matemático que se presenta a continuación está basado en la información referente al problema de calendarización de deportes, aunque no todas las restricciones que serán tomadas en este documento se encuentran presentes, sino las más relevantes que serán analizadas para el caso de estudio.

4.1. PROCEDIMIENTO PARA LLEGAR A LA SOLUCIÓN (CALENDARIO)

Este procedimiento está relacionado al trabajo realizado por D. Oliveri and F. Della Croce [4], en la elaboración del calendario de la Liga Italiana de Fútbol, los cuales lo definieron de la siguiente manera:

1. Búsqueda de conjuntos de esquemas factibles en base a los recursos de las televisoras.
2. Búsqueda de calendarios factibles basados a los esquemas factibles.
3. Emparejamiento de equipos a esquemas para la elaboración del calendario final.

4.1.1. PRIMERA FASE: BÚSQUEDA DE CONJUNTOS DE ESQUEMAS FACTIBLES

En la primera fase se busca un número suficientemente grande de diferentes conjuntos de esquemas formados por esquemas complementarios los cuales están balanceados respecto a la cobertura de televisión, además de minimizar el número total de quiebres.

Modelo de Programación Entera

p_i = Número de quiebres del esquema i (que toma valores de p_i = 0, 1 o 2).

NTV_k = Número total de equipos asignados a la televisora k.

$C_{i,l}$ = Conjunto que contiene los pares de esquemas complementarios.

$$A_{i,j} = \begin{cases} 1 \text{ Si el esquema } i \text{ juega de visitante en la semana } j \\ 0 \text{ Si no} \end{cases}$$

Y por último, definimos una variable binaria, la cual se describe de la siguiente manera:

$$X_{i,k} = \begin{cases} 1 \text{ Si el esquema } i \text{ es seleccionado a la televisora } k \\ 0 \text{ Si no} \end{cases}$$

En ésta programación matemática el objetivo se basa principal es minimizar el número de quiebres, la cual se la define de la siguiente manera:

$$Min \quad z = \sum_i \sum_k p_i x_{i,k}$$

Sujeto a las siguientes restricciones:

$$\sum_i A_{i,j} x_{i,k} = \frac{N_k}{2}; \quad k = 1,2 \forall j, \ (1)$$

$$\sum_i x_{i,k} = N_k; \quad k = 1,2, \quad (2)$$

$$x_{i,k} - x_{j,k} = 0; \quad \forall (i,j) \in C_{i,l}, \quad k = 1,2, \quad (3)$$

$$\sum_k x_{i,k} \le 1; \quad \forall i, \quad (4)$$

$$x_{i,k} \in \{0,1\}; \quad \forall i,k, \quad (5)$$

Para obtener un segundo conjunto de esquemas, como para generar un nuevo calendario o para buscar esquemas factibles para uso posterior, se repite esta fase agregando la siguiente restricción:

$$\sum_{(i,k) \in S} x_{i,k} \le \frac{n}{2}; \ (6)$$

donde S es el conjunto de esquemas seleccionados por la solución inmediatamente anterior. La restricción 6 asegura que al menos el 50% del conjunto de variables en la solución previa tomarán el valor de cero, es decir se obtendrá una nueva solución significativamente diferente a la anterior.

4.1.2. SEGUNDA FASE: BÚSQUEDA DE CALENDARIOS FACTIBLES

En esta fase, para cada conjunto de esquemas generados en la fase 1, se busca un calendario factible y consistente con este.

Modelo de Programación Entera

En ésta segunda fase interviene la siguiente variable binaria, la cual se la define de la siguiente manera:

$$X_{i,j,t} = \begin{cases} 1 \text{ Si el esquema } i \text{ es emparejado al esquema } j \\ \quad \text{en la semana } t \\ 0 \text{ Si no} \end{cases}$$

$d_{i,j,t}$ = Conjunto de esquemas que tienen juego de local (visitante) en la semana t donde el esquema i juega de visitante (local)

Como estrategia de modelización para esta fase usaremos como objetivo minimizar una constante (función de costos ficticia).

Sujeto a las siguientes restricciones:

$$\sum_{j \in d_{i,j,t}} x_{i,j,t} = 1; \quad \forall i,t, \qquad (7)$$

$$x_{i,j,t} = x_{j,i,t}; \quad \forall i, j > i, t, \qquad (8)$$

$$\sum_{t} x_{i,j,t} = 1; \qquad \forall i, j \neq i, \qquad (9)$$

$$x_{i,j,t} \in \{0,1\}; \quad \forall i, j, t \qquad (10)$$

4.1.3. TERCERA FASE: EMPAREJAMIENTO DE EQUIPOS A ESQUEMAS

En esta fase se recibe como datos de entrada cada calendario factible generado por la fase 2. Es posible que en la fase 2 no se generen calendarios factibles para algún conjunto de esquemas, para estos casos la fase 3 no se ejecutara. Se asignarán equipos reales a cada calendario factible cumpliendo las restricciones de los cabezas de serie y de aquellos que se localizan en la misma ciudad.

Modelo de Programación Entera

El modelo de programación entera para esta fase va definida de la siguiente manera:

Sea x una variable binaria descrita como:

$$X_{i,j} = \begin{cases} 1 \text{ Si el esquema } i \text{ es emparejado al equipo } j \\ 0 \text{ Si no} \end{cases}$$

$C_{i,h}$ Conjunto que contiene los pares de esquemas complementarios.

$D_{j,j}$ Conjunto que contiene los pares de equipos que juegan en la misma ciudad.

$E_{i,r}$ = Conjunto de esquemas que no pueden ser emparejados a un equipo cabeza de serie β si el esquema i ya ha sido emparejado con otro equipo cabeza de serie α. Esta variable ayuda a cumplir la restricción de que los partidos de los equipos cabeza de serie no pueden jugarse en las primeras y últimas γ semanas.

F_i = Conjunto de esquemas que no pueden ser emparejados a un equipo dado δ si el esquema i es emparejado a otro equipo dado el cual está localizado en la misma ciudad de δ. Esta variable ayuda a cumplir la restricción de que los partidos de los equipos cabeza de la misma ciudad no pueden jugarse en las primeras y últimas γ semanas.

$T(j)$ = Conjunto de equipos cabezas de serie.

$$min \quad z = \sum_i \sum_j x_{i,j}$$

Sujeto a las siguientes restricciones:

$$\sum_i x_{i,j} = 1; \quad \forall j, \qquad (11)$$

$$\sum_j x_{i,j} = 1; \quad \forall i, \qquad (12)$$

$$x_{i,j} + \sum_{r \in E_{i,r}} x_{l,m} \leq 1; \qquad \forall i, m \in (T/j), \qquad (13)$$

$$x_{i,j} + \sum_r x_{r,m} \leq 1; \qquad \forall i, (j,m) \in D, r \in F_{i,r}, \qquad (14)$$

$$x_{i,j} - x_{l,m} = 0; \quad \forall (i,l) \in C, (j,m) \in D, \qquad (15)$$

$$x_{i,j} \in 0,1 \quad \forall i, j, \qquad (16)$$

4.2. RECURSOS COMPUTACIONALES EMPLEADOS

En la realización para encontrar una solución al problema de la calendarización de deportes, hemos definido tres modelos matemáticos, los cuales están descritos en la sección anterior. Una vez definido el modelo procedemos al uso de las herramientas computacionales, para estas se uso las siguientes:

- Sistema Operativo: Microsoft Windows XP Profesional Service Pack 3.
- Equipo: Inter ® Pentium ® 4, CPU 2.80GHz, 704MB de RAM.
- Gams® (Motor de Optimización).
- Wolfram Mathematica 7 ®.

5. COMPARACIÓN DEL CALENDARIO PROPUESTO VS ORIGINAL

En esta sección se campara la eficiencia de este trabajo en base al calendario que fue utilizado en el año 2011. De antemano se presenta el Calendario Propuesto, bajo la metodología desarrollada en este trabajo (Ver Figura 1), correspondientes a las 22 fechas que se desarrolló; a partir de la fecha o semana 12 representa los partidos de revancha (de vuelta) de la primera ronda.

FIGURA 1

Desarrollo de una aplicación para calendarizar el campeonato Ecuatoriano de fútbol profesional por medio de una aproximación Heurística utilizando programación entera

Calendario Propuesto en este trabajo

Semana 1			Semana 2			Semana 3			Semana 4			Semana 5			Semana 6		
El Nac	vs	Espoli	D. Cuenca	vs	Indep	El Nac	vs	D. Cuenca	Manta	vs	L. Loja	El Nac	vs	D. Quito	Indep	vs	Bar
Bar	vs	Olm	Eme	vs	El Nac	Bar	vs	Eme	D. Cuenca	vs	LDU Q.	Bar	vs	D. Cuenca	Manta	vs	Espoli
LDU Q.	vs	Imba	D. Quito	vs	Manta	LDU Q.	vs	D. Quito	Eme	vs	Indep	LDU Q.	vs	Manta	D. Cuenca	vs	L. Loja
L. Loja	vs	Eme	Olm	vs	LDU Q.	L. Loja	vs	Olm	D. Quito	vs	Bar	L. Loja	vs	Indep	Eme	vs	LDU Q.
Indep	vs	D. Quito	Imba	vs	Bar	Indep	vs	Manta	Olm	vs	Espoli	Imba	vs	Olm	D. Quito	vs	Imba
Manta	vs	D. Cuenca	Espoli	vs	L. Loja	Espoli	vs	Imba	Imba	vs	El Nac	Espoli	vs	Eme	Olm	vs	El Nac
Semana 7			**Semana 8**			**Semana 9**			**Semana 10**			**Semana 11**			**Semana 12**		
El Nac	vs	Manta	L. Loja	vs	LDU Q.	El Nac	vs	L. Loja	Bar	vs	El Nac	El Nac	vs	LDU Q.	LDU Q.	vs	El Nac
Bar	vs	L. Loja	Indep	vs	El Nac	Bar	vs	LDU Q.	LDU Q.	vs	Espoli	Eme	vs	Manta	Manta	vs	Eme
LDU Q.	vs	Indep	Manta	vs	Bar	D. Quito	vs	Eme	L. Loja	vs	D. Quito	D. Quito	vs	D. Cuenca	D. Cuenca	vs	D. Quito
Olm	vs	Eme	D. Cuenca	vs	Espoli	Olm	vs	D. Cuenca	Indep	vs	Imba	Olm	vs	Indep	Indep	vs	Olm
Imba	vs	D. Cuenca	Eme	vs	Imba	Imba	vs	Manta	Manta	vs	Olm	Imba	vs	L. Loja	L. Loja	vs	Imba
Espoli	vs	D. Quito	D. Quito	vs	Olm	Espoli	vs	Indep	D. Cuenca	vs	Eme	Espoli	vs	Bar	Bar	vs	Espoli
Semana 13			**Semana 14**			**Semana 15**			**Semana 16**			**Semana 17**			**Semana 18**		
El Nac	vs	Bar	L. Loja	vs	El Nac	LDU Q.	vs	L. Loja	Manta	vs	El Nac	Bar	vs	Indep	D. Quito	vs	El Nac
Espoli	vs	LDU Q.	LDU Q.	vs	Bar	El Nac	vs	Indep	L. Loja	vs	Bar	Espoli	vs	Manta	D. Cuenca	vs	Bar
D. Quito	vs	L. Loja	Eme	vs	D. Quito	Bar	vs	Manta	Indep	vs	LDU Q.	L. Loja	vs	D. Cuenca	Manta	vs	LDU Q.
Imba	vs	Indep	D. Cuenca	vs	Olm	Espoli	vs	D. Cuenca	Eme	vs	Olm	LDU Q.	vs	Eme	Indep	vs	L. Loja
Olm	vs	Manta	Manta	vs	Imba	Imba	vs	Eme	D. Cuenca	vs	Imba	Imba	vs	D. Quito	Olm	vs	Imba
Eme	vs	D. Cuenca	Indep	vs	Espoli	Olm	vs	D. Quito	D. Quito	vs	Espoli	El Nac	vs	Olm	Eme	vs	Espoli
Semana 19			**Semana 20**			**Semana 21**			**Semana 22**								
L. Loja	vs	Manta	D. Cuenca	vs	El Nac	Indep	vs	D. Cuenca	Espoli	vs	El Nac						
LDU Q.	vs	D. Cuenca	Eme	vs	Bar	El Nac	vs	Eme	Olm	vs	Bar						
Indep	vs	Eme	D. Quito	vs	LDU Q.	Manta	vs	D. Quito	Imba	vs	LDU Q.						
Bar	vs	D. Quito	Olm	vs	L. Loja	LDU Q.	vs	Olm	Eme	vs	L. Loja						
Espoli	vs	Olm	Manta	vs	Indep	Bar	vs	Imba	D. Quito	vs	Indep						
El Nac	vs	Imba	Imba	vs	Espoli	L. Loja	vs	Espoli	D. Cuenca	vs	Manta						

En las figuras 2 y 3 se muestra la cantidad de quiebres de ambos calendarios, con el objetivo de escoger el mejor. En ambos calendarios se ha podido observar que poseen el mismo número de quiebres (partidos consecutivos), con un total de 10, con la ventaja que el calendario propuesto cubre los recursos de las televisoras (TVs), cuestión que se considera primordial por los posibles cuestionamientos que puedan generar los aficionados de cada equipo. En la Figura 4 podemos constatarnos que en un promedio de 0.361 seg nos tomaría generar los datos correspondientes a la fase 1, en un promedio de 0.149 seg nos tomaría para ejecutar y extraer los datos correspondientes de la Fase 2 y finalmente en 0.050 seg en promedio para obtener los datos de la fase 3; es decir solo generando los valores de entrada para cada fase nos tomaría 0.560 seg con una desviación

FIGURA 2

Desarrollo de una aplicación para calendarizar el campeonato ecuatoriano de fútbol profesional por medio de una aproximación heurística utilizando programación entera

Números de Quiebres presentados en el Calendario Original

Matriz de Esquemas	Quiebres del Calendario Original											# Quiebres
	1	2	3	4	5	6	7	8	9	10	11	
Olmedo	1	0	1	0	1	0	1	0	0	1	0	1
Barcelona	1	0	1	0	1	0	1	0	1	0	0	1
Liga de Loja	1	0	1	0	0	1	0	1	0	1	0	1
Dep. Quito	1	0	1	0	1	0	0	1	0	1	0	1
Espoli	1	0	0	1	0	1	0	1	0	1	0	1
Indepte	1	0	1	0	1	0	1	0	1	0	1	0
LDU Q	0	1	0	1	1	0	1	0	1	0	1	1
Imbabura	0	1	1	0	1	0	1	0	1	0	1	1
El Nacional	0	1	0	1	0	1	0	1	1	0	1	1
Manta	0	1	0	1	0	1	1	0	1	0	1	1
Emelec	0	1	0	1	0	1	0	1	0	1	1	1
Dep. Cuenca	0	1	0	1	0	1	0	1	0	1	0	0
											Numero Total de Quiebres	***10***

FIGURA 3

Desarrollo de una aplicación para calendarizar el campeonato ecuatoriano de fútbol profesional por medio de una aproximación heurística utilizando programación entera

Números de Quiebres presentados en el Calendario Propuesto

Matriz de Esperanza	Quiebres del Calendario Propuesto											
	1	2	3	4	5	6	7	8	9	10	11	# Quiebres
El Nacional	1	0	1	0	1	0	1	0	1	0	1	**0**
Barcelona	1	0	1	0	1	0	1	0	1	1	0	**1**
LDU Q	1	0	1	0	1	0	1	0	0	1	0	**1**
Liga de Loja	1	0	1	0	1	0	0	1	0	1	0	**1**
Indepte	1	0	1	0	0	1	0	1	0	1	0	**1**
Manta	1	0	0	1	0	1	0	1	0	1	0	**1**
Dep. Cuenca	0	1	0	1	0	1	0	1	0	1	0	**0**
Emelec	0	1	0	1	0	1	0	1	0	0	1	**1**
Dep. Quito	0	1	0	1	0	1	0	1	1	0	1	**1**
Olmedo	0	1	0	1	0	1	1	0	1	0	1	**1**
Imbabura	0	1	0	1	1	0	1	0	1	0	1	**1**
Espoli	0	1	1	0	1	0	1	0	1	0	1	**1**
											Numero Total de Quiebres	***10***

estándar de 0.036 seg en total.

FIGURA 4

Desarrollo de una aplicación para calendarizar el campeonato ecuatoriano de fútbol profesional por medio de una aproximación heurística utilizando programación entera

Tiempos de Ejecución de las fases en Wolfram Mathematica

	Ejecuciones realizadas en Wolfram Mathematica											
Tiempo de Corrida(seg)	Iter 1	Iter 2	Iter 3	Iter 4	Iter 5	Iter 6	Iter 7	Iter 8	Iter 9	Iter 10	Promedio	Desviación Estandar
FASE 1	0,422	0,329	0,312	0,343	0,374	0,344	0,406	0,344	0,391	0,344	0,361	0,036
FASE 2	0,157	0,141	0,156	0,156	0,157	0,125	0,156	0,156	0,125	0,157	0,149	0,013
FASE 3	0,032	0,047	0,047	0,062	0,032	0,063	0,063	0,062	0,047	0,047	0,050	0,012
Tiempo Total	*0,611*	*0,517*	*0,515*	*0,561*	*0,563*	*0,532*	*0,625*	*0,562*	*0,563*	*0,548*	*0,560*	*0,036*

6. CONCLUSIONES Y RECOMENDACIONES

6.1. CONCLUSIONES

1. Esta metodología, permite generar en cuestiones de minutos nuevos calendarios, disponibles para ser presentado como propuesta en la Federación Ecuatoriana de Fútbol (FEF), enfocadas en restricciones vistas en el Capítulo 3.

2. Basándonos a los recursos de las empresas de TV, podemos concluir que esta metodología les permite planificar sus actividades relacionadas a los derechos de transmisión sin ningún imprevisto de horarios.

3. Por medio de ésta heurística, se permite elaborar calendarios de juegos, dependiendo del número de equipos participantes en el torneo. Cabe señalar, que el nivel de dificultad depende del número de restricciones que esta contenga.

4. Esta aplicación generaría un impacto positivo tanto a nivel cuantitativo como cualitativo. A pesar de que la medición del impacto cuantitativo no es directa y es prácticamente imposible aislarlo de efectos exógenos, algunas observaciones pueden ser realizadas. Un factor relevante, que significaría ahorros para las TVs es en la logística que ellos plantearían cuando tiene que cubrir más de dos partidos de local en una determinada fecha (semana).

6.2. RECOMENDACIONES

1. Implementar esta metodología, inicialmente en torneos pequeños, para realizar un análisis del mismo, para luego, llegar a plasmarlo en la calendarización de la 1era y 2da División del Campeonato Nacional de Fútbol y otros torneos de interés.

2. Evaluar el impacto que genera esta metodología en base a los recursos de las empresas de TV, tomando en cuenta los ingresos que estas podrían generar por las transmisiones de los partidos y su aumento en competitividad.

3. Teniendo en cuenta, que cada equipo participante en el torneo (Clubes), presentan su propuesta en la calendarización en base a sus propios beneficios, se recomienda que se lleve a cabo una sesión ordinaria del Comité Ejecutivo de la Federación Ecuatoriana de Fútbol, junto a los representantes de cada club, con el _n de realizar en conjunto características propias del torneo, para luego, ingresarlas computacionalmente mediante el método propuesto en este documento.

REFERENCIAS BIBLIOGRÁFICAS Y ELECTRÓNICAS

[1]. **ADENSO DÍAZ** (1996). "*Optimización Heurística y Redes Neuronales Addison Wesley*".

[2]. **AIDA DÍAZ FERNÁNDEZ, JAVIER RODRIGO, MARÍA LUISA GUERRERO LERMA.** "*Emparejamientos aplicados a la elaboración de calendarios deportivos*".

[3]. **CHIN-YEN CHEN** (2008). "*Using integer programming to solve the school timetabling problem at chin-min institute of technology*". American Academy of Business.

[4]. **D. OLIVERI AND F. DELLA CROCE** (2004). "*Scheduling the italian football league: an ILP-based approach*". ELSEVIER.

[5]. **DURÁN, G., M. GUAJARDO, J. MIRANDA, D. SAURÉ, S. SOUYRIS, A. WEINTRAUB**, (2007). "*Scheduling the Chilean soccer league by Integer Programming*", Interfaces 37(6) 539-552.

[6]. **JIN-KAO HAO AND ZHIPENG LU** (2008). "*Adaptive tabú search for course timetabling*". ELSEVIER.

[7]. **KI-SEOK SUNG AND ENZHE YU** (2007). "A genetic algorithm for a university weekly courses *timetabling problem*". Blackwell Publishers.

[8]. **NEMHAUSER GL AND TRICK MA** (1998). "*Scheduling a major college basketball conference*". Operations Research.

[9]. **RIBEIRO, C., S. URRUTIA.** (2009). "*Scheduling the Brazilian soccer tournament by integer programming maximizing audience shares under fairness constraints*". 23rd European Conference on Operational Research, Book of Abstracts. Bonn, Germany, p.240.

[10]. **XAVIER CABEZAS G.** (2009). "*Calendarización de la Liga Italiana de Fútbol: Una aproximación (heurística) basada en ILP*". Escuela Superior Politécnica del Litoral.

matemática: Una publicación del ICM – ESPOL
2012, Vol. 10, No. 1

MANIPULACIÓN DEL ESPECTRO DE UNA FUNCIÓN BIDIMENSIONAL PARA REALCE DE DEFECTOS SUPERFICIALES EN PIEZAS METÁLICAS

González Javier[1], Calvo Camilo[2], Cruz José[3], Tolosa Jorge[4]

Resumen. *Este trabajo tiene el objetivo de mostrar los resultados obtenidos a través de la manipulación del espectro de una función bidimensional para realzar los defectos superficiales en piezas metálicas. La función bidimensional corresponde a una imagen de intensidades tomada sobre una pieza metálica y a través de la transformada discreta de Fourier bidimensional se ha trasladado al dominio de la frecuencia para ser alterada por un banco de filtros.*

Palabras Claves: Filtrado digital, dominio de la frecuencia, Piezas Metálicas.

Abstract. *This paper shows results obtained through the manipulation of two-dimensional spectrum from a function for enhance the surface defects in metal parts. Bidimensional function corresponds to an intensity image taken on a metal part and through the two-dimensional discrete Fourier transform is moved to the frequency domain to be changed by a filter bank.*

Keywords: Digital filtering, frequency domain, Metallic part.

Recibido: Abril, 2012
Aceptado: Junio, 2012

I. INTRODUCCION

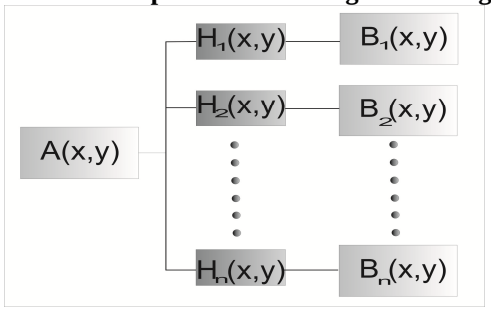
Una imagen digital es una función bidimensional *A(x,y)* en el dominio del espacio, que se describe a través de una matriz que posee un conjunto de valores discretos de intensidades de luz reflejada por un objeto. El procesamiento básico de una imagen digital se puede realizar, en el dominio del espacio, a través de la convolución discreta en la cual interviene la función A(x,y) y una función denominada máscara H(x,y) y se obtiene una función de salida B(x,y) [1]. Ver ecuación 1.

$$B(x,y) = A(x,y) * H(x,y) \quad (1)$$

El proceso basado en la convolución discreta, descrito en la ecuación 1 también se denomina filtrado digital y la función H(x,y) puede tomar diferentes valores que dependen de su aplicación. Tradicionalmente se han utilizado diferentes tipos de máscaras u operadores como: Prewit, Sobel, Kirsch y Gabor, en diversas aplicaciones del procesamiento digital de imágenes [2].

Como un método de aplicación del proceso de convolución, se han desarrollado los bancos de filtros, en el cual se tiene una función de entrada A(x,y) y es procesada por un arreglo en paralelo de filtros $H_n(x,y)$, como se ilustra en la figura 1.

FIGURA 1
Manipulación del espectro de una función bidimensional para realce de defectos superficiales en piezas metálicas
Estructura de un banco de filtros aplicado al contexto del procesamiento digital de imágenes



El arreglo de filtros $H_n(x,y)$, son un grupo de funciones diseñadas previamente con la finalidad de extraer o resaltar características, en el dominio del espacio, de la función de entrada A(x,y). Dentro de las principales características que son analizadas en el procesamiento digital de imágenes, es la textura, definida como la propiedad de los píxeles de presentar cambios en sus valores asociados a las irregularidades de la superficie del objeto [3].

Cada uno de los filtros $H_n(x,y)$ poseen una función base con sus respectivas expansiones en el dominio del espacio y de la frecuencia. Los bancos de filtros espaciales se utilizan en aplicaciones diversas como estrategias para eliminación de ruido (*denoising*) [4,5], análisis del proceso de diezmado [6], compresión de

[1] *González Barajas Javier, Magister en Ingenierías - Área Electrónica, UIS. Docente – Investigador. Facultad de Ingeniería Electrónica. Universidad Santo Tomás. Bogotá. Colombia. (e_mail: Javiere_gonzalez@yahoo.com.mx).*
[2] *Calvo Camilo, Fac. Ing. Electrónica. Universidad Santo Tomás. Bogotá Colombia. (e_mail: cam_calvo@ieee.org).*
[3] *Cruz José Manuel, Facultad de Ingeniería. Electrónica. Universidad Santo Tomás. Bogotá Colombia.* ***(e_mail: joseman61@hotmail.com).***
[4] *Tolosa Jorge Andrés, Facultad de Ingeniería Electrónica. Universidad Santo Tomás. Bogotá Colombia. (e_mail: georgetpr@hotmail.com).*

imágenes [7,8,9], codificación de imágenes [10], segmentación de imágenes[11], marcas de agua [12], restauración de imágenes [13].

En el análisis de texturas, a través del uso de banco de filtros, ha sido reportado en la literatura con resultados positivos para aplicaciones de segmentación [14], reconocimiento de objetos [15], extracción de información de la magnitud y fase del espectro [16] y recuperación de imágenes [17,18].

Para el estudio de defectos superficiales de piezas metálicas, se han desarrollado aplicaciones del procesamiento digital de imágenes, aplicadas principalmente para el pre-procesamiento de las imágenes adquiridas para la inspección superficial de elementos, a través de operaciones morfológicas en el dominio del espacio [19]. En aplicaciones industriales, como es el caso de la inspección de superficies, el estudio de las texturas ha sido de gran utilidad para la identificación de patrones [20], para la caracterización de regiones con patrones dinámicos [21]. En el caso del estudio superficial de metales, ya se ha evidenciado resultados del uso del procesamiento digital de imágenes para la caracterización de corrosión superficial en piezas metálicas [22].

Tradicionalmente las aplicaciones del procesamiento digital de imágenes, para inspección superficial de piezas metálicas, son implementadas en el dominio del espacio. Por lo cual se propone en este trabajo el diseño e implementación de una técnica basada en banco de filtros para la manipulación en el dominio de la frecuencia de piezas metálicas.

II. MATERIALES Y MÉTODOS

Para el diseño del banco de filtro, se debe tener en cuenta que la función de entrada $A(x,y)$ posee un espectro $A(u,v)$ el cual puede ser calculado a través de la transformada discreta de Fourier bidimensional (FFT2) (Ecuación.2).

$$A(u,v)=\frac{1}{MN}\sum_{x=0}^{N-1}\sum_{y=0}^{M-1}A(x,y)e^{-j2\pi\left(\frac{ux}{N}+\frac{vy}{M}\right)} \quad (2)$$

Para obtener una modificación del espectro de la función de entrada $A(x,y)$, se procede a multiplicar su función en el dominio de la frecuencia $A(u,v)$, por la función H(u,v) que desempeña el papel de una máscara en el dominio espectral. Ver Ecuación.3.

$$B(u,v)=A(u,v).H(u,v) \quad (3)$$

El nuevo espectro, contenido en la función $B(u,v)$, es el resultado del producto del espectro de la función de entrada y la máscara aplicada. La nueva imagen se obtiene a partir de la transformada bidimensional de Fourier Inversa (IFFT2). Ver Ecuación 4.

$$B(x,y)=\frac{1}{MN}\sum_{u=0}^{N-1}\sum_{v=0}^{M-1}B(u,v)e^{j2\pi\left(\frac{ux}{N}+\frac{vy}{M}\right)} \quad (4)$$

Teniendo en cuenta el procesamiento en el dominio de la frecuencia de una imagen digital, se diseña la arquitectura del banco de filtros que opera basados en máscaras $H_n(u,v)$, como se aprecia en la figura 2.

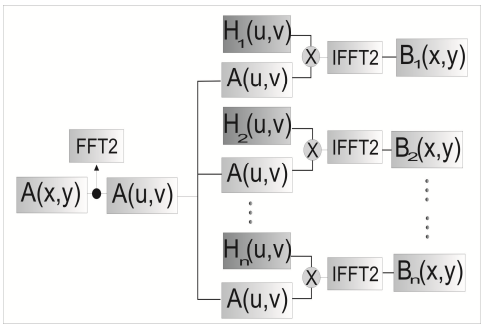
El sistema implementado (ver figura 2) se basa en el diagrama de bloques expuesto en la figura 1. El sistema toma una imagen adquirida en intensidades de grises $A(x,y)$ de dimensiones *N*x*N* y extrae su espectro $A(u,v)$ a través de la FFT2. El espectro $A(u,v)$ es sometido a un conjunto de máscaras $H_n(u,v)$ las cuales son funciones diseñadas para modificar el espectro. Se han diseñado dos tipos de máscaras básicas: pasa bajos y pasa altos.

La máscara pasa bajo se caracteriza por tener una función descrita en la ecuación 5. Donde *L* es el radio que delimita la banda de paso y *n* es un número entero.

FIGURA 2

Manipulación del espectro de una función bidimensional para realce de defectos superficiales en piezas metálicas

Diagrama de bloques del diseño del sistema de filtrado en el dominio de la frecuencia. El sistema toma una imagen de entrada A(x,y) y aplica, en el dominio de la frecuencia, un conjunto de máscaras $H_n(u,v)$



El índice n está definido para 1< *n*< *N/L*. Siendo N la dimensión de la matriz de entrada $A(x,y)$. El papel que juega el índice *n* es el de aumentar el área del ancho de banda de la máscara a medida que se incrementa su valor.

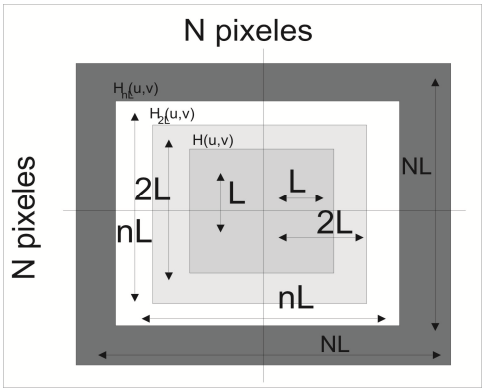
$$H_n(u,v)=\begin{cases}1 \text{ si } u<L*n \text{ y } v<L*n\\ 0 \text{ si } u>L*n \text{ y } v>L*n\end{cases} \quad (5)$$

La figura 3 ilustra el comportamiento de la máscara pasa bajos $H_n(u,v)$ a medida que se incrementa el valor *n*. La finalidad de la máscara pasa bajos es la de atenuar las altas frecuencias que posee la función $A(u,v)$ y solo permitir estudiar cómo influyen las bajas frecuencias en la imagen y determinar que características están relacionadas con las bajas frecuencias.

FIGURA 3

Manipulación del espectro de una función bidimensional para realce de defectos superficiales en piezas metálicas

Comportamiento de la máscara pasa bajos $H_n(u,v)$. Al inicio la máscara tiene una dimensión de 2L x 2L y se puede apreciar que, para cada iteración el área, de la máscara pasa bajos aumenta acorde a nL.



Para el caso de la máscara pasa altos, se ha diseñado una función descrita en la ecuación 6.

$$H_n(u,v) = \begin{cases} 0 \text{ si } u<L*n \text{ y } v<L*n \\ 1 \text{ si } u>L*n \text{ y } v>L*n \end{cases} \quad (6)$$

La finalidad de la máscara pasa altos consiste en permitir estudiar que características de la imagen se realzan con las altas frecuencias.

III. RESULTADOS

Las imágenes de prueba fueron adquiridas a través de un microscopio *Celestron 4302* con resolución de 1.3M pixeles y un zoom óptico de 150x. El radio inicial de la máscara pasa bajos fue *L*=1 y para cada iteración *L* aumentaba en una distancia de 10 pixeles. La dimensión de la imagen es de 1024 x1024 píxeles. En la figura 4 se aprecia una imagen de prueba adquirida de la superficie de una pieza metálica sin defectos, solo se aprecian los pequeños canales de la muestra.

FIGURA 4

Manipulación del espectro de una función bidimensional para realce de defectos superficiales en piezas metálicas

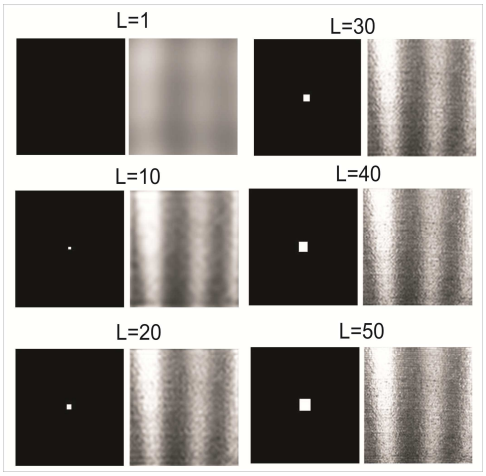
Imagen de prueba tomada de una superficie sin defectos.

En la figura 5 se aprecia el resultado de la aplicación de la máscara pasa bajos para cuatro diferentes valores de L.

FIGURA 5

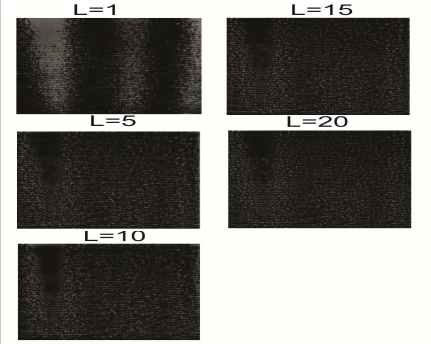
Manipulación del espectro de una función bidimensional para realce de defectos superficiales en piezas metálicas

Resultado obtenido con la máscara pasa bajos $H_n(u,v)$, para n=1, n=10, n=30 y n=50. A partir de n=40 se puede obtener una imagen clara del sujeto



Se puede ver en la figura 5, que a partir de *L*=50, la imagen es suficientemente clara. El resultado de este primer experimento permite determinar el conjunto de bajas frecuencias que están relacionadas con las características propias de la superficie de la pieza metálica sin defectos. El mismo experimento se realiza con la máscara pasa alto y el resultado se puede apreciar en la figura 6.

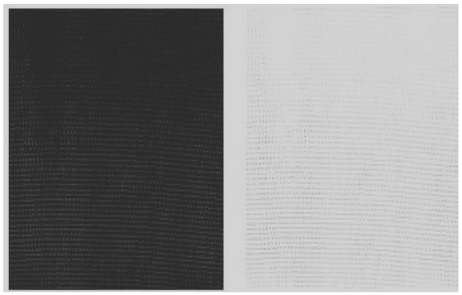
FIGURA 6
Manipulación del espectro de una función bidimensional para realce de defectos superficiales en piezas metálicas
Resultado obtenido con la máscara pasa altos $H_n(u,v)$, para n=1, n=2, n=3 y n=4



La figura 6 permite ver que las altas frecuencias están altamente relacionadas con los bordes característicos de los canales de la pieza. Teniendo en cuenta los resultados obtenidos con las máscaras pasa bajos y pasa altos, se pude determinar el rango de frecuencias que determinan las características propias de la pieza metálica. Al unir estos dos tipos de filtros se puede obtener un filtro pasa banda.

En la figura 7 se puede apreciar el resultado obtenido de procesar la imagen de prueba con un filtro pasa banda y su el resultado obtenido al extraer su versión negativa.

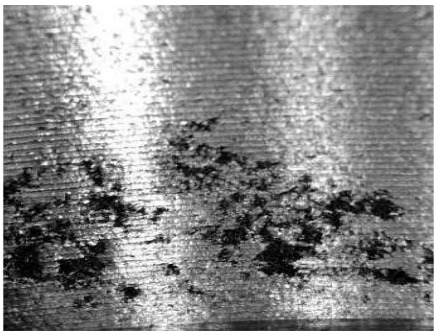
FIGURA 7
Manipulación del espectro de una función bidimensional para realce de defectos superficiales en piezas metálicas
Resultado obtenido de procesar la imagen de prueba con el filtro pasa banda construido a partir de las máscaras para alto y pasa bajo (izquierdo). Posteriormente se ha realizado a extraer la versión negativa de la imagen procesada (derecho)



El resultado obtenido con el filtro pasa banda permite obtener un realce de la textura asociada a los patrones propios de la pieza metálica sin defectos. Este proceso facilita la posterior realización de operaciones de binarización y segmentación de la imagen. Este resultado permite obtener el ancho de banda de las componentes espectrales asociadas a los patrones normales de la pieza bajo estudio.

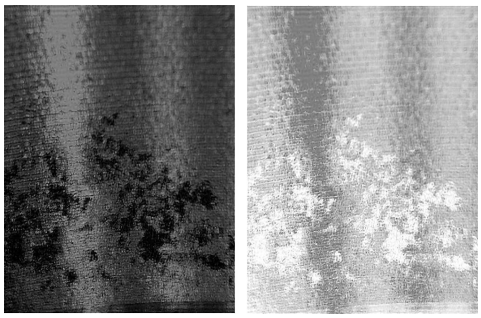
Esta información aportada por el filtro pasa banda, puede ser utilizada para detectar imperfecciones como las que se pueden apreciar en la imagen adquirida sobre la superficie de la pieza, como se ilustrada en la figura 8.

FIGURA 8
Manipulación del espectro de una función bidimensional para realce de defectos superficiales en piezas metálicas
Imagen tomada de una superficie con defectos



La imagen que contiene defectos superficiales se ha procesado por un filtro rechaza banda, diseñado a partir de las máscaras para altos y pasa bajos. EL filtro rechaza banda cumple la función contraria al pasa banda y se ha aplicado a la imagen con defectos para atenuar las frecuencias de los patrones normales de la pieza. El resultado obtenido se ilustra en la figura 9.

FIGURA 9
Manipulación del espectro de una función bidimensional para realce de defectos superficiales en piezas metálicas
Imagen tomada de una superficie con defectos procesada con un filtro rechaza banda, con la finalidad de atenuar los patrones normales y resaltar los defectos superficiales.



Como se pudo apreciar en la figura 9, se han atenuado las frecuencias de los patrones normales y como resultado se puede detallar los defectos superficiales. Esta operación permite la fácil detección de alguna textura que no pertenezca a los patrones de la pieza sin defecto.

IV. CONCLUSIONES

En este trabajo se ha generado una aplicación basada en la aplicación de bancos de filtros

basados en la generación de máscaras pasa bajos y pasa altos que procesan una imagen en el dominio de la frecuencia. Los resultados presentados han demostrado la utilidad del banco de filtros en el contexto de imágenes utilizadas para la inspección superficial de piezas metálicas.

La manipulación del espectro de la imagen digital, a través de los filtros para banda y rechaza banda, permitió poder resaltar los pixeles que hacen de los patrones normales de la pieza metálica, facilitando un posterior proceso de binarización y segmentación de la imagen. Al conocer las componentes espectrales propias de los patrones normales de la pieza, se ha facilitado la implementación de un filtro rechaza banda con la finalidad de atenuar estos patrones.

El filtro rechaza banda cumple con la labor de resaltar las componentes espectrales de los defectos superficiales, permitiendo posteriores procesos de binarización y segmentado.

V. AGRADECIMIENTOS

Los resultados obtenidos en este trabajo fueron logrados dentro de los objetivos del proyecto “Diseño e implementaciòn de una herramienta para la caracterización de defectos superficiales en piezas metàlicas a travès del procesamiento digital de imàgenes”, financiado por la convocatoria interna de proyectos de grupos de investigación, realizada por la unidad de investigación y postgrados de la Universidad Santo Tomás.

REFERENCIAS BIBLIOGRÁFICAS Y ELECTRÓNICAS

[1] **GONZALEZ , WOODS R.** (2008). "*Procesamiento Digital de Imágenes*". New Jersey, USA: Prentice Hall, pp120 - pp144.

[2] **PAJARES G.** (2008). "*Ejercicios Resueltos de Visión por Computador*". México D.F., México: Alfaomega-RaMa, pp115-pp142.

[3] **PAJARES G.** (2002). "*Visión por Computador*". México DF, México: Alfamaomega. p250.

[4] **ZAO-CHAO BAO; XIN-GE YOU; CHUN-FANG XING; QING-YAN HE.** (2007). "*Image Denoising by using Non-Tensor Product Wavelets Filter Banks*", Machine Learning and Cybernetics, 2007 International Conference on, vol.3, no., pp.1734-1738, 19-22.

[5] **WEIHUA LIU; MINGYI HE; PENGLANG SHUI; YUANYUAN CHENG.** (2009). "*Residue-based fusion of denoised images by different filter Banks*", Industrial Electronics and Applications, 2009. ICIEA 2009. 4th IEEE Conference on , vol., no., pp.2420-2423, 25-27.

[6] **KHAN, M.A.U.; KHAN, M.K.; KHAN, M.A.** (2005). "*Comparative Analysis of Decimation-Free Directional Filter Bank with Directional Filter Bank: In Context of Image Enhancement*", 9th International Multitopic Conference, IEEE INMIC 2005 , vol., no., pp.1-8, 24-25.

[7] **KOTTERI, K.A.; BELL, A.E.; CARLETTA, J.E.** (2004). "*Design of multiplierless, high-performance, wavelet filter banks with image compression applications*", Circuits and Systems I: Regular Papers, IEEE Transactions on , vol.51, no.3, pp. 483- 494.

[8] **QUIRK, M.D.; BRISLAWN, C.M.** (2004). "*Existence of optimal paraunitary finite impulse response filter banks for continuous objective functionals [JPEG-2000 image compression applications]*", Digital Signal Processing Workshop, 2004 and the 3rd IEEE Signal Processing Education Workshop. 2004 IEEE 11th , vol., no., pp. 44- 48, 1-4.

[9] **GORNALE, S.S.; HUMBE, V.T.; JAMBHORKAR, S.S.; YANNAWAR, P.; MANZA, R.R.; KALE, K.V.** (2007). "*Multi-Resolution System for MRI (Magnetic Resonance Imaging) Image Compression: A Heterogeneous Wavelet Filters Bank Approach*", Computer Graphics, Imaging and Visualisation, 2007. CGIV '07 , vol., no., pp.495-500, 14-17.

[10] **LOTFY, M.; RASHWAN, A.** (2006). "*A Comparative Study of Multirate Filter-Bank Structures for Wavelet Image Coding*", Radio Science Conference, 2006. NRSC 2006. Proceedings of the Twenty Third National, vol.0, no., pp.1-8, 14-16 March.

[11] **ERER, I.; KENT, S.; KARTAL, M.** (2008). "*SAR image segmentation using 2D four channel filter bank with lattice structure*", Radar Conference. RADAR '08. IEEE , vol., no., pp.1-4, 26-30 May.

[12] **MIYAZAKI, A.** (2003). "*On the evaluation of wavelet filter banks for wavelet-based image watermarking*", Image and Signal Processing and Analysis. ISPA 2003. Proceedings of the 3rd International Symposium on , vol.2, no., pp. 877- 882 Vol.2, 18-20 Sept. 2003.

[13] **ZHANG, X.; WANG, S.** (2006). "*Image restoration using truncated SVD filter bank based on an energy criterion*", Vision, Image and Signal Processing, IEE Proceedings - , vol.153, no.6, pp.825-836, December.

[14] **HONG, P.S.; KAPLAN, L.M.; SMITH, M.J.T.** (2003). "*Hyperspectral image segmentation using filter banks for texture augmentation*", Advances in Techniques for Analysis of Remotely Sensed Data, 2003 IEEE Workshop on, vol., no., pp. 254- 258, 27-28 October.

[15] **YOUSUN KANG; KIDONO, K.; NAITO, T.; NINOMIYA, Y.** (2008). "*Multiband image segmentation and object recognition using texture filter Banks*", Pattern Recognition. ICPR 2008. 19th International Conference on , vol., no., pp.1-4, 8-11 Dec. 2008.

[16] **VO, A.P.N.; ORAINTARA, S.; NGUYEN, T.T.** (2007). "*Using Phase and Magnitude Information of the Complex Directional Filter Bank for Texture Image Retrieval*", Image Processing. ICIP 2007.

IEEE International Conference on , vol.4, no., pp.IV-61-IV-64, Sept. 16 2007-Oct. 19 2007.

[17] VO, A.P.N.; NGUYEN, T.T.; ORAINTARA, S. (2006). "*Texture image retrieval using complex directional filter bank*". Circuits and Systems, 2006. ISCAS 2006. Proceedings. IEEE International Symposium on , vol., no., pp.4 pp.-5498, 0-0 0.

[18] ZHENYU HE; XINGE YOU; YUAN YAN TANG; WANG, P.; YUN XUE. (2006). "*Texture Image Retrieval Using Novel Non-separable Filter Banks Based on Centrally Symmetric Matrices*", Pattern Recognition. ICPR 2006. 18th International Conference on , vol.4, no., pp.161-164, 0-0 0.

[19] HASHIM, H.S.; PRABUWONO, A.S.; SHEIKH ABDULLAH, S.N.H. (2010). "*A study on pre-processing algorithms for metal parts inspection*". Energy, Power and Control (EPC-IQ), 1st International Conference on , vol., no., pp.195-198, Nov. 30 2010-Dec. 2 2010.

[20] NGAN, H.Y.T.; PANG, G.K.H. (2009). "*Regularity Analysis for Patterned Texture Inspection*". Automation Science and Engineering, IEEE Transactions on , vol.6, no.1, pp.131-144, Jan.

[21] CHUANZHEN LI; JINGLING WANG; LONG YE; HUI WANG. (2009). "*A Novel Method of Dynamic Textures Analysis and Synthesis*" Computational Sciences and Optimization. CSO 2009. International Joint Conference on , vol.2, no., pp.328-332, 24-26 April 2009.

[22] GARZÓN R., J., C. BARRERO, K. E. GARCÍA, F. PÉREZ, J. GALEANO, A. SALAZAR, AND H. LORDUY. (2006). "*Morphological analysis and classification of types of surface corrosion damage by digital image processing*". Revista Colombiana de Física 38, no. 2: 557-560.

matemática: Una publicación del ICM – ESPOL
2012, Vol. 10, No. 1

MECÁNICA CUÁNTICA: POSTULADOS

Iza Peter[1]

Resumen. *La Mecánica Cuántica describe de una manera completa el mundo microscópico, y para esto el físico británico P.A.M. Dirac hace uso de los elementos básicos del álgebra lineal para lograrlo. Este trabajo da a conocer los postulados fundamentales de la mecánica cuántica haciendo uso de los espacios vectoriales y operadores cuánticos.*

Palabras clave: Mecánica Cuántica, Postulados, Operadores.

Abstract. *The microscopic world is fully described by the quantum mechanics and P.A.M. Dirac uses the basic elements of* linear algebra to do so. In this work the postulates of quantum mechanics are presented using vector spaces and quantum operators.

Key words: Quantum Mechanics, Postulates, Operators.

Recibido: Abril, 2012
Aceptado: Junio, 2012

1. INTRODUCCIÓN

A finales del siglo XIX, la mecánica clásica creada por Newton en el siglo XVII, y complementada por las ecuaciones de Maxwell en la segunda mitad del siglo XIX, proporcionaba un marco teórico suficiente para la comprensión del mundo macrocósmico. Pero entre 1925 y 1930 la Mecánica Cuántica (MC) surge con el propósito de explicar los fenómenos que tenían lugar en condiciones pocos usuales, como velocidades muy altas o a escala microscópica; y su desarrollo ha dependido en gran medida de la exactitud de los resultados numéricos obtenidos en sus observaciones [1].

Los modelos matemáticos propuestos para describir los fenómenos microcósmicos y su posterior interpretación, fueron muy diversos. En algunos casos, las matemáticas usadas resultaban insatisfactorias y en absoluto rigurosas, lo que motivo en parte el desarrollo de algunas de las ramas mas activas e interesantes de las Matemáticas. Una formulación matemática rigurosa de la MC fue desarrollada por el físico Dirac [2].

La notación de Dirac, como es conocida hoy en día, proporciona una presentación abstracta del álgebra lineal que le da soporte a la mecánica cuántica basada en un conjunto de postulados. En este artículo se presenta una enumeración canónica de dichos postulados fundamentales.

[1] *Peter Iza, Ph.D., Profesor del Instituto de Ciencias Físicas, ESPOL. (e mail: piza@espol.edu.ec).*

2. BASE MATEMÁTICA

Los objetos básicos del algebra lineal son los espacios vectoriales y los elementos de un espacio vectorial se denominan vectores. En la mecánica clásica, la posición o "estado" de una partícula se describe por un vector que tiene tres números reales (x, y, z); por ejemplo: el vector posición, $\vec{r} = x\hat{i} + y\hat{j} + z\hat{k}$ está representado en un espacio tridimensional.

Dirac haciendo uso de los conceptos del álgebra lineal establece que el estado de un sistema cuántico es descrito por un elemento perteneciente a un espacio vectorial abstracto llamado espacio de estados o de Hilbert y denotado por $\mathcal{E}$. Por ejemplo, se tiene el siguiente espacio:

$$\mathcal{E} = \{|\Psi_1\rangle, |\Psi_2\rangle, ..., |\Psi_n\rangle, ...\}$$

Según la notación de Dirac, un elemento del espacio $\mathcal{E}$ se llama ket y se denota por el símbolo $|\Psi_n\rangle$, donde Ψ_n representa el n-ésimo estado del sistema cuántico. Matemáticamente, siempre que se tenga un conjunto de vectores kets, se puede construir un segundo conjunto de vectores, denominados vectores duales. Por lo tanto, para los vectores kets existen los vectores bra, representados por el símbolo $\langle\ |$, imagen simétrica del símbolo de un ket. El producto escalar del bra $\langle\Psi_1|$ y del ket $|\Psi_2\rangle$ se lo escribe como:

$$\langle\Psi_1|\Psi_2\rangle$$

y es conocido como un "braket".

Cualquier cantidad física experimentalmente medible como por ejemplo: la energía, el momento dipolar, momento angular orbital, el momento angular de espín o la energía cinética, cuyas expresiones en mecánica clásica puede escribirse en términos de las

posiciones cartesianas $\{q_i\}$ y los momentos $\{p_i\}$ de las partículas que componen el sistema de interés, se les asignan un correspondiente operador en la MC, conocido como un observable. Matemáticamente este observable es un operador lineal no comutable o Hermitiano para el que se puede encontrar una base ortonormal del espacio de estados, que consiste en los autovectores del operador.

Un operador es una instrucción que transforma un vector dado $|\Psi\rangle$ en otro vector $|\Psi'\rangle$. La acción del operador $\hat{\Omega}$ se la representa como:

$$\hat{\Omega}|\Psi\rangle = |\Psi'\rangle$$

El operador lineal obedece las siguientes reglas:

$$\hat{\Omega}\alpha|\Psi_i\rangle = \alpha\hat{\Omega}|\Psi_i\rangle$$

$$\hat{\Omega}\{\alpha|\Psi_i\rangle + \beta|\Psi_j\rangle\} = \alpha\hat{\Omega}|\Psi_i\rangle + \beta\hat{\Omega}|\Psi_j\rangle$$

donde α y β son constantes.

A un operador lineal $\hat{\Omega}$ se le puede asociar otro operador $\hat{\Omega}^\dagger$, llamado conjugado hermitiano de $\hat{\Omega}$, el cual debe satisfacer la siguiente igualdad, para cualquier vector Ψ_1 y Ψ_2:

$$\langle\Psi_1|\hat{\Omega}\Psi_2\rangle = \langle\Psi_1|\hat{\Omega}^\dagger\Psi_2\rangle$$

por lo tanto,

$$\hat{\Omega} = \hat{\Omega}^\dagger$$

Destaquemos cuatro propiedades importantes de estos operadores:

i) $\left(\hat{\Omega}^\dagger\right)^\dagger = \hat{\Omega}$;

ii) $\left(\hat{\Omega}\hat{\theta}\right)^\dagger = \hat{\theta}^\dagger\hat{\Omega}^\dagger$;

iii) $\left(\hat{\Omega}+\hat{\theta}\right)^\dagger = \hat{\Omega}^\dagger + \hat{\theta}^\dagger$;

iv) $\left(\lambda\hat{\Omega}\right)^\dagger = \lambda^*\hat{\Omega}^\dagger$, donde λ es un número complejo y λ^* es el complejo conjugado.

3. POSTULADOS

Los siguientes postulados de Mecánica Cuántica han sido adaptados de una manera sencilla; para esto se ha considerado como base algunos textos clásicos de la MC [2-5], ver bibliografía.

Postulado 1

El estado de un sistema físico en el tiempo *t* se define mediante la especificación de un ket $|\Psi(t)\rangle$ perteneciente al espacio de Hilbert $\mathcal{E}$. La función de onda $\Psi(t)$ es una representación del estado cuántico del sistema en una base particular del espacio de estados y contiene la información sobre el sistema en dicho instante. Es importante observar que, puesto que $\mathcal{E}$ es un espacio vectorial, este primer postulado implica un principio de superposición, es decir, si los vectores $|\Psi_1\rangle$ y $|\Psi_2\rangle$ representa posibles estados de un sistema cuántico, el vector $|\Psi_3\rangle = \alpha|\Psi_1\rangle + \beta|\Psi_2\rangle$ representa también un posible estado del sistema.

Postulado 2

Una cantidad física medible *A* es descrita por un observable $\hat{A}$ actuando sobre $\mathcal{E}$. Este observable es un operador lineal y satisface una ecuación de autovectores de la forma:

$$\hat{A}|\Psi_n\rangle = a_n|\Psi_n\rangle,$$

en la que los autovalores (a_n) son números reales y las funciones propias $|\Psi_n\rangle$ forman un conjunto ortogonal completo en el espacio $\mathcal{E}$. Los autovalores, pueden tomar valores discretos o puede existir un rango continuo de valores, son reales si el operador correspondiente es hermitiano. Los autovectores $|\Psi_n\rangle$ del operador $\hat{A}$ constituyen un conjunto completo o normalizado, es decir

$$\sum_n |\Psi_n\rangle\langle\Psi_n| = 1$$

Donde el 1 se entiende como el operador identidad, esta ecuación es conocida como relación de clausura.

Postulado 3

El único resultado posible de la medición de una magnitud física *A* es uno de los autovalores del correspondiente observable $\hat{A}$.

Postulado 4

Cuando la cantidad física *A* es medida sobre un sistema en el estado normalizado $|\Psi\rangle$, la probabilidad $P(a_n)$ de obtener el autovalor a_n del correspondiente observable $\hat{A}$ es:

$$P(a_n) = |\langle a_n|\Psi\rangle|^2$$

donde $|a_n\rangle$ es el autovector normalizado de $\hat{A}$ asociado al autovalor a_n. En mecánica clásica cuando un estado esta dado por (*x*, *p*), se puede decir

que si una variable ω es medida en ese estado, el resultado será $\omega(x,p)$. ¿Cuál será el planteamiento análogo en la mecánica cuántica? La respuesta se la puede hacer en base a los postulados anteriores:

1. Se construye el correspondiente operador cuántico $\Omega = \omega\left(x \to \hat{X}, p \to \hat{P}\right)$, donde $\hat{X}$ y $\hat{P}$ son operadores.

2. Se encuentra los autovectores ortonormales $|\omega_i\rangle$ y los autovalores ω_i de Ω.

3. Se expande $|\Psi\rangle$, o sea:

$$|\Psi\rangle = \sum_i |\omega_i\rangle\langle\omega_i|\Psi\rangle$$

4. La probabilidad $P(\omega)$ de que el resultado ω se obtenga es proporcional al cuadrado del módulo de la proyección de $|\Psi\rangle$ sobre el autovector $|\omega\rangle$, o sea $P(\omega) \propto \left|\langle\omega|\Psi\rangle\right|^2$. En términos del operador proyección $\hat{P}_\omega = |\omega\rangle\langle\omega|$,

$$P(\omega)\,\alpha\left|\langle\omega|\Psi\rangle\right|^2 = \langle\Psi|\omega\rangle\langle\omega|\Psi\rangle$$
$$= \langle\Psi|\hat{P}_\omega|\Psi\rangle$$

Cuando se realiza una gran cantidad de medidas de una variable dinámica en un sistema, sus resultados pueden ser diferentes, pero la media o valor esperado de todos los valores observados está dado por:

$$\langle\hat{A}\rangle = \langle\Psi|\hat{A}|\Psi\rangle$$

A la magnitud $\langle\hat{A}\rangle$ se la denomina valor esperado de la variable dinámica $\hat{A}$ en el estado cuántico $|\Psi\rangle$.

Postulado 5

La evolución temporal de un vector estado $|\Psi(t)\rangle$ de un sistema físico está descrita por la ecuación de Schrödinger:

$$i\hbar\frac{d|\Psi(t)\rangle}{dt} = \hat{H}(t)|\Psi(t)\rangle$$

donde $\hbar = h/2\pi$ la constante de Planck racionalizada y $\hat{H} = H\left(x \to \hat{X}, p \to \hat{P}\right)$ es un observable asociado a la energÍa total del sistema en estudio, constituÍdo por tÉrminos de la energÍa cinÉtica y potencial, y H es el Hamiltoniano clásica. Las funciones de onda que son solución de la ecuación de Schrödinger tiene que cumplir con las siguientes propiedades:

- Tiene que ser cuadráticamente integrable.
- Ser finita en todo el rango de definición.
- Continua, ya que estas son la representación del movimiento físico de un sistema.

4. CONCLUSIONES

Los postulados que se acaban de presentar proporcionan un marco formal para el desarrollo e interpretación de la mecánica cuántica; la cual ofrece apenas predicciones probabilísticas para n estado del sistema mediante la función de onda Ψ; contraria a la mecánica clásica que es totalmente determinista. Los observables o variables dinámicas que se puedan medir aparecen como operadores. La medida de un observable es una operación física que proporciona un número real, la medida del observable es un valor propio del operador.

La Mecánica Cuántica representa una de las mayores revoluciones de la Física y propone un cambio radical sobre nuestra concepción de la realidad; como por ejemplo, la combinación lineal de los estados que es la base conceptual que permite construir aplicaciones como la criptografía cuántica y ordenadores cuánticos.

REFERENCIAS BIBLIOGRÁFICAS Y ELECTRÓNICAS

[1]. **ESPINOZA M., IZA P.** (2011) Revista Investigación y Desarrollo, número 18, páginas 11-16.

[2]. **DIRAC P. A. M.** (1958). The Principles of Quantum Mechanics, Oxford University Press.

[3]. **SHANKAR R.** (1994) Principles of Quantum Mechanics, Kluwer Academic/ Plenum Publishers, Second Edition.

[4]. **COHEN-TANNOUDJI C.** (1977) Quantum Mechanics, John Willey & Sons.

[5]. **SAKURAI J.J.** (1994) Modern quantum Mechanics, Addison Wesley Pu. Co. Inc.

matemática: Una publicación del ICM – ESPOL
2012, Vol. 10, No. 1

APLICACIÓN DE ALGORITMOS EVOLUTIVOS A LA BÚSQUEDA DE MOTIVOS BIOLÓGICOS EN REGIONES PROMOTORAS DEL GENOMA

C. I. Jordán [1], C. J. Jordán[2]

Resumen. *El control en la producción de proteínas en las células es un problema importante de la biología molecular, del cual dependen un sinnúmero de aplicaciones en los campos de la medicina, la agricultura y ganadería. Esta producción está regulada al interior de la célula por un interruptor biológico basado en la fijación de una proteína (el factor de transcripción) sobre sitios determinados al interior de los genes portadores de la información genética. Este sitio de fijación se conoce como el motivo de una proteína. Existen métodos en la biología para identificar estos sitios de fijación, pero son procedimientos muy costosos y toman mucho tiempo. Los métodos basados en la computación evolutiva han demostrado ser algoritmos más eficientes y eficaces a la hora de identificar las posiciones del motivo que todo otro método desarrollado hasta la fecha. En este trabajo se implementaron dos métodos propios de búsqueda de motivos, uno basado en los algoritmos genéticos (MBMAG) y otro en la estimación de distribuciones (MBMEDA), los cuales evalúan el contenido de información de los individuos de la población para discriminar las mejores soluciones en cada generación. Los resultados obtenidos sobre bases de ADN fueron evaluados utilizando métricas estándar para medir el desempeño de métodos computacionales de búsquedas. Estos resultados muestran que los métodos evolutivos son superiores respecto a otros métodos conocidos, en cuanto a encontrar un mayor número de secuencias correctas que constituyen el motivo.*

Palabras Clave: *Bioinformática, TFBS, Dogma Central Biología Molecular, Computación Evolutiva, Algoritmo Genético, Algoritmo por Estimación de Distribuciones.*

Abstract. *The control in the production of proteins in cells is an important problem in molecular biology, which depend on a number of applications in the fields of medicine ,agriculture and livestock. This production is regulated within the cell by a biological switch based on binding of a protein on binding sites within certain genes carrying genetic information. These binding sites are known as the motif of a protein. There are methods in biology to identify these binding sites, but are very expensive procedures and time consuming. The methods based on evolutionary computation algorithms have proven to be more efficient and effective in identifying the positions of the sequences that every other method developed so far. In this work we implemented two search methods , one based on genetic algorithms (MBMAG) and the other in the estimation of distributions (MBMEDA), which assess the information content of the individuals in the population to discriminate the best solutions in each generation. The results of these methods in the search for motifs on DNA bases were evaluated using metrics to measure the performance of different search methods. These results demonstrate that evolutionary methods are superior in precision and recall to other methods in the task of finding the correct sequences of a motif.*

Keywords: *Bioinformatics, TFBS, Central Dogma of Molecular Biology, Evolutionary Computation, Genetic Algorithm, Estimation of Distribution Algorithm.*

Recibido: Agosto 2012
Aceptado: Septiembre 2012

1. INTRODUCCIÓN

Todo organismo depende de un número muy grande de proteínas para cumplir sus funciones vitales; se clasifican en dos tipos: proteínas esenciales y no esenciales; a las primeras se las llama así porque el organismo no las produce y requiere ingerirlas mediante la alimentación; las proteínas no esenciales, en cambio, son aquellas que se producen al interior de las células mediante un proceso de biosíntesis conocido como el dogma central de la biología molecular [7].

No obstante su denominación, las proteínas no esenciales son vitales a los organismos; por ejemplo: el colágeno, la insulina, gran parte de las hormonas y un sinnúmero de enzimas, son proteínas no esenciales, de las cuales depende la vida de manera significativa.

El dogma central de la biología molecular explica como la información genética se transcribe y traduce en cadenas de aminoácidos que son las proteínas. Un componente importante de este mecanismo es la transcripción, proceso por el cual la información contenida en los genes, originalmente en forma de molécula de ADN, pasa a una moléculas de ARN mensajero; este proceso depende de una clase de interruptor biológico conocido como el TFBS, por sus siglas en inglés (*transcription factor binding site*), cuyo funcionamiento requiere que una proteína conocida como factor de transcripción (TF) se fije en un

[1] *Jordán Carlos I., Facultad de Ingeniería en Electricidad y Computación. Escuela Superior Politécnica del Litoral (ESPOL); (e_mail: cjjordan@espol.edu.ec).*
[2] *Jordán Carlos J., Facultad de Ingeniería en Electricidad y Computación. Escuela Superior Politécnica del Litoral (ESPOL); (e_mail: cjordan@espol.edu.ec).*

conjunto de patrones de nucleótidos (BS, por binding site) ubicados en la región promotora del gen, en su cabecera. Por lo tanto, estos sitios de fijación deben poder diferenciarse de otras secuencias de nucleótidos, y permitir la fijación del factor de transcripción asociado a la producción de una cierta proteína. Este sitio de fijación recibe también el nombre de motivo de dicha proteína. En el campo de la biología molecular, la identificación de motivos en un genoma constituye uno de los problemas más importantes en la actualidad, debido a los enormes beneficios potenciales que esto tendría; por ejemplo: curar enfermedades de manera más natural, estimulando la producción de ciertas proteínas que un organismo hubiera dejado de producir por motivos hasta ahora desconocidos, y cuyo déficit sea responsable de la patología. La identificación o reconocimiento de motivos biológicos es, sin duda, un verdadero reto; esto debido a que se desconoce a priori cual es el patrón de similitud subyacente a las secuencias de nucleótidos que lo constituyen. Tomando en cuenta que el alfabeto genético está compuesto únicamente por los símbolos A, C, G y T, no es difícil reconocer que hallar diferencias y similitudes entre varias secuencias de nucleótidos constituya un desafío. Por otro lado, se desconoce también cual es la ubicación exacta de los motivos en las regiones promotoras, pues no siempre se encuentran en las mismas posiciones: pudieran ocurrir al inicio, al final o en el centro de la zona de regulación o promotora. En las ciencias biológicas existen métodos confiables y precisos para identificar los TFBS, por ejemplo: el análisis ADN footprint [5] y la electroforesis en gel [6]; sin embargo, estos métodos requieren mucho tiempo y su implementación es muy costosa. Por esta razón, en la actualidad los métodos computacionales han surgido como una alternativa viable para la búsqueda de motivos; los métodos informáticos clásicos pueden clasificarse en dos grupos: aquellos con base en secuencias de nucleótidos y los que se basan en modelos probabilísticos [2]. Los métodos del primer grupo garantizan que encuentran el motivo óptimo; no obstante, sus tiempos de ejecución exponenciales determinan que sólo sean útiles para motivos de tamaño pequeño. Un ejemplo de este grupo es el algoritmo MITRA [16]. Por otro lado, los métodos que utilizan en la búsqueda modelos probabilísticos no siempre encuentran la solución óptima, pero en cambio son más eficientes en cuanto a tiempos de ejecución, y los resultados son generalmente aproximadamente correctos. Ejemplos de estos métodos son los algoritmos MEME [17] y Gibbs Sampler [18]. Entre los métodos computacionales de búsqueda de motivos, aquellos que aplican computación evolutiva han ganado recientemente importancia debido a sus buenos resultados. Los métodos con base en los algoritmos genéticos [1] [13] y por estimación de distribuciones [9] a pesar de no ser muy conocidos estos últimos- presentan los mejores resultados. El principal objetivo de este trabajo es presentar dos métodos de búsqueda de motivos con base en la computación evolutiva. Cada uno se ha implementado utilizando un motor de búsqueda diferente: algoritmos genéticos y algoritmos por estimación de distribuciones. Estos métodos se llamarán en adelante según sus siglas: MBMAG (método de búsqueda de motivos con base en algoritmos genéticos) y MBMEDA (método de búsqueda de motivos basado en estimación de distribuciones), respectivamente. Una vez implementados, se probaron utilizando primero bases de datos sintéticas y luego cierto número de bases reales de ADN correspondientes a diferentes organismos; cada una de estas bases reales consiste de un conjunto de secuencias promotoras del genoma de un organismo donde se sabe que existe por lo menos una instancia del motivo que fija cierto factor de transcripción común. La calidad de los resultados se midió mediante métricas que fueron tomadas del campo de la recuperación de información (IR, por Information Retrieval), a saber: *precisión* y *exhaustividad* [9]. Lo que sigue de este documento tiene la siguiente estructura: en la Sección 2 se hace una hace una breve introducción a la computación evolutiva y a los paradigmas utilizados en el desarrollo de los métodos antes indicados; luego, en la Sección 3, se explican los detalles de la aplicación de estos métodos a la solución del problema concreto de la búsqueda de motivos biológicos; en la sección 4 se indicará que datos fueron utilizados para probar los métodos y que métricas fueron usadas para medir su desempeño; en la sección 5 se presentan los resultados obtenidos y se comparan con los de otros métodos que aparecen en la literatura; en la sección 6 se dan algunas conclusiones; y, finalmente, en la sección 7, se indican maneras en que este trabajo podría extenderse en el futuro.

2. MÉTODOS EVOLUTIVOS

Los métodos evolutivos son métodos metaheurísticos [3] que resuelven problemas haciendo búsquedas globales en un espacio de soluciones potenciales. La búsqueda se hace en base a poblaciones, que son subconjuntos del universo de soluciones potenciales del problema. Los

métodos evolutivos evalúan grupos de soluciones de forma paralela, lo que reduce el tiempo de ejecución del método al descartar simultáneamente grupos de soluciones que no son idóneas al problema, reduciendo el espacio sobre el cual realizar la búsqueda. Los métodos evolutivos son excelentes para resolver problemas con espacios de soluciones de gran cardinalidad. Generalmente, los problemas de optimización presentan este tipo de características, por lo que los métodos evolutivos se utilizan ampliamente en problemas de optimización en una variedad de campos. Un algoritmo evolutivo tiene tres componentes principales:

1. Una población de individuos.
2. Una función para evaluar la calidad de los de los individuos como soluciones.
3. Operadores de variación sobre los individuos.

Cada individuo de una población se representa por una estructura de datos que almacena las características que hacen única a cada solución. Se dice que la representación de cada individuo constituye un "cromosoma", que a su vez agrupa sus "genes".

La función de evaluación o función de fitness es una función que asigna a cada individuo de la población un valor numérico para representar el grado de aptitud como solución correcta del problema. La función de fitness es el criterio principal que conduce el proceso evolutivo hasta que el método converge a una solución. La convergencia del algoritmo genético ocurre cuando el valor de fitness del mejor individuo no mejora después de un cierto número de generaciones. La elección de una función de fitness apropiada depende de las características del problema. La función de fitness podría ser igual a la función objetivo en un problema de optimización, o para problemas complejos que no cuentan con una ecuación definida, existen funciones alternativas que quedan a criterio del investigador para su uso. Los operadores de variación modifican el contenido genético de los individuos de una población generando nuevos individuos para constituir una nueva población. Los individuos sobre los que se ejecutan los operadores de variación se denominan padres, mientras que los individuos resultantes se conocen como hijos. Se espera que individuos de la nueva población tengan mejores valores de función de fitness que de la generación anterior. La mecánica tras un método evolutivo es bastante simple: 1) se genera de forma aleatoria una población inicial P, 2) se evalúan los individuos de la población, 3) se ejecutan los operadores de variación sobre los individuos padres seleccionados de la población y 4) se genera una nueva población en base a operaciones de selección entre los individuos padres y los hijos. Los pasos 2) al 4) se ejecutan de manera iterativa hasta que el método converge, o hasta que se satisface una condición de terminación conveniente. Los algoritmos genéticos son métodos adaptivos de búsqueda sobre el espacio de soluciones del problema. El proceso de adaptación consiste en utilizar operadores de exploración y explotación para modificar los individuos de una población de tal manera que el método converge a la mejor solución del problema. La operación de exploración busca nuevos individuos dentro del espacio de búsqueda; generalmente recibe el nombre de mutación y es un operador unario, es decir, que tiene un solo operando: un individuo de la población actual, al que modifica de manera aleatoria en uno o varios genes. De esta manera se obtiene un individuo nuevo que normalmente no se encuentra en la población actual. La operación de mutación brinda diversidad a la población, evitando la convergencia pronta del algoritmo genético a soluciones óptimas locales. La operación de explotación combina las características genéticas de dos individuos con el propósito de producir nuevos individuos con características mejores que las sus padres tenían. La operación de explotación se conoce como de cruce o recombinación; en general, este operador selecciona de forma aleatoria uno o varios genes de dos individuos padres e intercambia su contenido genético entre ellos, generando otros dos nuevos individuos. Los operadores de mutación y cruce se ejecutan sobre una población en función de tasas de mutación y cruce que toman valores entre en el intervalo [0,1]. Estas tasas determinan la frecuencia con que se ejecutan estos operadores de variación. Los procesos de evaluación y las operaciones de cruce y mutación se ejecutan iterativamente hasta que se cumpla un criterio de parada apropiado, que puede depender del número de generaciones o de la convergencia del algoritmo a una solución. En la siguiente figura se describe mediante un pseudocódigo el funcionamiento básico de un algoritmo genético:

Pseudocódigo 1.
Código del funcionamiento de un algoritmo genético

P <- P ∪ **{**individuo generado al azar**}**
Repetir
para P_i e P **hacer**
FuncionFitness**(**Pi**)**
para childsize **hacer**
Padres P_a, P_b <- Seleccion**(**P**)**
Hijos Ha**,** Hb <- Cruce**(**P_a**,** P_b**)**

Q <- Q U **{**Mutar**(**H_a**),** Mutar**(**H_b**)}**
P <- Q
Best <- Mejor **(**P**)**
Hasta **(**Criterio de Parada**)**
Retornar Best

El algoritmo por estimación de distribución es una variación de un algoritmo genético [3]. A diferencia de este, donde los individuos se generan aplicando los operadores de cruce y mutación, el algoritmo por estimación de distribuciones (EDA) estima un modelo probabilístico a partir de los mejores individuos de la población, y genera los nuevos individuos de la siguiente generación tomando una muestra del modelo. Los métodos basados en la estimación de una distribución (ED) permiten mediante el modelo probabilístico expresar de manera explícita la relación entre las variables del problema, lo que permite encontrar soluciones más eficientes al problema que la mayoría de métodos evolutivos [8]. El siguiente pseudocódigo describe el funcionamiento básico de un algoritmo por estimación de distribución:

Pseudocódigo 2.
Código del funcionamiento de un algoritmo de estimación de distribuciones

P <- P ∪**{**individuo generado al azar**}**
Repetir
para cada individuo Pi e P **hacer**
Fitness**(**P_i**)**
Q <- Seleccionar**(**P**)**
M <- Generar_ModeloP**(**Q**)**
para childsize **hacer**
H <- Muestrear_Ind**(**M**)**
Q <- Q U **{**H**}**
P <- Q
Best <-Mejor**(**P**)** Hasta **(**Criterio de parada**)**
Retornar Best

El algoritmo por estimación de distribuciones genera al azar la primera población de individuos. Luego se evalúa el valor de la función de fitness para cada uno de los individuos de la población, y se eligen los mejores según cierta función de selección. El modelo probabilístico M se estima entonces a partir de los individuos seleccionados de P. La operación de muestreo genera nuevos individuos a partir de M. La nueva población estará conformada por la unión de los individuos padres y los nuevos individuos obtenidos muestreando el modelo. El criterio de parada suele ser similar al presente en los algoritmos genéticos. Finalmente, el algoritmo retorna la mejor solución de la última generación.

3. APLICACIÓN DE LOS MÉTODOS EVOLUTIVOS EN LA BÚSQUEDA DE MOTIVOS

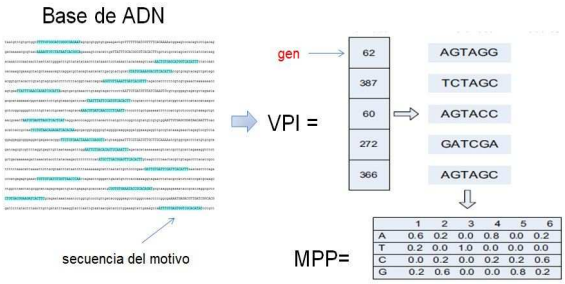
En este trabajo se implementaron dos métodos evolutivos de búsqueda de motivos: MBMAG y MBMEDA, basados respectivamente en los algoritmos genéticos y los algoritmos por estimación de distribución. Ambos métodos tienen elementos comunes, a saber:

1. La representación de los individuos
2. La función de fitness
3. El operador de selección de los mejores individuos

Como se mostró en la sección anterior, la diferencia esencial entre estos métodos estriba en la forma en la que generan los individuos de la siguiente generación. A continuación se muestra los componentes utilizados por ambos métodos evolutivos:

Representación de los individuos
Cada individuo de la población P se representa por un vector posiciones iniciales VPI. Cada celda en el VPI representa un gen, es decir, la posición donde inicia una secuencia del motivo en la fila correspondiente en la base de ADN. El valor que toma cada celda del vector se encuentra dentro del intervalo [0, n - w +1], donde w es el tamaño de las palabras del motivo y n el tamaño de la región promotora en la base de ADN. Esta representación se basa en el modelo de motivos 1-S que supone la presencia de una sola instancia del motivo por cada fila de la base de ADN. El vector de posiciones iniciales (VPI) determina a su vez un vector S cuyos elementos son las instancias del motivo correspondientes a dichas posiciones iniciales. A partir de S se calcula la matriz de pesos posicionales (MPP) $M_{b\ x\ l}$, donde l representa el tamaño de las instancias del motivo y b el número de símbolos del alfabeto genético que tiene cardinalidad 4, por lo tanto b = 4. Los elementos de la matriz representan la frecuencia con que cada símbolo de dicho alfabeto aparece en la posición correspondiente del motivo; de esta manera, S representa un modelo probabilístico de como se distribuyen los nucleótidos en el motivo. La matriz de pesos posicionales es una representación más completa del patrón detrás de las palabras del motivo candidato. La siguiente figura ilustra el proceso de creación de un nuevo individuo a partir de una base de ADN:

FIGURA 1
Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma
Construcción de un VPI a partir de la información seleccionada de la Base de ADN



3.1 FUNCIÓN DE FITNESS

La función de fitness utilizada para ambos métodos es la función de contenido de información [12]. Esta función mide el grado de similitud en la distribución de los nucleótidos del motivo con respecto a la distribución de los nucleótidos en la base de ADN. La utilidad de esta función como función de fitness estriba en el supuesto que los nucleótidos en las secuencias del motivo presentan una distribución diferente a la distribución de los nucleótidos en la Base. Por ello, mientras mayor sea la diferencia entre la distribución de los nucleótidos en el motivo con la base de ADN, mayor será la probabilidad que ese sea el motivo buscado. El contenido de información está definido como:

$$IC = \sum_{i=1}^{L} \sum_{b} \frac{fb(i)\log\left(fb(i)\right)}{pb} \qquad [1]$$

Donde f_b representa la frecuencia del símbolo b en la matriz de pesos posicionales MPP y p_b es la frecuencia del símbolo b en la base de ADN.

3.2 OPERADOR DE SELECCIÓN

El operador de selección utilizado para ambos métodos es la ***selección por torneo***, que consiste en realizar una competencia entre dos o más soluciones escogidas aleatoriamente; producto de esta competencia se escoge el individuo que tenga el mejor valor de fitness.

3.3 OPERADORES DE VARIACIÓN

El método con base en los algoritmos genéticos (MBMAG) utiliza como operadores de variación la mutación y el cruce en un punto, con tasas de mutación y cruce de 0.1 y 0.9, respectivamente. Asignar valores óptimos a estas tasas de mutación y de cruce consiste en sí mismo un problema de optimización; por esta razón los valores escogidos en este trabajo se fueron obtenidos en base a varios experimentos realizados, a partir de los cuales se eligieron las tasas donde los resultados fueron mejores.

El método basado en la estimación de distribuciones (MBMEDA) utiliza cuatro modelos probabilísticos basados en la distribución normal univariada para cada símbolo del alfabeto genético. Para estimar los modelos probabilísticos se utilizaron los estimadores estándares de la media y la varianza. Una vez estimados los parámetros de los modelos probabilísticos, el método MBMEDA genera una nueva población con el operador de muestreo sobre los cuatro modelos probabilísticos. Además de los operadores de variación clásicos, los métodos evolutivos que son objeto de este trabajo requieren de la aplicación posterior de otros operadores para mejorar los resultados obtenidos. Estos operadores son utilizados también por otros métodos descritos en la literatura, y reciben el nombre de operadores de desplazamiento [19] y filtrado local [19]. El operador desplazamiento modifica el contenido de los genes del mejor individuo de la población considerando la posibilidad de que el motivo buscado se encuentre desplazando unas cuantas posiciones del mejor individuo de la población. Este operador se aplica cada 10 generaciones, buscando siempre mejorar el valor de fitness del individuo al encontrar probablemente el individuo más cercano a la solución óptima. La operación de filtrado local modifica el vector de posiciones iniciales de un individuo en base al criterio de similitud entre las instancias del motivo definidas por VPI. Si una palabra es poco similar al resto, quiere decir que se necesita buscar una palabra con mayor similitud dentro de la fila correspondiente en la base de ADN. La operación de filtrado local se ejecuta sobre toda la población cada 10 generaciones.

Las tablas siguientes muestran valores de los parámetros utilizados en cada uno de los métodos de búsqueda implementados.

TABLA I

Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma

Parámetros del método MBMAG

Representación:		Vector Posición Inicial
Tamaño Población:		500
Número de Hijos:		250
Función Fitness:		Contenido de Información
Operador Selección:		Torneo
Operador Reproducción:	**Modelo P.**	Normal Univariado
	Muestreo	Func. Asociada al Modelo
Métodos Adicionales:		Transformación
		Filtrado Local
		Desplazamiento

TABLA II

Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma

Parámetros del método MBMEDA

Representación:		Vector Posición Inicial
Tamaño Población:		500
Número de Hijos:		250
Función Fitness:		Contenido de Información
Operador Selección:		Torneo
Operador Reproducción:	**Mutación**	Un punto(0.1)
	Cruce	1 Punto (0.9)
Métodos Adicionales:		Desplazamiento
		Filtrado Local

El pseudocódigo siguiente describe el funcionamiento del método con base en un algoritmo genético (MBMAG):

Pseudocódigo 3. Código del MBMGA

```
Repetir{
P<- {}
para popsize hacer
P <- P U {individuo generado al azar}
Best < valor inicial repeat
para cada individuo Pi e P hacer
Information_Content(Pi)
si(Best = valor incial O Fitness(Pi) > Fitness(Best))
Best <- Pi
para childsize hacer
Padre Pa <- Seleccionar_Tournament(P)
Padre Pb <- Seleccionar_Tournament(P)
Hijos Ha,Hb <- Cruce_1Punto(Pa,Pb)
Q <- Q U {Mutar_1Pto(Ha),Mutar_1Pto(Hb)}
para (%10)
Filtrado Local(Q)
Desplazamiento(Mejor(Q))
P <- Q
Hasta(condición de terminación) Retornar Best
}Hasta(const)
```

El siguiente pseudocódigo en cambio describe el funcionamiento del método con base en EDA (MBMEDA):

Pseudocódigo 4. Código del MBMEDA

```
Repetir{
P <- {} Pad <-{}
para popsize times hacer
    P <- P U {individuo generado al azar}
Filtrado_Local(P) Best
<- valor frontera repeat
    Para cada individuo Pi e P hacer
        Information_Content(Pi)
        si(Best = valor frontera o Fitness(Pi) >
        Fitness(Best)
        Best <- Pi
    Q <- {}
    Padres Pad <- Seleccionar_Tournament(P)
    M[4] <- Modelo_Normal_Univariado(Pad)
    para childsize hacer
        H <- Muestrear_MNU(M)
        I <- Mapeo(H)
        Q <- Q U {I}
    para (%10) Filtrado_Local(Q)
        Desplazamiento(Mejor(Q))
    P <- Q
  Hasta (condición de terminación)
  Retornar Best
}Hasta(const)
```

La variable const presente en ambos métodos representa el número de ejecuciones del algoritmo evolutivo en la búsqueda del motivo. En cada iteración se elige el mejor individuo de toda la población, representado por la variable Best, y una vez terminada la ejecución del algoritmo evolutivo, se elige al mejor de cada iteración como la solución final al problema. Este es un procedimiento estándar en varios métodos evolutivos de búsqueda de motivos [1][9][15], que tiene el fin de mejorar las posibilidades de encontrar un buen resultado.

4. DATOS Y MÉTRICAS

Los métodos evolutivos desarrollados en este trabajo se probaron utilizando dos tipos de bases de ADN: sintéticas y reales. Las bases sintéticas se utilizaron como casos de prueba elementales para medir el rendimiento de los métodos aun antes de aplicarlos sobre las bases de ADN reales. Las bases sintéticas fueron construidas según el método propuesto en [15], y consisten de un conjunto de t líneas de n nucleótidos generadas aleatoriamente;

luego, en cada línea se sobrescribe una secuencia de *l* nucleótidos generada al azar, que empieza en una posición aleatoria; finalmente, los nucleótidos de cada instancia del patrón sobrescrito sufren una mutación. El grado al que se modifican las instancias del patrón así "sembrado" depende de factores tales como: el tamaño del motivo, la conservación de los nucleótidos y la presencia de más de una instancia por fila en la base de ADN. Los métodos que aquí se presentan fueron probados con doce bases sintéticas generadas con este procedimiento.

Las bases de ADN reales son un conjunto de regiones promotoras de uno o varios genes regulados por el mismo factor de transcripción, para los que se ha determinado experimentalmente la ubicación de las instancias del motivo. La denominación de cada una de estas bases corresponde al factor de transcripción que se fija en los TFBS de estas regiones promotoras. En la siguiente tabla se listan las bases de ADN reales que fueron utilizadas en este trabajo, donde w representa el tamaño del motivo y N_t el número total de instancias que se encuentran en la base de ADN.

TABLA III

Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma

Bases reales de ADN

Base	Número de secuencias (T)	Tamaño de cada Secuencia(bp)	*w*	N_t
CRP	18	105	22	23
ERE	25	200	13	25
E2F	25	200	11	27
MYOD	17	200	6	17
ME2F	17	199	7	21

Con el propósito de medir el rendimiento de los métodos evolutivos se utilizaron dos métricas tomadas del campo de la recuperación de información (IR) y conocidas como: precisión y exhaustividad. La precisión mide la exactitud del método para encontrar las palabras correctas del motivo; por otro lado, la exhaustividad mide la capacidad del método para encontrar el mayor número posible de instancias correctas del motivo. Las métricas de precisión y exhaustividad contestan las siguientes preguntas a partir de la búsqueda en una base de ADN de los motivos: ¿están todos los que son?, para el caso de la precisión, y ¿son todos los que están? Para el caso de la exhaustividad. Estas métricas han sido utilizadas extensamente por otros investigadores [11], lo que facilitará comparar sus resultados con los aquí obtenidos.

Dichas métricas se definen de la manera siguiente:

$$\text{precisión} = N_c/N_p \qquad [2]$$

$$\text{exhaustividad} = N_c/N_t \qquad [3]$$

Donde N_c representa el número de motivos correctos encontrados por el método y N_p representa el número supuesto de palabras del motivo presentes en la base de ADN.

5. RESULTADOS

5.1 RESULTADOS DEL MÉTODO CON BASE EN ALGORITMOS GENÉTICOS (MBMAG)

Para probar el desempeño del MBMAG sobre las bases sintéticas se generaron 12 bases aplicando el método propuesto por [15]. Los resultados obtenidos para las métricas precisión y exhaustividad se encuentran en la siguiente tabla:

TABLA IV

Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma

Resultados del método MBMAG sobre las bases sintéticas

Número de Secuencias	Tamaño motivo	Conservación	Ruido			
			Sin Ruido		Con Ruido	
			Precisión	Exh.	Precisión	Exh.
100	16	1	1	1	0.99	0.97
20	16	1	0.99	0.99	0.98	0.97
100	8	1	0.99	0.99	0.97	0.93
20	8	1	0.98	0.98	0.91	0.89
100	16	0	0.95	0.95	0.88	0.80
20	16	0	0.94	0.94	0.89	0.85

En el caso de las bases sintéticas, el parámetro conservación es una medida de la resistencia a la mutación en las palabras del motivo en las diferentes filas de base. Una conservación de 1 representa que los nucleótidos en las palabras del motivo se conservan con una tasa del 90%, mientras que una conservación de 0 significa una tasa de menos del 50% de cohesión en las palabras del motivo.

Los resultados obtenidos demuestran que el método MBMAG encuentra el mayor número correcto de instancias del motivo. En base a estos resultados, se aplica la búsqueda de motivos sobre 5 bases reales de regiones promotoras de ADN. Los resultados se muestran a continuación:

TABLA V

Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma

Resultados del MBMAG sobre las bases reales

Base	Precision	Exhaustividad
CRP	0.88	0.69
ERE	0.76	0.76
E2F	0.76	0.70
MYOD	0.94	0.76
ME2F	0.94	0.94

TABLA VI

Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma

Comparación del desempeño entre el método **MBMAG con GAME Y GALF**

				MBMAG		GAME		GALF	
Base	**l**	**T**	**N_t**	**Pr.**	**Ex.**	**Pr.**	**Ex.**	**Pr.**	**Ex.**
CRP	22	18	23	0.88	0.69	0.94	0.70	**0.94**	**0.74**
ERE	9	25	25	0.76	0.76	0.73	0.76	**0.84**	**0.84**
E2F	11	25	27	0.76	0.70	**0.96**	**0.85**	0.80	0.74
MYOD	6	17	21	**0.94**	**0.76**	0.48	0.48	0.88	0.71
ME2F	9	17	17	0.94	0.94	0.88	0.88	**1.00**	**1.00**

En la Tabla VI se puede observar que de las cinco bases reales de ADN, en cuatro los métodos GAME y GALF son más precisos y exhaustivos que MBMAG, lo que se debe principalmente a que ambos métodos realizan una vez terminado el proceso evolutivo un procesamiento posterior sobre las instancias del motivo obtenido, lo que les permite reconocer un mayor número de palabras del patrón buscado. A pesar de que MBMAG no cuenta con este post-procesamiento, en promedio la diferencia entre sus valores de precisión y exhaustividad con los de los otros métodos es 0.05; lo que significa que los resultados obtenidos por el método MBMAG son suficientemente buenos para ser tomados en consideración.

5.2 RESULTADOS DEL MÉTODO BASADO EN LA ESTIMACIÓN DE DISTRIBUCIONES (MBMEDA)

Luego se procedió a aplicar el método MBMEDA a las bases de datos sintéticas como reales de ADN, midiendo en cada caso los valores de precisión y exhaustividad; los resultados obtenidos para las primeras se tabulan en la Tabla 7 siguiente:

TABLA VII

Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma

Resultados del método MBMEDA sobre las bases sintéticas

Número de Secuencias	Tamaño motivo	Conservación	Ruido			
			Sin Ruido		Con Ruido	
			Pr. (%)	Ex. (%)	Pr. (%)	Ex. (%)
100	16	1	1	1	0.99	0.97
20	16	1	0.99	0.99	0.98	0.95
100	8	1	1	1	0.99	0.93
20	8	1	0.99	0.99	0.98	0.92
100	16	0	0.97	0.97	0.84	0.79
20	16	0	0.95	0.95	0.93	0.86

En la Tabla VII se observa que los resultados obtenidos con este método sobre las bases sintéticas son superiores a los del método MBMAG, lo cual constituyó un estimulo para aplicar dicho método a las bases de ADN reales. De hecho, en la siguiente tabla se muestran los resultados obtenidos con las cinco bases de ADN anteriores:

TABLA VIII

Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma

Resultados del MBMEDA sobre las bases reales

Base	Precisión	Exhaustividad
CRP	0.83	0.65
ERE	0.80	0.80
E2F	0.80	0.74
MYOD	1.00	0.80
ME2F	1.00	1.00

Al comparar los valores con los de la Tabla V se observa que el método MBMEDA obtiene mejores resultados que el MBMAG cuando se buscan motivos reales con longitudes menores a 10 bps; prueba de ello son los resultados obtenidos para las bases MYOD y ME2F, en donde la precisión del método MBMEDA es 1 y su exhaustividad mayor a

0.8. Al comparar el desempeño del MBMEDA con el otro método de búsqueda de motivos con base en la estimación de distribuciones conocido como EDAMD [9] se obtienen los siguientes resultados:

TABLA IX

Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma

Comparación del desempeño del método MBMEDA con EDAMD

	MBMEDAM		EDAMD	
Base	**Pr.**	**Ex.**	**Pr.**	**Ex.**
CRP	0.83	0.65	**0.94**	**0.74**
ERE	**0.80**	**0.80**	0.76	0.76
E2F	**0.80**	0.74	0.71	**0.80**
MYOD	**1.00**	**0.80**	0.86	0.9
ME2F	1.00	1.00	1.00	1.00

En la Tabla IX se observa que el método MBMEDA presenta mayor precisión en los resultados que EDAMD; sin embargo, este último método presenta mejor exhaustividad, es decir, encuentra un mayor número de patrones correctos que el método implementado en este trabajo. Esto se debe a que el método EDAMD utiliza un modelo multivariado para estimar la distribución probabilística de los individuos de la población, lo que permite tomar en cuenta las interrelaciones entre las variables del problema.

5.3 COMPARACIÓN DEL DESEMPEÑO ENTRE LOS MÉTODOS EVOLUTIVOS Y OTROS MÉTODOS DE BÚSQUEDA

La Tabla X permite comparar los resultados obtenidos al aplicar los dos métodos de computación evolutiva desarrollados en este trabajo a las bases de ADN reales:

TABLA X

Aplicación de algoritmos evolutivos a la búsqueda de motivos ***biológicos en regiones promotoras del genoma***

Comparación entre MBMEDA y MBMA

	MBMEDA		MBMAG	
Base	**Pr.**	**Ex.**	**Pr.**	**Ex.**
CRP	0.83	0.65	**0.88**	**0.69**
E2F	**0.80**	**0.74**	0.76	0.70
ERE	**0.8**	**0.8**	0.76	0.76
ME2F	**1.00**	**1.00**	0.94	0.94
MYOD	**1.00**	**0.80**	0.94	0.76

Como se puede observar en la tabla anterior, en cuatro de las cinco bases reales de ADN, el método MBMEDA presenta mejores valores para la precisión y la exhaustividad en la búsqueda de motivos que el método MBMAG; esto se debe a que el desempeño de este último depende de que se utilicen valores apropiados para las tasas de mutación y de cruce; dependencia que no presenta el MBMEDA, que trabaja sobre un modelo más natural a la distribución de las variables del problema, lo que permite que las soluciones obtenidas por este método sean más próximas a las correctas que el MBMAG.

Finalmente, la Tabla XI permite comparar el desempeño de los métodos aquí desarrollados con los de dos métodos estadísticos que resuelven el mismo problema: MEME [17] y BioProspector; los métodos estadísticos están entre aquellos que mejores resultados obtienen en la búsqueda de motivos.

TABLA XI

Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma

Comparación del desempeño de los métodos desarrollados con los algoritmos MEME y Bioprospector

	MBMEDA		**MBMAG**		**MEME**		**BioProspector**	
Base	**Pr.**	**Ex.**	**Pr.**	**Ex.**	**Pr.**	**Ex.**	**Pr.**	**Ex.**
CRP	0.83	0.65	0.88	**0.69**	0.92	0.52	**1.00**	0.35
E2F	**0.80**	**0.74**	0.76	0.70	0.80	0.70	0.52	0.41
ERE	0.80	**0.80**	0.76	0.76	**0.88**	0.60	0.46	0.56
ME2F	**1.00**	**1.00**	0.94	0.94	0.93	0.82	0.71	0.71
MYOD	**1**	**0.80**	0.94	0.76	0.00	0.00	0.00	0.00

Con las bases CRP y ERE, tanto MEME como Bioprospector obtienen valores para Np y NT menores que los de métodos evolutivos. Esta es la razón por la que presentan mejores resultados en precisión y exhaustividad que los métodos evolutivos desarrollados. Para las demás bases reales, donde ambas variables tienen los mismos valores, los métodos evolutivos son más precisos y exhaustivos que los métodos estadísticos.

Por último, se muestra en las siguientes figuras representaciones gráficas de los motivos mediante los logos de secuencias [13]; en la Figura 2 se observa el logo para el motivo que corresponde al factor de transcripción CRP obtenido de manera experimental, mientras que en las figuras 3 y 4 se muestran los logos para los logos encontrados por los métodos aquí desarrollados: MBMAG y MBMEDA, respectivamente.

FIGURA 2
Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma
Logo de las secuencias reales donde se fija el factor de transcripción CRP

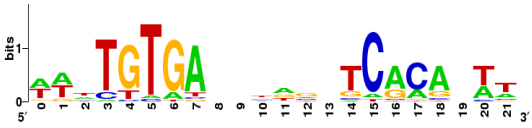


FIGURA 3
Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma
Logo de las secuencias encontradas por el método basado en AG

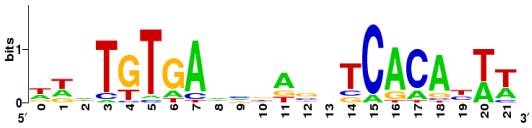
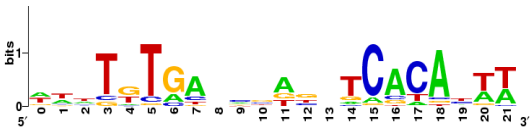


FIGURA 4
Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma
Logo de secuencias encontradas por el método basado en ED



Una comparación visual de estos logos permite concluir que el logo del método basado en AG es más similar al logo del método basado en ED para el caso de la base CRP. Este criterio concuerda con los valores de precisión y exhaustividad de ambos algoritmos para esta base real de ADN.

6. CONCLUSIONES

En este trabajo se han presentado dos métodos evolutivos para resolver el problema de encontrar motivos biológicos en una base de ADN. Se ha denominado a tales métodos MBMAG y MBMEDA, según el algoritmo que utilizan como motor de búsqueda: algoritmos genéticos y algoritmo por estimación de distribuciones, respectivamente.

Según las métricas utilizadas: precisión y exhaustividad, estos métodos son en general tan precisos y completos como otros métodos computacionales que resuelven el mismo problema, y en algunos casos mejores. Entre los dos, el que obtienen mejores resultados es aquel que utiliza estimación de distribuciones para generar la próxima población de soluciones potenciales, especialmente cuando la longitud del motivo buscado es pequeña, que es cuando son mas difíciles de reconocer. A fin de evitar la pronta convergencia de estos métodos evolutivos a mínimos locales, fue necesario introducir al proceso de búsqueda corrección y procesamiento posterior mediante los operadores desplazamiento y filtrado local, en adición a los ya clásicos de variación.

Los métodos aquí presentados utilizaron un modelo 1-S que limita a 1 el número de instancias del motivo por secuencia promotora, restricción que no está presente en las bases de ADN reales. No obstante, los resultados obtenidos fueron buenos si se los compara con los de otros métodos que no aplican tal restricción, pues en promedio la diferencia entre métricas correspondientes es inferior al 5%. Esta pequeña perdida en la calidad de los resultados se ve recompensada con el aumento en eficiencia en cuanto a tiempos de ejecución, debido a la mayor rapidez de los métodos aquí propuestos.

Finalmente, no existe una relación directa entre la calidad de los resultados obtenidos al probar un método con bases sintéticas o artificiales con aquella que se obtiene al utilizar bases reales de ADN; esto, debido a que el procedimiento aquí utilizado [15] para generar tales bases es completamente aleatorio, en tanto que se desconoce el método utilizado por la naturaleza en el diseño de motivos biológicos. Una muestra de ello ocurrió con motivos de gran tamaño; a pesar de que el método MBMEDA obtuvo mejores resultados que el MBMAG sobre bases de datos sintéticas, sobre la base CRP que tiene un motivo de 23 bps -considerado de gran tamaño- se obtuvieron mejores resultados con el método basado en los algoritmos genéticos.

7. TRABAJOS FUTUROS

El diseño de algoritmos de búsqueda de motivos biológicos con base en la computación evolutiva abre muchos frentes de investigación y desarrollo. Para el caso de los algoritmos aquí presentados debe estudiarse la posibilidad de trabajar con modelo M-S, que incluyen la posibilidad de que ocurra mas de 1 motivo por región promotora, con lo que podría mejorar significativamente los resultados para la métrica exhaustividad.

Este trabajo podría extenderse en el caso del método MBMEDA utilizando otros modelos probabilísticos para generar próximas poblaciones; por ejemplo, que tomen en cuenta de manera explícita la dependencia entre variables del

problema, que aquí fueron consideradas como independientes.

Finalmente, también existe la posibilidad de mejorar el desempeño de los métodos aquí propuestos, incorporando a la búsqueda nuevos operadores de variación y procesamiento posterior, para evitar la tendencia casi natural de estos procedimientos a converger a mínimos locales, lo que da lugar a encontrar muchos falsos positivos.

REFERENCIAS BIBLIOGRÁFICAS Y ELECTRÓNICAS

[1]. **CHAN, TALK MING, KWONG SAK LEUNG, AND KIN HONG LEE.** (2008). "*TFBS Identification Basen on Genetic Algorithm with Combined Representations and Adaptive Post-processing*". *Bioinformatics* 24.3 (2008): 341-49.

[2]. **DAS MK, DAI HK.** (2007). "*A survey of DNA motif finding algorithms*". *BMC Bioinformatics*. Nov 1;8 Suppl. 7:S21.

[3]. **EIBEN, AGOSTON E., AND J. E. SMITH.** (2003). "*¿What is an Evolutionary Algorithm?*" Introduction to Evolutionary Computing. New York: Springer, 2003. 15-35

[4]. **EIBEN, AGOSTON E., AND J. E. SMITH.** (2003). "*Genetic Algorithms*". Introduction to Evolutionary Computing. New York: Springer, 2003. 37-69

[5]. **GALAS, DAVID J., AND ALBERT SCHMITZ.** (1978). "*DNAase Footprinting a Simple Method for the Detection of Protein-DNA Binding Specificity*". Nucleic Acids Research 5.9 (1978): 3157-170.

[6]. **GARNER, MARK M., AND ARNOLD REVZIN.** (1981). "*A Gel Electrophoresis Method for Quantifying the Binding of Proteins to Specific DNA Regions: Application to Components of the Escherichia Coli Lactose Operon Regulatory System*" Nucleic Acids Research 9.13: 3047-060.

[7]. **JONES, NEIL C., AND PAVEL PEVZNER.** (2004). "*Molecular Biology Primer*" An Introduction to Bioinformatics Algorithms. Cambridge, MA: MIT, 2004. 57-68.

[8]. **LARRAÑAGA, PEDRO, AND JOSÉ A. LOZANO.** (2002). "*Estimation of Distribution Algorithms: a New Tool for Evolutionary Computation*". Boston: Kluwer Academic.

[9]. **LI, GANG, TAK MING CHAN, KWONG SAK LEUNG, AND KIN HONG.** (2008). "*An Estimation of Distribution Algorithm for Motif Discovery*". *Evolutionary Computation* (2008): 2411-418. .

[10]. **LUKE, SEAN.** (2009). "*Essencials of Metaheuristics*". 1st ed. Washington: Lulu.

[11]. **MANNING, CHRISTOPHER D., PRABHAKAR RAGHAVAN, AND HINRICH SCHUTZE.** (2008). "*Introduction to Information Retrieval*". New York: Cambridge UP. 151-158.

[12]. **SCHNEIDER, T., G. STORMO, L. GOLD, AND A. EHRENFEUCHT.** "Information Content of Binding Sites on Nucleotide Sequences" *Journal of Molecular Biology* 188.3 (1986): 415-31.

[13]. **SCHNEIDER, THOMAS D., AND R.MICHAEL STEPHENS.** (1990). "*Sequence Logos: a New Way to Display Consensus Sequences*". *Nucleic Acids Research* 18.20: 6097-100.

[14]. **USSERY, DAVID W., TRUDY M. WASSENAAR, AND STEFANO BORINI.** (2009). "*Sequences as Biological Information: Cells Obey the Laws of Chemistry and Physics*" *Computing for Comparative Microbial Genomics Bioinformatics for Microbiologists.* London: Springer. 3-17.

[15]. **WEI, Z.** (2006). "*GAME: Detecting Cis-regulatory Elements Using a Genetic Algorithm*". *Bioinformatics* 22.13: 1577-584.

[16]. **ESKIN, ELEAZAR, AND PAVEL A. PEVZNER.** (2002). "*Finding Composite Regulatory Patterns in DNA Sequences*". *Bioinformatics* 18 (2002): 354-63.

[17]. **BAILEY, TIMOTHY L., AND CHARLES ELKAN.** (1993). "*Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization*". La Jolla, CA: Dept. of Computer Science and Engineering, University of California, San Diego.

[18]. LAWRENCE, C., S. ALTSCHUL, M. BOGUSKI, J. (1993). "*Detecting Subtle Sequence Signals: a Gibbs Sampling Strategy for Multiple Alignment*". *Science*262.5131: 208-213

[19]. MANNING, CHRISTOPHER D., PRABHAKAR RAGHAVAN, AND HINRICH SCHUTZE. (2000). "*Introduction to Information Retrieval*" New York: Cambridge UP.

matemática: Una publicación del ICM – ESPOL
2012, Vol. 10, No. 1

DISCRETIZING THE HOPF–HOPF BIFURCATION

Páez Joseph[1]

Abstract. *In this presentation we study the one-step discretization of ODEs with a Hopf–Hopf Bifurcation. We analyze how the local bifurcation diagram is perturbed by the one-step methods. The numerical approximation of the critical eigenvalues is also discussed. We interpret the obtained results in terms of perturbation of branching points. We illustrate the main results by means of numerical experiments.*

Key–Words. bifurcation problems, ordinary differential equations, one-step methods.

Resumen. *En este artículo se estudia la discretización de un paso de EDOs con una bifurcación Hopf-Hopf. Nosotros analizamos como el diagrama de bifurcación local es perturbado por los métodos de un paso. Se discute también la aproximación numérica de los valores propios críticos. Los resultados son interpretados en términos de perturbación de puntos de ramificación. Los resultados principales se ilustran a traves de experimentos numéricos.*

Palabras claves: Problemas de bifurcación, ecuaciones diferenciales ordinarias, método 1-paso.

RECIBIDO: Marzo 2012
ACEPTADO: Junio 2012

1. INTRODUCTION

Consider the one-step method with step-size h

$$x_n = \psi^h\left(x_{n-1},\alpha\right), \quad n \in \mathbb{N} \qquad \textbf{(1)}$$

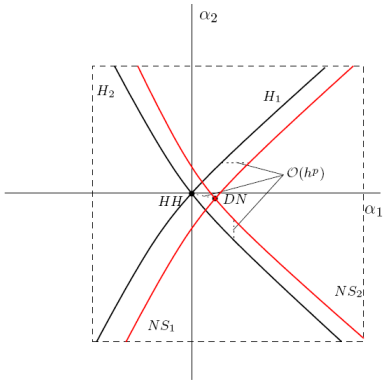
which approximates the evolution operator of the parametrized family of continuous-time dynamical systems

$$\dot{x}(t) = f\left(x(t),\alpha\right), \qquad \textbf{(2)}$$

where $f \in C^k\left(\mathbb{R}^N\times\mathbb{R}^2,\mathbb{R}^N\right)$ and $k \geq 1$ is sufficiently large. In many practical applications the only way to explore the dynamics of (2) is via numerical methods. For this purpose, we can use the numericaltime integration given by (1). This approach is often referred to as the *indirect method*, cf. [2]. Thus, it is important to investigate whether, and to what extent, the numerical data produced by (1) accurately represents the dynamics of (2). This analysis becomes more involved if the system undergoes bifurcations under variation of parameters.
In this work we assume that (2) has a Hopf–Hopf bifurcation, which occurs when the linearization of (2) about an equilibrium has two pairs of pure imaginary eigenvalues. Furthermore, we suppose that (2) is discretized via general one-step methods of order $p \geq 1$, see Section 2.

The results we want to show are illustrated in Figure 1. In this picture, the curves labeled $H_{1,2}$ and $NS_{1,2}$ represent paths of Hopf points of (2) and Neimark–Sacker points of (1), respectively. The labels *HH* and *DN* stand for a Hopf–Hopf point of (2) and for the resulting double Neimark–Sacker point of the one-step method, respectively.

FIGURE 1
Discretizing the hopf–hopf bifurcation
Discretization of the Hopf curves near an HH bifurcation



2. BASIC SETUP

In this presentation we discretize (2) via the one-step map

$$\psi^h(x,\alpha) := x + h\,\Phi\left(h,x,\alpha\right), \quad (3)$$

with $\Phi : \left[-h^*,h^*\right]\times\widetilde{\Omega}\times\widetilde{\Lambda} \to \mathbb{R}^N$ sufficiently smooth, $h^* > 0$, where $\widetilde{\Omega} \subset \mathbb{R}^N$ and $\widetilde{\Lambda} \subset \mathbb{R}^2$ are compact sets. We say that (3) is of order $p \geq 1$ if there exists a positive constant K (depending only on f) such that

$$\left\|\varphi^h(x,\alpha) - \psi^h(x,\alpha)\right\| \leq K\,|h|^{|p+1|}$$

holds for all $\left(h,x,\alpha\right) \in \left[-h^*,h^*\right]\times\widetilde{\Omega}\times\widetilde{\Lambda}$, where $\varphi^h(\cdot,\alpha)$ represents the t-flow of (2). It can be shown

[1]*Joseph Páez Chávez, Ph.D, Profesor del Instituto de Ciencias Matemáticas, ESPOL. (e_mail: jpaez@espol.edu.ec)*

that there exist smooth functions $\Upsilon, \Xi : [-h_0, h_0] \times \widetilde{\Omega} \times \widetilde{\Lambda} \to \mathbb{R}^N$ such that:

$$\begin{aligned} \psi^h(x,\alpha) &= \varphi^h(x,\alpha) + \Upsilon(h,x,\alpha)\, h^{p+1} \text{ and} \\ \Phi(h,x,\alpha) &= f(x,\alpha) + \Xi(h,x,\alpha)\, h \end{aligned} \tag{4}$$

hold for all $(h,x,\alpha) \in [-h_0, h_0] \times \widetilde{\Omega} \times \widetilde{\Lambda}$, where $0 < h_0 < h^*$, see [4].

An HH point lies generically at the transversal intersection of two curves of Hopf points. The presence of such a point in (2) may produce a very complex system response. Depending on the values of the normal form coefficients (cf. [3, Section 8.6.2]), the system can have invariant tori and chaotic dynamics, as well as Neimark–Sacker bifurcations of cycles and Shil'nikov homoclinic bifurcations. Our main concern in this presentation is to analyze the effect of onestep methods on an HH point and on the intersecting Hopf curves.

A generic HH bifurcation is a regular zero of the real form of the system (cf. [1])

$$\begin{cases} f(x,\alpha) = 0, \\ f_x(x,\alpha)\phi_1 - i w_1 \phi_1 = 0, \\ \langle l_1, \phi_1 \rangle - i = 0, \\ f_x(x,\alpha)\phi_2 - i w_2 \phi_2 = 0, \\ \langle l_1, \phi_1 \rangle - i = 0, \end{cases} \tag{5}$$

where $\phi_{1,2}$ are eigenvectors corresponding to the critical eigenvalues $i w_{1,2}$ and $l_{1,2}$ are suitably chosen normalizing vectors. In this system we use the standard inner product $\langle p,q \rangle := \langle p,q \rangle_{\mathbb{C}^N} = \bar{p}^T q$.

In a similar way, a double Neimark–Sacker point (the "discrete version" of an HH point) of the onestep map (3) can be seen as a solution of the following complex system

$$\begin{cases} \frac{1}{h}\left(\psi^h(x,\alpha) - x\right) = 0, \\ \frac{1}{h}\left(\psi_x^h(x,\alpha)\phi_1 - e^{ihw_1}\phi_1\right) = 0, \\ \langle l_1, \phi_1 \rangle - i = 0, \\ \frac{1}{h}\left(\psi_x^h(x,\alpha)\phi_2 - e^{ihw_2}\phi_2\right) = 0, \\ \langle l_2, \phi_2 \rangle - i = 0, \end{cases} \tag{6}$$

Define $z := \left(\mathrm{Re}(\phi_{1,2}), \mathrm{Im}(\phi_{1,2}), w_{1,2}\right) \in \mathbb{R}^{4N+2}$ and write the systems (5) and (6) as $F(x,\alpha,z) = 0$ and $G(x,\alpha,z) = 0$, respectively. By (4), we can show that

$$G(x,\alpha,z) = F(x,\alpha,z) + \mathrm{O}(h)$$

holds for all $(h,x,\alpha,z) \in [-h_0, h_0] \times \widetilde{\Omega} \times \widetilde{\Lambda} \times \mathbb{R}^{4N+2}$. This relation, combined with the implicit function theorem, allows us to prove that an HH point is turned into a DN point by the one-step methods.

Furthermore, the DN point of the one-step map varies smoothly with the step-size and remains close to the original HH point, provided the step-size is sufficiently small. This fact can also be discussed in the context of perturbation of simple branching points. According to [3, Lemma 10.3], a generic HH bifurcation of (2) is a simple branching point (in short SB point) of the system

$$Q(x,\alpha) := \begin{pmatrix} f(x,\alpha) \\ \det\left(2 f_x(x,\alpha) \odot I_N\right) \end{pmatrix} = 0 \tag{7}$$

Here, the symbol $\odot$ stands for the bialternate product of matrices (cf. [3, Section 10.7]). On the other hand, a DN bifurcation of (3) is an SB point of the system

$$P(h,x,\alpha) := \begin{pmatrix} \frac{1}{h}\left(\psi^h(x,\alpha) - x\right) \\ \det\left(\frac{1}{h}\left(\psi_x^h(x,\alpha) \odot \psi_x^h(x,\alpha) - I_m\right)\right) \end{pmatrix} = 0$$

where $m := \frac{1}{2} N (N-1)$. The functions Q and P defined above satisfy locally the relation

$$P(h,x,\alpha) = Q(x,\alpha) + \mathrm{O}(h).$$

This means that the SB point of (7) is stable under smooth $\mathrm{O}(h)$-perturbations produced by the one-step methods. It is important to point out that we have not assumed any special structure of the vector field (2), so the fact described above is not a consequence of the preservation of any symmetry but of the transversal intersection of the emanating Hopf curves.

REFERENCES AND ELECTRONIC

[1]. AMDJADI, F. The Calculation of the Hopf/Hopf Mode Interaction Point in Problems with Z2-Symmetry. *Internat. J. of Bif. and Chaos 12*, 8 (2002), 1925–1935.

[2]. BEYN, W.-J. Numerical methods for dynamical systems. In *Advances in Numerical Analysis*, Oxford Sci. Publ., Ed., vol. I. Oxford University Press, New York, 1991, pp. 175–236.

[3]. KUZNETSOV, Y. A. *Elements of Applied Bifurcation Theory*, third ed., vol. 112 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.

[4]. STUART, A., AND HUMPHRIES, A. R. *Dynamical Systems and Numerical Analysis*. Cambridge University Press, New York, 1998.

matemática: Una publicación del ICM – ESPOL
2012, Vol. 10, No. 1

ASYMPTOTIC DISTRIBUTION THEORY FOR CONTAMINATION MODELS

Vera Francisco[1], Dickey David[2] & Lynch James[3]

Abstract. *In many situations one is interested in identifying observations that come from sources of variation other than the normal background or baseline source. A simple model for such situations is a two point mixture model where one component in the mixture corresponds to the baseline model and the second to the other sources (the contamination component). Here the goal is two-fold: (i) detect the overall presence of Contamination and (ii) identify observations that may be contaminated. A locally most powerful test is presented which gives some insights on how to accomplish this. Surprisingly, the test statistic can have an asymptotic distribution that is based on a stable law that is not the normal distribution. Examples and simulations are given to illustrate the approach.*

Keywords: multiple testing, anomaly detection, stable law, false discovery rate AMS 2000 Subject Classification: Primary 62F03, 62-07 Secondary 62E17, 62F30.

Resumen. *En muchas situaciones se tiene interés en identificar las observaciones que provienen de fuentes de variación distintas de la normal de base o de la fuente de referencia. Un modelo simple para tales situaciones es un modelo de mezcla de dos puntos, donde uno de los componentes en la mezcla corresponde al modelo de línea de base y la segunda a los de otras fuentes (el componente de la contaminación). Aquí el objetivo es doble: (i) detectar la presencia global de la Contaminación y (ii) identificar las observaciones que puedan estar contaminados. Una prueba localmente más poderosa se presenta la cual da algunas ideas sobre cómo lograr el objetivo. Sorprendentemente, la estadística de prueba puede tener una distribución asintótica que se basa en una ley estable que no es la distribución normal. Ejemplos y simulaciones se dan para ilustrar el enfoque.*

Key words: múltiples pruebas, detección de anomalías, ley estable, tasa de falso descubrimiento.

Recibido: Mayo, 2012
Aceptado: Julio, 2012

1. INTRODUCTION

The point of this paper is to present a contamination detection method based on the two point mixture model

$$f_p(x) \equiv \bar{p} f_0(x) + p f_1(x), \text{ where } \bar{p} = 1 - p, \text{ and } 0 \le p \le 1 \quad \textbf{(1.1)}$$

and to investigate the asymptotic distribution of the maximum likelihood estimator (MLE) and the locally most powerful (LMP) test for the parameter *p*. The distribution, f_p, is the so called contaminated distribution model which is sometimes used to model outliers from the baseline model f_0. In this simple setting we shall see that, when $p = 0$, the MLE and the LMP test have asymptotic distributions that are non-standard. They exhibit the Chernoff phenomena (Chernoff, 1954) of being two point mixtures. These two points mixtures each have point masses at zero where the second component in the mixture is based on an α-stable law depending on the tail behavior of the likelihood ratio $f_1(X) / f_0(X)$ under f_0.

In low contaminated situations $(p \approx 0)$, these asymptotics suggest using the LMP test to detect the presence of contamination. If the LMP test rejects $p = 0$, then we can use the empirical posterior $\frac{p^* f_1(x)}{f_{p^*}(x)}$, where p^* is the mle to investigate what observations may be contaminated (from f_1). Confidence bounds for this posterior can also be constructed using confidence intervals for p^*. The LMP for $p = 0$ suggests using the ratio $\frac{f_1(x)}{f_0(x)}$ to identify observations from f_1.

The asymptotics indicate that the determination of contamination when *p* is small can be problematic using classical frequentist approaches, especially if parameters need to be estimated. In addition, this has similar implications for multiple testing problems. E.g., in the analysis of microarrays, a mixture model f_0 is the model for the expression levels of the nonexpressed genes and f_1 for the differentially expressed genes. In particular, there can be a justification for the use of a central *t* distribution where the degrees of freedom is determined by the amount of replication in the

[1] *Vera Francisco, Ph. D., Profesor de la Escuela Superior Politécnica del Litoral (ESPOL). (e_mail: fvera@espol.edu.ec).*
[2] *Dickey David,Ph.D., University of South Carolina.*
[3] *Lynch James, Ph.D.,North Carolina State University.*

experiment or a central normal if the degrees of freedom is large. A similar justification can be used to use noncentral *t*'s or noncentral normals to model the differentially expressed genes. Here *p* is the proportion of differentially expressed genes.
Next section shows the data analytic model for detecting contamination, while Section 3 introduces the LMP test for *p*. Section 4 considers the asymptotic distribution of the MLE for p and the test statistic for the LMP, along with the tail behaviors of the terms in the LMP test statistic for the normal and exponential distributions.

2. POOLING AND MIXTURES

In many data analytic problems the observations $X_1,...,X_n$ arise from pooling data from various sources of variation. In many cases, the pooling model has the following formulation for two sources of variation. In this formulation, a configuration C which is a subset of $\{1,2,...,n\}$ indicates which observations come from one source and C^c from the other. For example, such a pooling model might occur in a binary network where the network is modeled by a Markov random field. In the spread of an infectious disease over the network, the nodes are partitioned into two groups, C and C^c, where C^c is the Collection of sites that have elevated levels of infections and C is the Collection of sites which are normal. In the normal case the number of infections is governed by f_0 while for the elevated level by f_1. Then,

$$p\left(\left(C,C^c\right),X_1,...,X_n\right)$$
$$= K\exp\left\{E\left(\left(C,C^c\right)\right)\right\}\prod_{i\in C} f_0\left(X_i\right)\prod_{i\in C^c} f_1\left(X_i\right)$$

where $E\left(\left(C,C^c\right)\right)$ is related to the energy of the partition $\left(C,C^c\right)$ (Huang, 1963) and where we have suppressed parameters in $E\left(\left(C,C^c\right)\right)$ and the normalizing constant *K*. Here we have assumed the positivity condition that all partitions have positive probabilities. In general, the pooling model is given as follows.

- Generate a configuration C with probability $p(C)$
- Given C, for $i\in C, X_i$ are iid $\sim f_0$ and, for $i\in C^c, X_i$ are iid $\sim f_1$

  - C and C^c model a spatial or temporal (e.g., a change-point) pattern
  - You are "pooling" observations based on the configuration C where the configuration C is a hidden variable
  - The likelihood is then

$$\sum_C p(C)\prod_{i\in C} f_0\left(X_i\right)\prod_{i\in C^c} f_1\left(X_i\right)$$

Throughout we assume that all densities f are absolutely continuous with respect to a common measure m and absolutely continuous with respect to one another. The basic data analytic method is as follows:

- Envision that the data are the effects of pooling observations from f_0 and f_1 where f_0 is the background distribution and f_1 is the distribution of the contaminated observations.
- Treat the data as if it is from a mixture model and use a mixture model to estimate the mixing proportions for f_0 and f_1, that is, the proportions in C and C^c. Use the estimates to test the null hypothesis that one of the mixing proportions is equal to zero. If this hypothesis is rejected, see if the fitted mixture model can give insights into which observations came from f_0, that is, into the configuration C.

Formally, the basic data analytic model is the simple contaminated model

- $X_1,...,X_n$ iid $\sim f_p = (1-p)\,f_0 + pf_1$
  - f_0 is the density of the background mode.
  - f_1 models the contamination.
  - The likelihood is then.

$$\prod_{i=1}^{n}\left\{(1-p)\,f_0\left(X_i\right)+pf_1\left(X_i\right)\right\}$$
$$=\sum_{j=0}^{n}\sum_{C_j}(1-p)^j\,p^{n-j}\prod_{i\in C_j} f_0\left(X_i\right)\prod_{i\in C_j^c} f_1\left(X_i\right)$$

where C_j denotes a subset of size j from $\{1,...,n\}$.

For low contaminated models one approach is to calculate the mle, p^*, of p. Use p^* to test $H_0: p=0$ versus $H_1: p>0$. If H_0 is rejected see if the mixture model can give insights into the configuration C_j. For example, calculate the empirical Bayes posterior with prior $p\left(C_j\right)=\left(1-p^*\right)^j\,p^{*n-j}$. Then

$$p\left(C_j \middle| X_1,...,X_n\right) \propto \left(1-p^*\right)^j p^{*n-j} \prod_{i\in Cj} f_0\left(X_i\right) \prod_{i\in C_j^c} f_1\left(X_i\right)$$

(2.1)

Another approach is the following two stage multiple testing type of method for $p \approx 0$. This suggests using the locally most powerful (LMP) test statistic (discussed in the next section) for testing $H_0 : p = 0$ versus $H_1 : p > 0$ as a screening test to detect if contamination is present. If the null hypothesis is rejected, then further diagnostic tools are used to try to identify which observations are contaminated.

One was given in (2.1) and some others are given below.

For a mixed distribution $f_p, A_0\left(X_i\right) = \left(1-p\right)\frac{f_0(X_i)}{f_p(X_i)}$, and $A_1\left(X_i\right) = 1 - A_0\left(X_i\right)$ are referred to as the assignment function (or membership function), of X_i to f_0 and f_1, respectively. The assignment function can be interpreted as the posterior probability that an observation came from one of the components of the mixture, and can be used to decide which observations are contaminated. Related to the assignment function is the contamination assignment set measure, $p_1\left(B\right) = p\frac{F_1\left(B\right)}{F_p\left(B\right)}$ where $F_i\left(B\right) = \int_B f_i\left(x\right) dm\left(x\right)$ for $i = 0,1, p$. The functions $A_0\left(X\right)$ and $p_0\left(B\right) = 1 - p_1\left(B\right)$ with $B = (-\infty, x]$ or $B = [x, \infty)$ are also referred to as the local false Discovery rate (FDR) and the FDR in multiple testing situations (Efron, 2007). Note that when the null hypothesis is rejected, $p_1\left(B\right)$ (with p replaced by its mle estimator) could be interpreted heuristically as an empirical Bayes posterior probability that an observation is contaminated given that it is in B and gives some indication of the proportion of contamination in B among the background. Also note that

$$p_1\left([x,x+\varepsilon)\right) = p\frac{F_1([x,x+\varepsilon))/m([x,x+\varepsilon))}{F_p([x,x+\varepsilon))/m([x,x+\varepsilon))} \to A_1\left(X\right) \quad as\ \varepsilon \to 0$$

The LMP test (next section) suggests the use of $f_1/f_0\left(X_i\right)$ to detect the contaminated observations. A plot of this quantity should be centered around 1 when there is no contamination. To find a significant collection of spurious observations consider the following approach based on the LMP test statistic. Define $L_i = \left(f_1\left(X_i\right) - f_0\left(X_i\right)\right)/f_0\left(X_i\right)$. Let the order statistics be $L_{(1)} < L_{(2)} < ... < L_{(n)}$ and let $j\left(i\right)$ denote the inverse rank, i.e., $L_{(i)} = L_{j(i)}$. For mixture or scanning purposes, consider the sets

$$D_i = \left\{j\left(n\right),..., j\left(n-i+1\right)\right\} = \left\{k : L_{(n-i+1)} \le L_k\right\}$$

(2.2)

For mixtures with mle p^*, assign D_i to f_1 and D_i^c to f_0 where $i \approx np^*$. Look through the increasing sequence of sets D_i for a spatial pattern to emerge. Use (2.1) to determine which D_i^c is most probable.

3. THE LMP TEST

In this section we discuss the LMP and the MLE of the simple contaminated model (1.1). To obtain the LMP test we need the following. Let $\phi\left(f\left(X_1\right),..., f\left(X_n\right)\right) = \phi\left(f\right) = \log\prod_i f\left(X_i\right)$ denote the log likelihood of a set of observations from a common distribution f and let

$$\Phi_{p_0}\left(f_1; f_0\right) \equiv \lim_{p\to p_0} \frac{\phi\left(f_p\right) - \phi\left(f_{p_0}\right)}{p - p_0} = \frac{\partial}{\partial p}\log\prod_{i=1}^n f_p\left(X_i\right)\Big|_{p=p_0} = \sum_{i=1}^n \frac{f_p(X_i) - f_0(X_i)}{f_{p0}(X_i)}$$

From the generalized Neyman-Pearson lemma (cf., Ferguson, 1967, Sections 5.1 and 5.5), it is easy to show that the LMP test for testing $H_0 : p = p_0$ versus $H_1 : p > p_0$ is based on $\Phi_{p_0}\left(f_1; f_0\right)$ (see Ferguson, 1967, equation 5.78).

The LMP test statistic is related to the gradient plot introduced by Lindsay (1983a) in the study of mixed distribution models of which (1.1) is a special case. He uses the gradient plot to determine when the one point mixture mle (i.e., $p = 0$) is the global mixture mle. When it isn't, this suggests that some contamination is present. However, as shown in the next section, when the sample size is large and $p = 0$, the MLE p^* will be greater than 0 with probability 0.5. The function $\Phi p\left(f_1; f_0\right)$ plays a

prominent role in the análisis of data from mixtures models where it is the directional derivative $D(\theta;Q)=\Phi(f_\theta; f_Q)=\sum_{i=1}^{n}\left\{\frac{f\theta(X_i)}{f_Q(X_i)}-1\right\}$ defined below. Here the mixture is over a family of densities $\{f_\theta:\theta\in\Theta\}$. Let M denote the set of probability measures on Θ. For $Q\in M$ denote the mixed distribution over the family with mixing distribution *Q* by

$$f_Q=\int f_\theta dQ(\theta)$$

For $X_1,...,X_n$ being iid from f_Q, the likelihood and log likelihood are given by

$L(Q)=\prod_i f_Q(X_i)$ and $\phi(f_Q)=\log\prod_i fQ(Xi)$

where $f_Q=(f_Q(X_1),...,f_Q(X_n))$. The directional derivative of ϕ at f_{Q_0} towards f_{Q_1} is

$$\Phi(f_{Q_1};f_{Q_0})=\lim_{\varepsilon\to 0}\left(\phi\left((1-\varepsilon)f_{Q_0}+\varepsilon f_{Q_1}\right)-\phi(f_{Q_0})\right)/\varepsilon$$

$$=\sum_{i=1}^{n}\frac{f_{Q_1}(X_i)-f_{Q_0}(X_i)}{f_{Q_0}(X_i)}=\sum_{i=1}^{n}\left\{\frac{f_{Q_1}(X_i)}{f_{Q_0}(X_i)}-1\right\}$$

$$=\int D(\theta;Q)dQ_1(\theta)$$

The directional derivative *D* is used to identify when a *k*-point MLE, Q_k^*, for $L(Q)$ is the global mle Q^* (a k-point mle maximizes the likelihood function restricted to mixtures with *k* components). The basic idea is that $D(\theta;Q)=0$ at the support points of the *k*- point MLE Q^* and $D(\theta;Q)\le 0$ if and only if Q^* is the global MLE (Lindsay, 1983a,b).

4. ASYMPTOTIC CONSIDERATIONS

In this section, we determine the asymptotic distributions of the MLE p^* of *p* and the LMP test statistic for testing $H_0: p=0$. When testing $H_0: p=p_0$ and p_0 is in the interior of the parameter space, i.e., $0<p_0<1$, the usual asymptotics go through, since they are based on sums of bounded random variables (see Proposition 4.1). Therefore, we focus only in the case when testing $H_0: p=0$. Section 4.1 considers the case when the true value of the parameter $p=0$. Since $p=0$ is on the boundary, this leads to asymptotics under nonstandard conditions. In particular, the asymptotic distribution of the MLE p^* is a mixed distribution, where one of the components is degenerate at 0, and the other is either half normal when the Fisher information $I_0=E_0\left(\left[(f_1-f_0)/f_0\right]^2\right)<\infty$ or is a stable law when $I_0=\infty$.

Section 4.2 considers the distribution of the LMP test statistic for testing $H_0: p=0$ when the true value of the parameter $0<p<1$. The results therein can be used for power calculations. Section 4.3 gives the distributional properties of the ratio of two densities for the cases used in the examples and simulations. Though out this section, let $X_1,...,X_n$ be iid with density $f_p(x)=(1-p)f_0(x)+pf_1(x)$ where all the random variables are assumed to be defined on the same probability space. Also let $Z_i=f_1(Xi)/f_0(Xi)$ and $L_i=Z_i-1=\frac{f_1(Xi)-f_0(Xi)}{f_0(Xi)}$. The LMP test statistic from Section 3 corresponding to the null hypothesis $H_0: p=0$ is denoted by $T_n=\sum_{i=1}^{n}Li$. Let $I_0=E_0(L_i^2)$ and $W_i=E_0\left(|Li|^3\right), i=0,1$ where E_0 denotes expectation under H_0. Note that I_0 is the Fisher information under H_0. Also, throughout this section, G_α represents the cumulative distribution function of a stable law with parameter $\alpha\in(0,2]$, i.e., its characteristic function is (A.1). Define $\overline{G}_\alpha=1-G_\alpha$.

The next proposition is used in some parts of this section and is the basis for the claim that when p_0 is in the interior of the parameter space the terms in the LMP test statistic are all bounded.

PROPOSITION 4.1.

$$\frac{f_1(x)-f_0(x)}{(1-p)f_0(x)+pf_1(x)}$$

is bounded for $0<p<1$ (hence all its moments are finite).

Proof. Notice that

$$\frac{f_1(x)-f_0(x)}{(1-p)\,f_0(x)+p\,f_1(x)}=\frac{1}{p}\left(\frac{f_1(x)}{\frac{1-p}{p}\,f_0(x)+f_1(x)}\right)$$

$$-\frac{1}{1-p}\left(\frac{f_0(x)}{f_0(x)+\frac{p}{1-p}\,f_1(x)}\right)$$

Therefore,

$$\left|\frac{f_1(x)-f_0(x)}{(1-p)\,f_0(x)+p\,f_1(x)}\right|\le\frac{1}{p}+\frac{1}{1-p}$$

4.1 FIRST CASE: *p* = 0

The next few lemmas show the distribution of the MLE p^* when $p = 0$ under different conditions.

Lemma 4.2. Under $H_0 : p = 0$, p^* converges to 0 almost surely.

Proof. Let

$$l(p)\equiv\phi(f_p)$$

$$l'(p)=\frac{\partial}{\partial p}\phi(f_p)=\frac{\partial}{\partial p}\log\prod_{i=1}^{n}f_p(X_i)=\sum_{i=1}^{n}\frac{f_1(X_i)-f_0(X_i)}{f_p(X_i)}$$

and note that

$$l''(p)=-\sum_{i=1}^{n}\frac{\left[f_1(X_i)-f_0(X_i)\right]^2}{f_p(X_i)^2}\le 0$$

So $l(p)$ is concave and attains its maximum, p^*, either at 0 or 1 or on (0; 1). Let $U_n(p)=l'(p)$ where $U_n=U_n(0)$ and note that $U_n(p)$ is the sum of *n* iid random variables with

$$E_0\left(\frac{f_1(X_i)-f_0(X_i)}{f_p(X_i)}\right)=0 \text{ when } p=0 \text{ and}$$

$$<0 \text{ for } p>0 \qquad \textbf{(4.1)}$$

When $U_n\le 0, U_n(p)\le U_n$ since $l(p)$ is concave. Thus, $l(p)$ attains it maximum at 0 on $\{U_n\le 0\}$. When $U_n>0, l(p)$ attains its maximum on (0,1]. Since $U_n(p)$ has mean less than 0 for $0<p<1$, $U_n(p)/n$ converges almost surely to a negative number (because of Proposition 4.1). When $U_n(p)<0$ and $U_n>0, 0<p^*<p$ with $U_n(p^*)=0$ since $U_n(0)>0$ and $U_n(p)<0$. Thus, $\lim p^*<p$ almost surely on the set where $U_n(p)/n$ converges to its mean. Since $p>0$ is arbitrary, this with the previous paragraph implies that p^* converges to zero almost surely.

Lemma 4.3. If $I_0<\infty$ and $W_i<\infty, i=0,1$, then, under $H_0, \sqrt{n}\,p^*$ converges in distribution to X where $X=0$ with probability .5 and $=\left|N\left(0, I_0^{-1}\right)\right|$ with probability .5.

Proof. If $p^*\in(0,1)$, then $l'(p^*)=0$ and

$$l'(0)=l'(0)-l'(p^*)=-l''(0)(p^*)-\frac{l'''(p')(p^*)^2}{2} \qquad (4.2)$$

where p' is between 0 and p^* and

$$l'''(p)=2\sum_{i=1}^{n}\frac{\left[f_1(X_i)-f_0(X_i)\right]^3}{f_p(X_i)^3}$$

Note that since the derivative of $l'''(p)$ is nonpositive, $l'''(p)$ is nonincreasing and $l'''(1)\le l'''(p)\le l'''(0)$. Thus, since $W_i<\infty$ for $i=0,1$, the sequence $l'''(p)/n, n=1,2,...$ is bounded almost surely. It follows from (4.2) that when $U_n>0$ and $U_n(1)<0, p^*\in(0,1)$ and

$$\frac{l'(0)}{\sqrt{n}}=\frac{-l''(0)}{n}\sqrt{n}\,p^*-\frac{l'''(p')\sqrt{n}\,(p^*)^2}{2n}$$

$$=\frac{-l''(0)}{n}\left(\sqrt{n}\,p^*\right)(1+R_n) \qquad (4.3)$$

In (4.3) R_n goes to zero almost surely since p^* converges to zero almost surely, the sequence $l'''(p)/n, n=1,2,...$ is bounded almost surely and $-l''(0)/n$ converges almost surely to I_0. When $U_n\le 0, p^*=0$. Since $U_n/\sqrt{n}$ is asymptotically $N(0, I_0)$ and $U_n(1)/n$ converges almost surely to a negative number by (4.2), $P(U_n\le 0)$ and $P(U_n>0 \text{ and } U_n(p)<0)$ both converge to 1/2. The second part of this lemma

follows from this and from (4.3) since $-l''(0)/n$ converges almost surely to $E_0\left(\left[f_1 - f_0/f_0\right]^2\right) = I_0$ and $-l'(0)/\sqrt{n}$ converges in distribution to $N(0, I_0)$.

For the next lemma, $\{a_n\}$ is a sequence of real numbers and

$$V_n(p) = \frac{1}{a_n^2}\sum \frac{\left(f_1(X_i) - f_0(X_i)\right)^2}{f_0(X_i)\, f_p(X_i)}$$

Lemma 4.4. If Z_1 Satisfies (A.2) for some $1 < \alpha \le 2$ (i.e., is in the domain of attractation of an α-stable law) and a_n satisfies (A.4), then, under H_0, $a_n p^* V_n(p^*)$ converges in distribution to X where $X = 0$ with probability $G_\alpha(0)$ and $P(X > d) = \bar{G}_\alpha(d)$ for $d > 0$.

Proof. For $p^* \in (0,1), l'(p^*) = 0$ and

$$\begin{aligned} l'(0) &= l'(0) - l'(p^*) \\ &= \sum_{i=1}^{n}\left(f_1(X_i) - f_0(X_i)\right)\left(\frac{1}{f_0(X_i)} - \frac{1}{f_{p^*}(X_i)}\right) \\ &= p^*\sum \frac{\left(f_1(X_i) - f_0(X_i)\right)^2}{f_0(X_i)\, f_{p^*}(X_i)} \\ &= p^* a_n^2 V_n(p^*) \end{aligned} \tag{4.4}$$

Note that when $p^* \in (0,1)$,

$$\frac{l'(0)}{a_n} = a_n p^* V_n(p^*)$$

and $l'(0)/a_n$ converges in distribution to G_α by Lemma A.1 since $E_0(Z_1) = 1$. Since $p^* = 0$ when $l'(0) < 0$, the results follows.

From the proof of Lemma 4.4, by setting $a_n = \sqrt{n}$ we can get the asymptotic distribution of p^* without the third moment assumption given in Lemma 4.3. The next corollary states this result.

Corollary 4.5. If $I_0 < \infty$, then, under H_0, $\alpha = 2$ and $p^*\sqrt{n}V_n(p^*)$ converges in distribution to X where $X = 0$ with probability .5 $= \left|N\left(0, I_0^{-1}\right)\right|$ and with probability .5. Moreover, $V_n(0)$ converges almost surely to I_0.

Remark 4.6. When $1 < \alpha < 2, V_n(0)$ converges in distribution to a stable law with parameter $\alpha/2$ (by Corollary A.2). So, one Could replace $V_n(p^*)$ with $V_n(0)$ in (4.4), except that one could not justify this replacement without putting some condition on $l'''(p)$ The next few lemmas show the distribution of the LMP test statistic T_n for various cases.

Lemma 4.7. If $I_0 < 1$, then, under $H_0, T_n/\sqrt{n}$ converges in distribution to $N(0, I_0)$.

Proof. The proof follows by a direct application of the central limit theorem.

Lemma 4.8. If Z_1 and a_n satisfy conditions (A.2) and (A.4), respectively, for some $1 < \alpha \le 2$, then, under $H_0, T_n/a_n$ converges in distribution to G_α (a stable law with parameter α).

Proof. The proof follows by a direct application of Lemma A.1

If f_1 has an unknown parameter and $p = 0$, an identifiability issue surfaces that makes it impossible to estimate that parameter. In this case, if the parameter is estimated from the data and used to calculate T_n, it is not clear to what limit distribution T_n is converging. The simulations shown later illustrate this point.

4.2 Case 2: $p > 0$

The asymptotic distribution of T_n given in Lemmas 4.7 and 4.8 is for $p = 0$. The next two lemmas give the asymptotic distribution of T_n when $p > 0$. For this, assume that $X_1 \sim f_p = (1-p) f_0 + p f_1$ and let $W_0' = E_0(L_1^3)$.

Lemma 4.9. If $I_0 < \infty$ and $W_0 < \infty$, then $(T_n - npI_0)/\sqrt{n}$ converges in distribution to $N\left(0, I_0 + pW_0' - p^2 I_0^2\right)$.

Proof. It is easy to prove that $E_p(L_1) = pI_0$ and $E_p(L_1^2) = I_0 + pW_0'$. The result then follows from a direct application of the central limit theorem.

Lemma 4.10. Suppose Z_1 and a_n satisfy conditions (A.2) and (A.4), respectively, for some $0<\alpha\le 2$. If $1<\alpha\le 2$, then $I_0<\infty$ and $(T_n - npI_0)/a_n$ converges in distribution to G_α, while if $0<\alpha<1$, T_n/a_n converges in distribution to G_α. If $\alpha=1$, then $(T_n-\mu_n)/a_n$ converges in distribution to a stable law with parameter 1, where μ_n is defined as in (A.3).

Proof. The proof is a direct application of the results in Appendix A.

4.3 DISTRIBUTIONAL PROPERTIES OF DENSITY RATIOS

In this section, we consider the properties of $Z = f_1(X)/f_0(X)$ for some frequently used distributions. These properties are required to use the lemmas in Sections 4.1 and 4.2. Section 4.3.1 considers the case when both, f_0 and f_1 are exponential distributions, and Section 4.3.2 considers the case of the normal distribution.

4.3.1 EXPONENTIAL DISTRIBUTION

Suppose $f_\theta(x)=\frac{1}{\theta}e^{-x/\theta}1_{[x>0]}$ and let $f_j = f_{\theta_j}$ for $j=0,1$. If $X\sim f_\theta$ then

$$P_\theta(Z>z)=\begin{cases}1_{[z\le 0]}+\left(1-(z/\beta)^{-\alpha}\right)1_{[0<z<\beta]} & \\ \qquad \text{when } \theta_1<\theta_0 & \\ 1_{[z\le\beta]}+(z/\beta)^{-\alpha}1_{[z>\beta]} & \\ \qquad \text{when } \theta_1>\theta_0 & \end{cases}$$

Where $\beta=\frac{\theta_0}{\theta_1}$ and $\alpha=\frac{\theta_0\theta_1}{\theta(\theta_1-\theta_0)}$.

For $\theta_1>\theta_0$, $z^\alpha P_\theta(Z>z)\to c=\beta^\alpha$ as $z\to\infty$, which corresponds to the first row in Tables 5 and 6. In particular, if $0<\alpha<2$, condition (A.2) is satisfied and therefore Z is in the domain of attraction of an α-stable law. If $\alpha\ge 2$ then condition (A.6) is satisfied and Z would be in the domain of attraction of a normal distribution. The appropriate normalizing constant is $a_n=\left(\beta^\alpha s_\alpha n\right)^{1/\alpha}$ for $0<\alpha<2$, and $a_n=\sqrt{0.5\beta^2 n\log n}$ for $\alpha=2$, and $a_n=\sqrt{0.5n\,\mathrm{var}_\theta(Z)}$ for $\alpha>2$. The mean of the density ratio Z is

$$E_\theta(Z)=\begin{cases}\frac{\alpha\beta}{\alpha-1} & \text{when } \theta_1<\theta_0 \text{ or } (\theta_1>\theta_0 \text{ and } \alpha>1)\\ \infty & \text{when } \theta_1>\theta_0 \text{ and } \alpha\le 1\end{cases}$$

and the variance is

$$\mathrm{var}_\theta(Z)=\begin{cases}\frac{\alpha\beta^2}{(\alpha-1)^2(\alpha-2)} & \text{when } \theta_1<\theta_0 \text{ or}\\ & (\theta_1>\theta_0 \text{ and } \alpha>2)\\ \infty & \text{when } \theta_1>\theta_0 \text{ and } 1<\alpha\le 2\\ \text{undefined} & \text{when } \theta_1>\theta_0 \text{ and } \alpha\le 1\end{cases}$$

In particular, when $\theta=\theta_0$, $\alpha=\theta_1/(\theta_1-\theta_0)$ and the mean becomes $E_{\theta_0}(Z)=1$ and the variance, which is actually I_0, becomes

$$I_0=\mathrm{var}_{\theta_0}(Z)=\begin{cases}\frac{(\theta_1-\theta_0)^2}{\theta_1(2\theta_0-\theta_1)} & \text{when } \theta_1<2\theta_0\\ \infty & \text{when } \theta_1\ge 2\theta_0\end{cases}$$

For $\theta=\theta_1$ the mean becomes $E_{\theta_1}(Z)=I_0+1$, which is infinite when $\theta_1\ge 2\theta_0$, and the variance becomes

$$\mathrm{var}_{\theta_1}(Z)=\begin{cases}\frac{\theta_0^3(\theta_1-\theta_0)^2}{2\theta_1^2(2\theta_0-\theta_1)^2(1.5\theta_0-\theta_1)} & \text{when } \theta_1<1.5\theta_0\\ \infty & \text{when } 1.5\theta_0\le\theta_1<2\theta_0\\ \text{undefined} & \text{when } \theta_1\ge 2\theta_0\end{cases}$$

4.3.2 NORMAL DISTRIBUTION

Let $\varphi(y)=\frac{1}{\sqrt{2\pi}}e^{-y^2/2}$ denote the density of a standard normal distribution. Suppose now that $f_{\mu,\sigma^2}(x)=\varphi\left((x-\mu)/\sigma\right)/\sigma$ and let $f_j = f_{\mu j,\sigma_j^2}$ for $j=0,1$. Assume also that $X\sim f_{\mu j,\sigma_j^2}$. Before examining the tail probability of the density ratio, we need the tail probability of a normal distribution. Let Y be a standard normal

distribution with density φ □. It is well known that, for $y > 0$,

$$\frac{y}{1+y^2}\varphi(y) < P(Y>y) < \frac{1}{y}\varphi(y)$$

This implies that

$$\frac{y^2}{1+y^2} < \sqrt{2\pi}ye^{y^2/2}P(Y>y) < 1$$

and thus

$$\lim_{y\to\infty}\sqrt{2\pi}ye^{y^2/2}P(Y>y) \to 1 \qquad (4.5)$$

Let $\delta_j = \frac{\mu_1^j}{\sigma_1^2} - \frac{\mu_0^j}{\sigma_0^2}$ for $j = 1, 2$ (for j=0 let $\delta_0 = \frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}$). The tail probability of the density ratio when $\sigma_1^2 > \sigma_0^2$ is given by

$$P_{\mu,\sigma^2}(Z > z)$$

$$= P_{\mu,\sigma^2}\left(\frac{f_1(X)}{f_0(X)} > z\right)$$

$$= P_{\mu,\sigma^2}\left(\frac{\left|X - \frac{\delta_1}{\delta_0}\right|}{\sigma} > \sqrt{\alpha\left(\delta_2 - \frac{\delta_1^2}{\delta_0} - \log(\beta)\right) + 2\alpha\log z}\right)$$

$$= P\left(Y > -b + \sqrt{\alpha\left(\frac{(\mu_1-\mu_0)^2}{\sigma_1^2-\sigma_0^2} - \log(\beta)\right) + 2\alpha\log z}\right)$$

$$+ P\left(Y > b + \sqrt{\alpha\left(\frac{(\mu_1-\mu_0)^2}{\sigma_1^2-\sigma_0^2} - \log(\beta)\right) + 2\alpha\log z}\right)$$

where

$$b = \frac{\left|\sigma_1^2(\mu-\mu_0) + \sigma_0^2(\mu_1-\mu)\right|}{\sigma\left(\sigma_1^2-\sigma_0^2\right)}, \alpha = \frac{\sigma_0^2\sigma_1^2}{\sigma^2\left(\sigma_1^2-\sigma_0^2\right)},$$

and $\beta = \sigma_0^2/\sigma_1^2$.

This expression, combined with (4.5), gives us the tail behavior, i.e., as $z \to \infty$,

$$\begin{cases} z^{\alpha}\sqrt{2\alpha\log z}P_{\mu,\sigma^2}\left(\frac{f_1(X)}{f_0(X)} > z\right) \to c_1 \\ \qquad \text{when } b = 0, \text{and} \\ z^{\alpha}\sqrt{2\alpha\log z}\exp\left(-b\sqrt{2\alpha\log z}\right)P_{\mu,\sigma^2}\left(\frac{f_1(X)}{f_0(X)} > z\right) \to c_2 \\ \qquad \text{when } b > 0, \end{cases}$$

where $c_1 = 2\frac{\exp\left[\frac{\alpha}{2}\left(\log(\beta) - \frac{(\mu_1-\mu_0)^2}{\sigma_1^2-\sigma_0^2}\right)\right]}{\sqrt{2\pi}}$ and

$$c_2 = \frac{\exp\left[-\frac{b^2}{2} + \frac{\alpha}{2}\left(\log(\beta) - \frac{(\mu_1-\mu_0)^2}{\sigma_1^2-\sigma_0^2}\right)\right]}{\sqrt{2\pi}}.$$

Therefore, when $\sigma_1^2 > \sigma_0^2$, the tail behavior coincides with those in Tables 5 and 6, which give the appropriate normalizing constants for $0<\alpha<2$ and $\alpha = 2$ respectively. When $\alpha > 2$, then the appropriate normalizing constant is $a_n = \sqrt{.5n\,\mathrm{var}_\theta(Z)}$.

The mean of the density ratio is

$$E_{\mu,\sigma^2}(Z) = \begin{cases} e^{\left(\frac{1}{2}\left[\sigma^2\delta_1^2 + 2\mu\delta_1 - \delta_2\right]\right)} & \text{when } \sigma_1^2 = \sigma_0^2 \\ \sqrt{\frac{\alpha\beta}{\alpha-1}}e^{\left(\frac{1}{2\sigma^2}\left[\frac{\alpha}{\alpha-1}\left(\sigma^2\delta_1+\mu\right)^2 - \mu^2 - \sigma^2\delta_2\right]\right)} & \\ \quad \text{when } \sigma_1^2 < \sigma_0^2 \text{ or } \left(\sigma_1^2 > \sigma_0^2 \text{ and } \alpha > 1\right) & \\ \infty \qquad \text{when } \sigma_1^2 > \sigma_0^2 \text{ and } \alpha \le 1 & \end{cases}$$

The variance is given by

$$\mathrm{var}_{\mu,\sigma^2}(Z) = \begin{cases} \left(e^{\sigma^2\delta_1^2}-1\right)e^{\left(\sigma^2\delta_1^2+2\mu\delta_1-\delta_2\right)} & \text{when } \sigma_1^2=\sigma_0^2 \\ \sqrt{\frac{\alpha\beta^2}{\alpha-2}}e^{\left(\frac{1}{2\sigma^2}\left[\frac{\alpha}{\alpha-2}\left(2\sigma^2\delta_1+\mu\right)^2-\mu^2\right]-\delta_2\right)} & \text{when } \sigma_1^2<\sigma_0^2 \text{ or} \\ -\frac{\alpha\beta}{\alpha-1}e^{\left(\frac{1}{\sigma^2}\left[\frac{\alpha}{\alpha-1}\left(\sigma^2\delta_1+\mu\right)^2-\mu^2\right]-\delta_2\right)} & \left(\sigma_1^2>\sigma_0^2 \text{ and } \alpha>2\right) \\ \infty & \text{when } \sigma_1^2>\sigma_0^2 \text{ and } 1<\alpha\le 2 \\ \text{undefined} & \text{when } \sigma_1^2>\sigma_0^2 \text{ and } \alpha\le 1 \end{cases}$$

When $f_{\mu,\sigma^2} = f_0$ the mean reduces to $E_0(Z)=1$ and the variance, which is I_0, reduces to

$$I_0 = \mathrm{var}_0(Z) = \begin{cases} e^{\left([\mu_1-\mu_0]^2/\sigma_0^2\right)}-1 & \text{when } \sigma_1^2=\sigma_0^2 \\ \frac{\sigma_0^2}{\sigma_1\sqrt{2\sigma_0^2-\sigma_1^2}}e^{\left(\frac{(\mu_1-\mu_0)^2}{2\sigma_0^2-\sigma_1^2}\right)}-1 & \text{when } \sigma_1^2<\sigma_0^2 \\ \infty & \text{when } \sigma_1^2\ge 2\sigma_0^2 \end{cases}$$

For $f_{\mu,\sigma^2} = f_1$ the mean becomes $E_1(Z) = I_0 + 1$ and the variance becomes

$$\mathrm{var}_1(Z) = \begin{cases} \left(e^{\left([\mu_1-\mu_0]^2/\sigma_0^2\right)}-1\right)e^{\left(2[\mu_1-\mu_0]^2/\sigma_1^2\right)} & \text{when } \sigma_1^2=\sigma_0^2 \\ \dfrac{\sigma_0^3}{\sigma_1^2\sqrt{2\left(1.5\sigma_0^2-\sigma_1^2\right)}}e^{\left(1.5\frac{(\mu_1-\mu_0)^2}{1.5\sigma_0^2-\sigma_1^2}\right)} & \text{when } \sigma_1^2<1.5\sigma_0^2 \\ -\dfrac{\sigma_0^4}{\sigma_1^2\left(2\sigma_0^2-\sigma_1^2\right)}e^{\left(2\frac{(\mu_1-\mu_0)^2}{2\sigma_0^2-\sigma_1^2}\right)} & \\ \infty & \text{when } 1.5\sigma_0^2\le\sigma_1^2<2\sigma_0^2 \\ \text{undefined} & \text{when } \sigma_1^2\ge 2\sigma_0^2 \end{cases}$$

5. SIMULATIONS

For the first set of simulations, background and contamination data are generated from exponential distributions with means $\theta_0 = 166.206$ and $\theta_1 = 592.922$, respectively, which are the estimated means for the 2-point mle for the Mining Data in the next section. Samples of sizes *n* = 100, 500, 1000 are generated, with 0, 1 and 5 percent of contamination (*p* = 0, 1.01, .05). With each sample we calculate $T_n = \sum_{i=1}^{n} \dfrac{f_1(X_i) - f_0(X_i)}{f_0(X_i)}$ and $S_n = \sum_{i=1}^{n} \dfrac{\hat{f}_1(X_i) - \hat{f}_0(X_i)}{\hat{f}_0(X_i)}$, where $\hat{f}_0$ and $\hat{f}_1$ are estimates of the densities based on the maximum likelihood estimators of θ_0 and θ_1. These estimates are used as if they were the true parameters and a normalizing constant and critical value are calculated based on these estimates. The process is repeated *N* = 10000 times and the number of rejections of the null hypothesis $H_0 : p = 0$ at the .05 level are recorded.

Following the results from Section 4.3.1, the variance of the terms in T_n corresponding to f_0 is infinite, but the tail behavior of the density ratio, under H_0, follows that of the first line of Table 5 with α = 592.922/(592.922-166:206) = 1.3895 and *c* = (166.206/592.922)$^{1.3895}$ =0.1708081. Hence, the normalizing constant is $a_n = \left(0.1708081 s_{1.3895} n\right)^{1/1.3895} = .4881128 n^{0.7196832}$. From Lemma A.1, the rejection region defined by $T_n / \left(.4881128 n^{0.7196832}\right) > 4.40186$ would reject the null hypothesis with probability 0.05 if there are no anomalies. A similar process is done to calculate the rejection region $S_n / a_n > d_{.05}$ in each sample, where the normalizing constant a_n and the critical value $d_{.05}$ change from sample to sample, and are calculated based on either the normal or the stable distribution[4]. The results of these simulations are shown

[4] *if* $\hat{\theta}_1 < 2\hat{\theta}_0$, *and these are assumed to be the actual parameters, the variance of the in Table 1.*

TABLE I

Asymptotic distribution theory for contamination models

Proportion of rejections of H_0, no anomalies, out of 10000 simulations with background and contamination generated from exponential distributions with means 166.206 and 592.922, respectively, with *p* = 0, 0.01, 0.05.

	Proportion of anomalies					
	Based on T_n			Based on S_n		
Sample sizes	0	0.01	0.05	0	0.01	0.05
100	0.0492	0.1646	0.5316	0.2853	0.3898	0.6340
500	0.0483	0.3836	0.9361	0.3248	0.5725	0.9252
1000	0.0544	0.5454	0.9935	0.3387	0.6898	0.9786

The simulations show that when using the true parameters to calculate T_n the proportion of samples that rejected the null hypothesis when it is true is about 0.05, as expected. Notice that with the background and contamination means fixed at θ_0 = 166.206 and θ_1 = 592.922, the power increases as the proportion of anomalies, *p*, increases and as the sample size increases. The exact asymptotic power for the LMP test can be calculated using Lemma 4.10 and the results from Section 4.3.1 and Table 5, with $\alpha = \theta_0/(\theta_1 - \theta_0) = 0.3895003$ and $c = p(\theta_0/\theta_1)^{\alpha} = 0.6093393\,p$, which gives $(cs_\alpha)^{1/\alpha} = 0.4473926\,p^{2.567392}$. Rejection occurs when $T_n/(.4881128n^{0.7196832}) > 4.40186$, which is equivalent to rejecting if

$$T_n/\left(0.4473926p^{2.567392}n^{2.567392}\right) > 4.802503/\left(p^{2.567392}n^{1.847709}\right)$$

The left hand side of this inequality converges to a stable law with parameter $\alpha =$ 0.3895003. The power can be obtained for each value of *n* and *p*. For instance, if *n* = 100 and *p* = 0.05, the power is the probability that a value from a stable law is larger than 2.119844, that is, 0.510141. In the case of *n* = 500 and *p* = 0.01, the tail starts at 6.750576, which gives a power of 0.3517792. For *n* = 1000 and *p* = 0.05, the power is 0.996379. The simulations confirm these values. If estimates are used as if they were the true parameters, rejection occurs 28.5% of the time when there are no anomalies present (*p* = 0) and *n* = 100. This is troubling and indicates that false discovery is a serious problem in this case. This is not the case when *p* > 0 and Appendix B indicates the necessary adjustments that need to be made when estimating parameters. For the second set of simulations data are generated from normal distributions, where the background consists of standard normal variables $(\mu_0 = 0$ and $\sigma_0^2 = 1)$ and the anomalies consist of a normal with mean $\mu_1 = 0$ and variance $\sigma_1^2 = 3$. Samples of sizes *n* = 100, 500, 1000 are generated, with 0, 1 and 5 percent of the observations being anomalies.

The results from Section 4.3.2 and Table 5 are used to determine the rejection region for each sample. Since the variance of f_1/f_0 is infinite under f_0, the tail behavior of the distribution of the ratio needs to be taken into consideration to see what stable law applies when $\mu = \mu_0 = 0$ and $\sigma^2 = \sigma_0^2 = 1$. This is described in Section 4.3.2 with

$$\alpha = 3/(3–1) = 1.5, \beta = 1/3, b = 0, c_1$$
$$= 2\beta^{\alpha/2}/\sqrt{2\pi} = 2\left(\tfrac{1}{3}\right)^{1.5/2}/\sqrt{2\pi} = 0.350025$$

The rejection region is now found using the results from Table V: reject the null hypothesis $H_0 : p = 0$ if

$$\frac{T_n}{0.7274158\left(n/\sqrt{\log n}\right)^{2/3}} > 3.824235$$

The results of these simulations are found in Table II. The LMP test seems somewhat conservative for *n* = 100, possibly because this sample size is too small to observe convergence to the stable law. This resulted in a test with somewhat poor power. For *n* = 500 and *n* = 1000, the test seems to perform better.

TABLE II

Asymptotic distribution theory for contamination models

Proportion of rejections of H_0, no anomalies, out of 10000 simulations with background and anomalies generated from normal distributions with mean 0 and variances 1 and 3, respectively

	Proportion of anomalies		
Sample sizes	0	0.01	0.05
100	0.0371	0.0788	0.2331
500	0.0447	0.1533	0.5649
1000	0.0408	0.2295	0.7885

The exact asymptotic power can be calculated for these tests by using Lemma 4.10, Table V and Section 4.3.2. Suppose the true proportion of anomalies p is positive $p > 0$. Then $P(Z_1 > z) = (1-p)P_0(Z_1 > z) + pP_1(Z_1 > z)$. Let $\alpha = 1/(3-1) = 0.5$ and $c = 2\left(\tfrac{1}{3}\right)^{0.5/2} p/\sqrt{2\pi} = 0.6062612\,p$ Thus, to get a stable law we need to normalize T_n by $0.2886751p^2\left(n/\sqrt{\log n}\right)^2$. A rearrangement of the rejection region (which was normalized originally by $0.7274158\left(n/\sqrt{\log n}\right)^{2/3}$) would result in rejections when

$$\frac{T_n}{0.2886751p^2\left(n/\sqrt{\log n}\right)^2} > 3.824235\frac{0.7274158\left(n/\sqrt{\log n}\right)^{2/3}}{0.2886751p^2\left(n/\sqrt{\log n}\right)^2}$$

$$= \frac{9.636468(\log n)^{2/3}}{p^2 n^{4/3}}$$

The exact asymptotic power when n = 100 and $p = 0.5$ is the probability that a stable law variable with parameter 0.5 is greater than $\frac{9.636468\left(\log 100\right)^{2/3}}{.5^2 100^{4/3}} = 22.98661$. This probability is 0.1652201. Similarly, for n = 500 and proportions p = .01 and p = .05, the probabilities of rejecting the null hypothesis are 0.08789053 and 0.4189768, respectively, whereas the simulations estimated these numbers as 0.1533 and 0.5649, respectively.

6. DATA EXAMPLES

Following the analysis from Grego et al. (1990) of the mining accident data, Figure 1 has the gradient functions for the 2 and 3-point mixture mle's where the mixing is over the mean of an exponential distribution and Figure 2 has the assignment function for the second component in the 3-point mle. The estimates of the means and mixing proportions are given in Table III. The gradient plot indicates that the 2-point mle is not the global mle but the 3-point is. The assignment function indicates a distinct difference in the first 53 times and rest of the times. Further analysis by Grego et al indicates that the first 53 are well fit by a single exponential and the rest by a 3- point mixture.

TABLE III

Asymptotic distribution theory for contamination models

Maximum likelihood estimates for the mining data

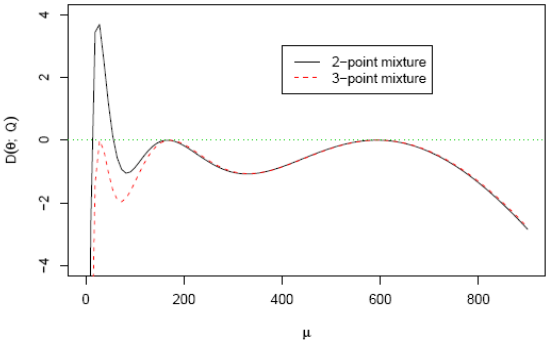
	$\mu_1(p_1)$	$\mu_2(p_2)$	$\mu_3(p_3)$
2-point mle	592.922 (.175149)	166.206 (.824851)	
3-point mle	595.495 (.171379)	171.587 (.805528)	29.0972 (.023093)

For the mining data we will use an exponential with mean 171.587 as f_0 and a 2-point mixed exponential with means 595.495 and 29.0972 and mixing proportions proportional to .171379 and .023093, respectively, as f_1. That is, $f_1 = f_{Q_1}$ where Q_1 has point masses at 595.495 and 29.0972 with mixing proportions .881253 and .118747 and the

FIGURE 1

Asymptotic distribution theory for contamination models

Gradient plots of a 2- and 3-point mixtures (mle) of exponentials for the Mining Data



family, $\{f_\mu\}$, being mixed over is the exponential with its mean parameterization. These are assumed as the true parameters and Lemma 4.8 along with Table V can be used to calculate critical values for the LMP test statistic. The LMP test statistic, T_n, assuming all the parameters are known, is then given by

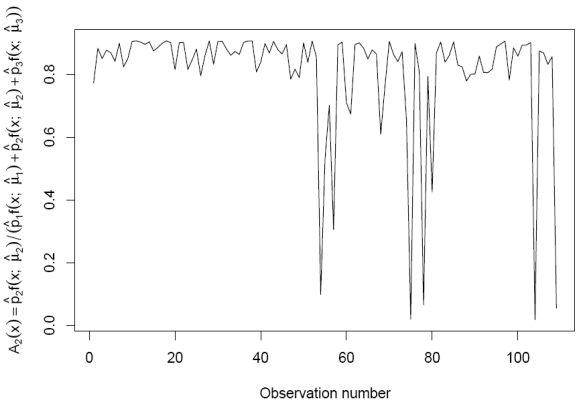
$$T_n = 0.8812528\sum_{i=1}^{n}\left(\frac{\frac{1}{595.495}e^{-X_i/595.495}}{\frac{1}{171.587}e^{-X_i/171.587}} - 1\right)$$

$$+ 0.1187472\sum_{i=1}^{n}\left(\frac{\frac{1}{29.0972}e^{-X_i/29.0972}}{\frac{1}{171.587}e^{-X_i/171.587}} - 1\right)$$

Under the null hypothesis, i.e., $X_i \sim f_0$, the terms in the first sum have infinite variance whereas the terms in the second sum have finite variance (see Appendix 4.3.1 for details). Using the notation of Appendix 4.3.1 with $\theta_0 = 171.587$ and $\theta_1 = 595.495$, let $\alpha = 595.495/(595.495\text{-}171.587) = 1.404774$, $c = (171.587/595.498)^{1.404774} = 0.1741281$ and $a_n = 0.5052435n^{0.7118582}$. If T_n is normalized by a_n, the second sum will quickly converge to zero as $n \to \infty$. The first sum converges in distribution to a stable law with parameter $\alpha = 1.404774$. Therefore $T_n/(.8812528a_n)$ converges in distribution to the same stable law.

For the mining data, $T_n = 574.871$ and $T_n/(.8812528a_n) = 45.77311$. Using Table IV for $\alpha = 1.4$ we can see that the p-value is between .005

FIGURE 2
Asymptotic distribution theory for contamination models
Assignment function for the second component of the 3-point mixture (mle) of exponentials for the Mining Data



and .001. The actual p-value is 0.002102145 (calculated computationally with $\alpha = 1.404774$). This indicates that there is strong evidence that some observations come from f_1. Note that parameters in both f_0 and f_1 are being estimated based on the 3-pt global mle. These estimates have to be taken into consideration in using the LMP test statistic to determine if spurious observations are present. As pointed out in the appendix, it would be impossible to estimate f_1 if $p = 0$, and hence the distribution of the LMP is not clear in this case. If $p > 0$, then we only need to check the regularity conditions discussed in Lemma B.1 and Remark B.2. For the exponential distribution these conditions reduce to the finiteness of the first three moments. We now illustrate some of these ideas using gene expression data. The approach here will start by the assignment function to identify possible anomalies (expressed genes) to get a pooled model. After that, we do the LMP test. Efron (2007) compared prostate data of $m_1 = 50$ non-tumor subjects with $m_2 = 52$ tumor patients for each of *n* = 6033 genes (see Singh et al., 2002). For each gene they perform a two-sample t-test to compare the mean gene-expression between cancer and noncancer subjects. Let t_i for $i = 1,...,n$ denote the test statistics used for each gene. For genes that have the same mean expression values for both groups ti will follow a central t-distribution with $m_1 + m_2 - 2 = 100$ degrees of freedom. Efron (2007) defines $z_i = \phi^{-1}\left(F_{100}(t_i)\right)$, where F_v denotes the cumulative distribution function (cdf) of a t-distribution with v degrees of freedom and ϕ denotes the cdf of a standard normal distribution.

Then the distribution of z_i is standard normal for those genes that have the same mean expression for both groups of subjects. Efron then fits the mixture $f = (1-p) f_0 + p f_1$ as follows. Suppose f is a 7-parameter exponential family and estimate this density from the z-values, obtaining $\hat{f}$. Suppose f_0 is the standard normal density and estimate *p* by using $\log\left((1-p) f_0\right)$ as a "quadratic approximation" of $\hat{f}$. From this he estimates the assignment function A_0 (false Discovery rate). These calculations can be done using the R package locfdr. He discovered 51 genes using false discovery rate, declaring an anomaly when $A_0 < 0.2$.

Another approach for this data is to work directly with the t-values. Let μ_{i1} and μ_{i2} be the mean expressions of gene i for tumor and non-tumor subjects respectively. Suppose the variance of the gene-expression for gene i is σ_i^2 is the same for both groups. Then t_i follows a central t-distribution if $\mu_{i1} = \mu_{i2}$. When the means are diferent t_i follows a non-central t-distribution with non-centrality parameter $\delta_i = \frac{\mu_{i1} - \mu_{i2}}{\sqrt{\sigma_i^2\left(\frac{1}{m_1} + \frac{1}{m_2}\right)}}$. In either case the degrees of freedom are $m_1 + m_2 - 2$.

Assume that all non-centrality parameters have the same magnitude, i.e., $|\delta_i| = \delta$. For simplicity also assume that half of the non-centrality parameters are positive and half are negative. If the proportion of t-values that follow a non-central t-distribution is *p*, then the distribution of each t_i is $f = (1-p) f_0 + p f_1$ where f_0 denotes the density of central t-distribution and $f_1 = .5g_{\delta} + .5g_{-\delta}$ denotes the density of the genes with different mean expression and g_{δ} denotes a t-distribution with non-centrality parameter δ. We shall comment on this choice of f_1 at the end

of this section. This model only has two parameters to estimate p and δ. Maximum likelihood estimation (mle) could be accomplished using the Expectation-Maximization (EM) algorithm, however it is hard to work with the density of the t-distribution. Instead, we use a modified version of the EM-algorithm with an ad-hoc M-step. The idea is that for each gene, the mle of δ_i is simply t_i by the invariance property of the mle (replace μ_{i1} and μ_{i2} with their mle's, the mle σ_i^2 is approximated by its unbiased version, the pooled variance estimate). One ad-hoc estimate of δ could be the average of $|t_i|$. A better estimate uses the weighted average $\sum_i w_i |t_i| / \sum_i w_i$ where w_i is the posterior probability of coming from f_1 given t_i. For the M-step this weighted average was used, calculating the w_i for one iteration using the estimates from the previous iteration as true parameters. For the prostate data, the proportion of t-values with $|t_i| > 2$ and the average of the absolute t-values (i.e., $(|t_1| + ... + |t_n|)/n$) were used as initial values for p and δ, respectively. The iterations were stopped when the change in both δ and p was no greater then 10^{-8}. Convergence was attained in 174 iterations, giving $\hat{\delta}$ = 2.473228 and $\hat{p}$ = 0:04612997.

FIGURE 3

Asymptotic distribution theory for contamination models

Histogram of *t* - values corresponding to the prostate gene-expression data

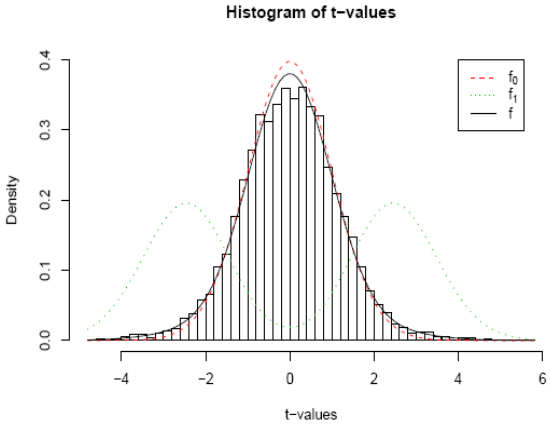


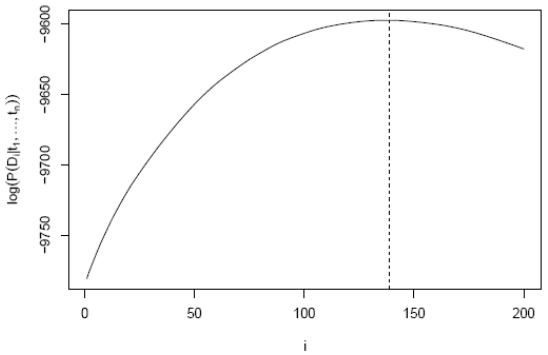
Figure 3 shows a histogram of the *t*-values with the estimated densities superimposed. It can be seen that the central t-distribution f_0 (dashed line) does not fit the *t*-values very well since the t-values have a heavier tail. The mixture of the two non-central t-distributions with parameters δ and $-\delta$, f_1 (dotted line), help to explain the tails. When these two distributions f_0 and f_1 are mixed with p as the proportion for f_1, then the fitted distribution f (solid line) fits the histogram quite nicely using only two parameters (compared to fitting 8 parameters). To do the LMP test, we need to explore the distribution of the density ratio f_1/f_0, and this is quite hard to do with the non-central t-distribution. To work around this, suppose that the variance of the ratio is finite and just use the regular central limit theorem. A random sample of one million values from a central t-distribution with 100 degrees of freedom was generated and the ratio f_1/f_0 was calculated for each value, where the sample variance was 102.2010. This estimate of the variance of f_1/f_0 is assumed to be the true variance. For the prostate data $T_n = \sum_i \frac{f_1(t_i) - f_0(t_i)}{f_0(t_i)} = 27186.5$ and assume that the variance of f_1/f_0 is 102.201. Then, if all observations are from f_0 and the observations were independent then $T_n/\sqrt{n\sigma^2} = 34.62255$. When compared to the quantiles of a standard normal distribution, this value indicates very strong evidence that some of the genes have different mean expression values for tumor and non-tumor patients. Next, we created the sets D_i, as in (2.2), and calculated their empirical posterior probabilities using an expression similar to (2.1), i.e., $P(D_i|t_1,...,t_n) \propto (1-p^*)^{n-i} p^{*i} \prod_{j \in D_i^c} f_0(t_j) \prod_{j \in D_i} f_1(t_j)$. Figure 4 shows the log of these posterior probabilities, giving a maximum at 139, indicating that 139 genes have significant difference in their expression number. Since Efron (2007) found 51 anomalies, the LMP test is used to verify the hypothesis $H_0 : p = p_0 = 51/6033 = 0.00845$. In this case the test statistic is

$$T_n = \sum_{i=1}^{n} \frac{f_1 - f_0}{(1-p_0) f_0 + p_0 f_1} = 5081.583$$

Since T_n is the sum of bounded r.v.'s (Proposition 4.1), the regular central limit theorem can be used to

decide on a rejection region. As before, we simulated one million samples from a central *t* and another million from the noncentral t with noncentrality parameter $\delta = 2.473228$ (using $-\delta$ gives the same variance). Then we calculated the average of these two, weighted by $1-p_0$ and p_0 respectively. This gives a sample variance for $(f_1 - f_0)/((1-p_0)f_0 + p_0 f_1)$ of $\sigma^2 = 18.20424$. So, $T_n/\sqrt{n\sigma^2} = 15.33367$, which is quite significant when compared to quantiles of a standard normal distribution. Therefore, we concluded that the proportion of anomalies is greater than 0.00845.

FIGURE 4
Asymptotic distribution theory for contamination models
Log-probability of the number of "anomalies" corresponding to the prostate gene-expression data



Since the use of *t*-values gave 139 anomalies, we now repeat the exercise from the previous paragraph to test the hypothesis $H_0 : p = p_0 = 139/6033 = 0.02304$. The estimated variance of $(f_1 - f_0)/((1-p_0)f_0 + p_0 f_1)$ is 9.111389. The test statistic is T_n = 1779.638, which normalized gives $T_n/\sqrt{n\sigma^2} = 7.590539$. This indicates that the proportion of anomalies is greater than 0.02304. This suggests a better method to identity anomalies is needed, possibly one based on a cutoff for $(f_1 - f_0)/f_0$.

It is worthwhile to mention that the independence assumption between genes may not be realistic. With regard to the choice of $f_1 = .5g_\delta + .5g_{-\delta}$, note that for this data one is only interested in determining what genes have different expression numbers and not the direction of the difference. Thus, one could consider $|t_i|$. For this symmetric situation f_1 seems appropriate to model the anomalies and what we would recommend for future data analysis. This Choice worked well for the original data set, but the biological justification for equal proportions of positive or negatively expressed genes is not clear to us.

Appendices

The first appendix simply states some stable distribution results used in Section 4, while the second appendix covers the asymptotics for the standard case when the parameters of a mixture model are in the interior of the parameter space.

A Generalized central limit theorems

The lemmas in this section are well known results which give us the limiting distributions of $\sum Z_i$, properly normalized (see Geluk and de Haan, 2000, and the references therein). A stable distribution with parameter α, $0 < \alpha \le 2$, is defined by its characteristic function $\phi_\alpha(t)$ given by

$$\phi_\alpha(t) = \exp\left\{-t^\alpha\left[1+i\,\mathrm{sign}(t)\tan\left(\frac{\alpha\pi}{2}\right)\left(|t|^{1-\alpha}-1\right)\right]\right\} + it\tan\left(\frac{\alpha\pi}{2}\right) \qquad \text{(A.1a)}$$

when $\alpha \neq 1$ and by

$$\phi_1(t) = \exp\left\{-t\left[1+i\,\mathrm{sign}(t)\frac{2}{\pi}\log|t|\right]+i2t\Gamma'(1)/\pi\right\} \qquad \text{(A.1b)}$$

when $\alpha = 1$. $\Gamma'(1)$ is the derivative of the gamma function evaluated at 1 (the negative of Euler's constant). For $\alpha = 2$ the stable distribution becomes a normal distribution with variance 2.

To calculate quantiles and probabilities for this distribution one could use the R functions qstable and pstable (from package fBasics) setting the parameters as alpha = α, beta = 1, gamma = 1, delta = $\tan(\alpha\pi/2)$ and pm = 0. In the case of $\alpha = 1$ the parameter delta should be set to $2\Gamma'(1)/\pi \approx -0.3674669$. Table IV shows some quantiles for selected values of α.

Let $Z_1, ..., Z_n$ be iid with $P(Z_1 > a) = 1$ for some finite *a*. We say that Z_1 is in the domain of attraction of an α-stable law, denoted by $Z_1 \in D_\alpha$, if there exist real sequences a_n and μ_n for which

$$\frac{1}{a_n}\sum_{i=1}^{n}(Z_i - \mu_n)$$

converges in law to a stable distribution with parameter α. The following two lemmas give necessary and sufficient conditions for $Z_1 \in D_\alpha$.

TABLE IV
Asymptotic distribution theory for contamination models
Quantiles of the stable distribution

	Right tail probabilities				
α	0.1	0.05	0.01	0.005	0.001
1	6.7612	13.6373	65.653	129.7645	NA
1.05	-6.3069	-0.5517	40.1627	87.9868	445.4734
1.1	-0.5213	4.355	36.8106	73.2907	330.9228
1.15	1.1151	5.2897	31.5672	59.932	250.189
1.2	1.7651	5.3706	26.929	49.3496	192.8381
1.25	2.0504	5.1874	23.0733	41.0467	151.2448
1.3	2.1718	4.9181	19.8967	34.4791	120.4426
1.35	2.2118	4.628	17.2687	29.2202	97.1814
1.4	2.2089	4.3431	15.0761	24.9537	79.2942
1.45	2.1832	4.074	13.2282	21.4463	65.3058
1.5	2.1457	3.8242	11.6541	18.5251	54.1916
1.55	2.1028	3.5946	10.2983	16.0605	45.2258
1.6	2.0584	3.3845	9.1171	13.9535	37.8839
1.65	2.0147	3.1931	8.0759	12.1273	31.7795
1.7	1.9733	3.0195	7.1466	10.5209	26.6205
1.75	1.9352	2.8631	6.3068	9.0842	22.179
1.8	1.9011	2.7234	5.5394	7.774	18.2669
1.85	1.8716	2.6002	4.8357	6.5525	14.7128
1.9	1.8468	2.4931	4.2054	5.3942	11.3245
1.95	1.827	2.402	3.6825	4.361	7.787
2	1.8124	2.3262	3.29	3.6428	4.3702

Lemma A.1. For $0 < \alpha < 2$, $Z_1 \in D_\alpha$ if and only if $R(z) = P(Z_1 > z)$ is regularly varying with index $-\alpha$, i.e.,

$$\lim_{t\to\infty} \frac{R(tz)}{R(t)} = z^{-\alpha} \qquad \text{(A.2)}$$

What are suitable choices for μ_n and a_n? When $\alpha > 1$, the regular variation condition ensures that $E(Z_1)$ exists and is finite, so an appropriate choice of μ_n when $1 < \alpha < 2$ is $\mu_n = E(Z_1)$. When $0 < \alpha < 1$ a fitting option is $\mu_n = 0$. For $\alpha = 1$ a suitable centering sequence is

$$\mu_n = \int_{\min(a,0)}^{0} P(Z_1 \le t)dt + \int_{\max(0,a)}^{a_n} P(Z_1 > t)dt \,. \qquad \text{(A.3)}$$

For an appropriate a_n, let $s_\alpha = \Gamma(1-\alpha)\cos(\alpha\pi/2)$ for $0 < \alpha < 1$ and $s_\alpha = \dfrac{\pi}{2\Gamma(\alpha)\sin(\alpha\pi/2)}$ for $1 \le \alpha < 2$. A fitting sequence a_n of normalizing constants would satisfy the condition

$$\lim_{n\to\infty} n s_\alpha P(Z_1 > a_n) = 1 \qquad \text{(A.4)}$$

Table V shows suitable choices for a_n given certain tail behaviors. Condition A.4 also makes a_n an appropriate normalizing constant for $M_n = \max\{Z_1,...,Z_n\}$ since

$$P(M_n / a_n \le m) \to e^{-m^{-\alpha}/s_\alpha} \qquad \text{(A.5)}$$

i.e., the Fréchet distribution.

TABLE V
Asymptotic distribution theory for contamination models
Normalizing constants for some specific tail behaviors
$(0 < c < \infty, b > 0, 0 < \alpha < 2)$

Tail behavior as $z \to \infty$	Normalizing constant
$z^\alpha P(Z_1 > z) \to c$	$a_n = (cs_\alpha n)^{1/\alpha}$
$z^\alpha \sqrt{2\alpha \log z}\, P(Z_1 > z) \to c$	$a_n = \left(\frac{cs_\alpha n}{\sqrt{2\log n}}\right)^{1/\alpha}$
$z^\alpha \sqrt{2\alpha \log z}\exp\left(-b\sqrt{2\alpha \log z}\right) P(Z_1 > z) \to c$	$a_n = \left(\frac{cs_\alpha n}{\sqrt{2\log n}}\exp\left[b^2 + b\sqrt{2\log n}\right]\right)^{1/\alpha}$

Corollary A.2. If $Z_1 \in D_\alpha$ for $1 < \alpha < 2$, then

$$\left(\frac{s_\alpha}{s_{\alpha/2}}\right)^{2/\alpha} \frac{1}{a_n^2}\sum_{i=1}^{n} Z_i^2$$

converges in distribution to a stable law with parameter $\alpha/2$.

For $\alpha/2$, the normal case, let $\mu = E(Z_1)$ and $h(t) = \int_\mu^t (z-\mu)P(Z_1 > z)\,dz - \int_a^\mu (z-\mu)P(Z_1 \le z)\,dz$ and let a_n be a sequence of real numbers satisfying

$$\lim_{n\to\infty} \frac{nh(a_n)}{a_n^2} = 1 \qquad \text{(A.6)}$$

Note that if the variance of Z_1 exists and is finite then $h(t) \to \text{var}(Z_1)/2$ as $t \to \infty$, in which case $a_n = \sqrt{.5n\,\text{var}(Z_1)}$.

Lemma A.3. $h(t)$ is slowly varying, i.e., regularly varying of order 0, (therefore $\mu = E(Z_1) < \infty$) if and only if

$$\frac{1}{a_n}\sum_{i=1}^{n}\left(Z_i - \mu\right)$$

converges in law to a normal distribution with mean 0 and variance 2.

If $h(t)$ is slowly varying, we say that Z_1 is in the domain of attraction of a 2-stable law (normal distribution with variance 2). Table 6 gives suitable choices for certain tail behaviors with slowly varying $h(t)$.

TABLE VI

Asymptotic distribution theory for contamination models

Normalizing constants for some specific tail behaviors of distributions in the domain of attraction of the normal law

$(0 < c < \infty, b > 0)$

Tail behavior as $z \to \infty$	Normalizing constant
$z^2 P(Z_1 > z) \to c$	$a_n = \sqrt{.5cn\log n}$
$2z^2\sqrt{\log z}\,P(Z_1 > z) \to c$	$a_n = \sqrt{cn\sqrt{.5\log n}}$
$2z^2\sqrt{\log z}\exp\left(-2b\sqrt{\log z}\right)P(Z_1 > z) \to c$	$a_n = \sqrt{\frac{c}{2b}n\exp\left[b^2 + 2b\sqrt{.5\log n}\right]}$

B Asymptotics for the MLE of a *k*-point mixture

The next lemma states the asymptotic distribution for the *k*-point MLE of the mixing distribution *Q* when *Q* is a discrete probability measure on Θ with *k* distinct mass points, $\theta_1, ..., \theta_k$, and respective masses, $p_1, ..., p_k$. Here we assume that $\theta_1, ..., \theta_k$ are in the interior of Θ and that all the masses are positive and less than 1. The following notation will be used. Let $\underline{\theta} = \left(\theta_1, ..., \theta_k\right)$, $\underline{p} = \left(p_1, ..., p_{k-1}\right)$ and $\underline{\eta} = \left(\eta_1, ..., \eta_{2k-1}\right) = \left(\theta_1, ..., \theta_k, p_1, ..., p_{k-1}\right)$. Then,

$$f\left(x;\underline{\eta}\right) = f\left(x;\underline{\theta}, p\right) \equiv f_Q(x) = \sum_{i=1}^{k} p_i f_{\theta_i}(x)$$

$$= f_{\theta_k}(x) + \sum_{i=1}^{k-1} p_i\left(f_{\theta_i}(x) - f_{\theta_k}(x)\right)$$

Let $X_1, ..., X_n$, be iid from $f\left(x;\underline{\eta}\right)$ and let $\hat{\underline{\eta}}$ denote the mle based on $X_1, ..., X_n$. For $i, j = 1, ..., 2k-1$ let

$$I_{ij}\left(\underline{\eta}\right) = \text{cov}_{\underline{\eta}}\left(\frac{\partial}{\partial\eta_i}\log f\left(X_1;\underline{\eta}\right), \frac{\partial}{\partial\eta_j}\log f\left(X_1;\underline{\eta}\right)\right)$$

and let $I\left(\underline{\eta}\right)$ denote the information matrix whose $i^{th} - j^{th}$ entry is $I_{ij}\left(\underline{\eta}\right)$.

Lemma B.1. Under suitable regularity conditions,

$$\sqrt{n}\left(\hat{\underline{\eta}}_n - \underline{\eta}\right) \xrightarrow{D} MN\left(\underline{0}, I^{-1}\left(\underline{\eta}\right)\right)$$

Proof. See Lehmann (1983, Section 6.4) regarding suitable regularity conditions and a proof under those conditions.

Remark B.2. The "suitable regularity conditions" alluded to in Lemma B.1 involve the usual differentiability assumptions on $f\left(x;\underline{v}\right)$ and passing derivatives through expectations as well as the usual assumptions of positive definiteness of the information matrix. These hold for the examples we consider here. The main regularity condition that we need to verify is that

$$\left|\frac{\partial^3 \log f\left(x;\underline{v}\right)}{\partial v_a \partial v_b \partial v_c}\right| \le M_{abc}(x); a, b, c = 1, 2, ..., 2k-1$$

where $E\left(M_{abc}(X_1)\right) < \infty$

For mixtures these third order derivatives have been derived in a separate document (See http://people.clemson.edu/~veraf/docs/ThirdDerivativeEquations.pdf) If $p_i > 0$ for $i = 1, ..., k$, then some of the quantities envolved in these third order partial derivatives are bounded (see Proposition B.3 below). Therefore, a suficient regularity condition is that the absolute value of functions such as

$$\frac{\frac{\partial^3}{\partial\theta_i^3} f_{\theta_i}(x)}{f_{\theta_i}(x)}, \frac{\frac{\partial^2}{\partial\theta_i^2} f_{\theta_i}(x)}{f_{\theta_i}(x)}, \frac{\frac{\partial}{\partial\theta_j} f_{\theta_j}(x)}{f_{\theta_j}(x)},$$

$$\frac{\frac{\partial}{\partial\theta_i} f_{\theta_i}(x)}{f_{\theta_i}(x)}\frac{\frac{\partial}{\partial\theta_j} f_{\theta_j}(x)}{f_{\theta_j}(x)}\frac{\frac{\partial}{\partial\theta_h} f_{\theta_h}(x)}{f_{\theta_h}(x)}, \frac{\frac{\partial^2}{\partial\theta_i^2} f_{\theta_i}(x)}{f_{\theta_i}(x)},$$

$$\frac{\frac{\partial}{\partial\theta_i} f_{\theta_i}(x)}{f_{\theta_i}(x)}\frac{\frac{\partial}{\partial\theta_j} f_{\theta_j}(x)}{f_{\theta_j}(x)}, \frac{\frac{\partial}{\partial\theta_i} f_{\theta_i}(x)}{f_{\theta_i}(x)} \text{ are bounded}$$

by functions of *X* with finite expectation.

The next proposition shows that the derivative of $\log f\left(x;\underline{\theta},\underline{p}\right)$ with respect to p_i is bounded.

Proposition B.3.

$$\frac{f_{\theta_i}(x)-f_{\theta_k}(x)}{p_1 f_{\theta_1}(x)+\ldots+p_k f_{\theta_k}(x)}, i=1,\ldots,k-1$$

is bounded if $p_j>0$ for $j=1,\ldots,k$ (hence all its moments are finite).

Proof. Notice that

$$\frac{f_{\theta_i}(x)-f_{\theta_k}(x)}{p_1 f_{\theta_1}(x)+\ldots+p_k f_{\theta_k}(x)}=\frac{1}{p_i}\left(\frac{f_{\theta_i}(x)}{f_{\theta_i}(x)+\sum_{j\neq i}\frac{p_j}{p_i}f_{\theta_j}(x)}\right)-\frac{1}{p_k}\left(\frac{f_{\theta_k}(x)}{f_{\theta_k}(x)+\sum_{j\neq k}\frac{p_j}{p_k}f_{\theta_j}(x)}\right)$$

Therefore,

$$\left|\frac{f_{\theta_i}(x)-f_{\theta_k}(x)}{p_1 f_{\theta_1}(x)+\ldots+p_k f_{\theta_k}(x)}\right|\leq\frac{1}{p_i}+\frac{1}{p_k}$$

The essence of the proof of Lemma B.1 is a consequence of the following two Lemmas.

Lemma B.4.

$$\sqrt{n}\left(\underline{\eta}-\hat{\underline{\eta}}_n\right)=\frac{1}{\sqrt{n}}J_n\left(\underline{\eta}\right)\left(\frac{H_n\left(\underline{\eta}\right)}{n}\right)^{-1}+R_n$$

Where

$$R_n\xrightarrow{P}0, J_n\left(\underline{\eta}\right)=\left[\sum_{i=1}^{n}\frac{\partial\log f\left(X_i;\underline{\eta}\right)}{\partial\eta_j}: j=1,2,\ldots,2k-1\right]$$

, and

$$H_n\left(\underline{\eta}\right)=\left[\sum_{i=1}^{n}\frac{\partial^2\log f\left(X_i;\underline{\eta}\right)}{\partial\eta_j\partial\eta_l}: j,l=1,2,\ldots,2k-1\right]$$

Since $\frac{H_n\left(\underline{\eta}\right)}{n}\xrightarrow{P}-I\left(\underline{\eta}\right)$ it is immediate that.

Lemma B.5.

$$\sqrt{n}\left(\hat{\underline{\eta}}_n-\underline{\eta}\right)=\frac{1}{\sqrt{n}}J_n\left(\underline{\eta}\right)\left(I\left(\underline{\eta}\right)\right)^{-1}+Q_n$$

Where $Q_n\xrightarrow{P}0$.

This representation for $\hat{\underline{\eta}}_n$ will be needed in the final lemma.

Remark B.6. Note that if $n^{-.5}J_n\left(\underline{\eta}\right)$ converges in law to a multivariate normal distribution then $n^{-q}J_n\left(\underline{\eta}\right)$ converges in probability to 0 for $q>.5$, but convergence may be slow.

We now determine the asymptotic distribution of the LMP test statistic when parameters in the statistic are estimated by mle's. To do this we adopt the following notation. Let $\underline{\theta}=\left(\underline{\theta}_0,\underline{\theta}_1\right)$ where $\underline{\theta}_0=(\theta_{01},\ldots,\theta_{0l})$ indicates the parameters governing the background and $\underline{\theta}_1=(\theta_{11},\ldots,\theta_{1m})$ denotes the parameters activating the spurious observations. The vector $\underline{p}=\left(\underline{p}_0,\underline{p}_1\right)$ denotes the vector of mixing proportions and $p=\sum_{j=1}^{m}p_{1j}$ denotes the proportion assigned to the spurious distributions. Then $Q=(1-p)Q_0+pQ_1$. The *k*-point mle of Q is $\hat{Q}=(1-\hat{p})\hat{Q}_0+\hat{p}\hat{Q}_0$ where $\hat{Q}_0$ and $\hat{Q}_1$ are the discrete probability measures putting masses at the mle's of $\underline{\theta}=\left(\underline{\theta}_0,\underline{\theta}_1\right)$ and the respective vector of mle masses are $\hat{\underline{m}}=\left(\frac{\hat{\underline{p}}_0}{1-\hat{p}},\frac{\hat{\underline{p}}_1}{\hat{p}}\right)$. If Q_0 and Q_1 were known, then the LMP test statistic for testing $H_0: p=0$ versus $H_1: p>0$ would be given by

$$T_n=\sum_{i=1}^{n}\left(\frac{f_{Q_1}}{f_{Q_0}}(X_i)-1\right)\equiv\sum_{i=1}^{n}R(\underline{v})(X_i)$$

where $\underline{v}=\left(\underline{\theta}_0,\underline{\theta}_1,\underline{p}_0,(p_{11},\ldots,p_{1m-1})\right)$. This suggest using

$$S_n=\sum_{i=1}^{n}\left(\frac{f_{\hat{Q}_1}}{f_{\hat{Q}_0}}(X_i)-1\right)=\sum_{i=1}^{n}R(\hat{\underline{v}})(X_i)$$

as the test statistic for the more general formulation of the problem. For each *x*, let $G(\underline{v})(x)$ and $H(\underline{v})(x)$ denote the gradient vector and Hessian matrix of $R(\underline{v})(x)$ at $\underline{v}$ and assume that the entries in the Hessian matrix satisfies $\left|H_{ab}\left(\underline{\eta}'\right)(x)\right|\leq N_{ab}(x)$ for $a,b=1,2,\ldots,2k-1$ and for all $\underline{\eta}'$ in an open neighborhood of $\underline{\eta}$. Then a second order Taylor's series expansion gives

$$R(\hat{\underline{v}})(x)=R(\underline{v})(x)+G(\underline{v})(x)\bullet(\hat{\underline{v}}-\underline{v})^t+\frac{1}{2}\sum_{a,b}(\hat{v}_a-v_a)(\hat{v}_b-v_b)\gamma_{ab}(x)N_{ab}(x)$$

where $\left|\gamma_{ab}(x)\right|\leq 1$ and $\bullet$ is the inner product. So,

$$\frac{1}{\sqrt{n}}\left(S_n - T_n\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{R(\hat{\underline{\nu}})(X_i) - R(\underline{\nu})(X_i)\right\}$$

$$= \frac{1}{n}\sum_{i=1}^{n} G(\underline{\nu})(X_i) \bullet \sqrt{n}\left(\hat{\underline{\nu}} - \underline{\nu}\right)^t$$

$$+ \frac{1}{2}\sum_{a,b}\sqrt{n}\left(\hat{\nu}_a - \nu_a\right)\left(\hat{\nu}_b - \nu_b\right)\frac{1}{n}\sum_{i=1}^{n}\gamma_{ab}(X_i)N_{ab}(X_i)$$

$$\equiv G_n + R_n$$

If $E\left|G_a\left(\underline{\nu}\right)\left(X_i\right)\right|$ and $E\left|N_{ab}\left(X_i\right)\right|$ are finite for all *a* and *b*, then

$$\frac{1}{\sqrt{n}}S_n = \frac{1}{\sqrt{n}}T_n + E\left(G(\underline{\nu})(X_1)\right) \bullet \sqrt{n}\left(\hat{\underline{\nu}} - \underline{\nu}\right)^t + R_n$$

where $R_n \xrightarrow{P} 0$. Since $\hat{\underline{\nu}}$ is just a relabeling of $\hat{\underline{\eta}}$, it follows from Corollary B.5 that

$$\frac{1}{\sqrt{n}}S_n = \frac{1}{\sqrt{n}}T_n + E\left(G(\underline{\nu})(X_1)\right) \bullet \left(\frac{1}{\sqrt{n}}J_n\left(\underline{\nu}\right)\left(I(\underline{\nu})\right)^{-1}\right)^t + R_n' \quad \text{(B.1)}$$

where $R_n' \xrightarrow{P} 0$.

Next, we develop the distribution theory for S_n assuming that $I_0 < \infty$ and $W_0 < 1$. Recall that

$$T_n = \sum_{i=1}^{n}\frac{f_{Q_1}}{f_{Q_0}}\left(X_i\right) \quad \text{and}$$

$$J_n\left(\underline{\eta}\right) = \left[\sum_{i=1}^{n}\frac{\partial \log f\left(X_i;\underline{\eta}\right)}{\partial \eta_j} : j = 1,2,\ldots,2k-1\right] \quad \text{where}$$

$$E_{\underline{\nu}}\left(\frac{f_{Q_1}}{f_{Q_0}}\left(X_1\right)\right) \equiv \mu\left(\underline{\nu}\right) \quad \text{and}$$

$$E_{\underline{\nu}}\left(\frac{\partial \log f\left(X_1;\underline{\nu}\right)}{\partial \nu_j} = 0\right) \quad \text{for} \quad j = 1, 2, \ldots, 2k-1.$$ Let

$$c_{0,0}\left(\underline{\nu}\right) = \mathrm{var}_{\underline{\nu}}\left(\frac{f_{Q_1}}{f_{Q_0}}\left(X_1\right)\right),$$

$$c_{0,j}\left(\underline{\nu}\right) = c_{j,0}\left(\underline{\nu}\right) = \mathrm{cov}\left(\frac{f_{Q_1}}{f_{Q_0}}\left(X_1\right), \frac{\partial \log f\left(X_1;\underline{\nu}\right)}{\partial \nu_j}\right)$$

for $j = 1, 2, \ldots, 2k-1$ and $c_{i,j}\left(\underline{\nu}\right) = I_{i,j}\left(\underline{\nu}\right)$ for $i, j \neq 0$. The matrix $C\left(\underline{\nu}\right)$ of the $c_{i,j}\left(\underline{\nu}\right)$'s is the covariance matrix of the row vector

$$\left(\frac{f_{Q_1}}{f_{Q_0}}\left(X_1\right), \frac{\partial \log f\left(X_1;\underline{\nu}\right)}{\partial \nu_j}; j = 1,2,\ldots,2k-1\right) \equiv \left(\frac{f_{Q_1}}{f_{Q_0}}\left(X_1\right), \underline{V}\right)$$

. From (B.1) we need the covariance matrix of

$$\left(\frac{f_{Q_1}}{f_{Q_0}}\left(X_i\right), \underline{V}I^{-1}\left(\underline{\nu}\right)\right).$$ A straight-forward calculation shows that this matrix is $\sum\left(\underline{\nu}\right)$ of $\sigma_{i,j}\left(\underline{\nu}\right)$'s where

$\sigma_{0,0}\left(\underline{\nu}\right) = c_{0,0}\left(\underline{\nu}\right), \sigma_{i,j}\left(\underline{\nu}\right) = I_{i,j}^{-1}\left(\underline{\nu}\right)$ for $i, j \neq 0$, and the row vector

$$\underline{\sigma}_0 = \left(\sigma_{0,j}(\underline{\nu}); j = 1,2,\ldots,2k-1\right) = \left(c_{0,j}(\underline{\nu}); j = 1,2,\ldots,2k-1\right) I^{-1}\left(\underline{\nu}\right)$$

determines $\sigma_{0,j}\left(\underline{\nu}\right) = \sigma_{j,0}\left(\underline{\nu}\right)$ for $j = 1, 2, \ldots, 2k-1$. Thus, by the multivariate central limit theorem,

Lemma B.7.

$$\frac{1}{\sqrt{n}}\left(T_n - n\mu\left(\underline{\nu}\right), J_n\left(\underline{\nu}\right)I^{-1}\left(\underline{\nu}\right)\right) \xrightarrow{D} MN\left(\underline{0}, \sum\left(\underline{\nu}\right)\right)$$

.

The asymptotic distribution of S_n is immediate from Lemma B.7 and identity (B.1).

Lemma B.8.

Let $\underline{a} = \left(1, E\left(G(\underline{\nu})(X_1)\right)\right)$. Then,

$$\frac{1}{\sqrt{n}}\left(S_n - n\mu(\underline{\nu})\right) \xrightarrow{D} N\left(0, \underline{a}\sum(\underline{\nu})\underline{a}^t\right)$$

Next, we develop the distribution theory for S_n when $Z_1 = \left(f_{Q_1}(X_1) - f_{Q_0}(X_1)\right) \Big/ f_{Q_0}\left(X_1\right)$ has $\mathrm{var}\left(Z_1\right) = \infty$. We assume that Z_1 is in the domain of attraction of an α-stable law for some $0 < \alpha \leq 2$ under $\underline{\nu}$. Let a_n be defined as in Appendix A. Note that $n/a_n^2 \to 0$ since $\mathrm{var}\left(Z_1\right) = \infty$.

Identity (B.1) can be rewritten as

$$\frac{S_n}{a_n} = \frac{T_n}{a_n} + \frac{\sqrt{n}}{a_n}E\left(G(\underline{\nu})(X_1)\right) \bullet \left(\frac{1}{\sqrt{n}}Jn(\underline{\nu})\left(I(\underline{\nu})\right)^{-1}\right)^t + R_n'$$

For the next lemma, let $b_n = nE_p\left(Z_1\right)$ if $1 < \alpha \leq 2$, $b_n = \mu_n$ (as in (A.3)) if $\alpha = 1$, and $b_n = 0$ if $0 < \alpha < 1$.

Lemma B.9. If Z_1 is in the domain of attraction of an α-stable law, then $\frac{S_n - b_n}{a_n}$ converges in distribution to a stable law with parameter α.

Proof. The result follows since $\sqrt{n}/a_n \to 0$.

We now close Appendix B by considering the consistency of the mle at Q_0. To do this, we

consider the parameter space P of all discrete probability measures on Θ with at most k mass points. Endow P with the Levy metric and note that with this metric P is closed. Also assume that $f(x,\theta)$ is continuous in θ. Thus, $f_Q(x)$ is continuous in Q in the Levy metric. With this framework, one can apply Wald (1949) proof with minor modifications (in particular, see Section 4 of that paper) to show that the k-point mle converges almost surely to Q_0 in the Levy metric, d_L, when Q_0 obtains. Wald's parameter space (denoted by Ω in his paper while points in Ω are denoted there by θ not ω) is a subset of a Cartesian product space but his proof holds for the parameter space P considered here under his Assumptions 1, 2, 4-6 with his θ's replaced by our Q's.

Regarding Wald's assumptions with regard to the problem here, note that his Assumptions 3 and 8 hold since $f(x,\theta)$ is continuous in θ (supremums of lower semi-continuous functions are lower semicontinuous, and hence, measurable). Also Assumption 7 holds since P is closed. Note that the family $\{f(x,\theta)\}$ being identifiable does not imply that $\{f_Q(x)\}$ is identifable. E.g, mixtures over p of n-trial binomials with $k > 2n-1$ have an infinite number of representations. The limiting condition in Assumption 5 for $\{f(x,\theta)\}$ implies that $f(x,\theta)$ goes to zero as the points in the support of Q goes to infity or minus infity. This limiting condition lets Wald truncate the parameter pace to a bounded set which is then compact because of Assumption 7. Here the limiting condition implies tightness of a subset P' of P, and hence, compactness of this subset. The essence of Wald's proof is to use the compactness to get construct a finite open cover, $I_0, I_1, ..., I_m$ for P'. Without loss of generality, assume $Q_0 \in I_0$ but not in I_j for $j = 1, 2, ..., m$. This cover, defined for $\rho > 0$; is done in such a way that

1. $d_L(Q_1, Q_2) < \rho$ whenever Q_1 and Q_2 are in the same I_j,

2. $E_{Q_0} \sup\{\log f_Q(x) : Q \in I_0\} > E_{Q_0} \sup\{\log f_Q(x) : Q \in I_j\}$ for $j > 0$,

3. $E_{Q_0} \sup\{\log f_Q(x) : Q \in I_0\} > E_{Q_0} \sup\{\log f_Q(x) : \text{for } Q \text{ ouside the cover}\}$

for Q outside the cover g.

It follows from the strong law and (1-3) that $P(\lim d(\hat{Q}_n, Q_0) < \rho) = 1$ where $\hat{Q}_n$ is the k-point mle. Since ρ is arbitrary, $\hat{Q}_n$ converges almost surely to Q_0.

7. ACKNOWLEDGEMENTS

This work was initiated when all three authors were involved in the Anomaly Detection Working Group at the Statistical and Applied Mathematical Sciences Institute (SAMSI). The first author was a post-doctoral fellow at SAMSI and the National Institute of Statistical Science (NISS). The authors would like to thank SAMSI and NISS for a stimulating environment and support while there. In particular, support from SAMSI for all three authors was funded by NSF under Grant DMS-0112069 and support from NISS for the first author was funded under grant EIA0131884. In addition, the third author was partially supported by NSF Grants DMS 0243594 and DMS 0805809.

REFERENCES AND ELECTRONIC

[1]. **CHERNOFF, H.** (1954). On the distribution of the likelihood ratio. Annals of Mathematical Statistics 25 (3), 573-578.

[2]. **EFRON, B.** (2007). Size, power and false discovery rates. Annals of Statistics 35 (4), 1351-1377.

[3]. **FERGUSON, T. S.** (1967). Mathematical Statistics: A Decision Theoretical Approach. Academic Press, NY.

[4]. **GELUK, J. L. AND L. DE HAAN** (2000). Stable probability distributions and their domains of attraction: a direct approach. Probability and Mathematical Statistics 20 (1), 169-188.

[5]. **GREGO, J., H.-L. HSI, AND J. D. LYNCH** (1990). A strategy for analyzing mixed and pooled exponentials. Applied Stochastic Models and Data Analysis 6 (1), 59-70.

[6]. **HUANG, K.** (1963). Statistical mechanics. Wiley, NY.

[7]. **LEHMANN, E. L.** (1983). Theory of Point Estimation. Probability and mathematical statistics. Wiley, NY.

[8]. **LINDSAY, B. G.** (1983a). The geometry of mixture likelihoods: A general theory. Annals of Statistics 11 (1), 86-94.

[9]. **LINDSAY, B. G.** (1983b). The geometry of mixture likelihoods, part ii: The exponential family. Annals of Statistics 11 (3), 783-792.

[10]. **SINGH, D., P. G. FEBBO, K. ROSS, D. G. JACKSON, J. MANOLA, C. LADD, P. TAMAYO, A. A. RENSHAW, A. V. D'AMICO, J. P. RICHIE, E. S. LANDER, M. LODA, P. W. KANTOFF, T. R. GOLUB, AND W. R. SELLERS** (2002). Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1, 203-209.

[11]. **WALD, A.** (1949). Note on the consistency of the maximum likelihood estimate. The Annals of Mathematical Statistics 20 (4), 595-6

matemática
UNA PUBLICACIÓN DEL ICM – ESPOL

Octubre 2012 Volumen 10 Número 1

CONTENIDO