UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS FÍSICAS

Transiciones orden-desorden en dinámicas en conflicto vía la función beta generalizada en rango-frecuencia.

TESIS

QUE PARA OPTAR POR EL GRADO DE:

DOCTOR EN CIENCIAS (FÍSICA)

P R E S E N T A:

Roberto Carlos Álvarez Martínez.

DIRECTOR DE TESIS: Gustavo Carlos Martínez Mekler

MIEMBRO DEL COMITÉ TUTORAL: Adonis Germinal Cocho Gil

MIEMBRO DEL COMITÉ TUTORAL: Carlos Villarreal Luján

Ciudad Universitaria Mayo, 2012

A mi papá, Pedro.
A mis hermanos, Juan Manuel y Marlene.
A mi sobrino y ahijado, Emiliano.
A Etzel.

Resumen

Las leyes de potencia han sido aplicadas con éxito variado en una gran cantidad de campos para ajustar conjunto de datos en forma descendente de acuerdo a una propiedad, por ejemplo frecuencia de aparición de palabras en un texto, ingresos económicos, etc. Sin embargo, en la mayor parte de estos casos la ley de potencia no es válida para todo el conjunto de valores. En 2009 Germinal Cocho propuso una generalización de la ley de potencia, que involucra dos parámetros ajustables a cada conjunto de datos (a,b). La función Beta-Cocho ha mostrado un éxito impresionante en distintos conjuntos de datos que van desde la física, biología, pintura, arquitectura y redes sociales.

En este trabajo, estudié dinámicas en conflicto (permanencia-cambio), en tres distintos sistemas: un modelo estocástico, uno caótico-determinista y uno de evolución molecular. Determiné la importancia de un proceso sobre otro en los valores de los exponentes de (a,b) . Demostré que estas dinámicas ayudan a clasificar y entender los valores de a y b: con orden y correlaciones largas, el exponente a es mayor que b, después se atraviesa un transitorio, y con caoticidad, entropía máxima,desorden, se invierte esta relación, es decir; b es mayor que a. He demostrado que existen transiciones de fase de primer orden y segundo orden asociados a lo sistemas antes descritos, con ello podemos asignarle significado a los parámetros de ajuste e inclusive clasificar un conjunto de datos sólo ordenando y calculando su expresión en términos de la función Beta-Cocho.

Summary

The power laws have been applied with varying success in a number of fields to adjust the data set in descending order according to a property, such as frequency of words in texts, incomes, etc. However, in most of these cases the power law is not valid for the entire set of values. In 2009 Germinal Cocho proposed a generalization of the power law, which involves two adjustable parameters for each data set (a, b). Beta-Cocho function has shown an impressive success in different data sets coming from physics, biology, painting, architecture and social networks.

In this thesis, I studied conflict dynamics (persistence-change) in three different systems: a stochastic model, a chaotic-deterministic and one of molecular evolution. I stablished the importance of a process on another in the values of the exponents of (a, b). I showed that these dynamics help us to classify and understand the values of a and b: with order and long correlations, the exponent is greater than b, then through a transitional and caoticity, maximum entropy, and disorder, this relationship is reversed, ie, b is greater than a. I have shown that first and second-order phase transitions could be associated with the systems described above, thus we can assign meaning to the setting parameters and even classify a data set and only represented in terms of the Beta-Cocho function.

Índice general

Introducción

En los últimos años se han dedicado una gran cantidad de esfuerzos a entender las leyes de potencia, en particular áquellas relacionadas con las redes complejas y los fenómenos críticos. Sin embargo, cuando los datos se analizan, en la mayoría de los casos, las leyes de potencia sólo son válidas en un cierto rango de valores; en general, existe una caída pronunciada en la cola de la distribución. Se han esgrimido distintas explicaciones a este fenómeno, tales como efectos de tamaño finito, constricciones en el crecimiento de la red, los fénomenos cerca del punto crítico ($|T - T_c| = \epsilon \neq 0$), etc. Por ello se han propuesto diferentes correcciones a la ley de potencia: de tipo exponencial gaussiana, exponencial estirada, ley de potencia con corte exponencial, entre otras. En esta tesis, nos dedicamos al estudio de distribuciones orden-tamaño u orden-frecuencia, ya que muchas de éstas distribuciones siguen una ley de potencia.

Hemos encontrado que una cantidad sorprendentemente grande de datos graficados descendentemente de acuerdo a alguna propiedad (frecuencia de aparición, tamaño, etc), se pueden ajustar con una distribución bi-paramétrica que está constituida por el producto de dos leyes de potencia definidas sobre todo el rango de datos. Una de estas leyes determina el comportamiento para valores pequeños y la otra, se ajusta para valores cercanos al número máximo de los elementos ordenados. Esta función que llamaremos función Beta-Cocho (FBC) en general mejora los ajustes de muchas de las correcciones propuestas anteriormente en la literatura.

El principal resultado de esta tesis consiste en asignar significados a los valores de ajuste de los parámetros de la FBC mediante la consideración de dinámicas en conflicto, que genéricamente se caracterizan por la dualidad permanencia-cambio. Cuando la permanencia domina al sistema, entonces el valor de un parámetro domina sobre el otro, existe una región de transición y después, cuando se invierten los procesos, se cambian los valores relativos

de los ajustes.

En el primer capítulo presentamos la FBC y la comparamos con otras funciones que han sido propuestas en la literatura para corregir la ley de potencia como primera, y a veces única, aproximación a un conjunto de datos ordenados descendentemente de acuerdo a una propiedad (frecuencia, tamaño,etc). Este comparativo lo realizamos mediante una medida de información que permite establecer, que en un amplio conjunto de datos, la FBC mejora a las usadas en la literatura.

El segundo capítulo muestra un proceso de generación de secuencias binarias propuesto por W. Li, que contiene dos sub-procesos: expansión y modificación. Con ellos, controlados por la probabilidad de mutación, es posible hacer una estadística sobre las secuencias de n-elementos consecutivos y hacer un ajuste orden-frecuencia. Se observa una transición orden-desorden cuando los valores de los parámetros se igualan.

El tercer capítulo presentamos familias de mapeos unimodales caóticos, en los cuales la variación de un parámetro le confiere una transición de estados intermitentes a caóticos. Se utiliza la dinámica simbólica y el formalismo termodinámico de los sistemas dinámicos para demostrar la existencia de una transición de fase relacionada a los intercambios en los valores relativos de los parámetros de ajuste de la FBC y a la divergencias en la densidad invariante.

En el cuarto capítulo, a manera de ilustración, se estudia un modelo de evolución molecular propuesto por Eigen y Schuster que contiene dos ingredientes de evolución para polímeros: mutación y replicación; con ellos es posible abordar, mediante una transformación de coordenadas, el equivalente a un sistema de espínes y la existencia de transiciones de fase se demuestra fehacientemente y se revela la relación entre este proceso y los parámetros de ajuste de la FBC, apuntalando así los resultados previos sobre los significados de éstos.

Capítulo 1

La función Beta-Cocho

En este capítulo se presentan distintas opciones para corregir las discrepancias entre los datos y las distribuciones orden-frecuencia u orden tamaño, ajustados en primera instancia mediante leyes de potencia. Definimos la función de distribución Beta-Cocho (FBC), para modelar las relaciones clasificación-tamaño o clasificación-frecuencia en un amplio conjunto de orígenes muy diversos de datos. La función FBC generaliza la ley de potencia, la función de distribución uniforme, entre otras.

Realizamos un análisis comparativo de distintas funciones de ajuste reportados en la literatura para establecer la prevalencia de la FBC sobre el resto. Para discriminar entre modelos de ajustes con distinto número de parámetros utilizamos el criterio de información de Bayesiana (CIB), de esta forma, se considera, no sólo la bondad del ajuste, sino que también se toma en cuenta el número de parámetros involucrados en los modelos propuestos.

1.1. Leyes de potencia e invariancia de escala

Se denomina ley de potencia a la relación existente entre dos variables y y x cuando se satisface la siguiente expresión $y \approx x^{\alpha}$, con α una constante. En la física básica existen muchas relaciones de este tipo entre ellas: la ley de gravitación universal, la ley de Hooke, la ley de Coulomb, sólo por mencionar algunas. Sin embargo, en este trabajo las leyes de potencia que consideraremos serán las que están asociadas a variables estocásticas, es decir a la relación entre una variable y su función de distribución de probabilidad

.[1] Por lo que, para este caso, una ley de potencia se establece si la relación existente entre una variable estocástica x y su función de distribución de probabilidad (fdp) $f(x)$, satisface que

$$f(x) \sim x^{\alpha}, \tag{1.1}$$

con α el parámetro de escalamiento.

A este tipo de relaciones, a menudo, se les denomina también distribuciones con "colas largas"; debido a que, a diferencia de las funciones de distribución gaussiana o normal, poisson etc. (unimodales), la ley de potencia tiene una probabilidad significativamente distinta de cero para valores grandes de la variable x .

Una de las propiedades más importantes de este tipo de distribuciones estadísticas es que no está determinada por una escala "característica", es decir que el promedio $\langle x \rangle$ (o algún momento de orden superior) no es representativo de la variable aleatoria. Debido a la ausencia de esta escala, comúnmente también a estas distribuciones se les denomina como "sin escala" (*scale-free*).

Las leyes de potencia se presentan en fenómenos diversos y se han utilizado para caracterizar una miríada de fenómenos naturales [1, 2, 3, 4, 5]. La propiedad de ausencia de escala es una de las leyes de invariabilidad en la naturaleza, con frecuencia subestimada en ramas distintas de la física; es precisamente está característica que la hace atractiva y propicia para la modelación de fenómenos que involucran "memoria de largo alcance", puntos críticos, transiciones de fase, entre otros.

1.2. Importancia de las leyes de potencia

1.2.1. Aparición y ejemplos

Históricamente el interés en el estudio de la leyes de potencia se inició en 1913 con el geógrafo alemán Auerbach [6] quien publicó un trabajo sobre las concentraciones de poblaciones humanas. Auerbach detectó que existe una relación entre la ordenación descendente (en inglés *ranking*) de las poblaciones de ciudades y su número de habitantes. De esta manera, le asignó el el número 1 a la más poblada, el número 2 a las segunda más poblada y así sucesivamente. Siguiendo este proceso de ordenación conforme al tamaño

[1] Por extensión también consideraremos leyes de potencia en la función acumulada de probabilidad y en orden-frecuencia u orden-tamaño.

y gratificándolo con respecto a su *ranking* se obtiene una ley de potencia. A este tipo de gráficas se les llama orden-tamaño (*rank-size*). El matemático norteamericano Lotka en 1924 [7], aplicó este método para poblaciones en los Estados Unidos y obtuvo resultados concordantes con el trabajo de Auerbach. Posteriormente, Singer [8], aplicó el ordenamiento descendente para encontrar que las rentas de los individuos se ajustaban bien a un caso más general propuesto por Pareto [9] . Los franceses, por su parte, acostumbran a citar la obra de R. Gibrat [10] como un claro precedente de la formulación posterior y más concoida de Zipf. Es por ello que en Francia a las leyes de potencia que surgen de estos estudios se les conoce como "Ley de Gibrat".

En el campo de la lexicografía, el lingüista estadounidense Georg Kingsley Zipf en 1949 [11], encontró otra ley de potencia. Zipf estudió textos extensos y escogió como propiedad a estudiar la frecuencia de aparición de palabras distintas. Clasificó en orden descendente y encontró una relación entre la frecuencia y su clasificación: así en el idioma inglés la palabra más utilizada es *the*, la segunda es *it*, que les corresponden respectivamente las clasificaciones 1 y 2. Debido a la amplia difusión del trabajo de Zipf, la relación entre frecuencia de repetición y clasificación se le denomina *ley de Zipf*. La formulación original fue la siguiente:

$$f(r) \approx \frac{1}{r \ln(1.78R)} \tag{1.2}$$

con $r = 1, 2, 3, \cdots, R$, R el número máximo de palabras distintas y $f(r)$ la frecuencia de aparición de la palabra r-ésima. En el caso de otros idiomas el parámetro de escalamiento es cercano a $\alpha \approx 1$. Zipf, a pesar de no ser el primero en establecer relación orden-frecuencia u orden-tamaño, ejerció una influencia decisiva para popularizar el concepto. Es por ello que generalmente se asocia su nombre a la relación entre frecuencia y orden; que se ha extendido para considerar cualquier relación entre el orden-frecuencia u orden-tamaño que sea descrita mediante una ley de potencia a pesar de que el exponente de escalamiento difiera de uno.

En las décadas más recientes las leyes de potencia han cobrado una mayor relevancia debido a su presencia en redes complejas [12], por ejemplo aparecen en la distribución de conectividades (de entrada y salida –para redes dirigidas–), en redes moleculares, de regulación génica, www, el internet, la red de contactos sexuales de seres humanos, redes ecológicas, entre otras [13, 14]. Es tal la robustez de este resultado, que es considerado el estándar de facto para la construcción de cierto tipo redes de coexpresión de tal forma

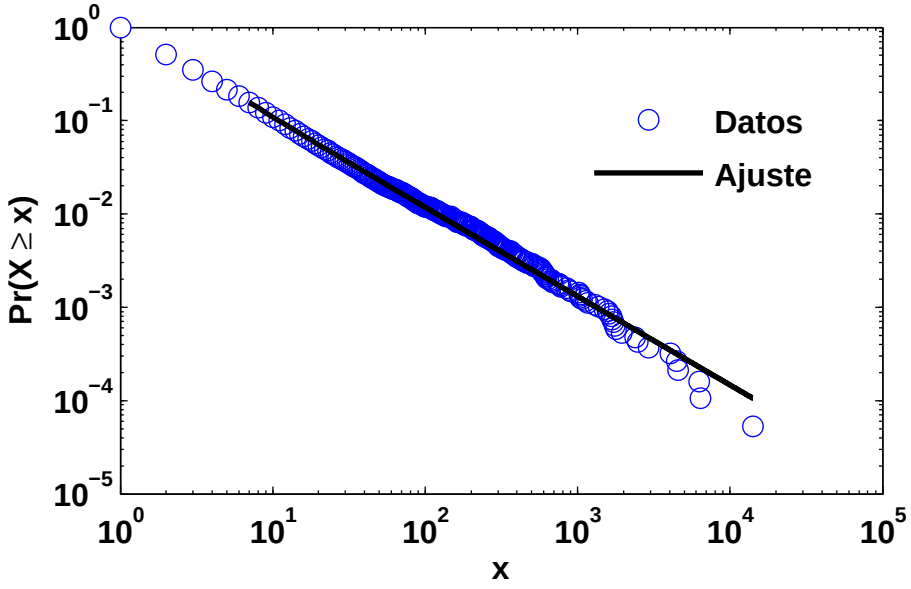


Figura 1.1: Ley de Zipf de las palabras en la novela en inglés de Herman Mellville "Moby Dick" Datos tomados de: http://tuvalu.santafe.edu/ aaronc/powerlaws/data/words.txt.

que el algoritmo de creación utiliza el hecho de que la distribución de enlaces debe ser una ley de potencia [15] . En la mayor parte de los fenómenos de la naturaleza, en donde se obtienen invarianzas de escala, el parámetro de escalamiento se encuentra entre $1 < \alpha < 3$ [16].

1.2.2. Algunos mecanismos de generación

A pesar de su, recientemente incrementada, importancia no se conocen hasta ahora mecanismos universales –válidos para todos los fenómenos– para la generación de leyes de potencia.

Sin embargo, es posible resaltar dos procesos como los más importantes en la generación de leyes de potencia, en lo que se refiere a fenómenos físicos:

- Conexión preferente (*Preferencial attachment*) Es un proceso mediante el cual unidades discretas se agregan de manera aleatoria o parcialmente aleatoria a elementos previos, de forma que se acumulan proporcionalmente al número que ya pertenecen a ellas. Este mecanismo fue propuesto originalmente para determinar el crecimiento de especies que pertenecen a un género más extenso por Yule en 1925 [17]. Este proceso también es conocido por algunas frases que describen, ya sea el mecanismo de aglomeración o el conjunto de datos involucrados. Entre ellas se encuentran las siguientes: "el rico se vuelve más rico", "Proceso San Mateo"[2], "Proceso de Yule" , "Ley de Gibrat " o "ventaja acumulativa". La distribución de probabilidad obtenida por el proceso de Yule lineal es una función de distribución beta

$$P(k) = \frac{\mathrm{B}(k + a, \gamma)}{\mathrm{B}(k_0 + a, \gamma - 1)}$$

para $k \geq k_0$, con $B(x, y)$ la función beta:

$$\mathrm{B}(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x + y)}$$

,

en donde $\Gamma(x)$ es la función gama estándar, y

[2]Proviene de una cita del Evangelio de Mateo :"Porque al que tiene, le será dado, y tendrá más; y al que no tiene, aún lo que tiene le será quitado."

$$\gamma = 2 + \frac{k_0 + a}{m}$$

.

La función beta asintóticamente se comporta como $B(x, y) \sim x^y$ para $x >> 1$ y y fija [18], con ello.

$$P(k) \propto k^{-\gamma}.$$

Yule y después Simon demostraron que este mecanismo produce una ley de potencia en la cola de la distribución. El término *preferential attachment*, fue acuñado por Réka Albert y Albert-Lazlo Barábasi al encontrar distribuciones de ley de potencia en redes complejas asociadas al internet, la www, la red de colaboraciones de actores, redes ecológicas, redes metabólicas, entre otras, en donde en particular, la distribución de la probabilidad de que un nodo al azar tenga k conexiones es una ley de potencia $P(k) \propto k^{\alpha}$ [13, 14].

Este proceso ha sido aplicado para explicar las distribuciones de crecimiento en especies, ciudades, libros más vendidos, ingresos económicos, entre otros.

- El segundo mecanismo consiste en los fenómenos críticos asociados al concepto de criticalidad y criticalidad auto-organizada [19], en el cual la longitud de correlación del sistema diverge, ya sea debido a que se le ha conducido externamente al sistema al punto crítico en su espacio de parámetros o debido que el sistema automáticamente se dirige por si mismo a ese punto mediante un proceso dinámico (auto-organizado). La divergencia deja al sistema sin un factor de escala característico de alguna cantidad medible y eso determina un punto ausente de escala característica y con ello la aparición de leyes de potencias que caracterizan estas divergencias alrededor del punto crítico.

Las leyes de potencia establecen una simetría del sistema, la autosimilaridad, es decir invariabilidad ante cambios de escala del sistema. Es decir, que no se puede definir una longitud representativa del sistema sino que existen en este estado todas las escalas posibles. El origen y generación de fenómenos de multiescalidad [16] no está aún aclarado.

1.2.2.1. Leyes de potencia en un rango acotado

El uso de las leyes de potencia es extenso pero, no siempre justificado. Se aduce, por ejemplo, que es sólo válida en la cola de la distribución, es decir, que existe un límite mínimo en la variable de orden x_{min} a partir del cual se obtiene una ley de potencia. Es por ello que a menudo las leyes de potencia se encuentran y se calculan en la cola de la distribución. En este proceso, frecuentemente, se omiten los primeros n datos en la en el cálculo del parámetro de escalamiento α, haciendo de ello un proceso artesanal y por lo tanto altamente subjetivo [20, 16].

Encontrar evidencia de la existencia de leyes de potencia se ha convertido en un proceso que, muchas veces, se realiza inadecuadamente, con descuido y sin fundamentos claros que demuestren su existencia para un conjunto de datos. Ejemplo de ello son los datos que han sido publicados [20, 16] y que pueden verse en la figura 1.2 en los que se señala la presencia de una ley de escala, sin embargo resulta, cuando menos, poco justificada y, a la vista de los resultados mostrados, poco convincente. En estos casos no hemos omitido dato alguno, por lo que los ajustes se realizan sobre todo el conjunto completo.

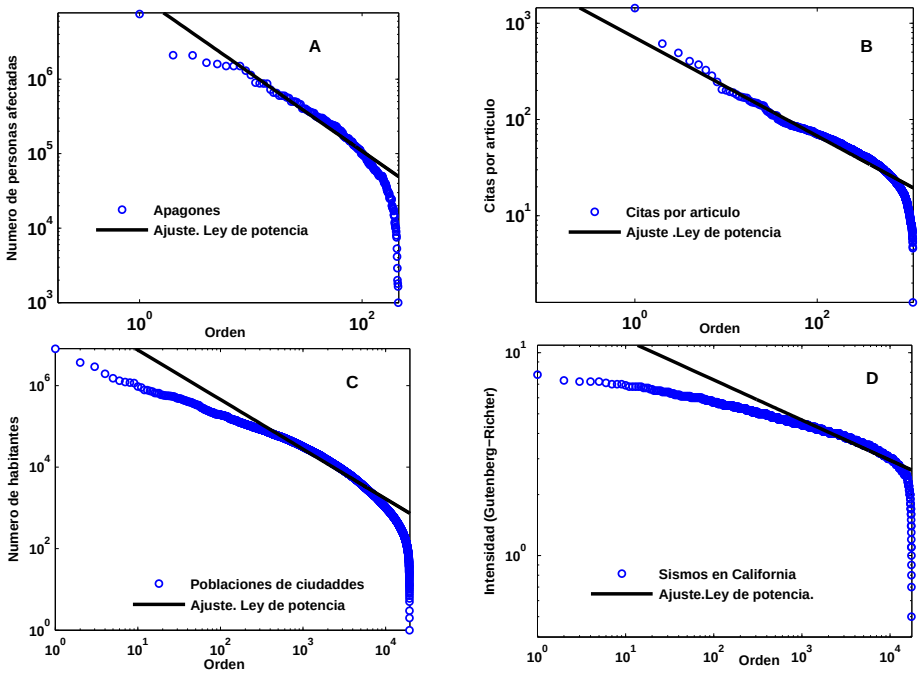


Figura 1.2: Ajustes de orden-tamaño en escala log-log. **A**. Número de personas afectadas por fallas de energía eléctrica en Estadros Unidos de América entre 1984 y 2002 [16]. Fuente: http://tuvalu.santafe.edu/aaronc/powerlaws/data/quakes.txt. **B**. Número de citas por artículo recibidas entre 1981 y Junio de 1997 reportados en el *Science Citation Index* [21] Fuente: http://physics.bi.edu/redner/projects/citation/physicsbypersons.html. **C**. Número de habitantes de poblaciones registradas en el censo 2000 en Estados Unidos de América. Fuente: http://www.census.gov/main/www/cen2000.html. **D**. Intensidad de sismos en California (EUA) reportados entre 1910 y 1992[16].

A pesar del uso muy extendido de los ajustes a una ley de potencia de un conjunto de datos, en una buena parte de estos trabajos no se demuestra cabalmente la pertinencia de dicho ajuste mediante un criterio bien establecido[3]. En muchas ocasiones se supone como la mejor (y con frecuencia la única)

[3]Si bien es cierto, que no existe un criterio perfecto, el hecho de no adoptar ninguno,

manera de comprobar la presencia de una ley de potencia la visualización de una gráfica en log-log y el ajuste de una recta en un intervalo acotado del conjunto de datos y la comprobación de que esta recta tiene un coeficiente de correlación "suficientemente cercano" a la unidad ($R^2 \approx 1$), sin embargo este método no es suficiente, y la estimación del exponente lleva a errores importantes [16].

Con frecuencia, también se apela a una modificación de la ley de potencia, introduciendo, otro(s) parámetro(s) para tomar en cuenta, la caída de la distribución para valores grandes de x .

La generalización de leyes de potencia ha sido estudiada ampliamente, una aportación de este trabajo consiste en estudiar los mecanismo generatrices de una función que, entre otras cosas, generaliza la ley de potencia:

la función de distribución Beta-Cocho.

1.3. La FPA y la orden-frecuencia

La relación entre función orden-frecuencia(tamaño) y las FDP se obtiene mediante la función de probabilidad acumulada (FPA). Dada una FDP, $\rho(x)$, la FPA denotada por Y(x) define mediante la siguiente expresión:

$$Y(x) = \int_x^\infty \rho(z)dz \tag{1.3}$$

de esta forma $Y(x)$ mide la probabilidad de que la variable estocástica X al menos sea x, $(P(X \geq x) = Y(x))$. Una de las ventajas de la FPA es que no es necesrio encontrar un tamaño de *bin* típico, a diferencia de las FDP que se encuentran . Si consideramos el arreglo orden-frecuencia descendentemente, entonces al considerar la r-ésimo elemento más frecuente, entonces hay r elementos con mayor o gual frecuencia que ésta. De esta forma, la relación entre las funciones orden-frecuencia y la FDA es clara: existe una relación de proporcionalidad entre la FDA y el orden (*ranking*), como se formaliza con la siguiente ecuación:

La parte entera de $NY(x)$ es el número esperado de valores más grandes o iguales a x, de tal forma que x es de hecho el r-ésimo valor más grande (*ranking*):

---

más allá del visual , induce muchos errores tanto la existencia de una ley de potencia como en la determinación del exponente de escalamiento

$$NY(x_r) = r \tag{1.4}$$

con $x_1 > x_2 > x_3 > \ldots N$ y $r = 1, 2, 3, \ldots N$

con ello podemos encontrar la función orden-frecuencia F invirtiendo la relación anterior

$$x \propto Y^{-1}(r) \tag{1.5}$$

En este trabajo utilizaremos la funciones orden-frecuencia para ajustar los datos subsiguientes.

1.3.1. Ley de potencia en orden-fecuencia

Por ejemplo, para el caso de la FDP en forma de ley de potencia, encontraremos, la FDP está dada por la ecuación 1.1, de tal forma que la FPA entonces quedá definida como

$$Y(x) = A \int_x^\infty z^\alpha dz \tag{1.6}$$

se observa que para que esta integral converja en el intervalo $-\infty < z < \infty$ $\alpha < -1$, amneos que se establezca un z_{min} a partir del cual sea válida la ley de potencia. Con esto $Y(x)$ está determinada por

$$Y(x) = Ax^{\alpha+1} \tag{1.7}$$

de tal forma que la función orden-frecuencia para la ley de potencia $f(r;\alpha)$,consiste en invertir la relación anterior , por lo que va como

$$x \propto f(r;\alpha) \propto r^{\frac{1}{\alpha+1}} \tag{1.8}$$

la ecuación anterior 1.8 muestra que la ley de potencia está relacionada con una ley de potencia para orden-frecuencia, pero con exponentes distintos $\alpha \to \frac{1}{\alpha+1}$[4]

[4]La FPA también es una ley de potencia con exponente $\alpha + 1$

1.4. Función Beta-Cocho. Definición

Germinal Cocho [22] propuso una función de ajuste bi-paramétrica, que debido a que en su forma funcional recuerda a la función de distribución beta, la llamaremos función Beta-Cocho(FBC) y está definida de la siguiente manera [23]:

$$f(r) = A\frac{(N+1-r)^b}{r^a} \tag{1.9}$$

Aquí N es el número máximo de valores por clasificar , f es la propiedad medida (frecuencia, tamaño, áreas,etc.), r la clasificación $r = 1, 2, \cdots, N$, (usamos la letra r por el *ranking*), A es una constante de normalización y a, b son los parámetros libres cuyos valores se ajustan a cada conjunto de datos.

En la figura 1.3 se muestra la forma de la función beta para distintos valores de los parámetros a y b, con ello mostramos la singular forma sigmoide de la beta que se ha encontrado en un conjunto de datos provenientes de una variedad muy amplia de campos.

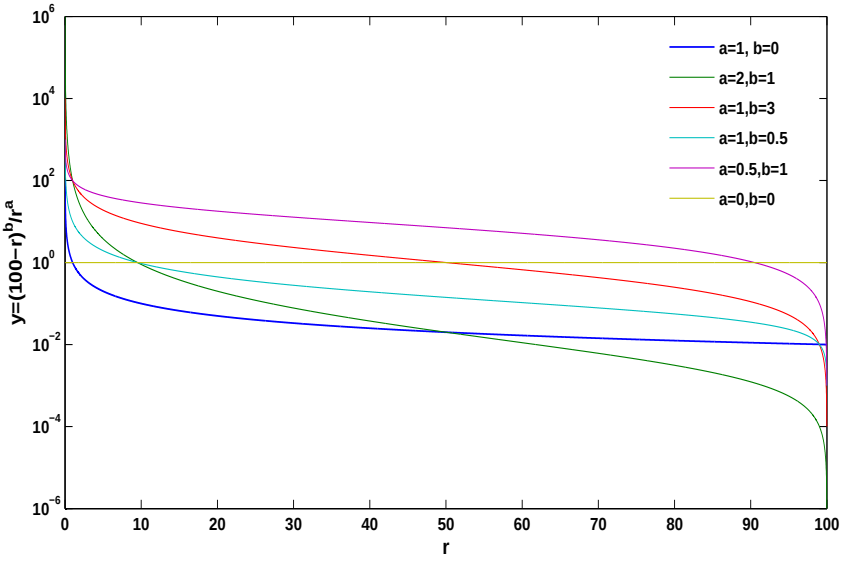


Figura 1.3: Función beta para distintos valores de los párametros a y b

1.4.1. Propiedades

La normalización de la FBC conduce a:

$$A = [\sum_{r=1}^{N} \frac{(N+1-r)^b}{r^a}]^{-1}, \tag{1.10}$$

y si $N \gg 1$, entonces $A = \frac{\Gamma(a+b+2)}{\Gamma(a+1)\Gamma(b+1)}$

Como se mostró en la figura anterior, la FBC tiene una gran flexibilidad para producir distintas distribuciones (orden-frecuencia) conocidas, por ejemplo:

- Con $b \approx 0$ la ecuación 1.9 se reduce a una ley de potencia con $\alpha = a$, este es el comportamiento que domina para $r \to 0$, por otra parte, cuando el rango se acerca a su máximo $r \to N$, el comportamiento refleja un efecto del tamaño finito de los elementos distintos del sistema (palabras, letras, número de personas, etc.).

- Si $a = b = 0$, tenemos enemos una distribución uniforme .

- Con $a = b$, se obtiene la distribución rango-frecuencia de Lavalette

$$F(r) = C(\frac{Nr}{N-r+1})^{\alpha} \tag{1.11}$$

  Esta distribución se ha utilizado para ajustar los factores de impacto de revistas científicas, la intensidad de terremotos, la frecuencia de aparición de dígitos en series numéricas, entre otras. Así, la función anterior es sólo un caso particular de la FBC [24].

La forma de la FBC tiene una gran flexibilidad que depende de los valores de los parámetros a, b.

Se han observado excelentes ajustes a la FBC en conjuntos de datos de orígenes muy diversos, por ejemplo, en el factor de impacto de revistas científicas [22], frecuencia de aparición de las notas musicales en las partituras de obras de distintos géneros, [25], frecuencia de ideogramas en el mandarín [26] , frecuencia de palabras en discursos presidenciales [27, 28], tamaños de los motivos en pinturas, tamaños de poblaciones de ciudades, número de colaboradores de actores de cine, clasificación de universidades,

redes genéticas, frecuencia de aparición de nucleótidos en el ADN, distancia por carretera entre ciudades, etc. [23]

En la figura 1.4 se observa una muestra de los ejemplos en donde hemos encontrado que la FBC ajusta de manera muy satisfactoria un conjunto de datos variados tanto en su origen como en las dinámicas subyacentes (cuando existen o se conocen). Los coeficientes de determinación en la mayor parte de los casos son muy buenos ($R^2 \approx 1$. Al final del capítulo empleamos un método para comparar la "bondad" de los ajustes entre distintos modelos).

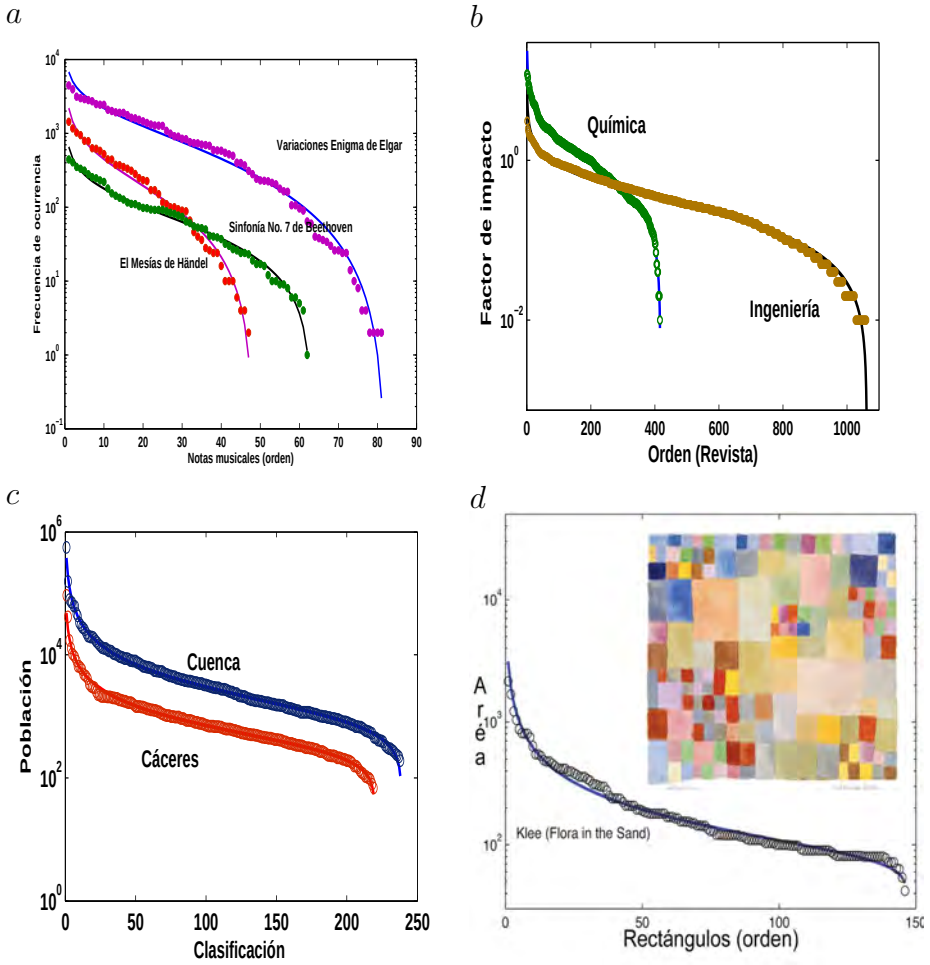


Figura 1.4: Distribuciones orden-frecuencia. **a**. Frecuencia de aparición de las notas musicales en las partituras de: "Variaciones Enigma" de E. Elgar (círculos violeta), la "Sinfonía número 7" de L. Beethoven (círculos verdes) y el "Mesías" de F. Händel (círculos rojos). **b**. Factor de impacto de las revistas científicas en el área de Química e ingeniería. **c**. Número de habitantes de las poblaciones de los municipios de Cuenca, círculos azules y Cáceres círculos rojos (fuente: Instituto Nacional de Estadística de España 2009). **d**. Áreas de los rectángulos de la pintura *"Flora in the Sand"* ordenados descendentemente. En todos los casos las curvas continuas son los ajustes del conjunto de datos correspondiente con la FBC.

En la tabla 1 se muestran otros ejemplos para los cuales se ha aplicado

la FBC para distintos conjuntos de datos.

Tabla 1. Ajustes de la FBC

Datos	**a**	**b**	$\mathbf{R}^2$
Frecuencia de letras en el idioma inglés	0.18	1.31	0.97
Potencial eléctrico en la corteza cerebral de un gato	0.08	0.24	0.970
Red de colaboradores de actores de cine	0.71	0.61	0.99
Clasificación académica de universidades	0.37	0.43	0.99
Factor de impacto de revistas de ciencias de la vida	0.59	0.83	0.99
Población de la municipalidad de Zaragoza (España)	0.95	0.54	0.99
Población de la municipalidad de Valladolid (España)	0.98	0.42	0.99
Red genética regulatoria de E.Coli	0.99	0.39	0.98
Distancia autopistas de Guanajuato a las principales ciudades de México	1.52	3.87	0.99
Magnitud de los desplomes (*crashes*) del mercado de valores de los E.U.	3.56	0.11	0.98
Diámetros de los círculos de la pintura "Several circles" de Kadinsky	0.62	0.32	0.98
Frecuencia de aparición de las notas en la " 5 Sinfonía de Beethoven	0.42	1.25	0.99
Frecuencia de aparición de las notas en "Million Dollar Baby"	0.23	1.54	0.99
Áreas de ocupación de especies de plantas en campos de Illinois	0.88	0.76	0.98

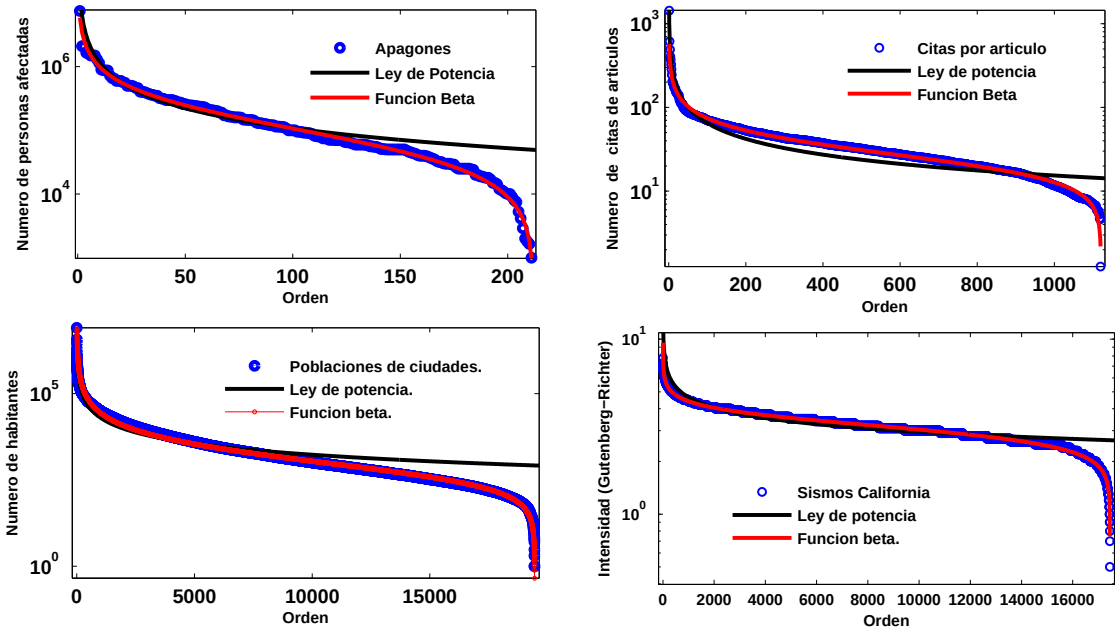


Figura 1.5: Comparativo de ajustes entre la ley de potencia y la función beta. Los datos son los mismos – para efectos comparativos– que los reportados en la figura 1.2

1.5. Método de ajuste

Para ajustar un conjunto de datos a la FBC utilizamos el método de los mínimos cuadrados no lineales [29] para calcular los valores de los parámetros a y b. Este método es iterativo por lo que la convergencia del método depende de los valores iniciales. La forma de elegir éstos es la siguiente, tomamos la expresión de la FBC y calculamos el logaritmo:

$$\ln f(r) = \ln A + b \ln(N + 1 - r) - a \ln r$$

entonces la expresión anterior se puede escribir de la siguiente manera que sugiere que se trata de una regresión lineal múltiple mediante el método de los mínimos cuadrados no lineales:

$$y = B + bx_1 + ax_2 \tag{1.12}$$

en donde $y = f(r)$, $B = \ln A$, $x_1 = \ln(N + 1 - r)$ y $x_2 = \ln r$ así que se hace una regresión estándar para calcular los valores de los valores estimados a y b, que denotamos $\hat{a}_{lineal}$ y $\hat{b}_{lineal}$

$$\hat{a}_{lineal} = (N(\sum_{i=1}^{N} x_{1i}y_i \sum_{i=1}^{N} x_{2i} - \sum_{i=1}^{N} x_{1i}x_{2i} \sum_{i=1}^{N} x_{2i}y_i) - \sum_{i=1}^{N} y_i(\sum_{i=1}^{N} x_{1i} \sum_{i=1}^{N} x_{2i}^2 - \sum x_{1i}x_{2i} \sum_{i=1}^{N} x_{2i}) + \sum_{i=1}^{N} x_{2i}(\sum_{i=1}^{N} x_{1i} \sum_{i=1}^{N} x_{1i}y_i - \sum_{i=1}^{N} x_{2i} \sum_{i=1}^{N} x_{1i}y_i))/\delta$$

$$\hat{b}_{lineal} = (N(\sum_{i=1}^{N} x_{1i}^2 \sum_{i=1}^{N} x_{2i}y_i - \sum_{i=1}^{N} y_i \sum_{i=1}^{N} x_{1i}x_{2i}) - \sum_{i=1}^{N} x_{1i}(\sum_{i=1}^{N} x_{1i} \sum_{i=1}^{N} x_{2i}y_i - \sum_{i=1}^{N} x_{1i}y_i \sum_{i=1}^{N} x_{2i}) + \sum_{i=1}^{N} y_i(\sum_{i=1}^{N} x_{1i} \sum_{i=1}^{N} x_{1i}x_{2i} - \sum_{i=1}^{N} x_{1i}^2 \sum_{i=1}^{N} y_i))/\delta$$

en donde x_{1i} el i-ésimo dato correspondiente a la variable x_1, significado análogo tiene x_{2i} ,y_i es el i-ésimo dato de la variable y y δ es el determinante del sistema de ecuaciones lineales generadas al minimizar el cuadrado de los errores. En nuestro caso $x_{1i} = \ln(N+1-i)$ y $x_{2i} = \ln i$, para $i = 1, 2, 3 \ldots N$, con estos valores estimamos $\hat{a}_{lineal}$ y $\hat{b}_{lineal}$ que se usan como valores iniciales en la regresión no lineal. En particular, usamos la función *nls()*, disponible en R para calcular los valores de $\hat{a}_{nolineal}$ y $\hat{b}_{nolineal}$. Una instrucción típica es la siguiente

$$nls((N + 1 - r)^{**}b/r^{**}a, data, start = \{a = al, b = bl\})$$

con $al = \hat{a}_{lineal}$ y $bl = \hat{b}_{lineal}$ y *data* el conjunto de datos ordenados descendente que serán ajustados. El método no lineal es más robusto que el lineal, ya que impide la influencia de relaciones de colinealidad y otros efectos asociados a la probable co-dependencia entre las variables x_1 y x_2.

1.6. Hipótesis general

La diversidad de los orígenes de los datos que se presentan con ajustes muy cercanos a la FBC, sugieren dos cosas: la primera de ellas –muy poco

probable– es la existencia de un mecanismo único o un meta- mecanismo subyacente a todos estos datos [30][5]; lo segunda es que, dado un fenómeno existen dominios de ellos, que por sus similitudes en sus procesos generativos son explicados por un mecanismo común. En este trabajo adoptaremos este enfoque. Esta perspectiva es análoga a la adoptada en leyes de potencia en donde, no existe un mecanismo único que explique la aprición de todas las leyes de potencia, sino que se pueden agrupar a distintos fenómenos a partir de sus mecanismos generatrices. Por ejemplo: multiplicación de exponenciales (tamaño de grupos poblacionales, tamaños de archivos computacionales [31]), caminatas aleatorias (tiempo de primer retorno,distribución de los tiempos de vida del registro fósil [32, 33]), acumulación (distribución de especies biológicas [17], tamaño de ciudades [34],citas a artículos científicos [35, 36], enlaces electrónicos al www [37, 13, 38]), y la críticalidad auto-organizada (incendios forestales [39], temblores [40, 41], erupciones solares [42], modelos de evolución biológica [43], distribución de tamaños de avalanchas [19]). En el caso de este trabajo consideraremos el mecanismo de dos proceso antagónicos que, genéricamente, se pueden caracterizar como permanencia-cambio, en tres modelos distintos.

1.7. Distribuciones alternativas a la ley de potencia

En esta sección mostraremos las principales alternativas propuestas en la literatura científica para la corrección de la ley de potencia cuando ésta es un modelo, cuando menos visiblemente, no adecuado. Es decir, las opciones sugeridas cuando los datos se alejan de un comportamiento lineal en una gráfica log-log y se tiene una curvatura pronunciada en los extremos. En particular, tres distribuciones han sido muy utilizadas como opciones a la ley de potencia: corte exponencial (*"exponential cut-off"*) , la exponencial estirada (*"streched exponential"*) y la distribución en la cual el logaritmo de una variable aleatoria está distribuida normalmente: la log-normal. En general las correcciones o generalizaciones a la ley de potencia se enfocan a la prominente curvatura que aparece en los extremos de las distribuciones

[5]En esta referencia se considera un proceso de adición de variables estocásticas obteniendo como caso límite una FDP tipo beta; como lo hemos señalado previamente la relación entre las FDP y las orden-frecuencia no es inmediata y por lo tanto, no es posible señalar una relación directa entre ellas.

[44], utilizando, por ejemplo, escalamiento finito [45], dilución de redes y constantes de crecimiento [46, 12] con un variado éxito.

1.7.1. Distribución log-normal

Esta distribución de probabilidad está definida de la siguiente forma sea X una variable estocástica, con distribución normal, entonces la variable Y, relacionada con X de la forma $Y = e^X$ tiene una distribución log-normal[47]. Es decir que es una variable aleatoria cuyo logaritmo está distribuido normalmente. La distribución log-normal es producida por un proceso multiplicativo de un número grande de variables independientes aleatorias positivas [48, 49] La expresión de la FDP log-normal es la siguiente:

$$f(x)_{\mu,\sigma} = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}} \tag{1.13}$$

La función de distribución acumulada (FDA) está entonces determinada por

$$FDA = \int_x^\infty f(u)_{\mu,\sigma} du = \frac{1}{2} + \frac{1}{2} erf(\frac{\ln x - \mu}{\sigma\sqrt{2}}) \tag{1.14}$$

con $erf(x)$ es la función de error definida por

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt$$

La pregunta es cómo es posible, si las formas funcionales de la ley de potencia y la log-normal son tan distintas, que puedan ser ambas distribuciones aplicadas a un mismo conjunto de datos, con resultados, a simple vista, satisfactorios. Como se ha señalado previamente uno de los métodos más usuales para determinar si un conjunto de datos satisface o no una ley de potencia, consiste en graficarlos en escala log-log, si en una región significativa de ésta se observa una recta, entonces esto sugiere la existencia de un ley de potencia. [6]

La respuesta se encuentra en la siguiente expansión en serie de Taylor del logaritmo de la FDP

$$\ln p(x) = -\ln(x) - \frac{(\ln x - \mu)^2}{2\sigma^2} - \ln(\sigma\sqrt{2\pi}) \tag{1.15}$$

$$= -\frac{(\ln x)^2}{2\sigma^2} - (1 - \frac{\mu}{\sigma^2}) \ln x - (\frac{\mu^2}{2\sigma^2} + \ln(\sigma\sqrt{2\pi})) \tag{1.16}$$

[6] Sin embargo este "método" ha sido cuestionado por la poca fiabilidad en la determinación de la existencia o no de una ley de potencia [20].

Se observa que esta función es cuadrática en la variable $\ln x$, pero una parábola "vista" de muy cerca se ve como una recta, por lo que con frecuencia, se asume que en ciertas regiones puede ser considerada, en una primera aproximación, como una tendencia lineal. Podemos observar una buena aproximación a una recta con las siguientes condiciones se satisfacen, si la varianza σ es grande entonces los términos que contienen a σ^2 en el denominador de la ecuación 1.7 tienden a cero, y por lo tanto se obtiene una recta con pendiente negativa. Obtener una varianza grande no es fenómeno poco común, ya que mediante el propio mecanismo de generación de la FDP lognormal, es decir el resultado de algún proceso multiplicativo, es posible obtener varianzas, que debido a su valor relativo con la media, se observan en log-log, como rectas. De hecho con este mismo proceso, con modificaciones menores, es posible producir aproximaciones a leyes de potencias [50, 51, 52].

García Naumis y Cocho a su vez, [53] demostraron que mediante un proceso multiplicativo, iniciando con k números al azar y formando todos los posibles productos, de k factores, entre ellos, es posible obtener, en el límite $k \to \infty$ la FBC en orden-frecuencia.

1.7.2. Corte exponencial

La distribución con corte exponencial (*cut-off exponential*), es de hecho, una variación de la ley de potencia, ya que en la región con $x \gg 1$, el término dominante se encuentra dado por una exponencial decreciente. Sea X una variables estocástica, y α y β dos parámetros, entonces

$$p(x;\alpha,\lambda) = \frac{\lambda^{1-\alpha}}{\Gamma(1-\alpha,\lambda x_{min})} x^{-\alpha} e^{-\lambda x} \tag{1.17}$$

Se ha utilizado con frecuencia, para corregir el comportamiento de la gráfica log-log cerca del límite superior de la variable de orden x, con ello se ajusta la curvatura en las gráficas, dada por el efecto de tamaño finito del número de datos considerados. Se ha utilizado está distribución, en los modelos de redes de interacción de proteínas [54], en la distribución de ingresos económicos [55], redes ecológicas [56], entre otras.

1.7.3. Exponencial estirada

Otra distribución, ampliamente utilizada en los ajustes de datos, es la exponencial estirada (*stretched exponential*), en donde uno de los parámetros

de la distribución β incrementa la influencia de la exponencial cuando $x \to N$, así, que puede modelar eventos con curvaturas más acentuadas cerca del número máximo de datos. La exponencial estirada se define de la siguiente manera sea X un variable estocástica y λ y β parámetros libres de ajuste.

$$p(x;\lambda,\beta) = \beta\lambda e^{\lambda x^{\beta}_{min}} x^{\beta-1} e^{-\lambda x^{\beta}} \tag{1.18}$$

Está distribución puede confundirse fácilmente con la ley de potencia para ciertos valores de los parámetros, por ejemplo $\lambda \to 0$. La exponencial estirada se ha utilizado para modelar distribuciones en economía, las intensidades de señales de radio y luz emitidas por las galaxias, la cantidad de reservas petroleras, poblaciones de países [57], y esta bien establecida en la modelación de tiempos de relajación en vidrios [58]. En [59, 53] Naumis y Cocho demostraron que la FBC proviene de sumar exponenciales estiradas.

1.7.4. Otras distribuciones orden-frecuencia

Existen otras distribuciones, menos conocidas, que han sido empleadas sobre todo en lingüística en donde las propuestas para la modificación a la ley de Zipf (en esta caso aplicada a orden-frecuencia) abundan. La siguiente lista muestra las distintas funciones –la mayoría de ellas listadas por W. Li y P. Miramontes [27]–, que adicionalmente a la exponencial estirada, la ley de potencia con corte exponencial, la ley de potencia y por supuesto la FBC compararemos como ajustes a un conjunto heterogéneo de datos para determinar si la función beta es singular (con relación a los ajustes de dichos datos), debido a que es posible determinar mediante algún criterio de información su superioridad con respecto al resto de las distribuciones estudiadas.

Gusein- Zade	$f = C\log(\frac{n+1}{r})$
Exponencial	$f = Ce^{-ar}$
Logarítmica	$f = C - a\log(r)$
Weibull	$f = C(\log\frac{n+1}{r})^{a}$
Logarítmica cuadrática	$f = C - a\log(r) - b(\log(r))^2$
Yule	$f = C\frac{b^r}{r^a}$
Menzerath-Altmann	$f = C\frac{e^{-\frac{b}{r}}}{r^a}$
Frappat	$f = C + br + Ce^{-ar}$
q-estadística	$f = CN(1 + (a-1)\frac{N^{a-1}}{M}r)^{\frac{1}{1-a}}$

1.8. Comparativo entre modelos

En esta sección mostraremos las distintas funciones que han sido propuestas en la literatura como alternativas a la función de distribución ley de potencia. Ajustamos los datos presentados en la sección previa para estos modelos y mostramos un comparativo gráfico entre varios ajustes, además hacemos un análisis entre las distintas funciones utilizando medidas de información. Para ello, utilizamos los criterios de información de Akaike y Bayesiano. Estos indicadores toman en cuenta, además de la bondad del ajuste, el número de parámetros del modelo, penalizándolo cuando éste aumenta. En esta sección consideraremos modelos con $n = 1, 2$ y 3 parámetros. La conclusión que parece ser obvia es que los ajustes mejoran con el número de parámetros del modelo, pero la utilización de estos criterios nos permite establecer un balance entre el ajuste de los datos y el número de parámetros, por lo que mostraremos,que , cunado menos en estos casos, no es cierta la aseveración anterior.

1.8.1. Criterio de información de Akaike

El coeficiente de determinación R^2, cuyo uso está muy extendido, es una medida sobre la bondad de un ajuste. Si este valor es cercano a $R^2 \approx 1$ se considera que el modelo propuesto es satisfactorio como ajuste a los datos. Sin embargo, cuando tenemos modelos con distintos tamaños de muestra y distinto número de parámetros este coeficiente no toma en cuenta estos factores. Es por ello que se han buscado índices que midan cantidades de información. El criterio de información de Akaike (CIA) propuesto por Hirotsugu Akaike en 1974 [60], está basado en el concepto de entropía de información y es una medida de la información perdida cuando un modelo se usa para ajustar una serie de datos. La expresión del CIA es la siguiente:

$$CIA = 2k - 2\log(L) \tag{1.19}$$

con k el número de parámetros del modelo y L el valor máximo de la función de verosimilitud (*maximum likelihood*).

Bajo el supuesto –usual– de que los errores del modelo son independientes y están idénticamente distribuidos, lo cual significa que satisfacen el teorema del límite central, la distribución de los errores es normal, de esta forma la función $\log(L)$ depende sólo del número de datos y de una medida de las distancias de los datos con respecto al modelo propuesto.

$$\log(L) = C - \frac{n}{2}\log\frac{SSE}{n} - \frac{n}{2} \tag{1.20}$$

C es una constante, SSE es la suma de los errores al cuadrado (*sum of squared errors*) definidos de la manera usual (ecuación 1.21) para una función orden-frecuencia(tamaño) $F(i;\alpha_1,\alpha_2,\ldots,\alpha_n)$ que depende de n parámetros α_i y el orden de los datos denotado por el índice i.

$$SSE = \sum_{i=1}^{n}(x_i - F(i;\alpha_1,\alpha_2,\ldots,\alpha_n))^2 \tag{1.21}$$

Sustituyendo 1.20 en la ecuación 1.19,

$$CIA = \underbrace{n\log\frac{SSE}{n}}_{\text{Bondad del ajuste}} + \underbrace{2k}_{\text{Penalización}} + \underbrace{n-2C}_{\text{Término constante}} \tag{1.22}$$

Es importante resaltar que la ecuación 1.22 contiene un término que crece linealmente con el número de parámetros, con ello penaliza la presencia de un mayor número de éstos, por lo cual, es posible comparar entre modelos con distinto número de éstos.

Sean $F_1(i;\alpha_1,\alpha_2,\ldots,\alpha_n))$ y $F_2(i;\alpha_1,\alpha_2,\ldots,\alpha_m)$ dos modelos, con n y m número de parámetros, el más adecuado para un conjunto de datos es aquel que tiene el menor CIA.

La diferencia ΔCIA entre ambas medidas es:

$$\Delta CIA = CIA(F_1) - CIA(F_2) = n\log\frac{SSE_1}{SSE_2} + 2(k_1 - k_2) \tag{1.23}$$

Si $\Delta CIA < 0$ se selecciona el primer modelo, en caso contrario, el segundo.

Cuando el primer modelo tiene un parámetro más que el segundo, el criterio $\Delta CIA < 0$ se traduce en:

$$n\log\frac{SSE_1}{SSE_2} + 2 < 0 \tag{1.24}$$

$$\frac{SSE_1}{SSE_2} < e^{-\frac{2}{n}} \tag{1.25}$$

De esta manera el criterio se reduce a calcular los SSE de cada modelo, tomar su cociente y verificar cuándo es menor que el valor de la exponencial de la ecuación anterior.

1.8.2. Criterio de información Bayesiano

Consideramos otro índice reportado en la literatura: el criterio de información bayesiano (CIB), que fue desarrollado por Akaike y Schwarz [61, 62] usando el formalismo bayesiano.

Para el caso de dos modelos con el CIB, sean dos funciones m_1 y m_2 con $m_1(a_1, a_2, \ldots, a_n)$ y $m_2(a_1, a_2, \ldots a_m)$, con n y m número de parámetros respectivamente, entonces para un conjunto de datos ajustados con ambos modelos, se selecciona el modelo m_i si el $CIB(m_i) < BIC(m_j)$ [63].

El CIB se define como:

$$CIB(m(a_1, a_2, \ldots a_k)) = -2\log(L) + k\log(n) \tag{1.26}$$

Nuevamente, con la suposición de la SSE tiene una distribución normal, podemos sustituir el valor del $\log(L)$ dado por la ecuación 1.20 en la ecuación 1.26, con lo que la expresión para el CIB queda como:

$$CIB = \underbrace{n\log\frac{SSE}{n}}_{\text{Bondad del ajuste}} + \underbrace{k\log(n)}_{\text{Penalización}} + \underbrace{(n-2C)}_{\text{Término constante}} \tag{1.27}$$

La diferencia entre los índices de ambos modelos es ΔCIB

$$\Delta CIB = CIB(F_1) - CIB(F_2) = n\log\frac{SSE_1}{SSE_2} + (k1 - k2)\log n \tag{1.28}$$

Así, bajo las condiciones en donde el primer modelo tiene un parámetro más que el segundo, se selecciona el primero con la siguiente condición

$$\frac{SSE_1}{SSE_2} < e^{\frac{-\log n}{n}} \tag{1.29}$$

Debido a que $\log(n) > 2$ para $n > e^2 \approx 9$ (todos los datos considerados en este trabajo satisfacen esta desigualdad), entonces $e^{\frac{-\log n}{n}} < e^{-(2/n)}$ para $n > 9$ por lo que la condición del CIB es más estricta que la del CIA. Por ello, presentaremos las comparaciones con respecto al este último índice en esta tesis. El anexo 3 muestra el programa escrito en R (http://www.r-project.org/), que compara el CIB para la serie de datos.

En la gráfica 1.6 y la tabla 1.1 se muestran los ajustes para las funciones mostradas, comparadas con el ajuste de la FBC para el conjunto de datos mostrados en la figura 1.2, mostramos las tres mejores al considerar el ΔCIB de cada función con el conjunto de datos. En los 4 casos considerados, así cómo los datos considerados en este trabajo, la mejor función que ajuste estos datos es la FBC [7]

Comparativo		
	Función	Δ CIB
Sismos en California	FBC	0.00
	Weibull	33009.93
	Exponencial estirada	50006.24
	Corte exponencial	51910.00
Citas por artículo	FBC	0.00
	Weibull	2017.488
	Exponencial estirada	2566.467
	Yule	2571.153
Población de E.U.A.	FBC	0.00
	Corte exponencial	10112.23
	Logarítmica cuadrada	29992.84
	Gusein	33307.88
Apagones	FBC	0.00
	Corte Exponencial	219.9555
	Yule	219.9555
	Weibull	652.3870

Cuadro 1.1: La tabla muestra los comparativos ente distintos modelos. Se calcula la diferencia entre los CIB con respecto al mejor, en los cuatro casos fue la FBC.

[7]Esto no significa que para cualquier conjunto de datos la mejor sea la FBC, ya que existen una buena cantidad de datos en los que los ajustes son otras funciones más conocidas, pero el presente trabajo se centra en la abundante aparición de la función FBC en ajustes orden-frecuencia.

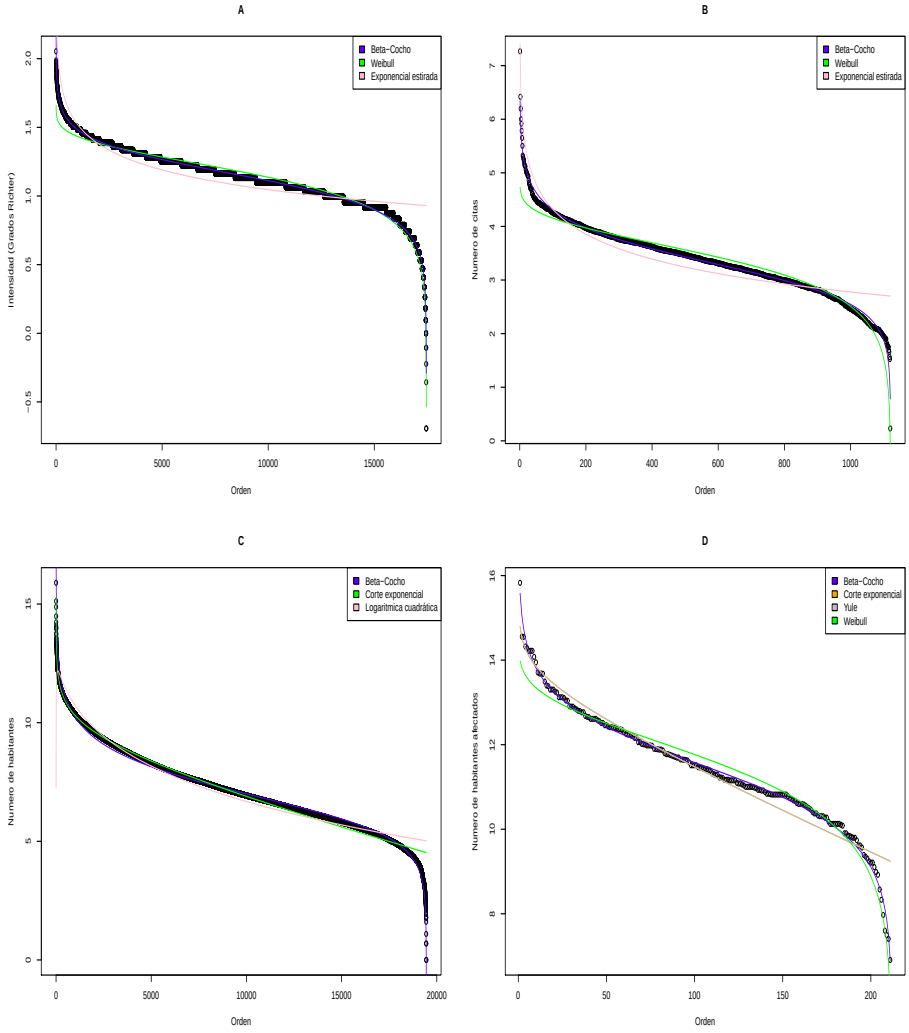


Figura 1.6: Distribuciones orden-frecuencia. En todos los casos se muestran para efectos comparativos los mismo datos reportados en la figura 1.2 . Se muestran los ajustes de las distribuciones que mejor correspondencia tienen con los datos en términos del criterio de información. **A**. Sismos registrados en California, EUA. **B**. Citas de artículos. **C**. Número de habitantes de ciudades de Estados Unidos registrados por el censo del 2000. **D**. Número de personas afectadas por cortes de energía eléctrica.

Capítulo 2

Modelos de expansión-modificación

"Si un eterno viajero la atravesara en cualquier dirección, comprobaría al cabo de los siglos que los mismos volúmenes se repiten en el mismo **desorden** (que, repetido, sería un orden: el **Orden**). Mi soledad se alegra con esa elegante esperanza"
La biblioteca de Babel
Jorge Luis Borges

En este capítulo se muestra un modelo propuesto por Wentian Li [64] con el objetivo de generar secuencias de símbolos con correlaciones largas mediante la competencia de dos procesos opuestos: inserción y mutación, con ello es posible generar secuencias determinadas por un parámetro p que modula las mutaciones en este proceso [65]. Estudiamos la dependencia de p, a, b (p es la probabilidad de cambio entre símbolos y a, b los parámetros de ajuste de la FBC) desde un estado en el cual existen correlaciones de largo alcance a otro en donde se extinguen, mediante el estudio de la frecuencia de n símbolos consecutivos n-ómeros. Modificamos y estudiamos otras implementaciones del proceso original para investigar la relación entre los exponentes de ajuste a y b y los parámetros del modelo.

2.1. FBC en la distribución de codones

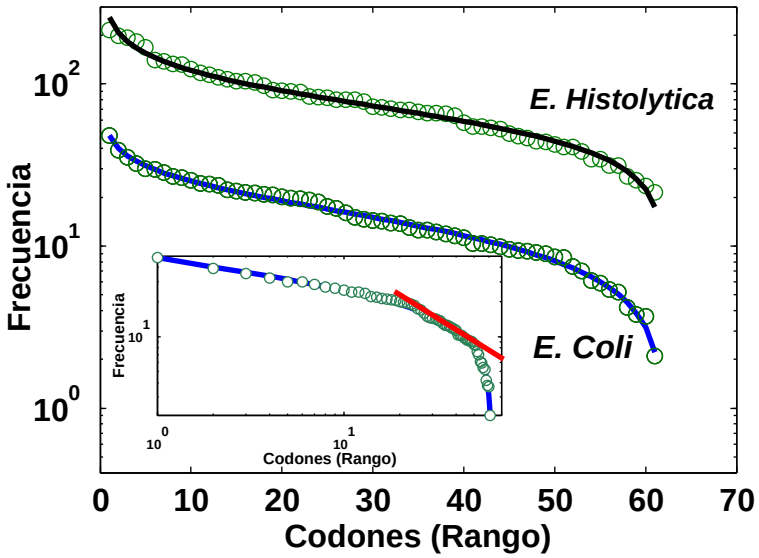


Figura 2.1: Función beta para distintos valores de los parámetros

La aparición de la FBC en la distribución de codones (triplete de nucleótidos consecutivos en la secuencia de DNA) permite aventurar una hipótesis sobre la naturaleza dinámica en la generación de secuencias simbólicas y su relación con la FBC. Esquemáticamente, y por lo tanto de manera simplificada, en el ADN intervienen entre otros muchos, dos procesos que determinan fuertemente la secuencia, a saber, la mutación puntual y la replicación. Éstos dos ingredientes esenciales, además de la característica hereditaria de los de la información genética, son fundamentales para la evolución de secuencias genéticas. De hecho se ha establecido que bastan estos tres ingredientes para que tenga lugar la evolución biológica [66]. La dinámica de generación de secuencias que pueden replicarse y en este proceso mutar es una dinámica que enfrenta dos procesos opuestos entre sí.

Con esta motivación, el principal objetivo de este trabajo consiste en generar mecanismos –si es posible– simples para la obtención de secuencias simbólicas sobre las cuales, al hacer una estadística clasificación-frecuencia,

se obtiene una función de distribución beta. El proceso, inspirado en la aparición de la FBC en el DNA, constituirá en incorporar dos mecanismos con dinámicas en conflicto que generen secuencias simbólicas.

Este ejemplo, sugiere que la dinámica de dos procesos compitiendo entre si permitiría generar secuencias que satisfacen la FBC, como resultado de su ordenamiento descendente. Esta hipótesis es la que exploramos en este trabajo utilizando para ellos diversos sistemas con dinámicas en oposición.

En la búsqueda del significado de los parámetros de la FBC es particularmente interesante el caso de las distribuciones orden-frecuencia de codones. La fig. 2.1 muestra una gráfica en semi-log de la frecuencia con la que los 61 codones (las señales de alto en la codificación no son consideradas, es decir los codones detienen la codificación a aminoácidos: UAG, UAA, y UGA) aparecen en el genoma de la bacteria *Escherichia coli* y de la ameba *Entamoeba histolytica* graficadas en orden decreciente. El recuadro corresponde a una gráfica log-log para *E. coli*, y la línea recta es una guía visual ,en la que se observa una clara desviación de una tendencia lineal en los valores iniciales y finales de esta distribución(Ver fig. 2.1). El ajuste con la FBC es igualmente bueno para las secuencias genéticas de decenas de organismos que hemos analizado desde arqueas, bacterias y eucariontes, así como para aminoácidos y distribuciones de codones codificantes [65], las secuencias fueron obtenidas de la base de datos GenBank (http://www.ncbi.nlm.nih.gov/genbank/).

Wentian Li et al. ([67]) observó que el espectro de frecuencias en la secuencia lineal de la molécula del ARN(ácido ribonucleico) decae lentamente, más que exponencialmente, independientemente, otros grupos de investigación reportaron el mismo hecho [68].En particular Li, calculó el espectro de potencias $P(f) = N|A(f)|^2$ para analizar las secuencias de nucleótidos en donde $A(f)$ son lo coeficientes discretos de Fourier definidos como:

$$A(f) \equiv \frac{1}{N} \sum_{j=0}^{N-1} x_j exp(-i2\pi(f/N)j) \tag{2.1}$$

con $x_j = 1, 2, 3, \ldots, 2^n$

El proceso dinámico para la creación de secuencias consiste en dos subprocesos opuestos en competencia: uno de ellos favorece las correlaciones largas y el otro, se encarga de destruirlas, la dinámica resultante consiste en objetos de distintas escalas, generando, de esta manera, multiescalidad y por lo tanto – bajo ciertas condiciones– un distribución de ley de potencia.

Nuestro interés se encuentra en la generación de secuencias simbólicas que nos permitan estudiar la estadística de elementos consecutivos de símbolos (

0 y 1) mediante el análisis de su frecuencia de aparición. Usamos el caso del DNA como una inspiración para generar secuencias, que no necesariamente representarán los múltiples y complejos procesos presentes en esta molécula, sino que tratamos de encontrar modelos suficientemente simples pero que conserven las características deseadas. En este caso, queremos tener un proceso con dos dinámicas en competencia entre si y que puedan generar, para ciertas condciones, correlaciones de largo alcance. Si queremos mantener la referencia al DNA podemos pensar que la elección de dos símbolos en lugar de cuatro, corresponde a la elección estándar de purinas (A y G) y pirimidinas (C y T/U).

En los modelos de expansión-modificación bajo el contexto de la "teoría neutral de la evolución molecular" propuesto por Kimura [69, 70, 71], la probabilidad de tener una modificación puede ser asociada a la probabilidad de mutación para secuencias genéticas, y las probabilidades der expansión están relacionadas con las duplicaciones o inserciones.

2.2. El modelo de expansión-modificación.

En el modelo booleano de expansión-modificación propuesto originalmente por Li [64] se elige un 0 ó 1 como condición inicial al azar y se somete esta semilla al siguiente algoritmo [1].

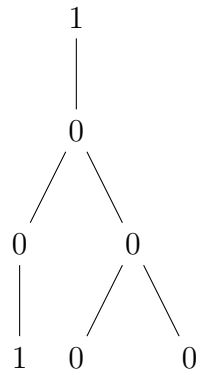
$$\begin{aligned} 1 &\rightarrow \begin{cases} 0 & p \\ 11 & 1-p \end{cases} \\ 0 &\rightarrow \begin{cases} 1 & q \\ 00 & 1-q \end{cases} \end{aligned} \tag{2.2}$$

En la expresión anterior p es la probabilidad de mutar de 1 a 0 mientras que q es la probabilidad de que la mutación inversa se llevo a cabo. Las cantidades $1-p$ y $1-q$ son las probabilidades de duplicación para 0 y 1, respectivamente. Los valores relativos de p y q modulan la prevalencia de los procesos involucrados, es decir, cuando p (o q) son pequeñas, la duplicación domina y se obtienen regiones de ceros o unos extensas, a diferencia del caso en el que p (o q) son grandes en donde se obtienen regiones con ceros y

[1] Este proceso está relacionado con lo que se concoe como "random tiling" (teselaciones aleatorias) en cuasicristales [72, 73], en este formalismo es posible calcular correlaciones de largo alcance [74] y matrices de transferencia [75].

unos alternados. La aplicación sucesiva del algoritmo genera una secuencia de ceros y unos tan grande como se desee. Para nuestro análisis de distribuciones, nos enfocaremos en la frecuencia con la cual los n-ómeros (grupos de **n** elementos consecutivos, no traslapados) aparecen en orden descendente. El *n-ómero* que aparece con mayor frecuencia se coloca en primera posición con $r = 1$, el segundo más frecuente $r = 2$ y así sucesivamente. Así, nuestra atención se dirigirá al estudio de las propiedades orden-desorden [2] de este proceso, así como, en la forma en la que los cambios en la probabilidad de modificación p afectan a los parámetros del ajuste a y b, y finalmente, en cómo estos parámetros se convierten en indicadores para la caracterización de una transición orden-desorden.

La siguiente figura muestra una instancia del proceso de expansión-modificación. Iniciando con el elemento 1 a $t = 0$, al siguiente tiempo $t = 1$, 1 muta al símbolo 1 ($1 \to 0$), al siguiente paso de iteración $t = 2$ se tiene el símbolo 0 se expande ($0 \to 00$); por último a $t = 3$ el primer símbolo muta ($0 \to 1$), mientras que el segundo, se expande ($0 \to 00$). Mediante este proceso generamos secuencias simbólicas S de $0, 1$ tan grandes como se desee.



$$S = s_0 s_1 s_2 \dots s_n \dots \tag{2.3}$$

con $s_i \in \{1, 0\}$.

A modo de ilustración, para la siguiente secuencia se consideran 6 elementos consecutivos (hexámeros), se incluyen los valores de estos hexámeros

[2] Podemos caracterizar esta transición a partir de las propiedades de estructura: cuando se tiene invarianza de escala en el espectro de potencias nos referiremos a una fase ordenada, en el caso opuesto hablaremos de una fase desordenada, es decir entropía máxima, ausencia de correlaciones, etc.

en su representación decimal a partir del binario.

$$S = \underbrace{100010}_{34}\underbrace{010010}_{18}\underbrace{010001}_{17}\underbrace{001001}_{9}\underbrace{011110}_{30}\ldots \tag{2.4}$$

A partir de este análisis se puede obtener información sobre el significado de los parámetros del ajuste. Para este estudio seguimos la siguiente secuencia:

1. Analizamos el comportamiento de la correlaciones de largo alcance de los *n-ómeros*.

2. Consideramos *n-ómeros* de diferente tamaños.

3. Consideramos los eventos inusuales (*n-ómeros* de frecuencia de aparición muy baja o nula).

4. Introducimos un retraso que permite que las mutaciones se lleven cabo cada τ pasos y estudiamos el impacto de este retraso en los valores de ajuste de la FBC.

5. Modificamos los valores relativos de p y q.

6. Finalmente exploramos los resultados de una implementación aproximada de la dinámica y comparamos esta aproximación con el resultado computacional.

2.2.1. Transición orden-desorden para $p = q$

Para el cálculo de las distribuciones orden-frecuencia empezamos con un 0 o un 1 como semilla elegida al azar. Entonces iteramos el algoritmo hasta obtener una secuencia de $1\ X\ 10^6$ símbolos. Con ello estudiamos la frecuencia de aparición de los 2^n *n-ómeros*. En la figura 2.2 mostramos la distribución orden-frecuencia para los $2^8 = 256$ octámeros con $p = q$. En la figura 2.3a mostramos el promedio de los valores de los parámetros de ajuste a y b, determinados a partir de 10 realizaciones independientes, graficadas como función de la probabilidad de mutación p.

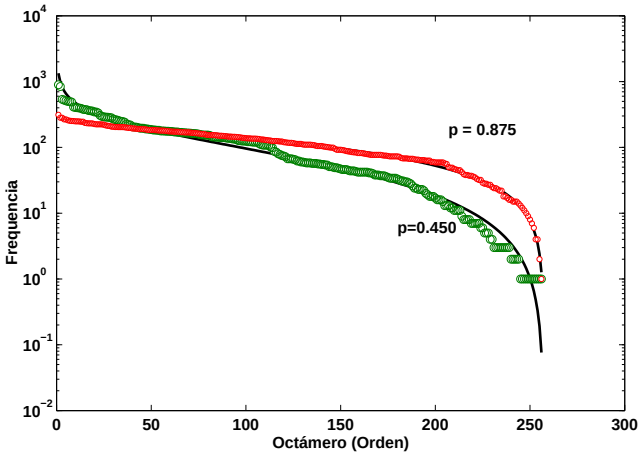


Figura 2.2: Distribuciones orden-frecuencia para los octámeros generados por el algoritmo de expansión-modificación 2.3. Las circunferencias en rojo están determinadas con una probabilidad de mutación $p = 0.875$, la línea sólida corresponde al ajuste de la FBC con $(a, b, R^2) = (0.36, 1.55, 0.96)$. Los circunferencias verdes corresponden a $p = 0.475$ y$(a, b, R^2) = (0.11, 1.28, 0.96)$.

Para p muy pequeña, se tiene que $a > b$, en este caso las mutaciones puntuales son poco probables y la expansión es favorecida conduciendo a intervalos extensos de ceros o unos y al dominio de unos pocos octámeros con una alta incidencia, en particular los n-ómeros con ceros o unos consecutivos (por ejemplo, para el caso de $n = 8$ las secuencias 00000000 y 11111111 dominan al resto). Cuando p aumenta, a disminuye y b se incrementa de tal forma que en las curvas para a y b como función de p existe una intersección en un valor umbral p_t por encima del cual a es siempre más pequeño que b.

Si graficamos la dependencia del siguiente parámetro $(a+b)/(a-b)$ con p promediado sobre 10 realizaciones observamos en la figura 2.3b, que existen tres regiones bien delimitadas: una antes de p_t, la siguiente entre p_t y el valor a partir del cual el espectro de potencias tiene una pendiente nula, y por último, la región con pendiente cero. Mostraremos, más adelante, que p_t es un punto de transición que determina un cambio en la dependencia de las correlaciones de largo alcance con respecto a la probabilidad de mutación p. Estas correlaciones decrecen y desaparecen en la región *III*. En la figura 2.4 se muestra la dependencia de la variación del tamaño del *n-ómero* con

respecto de a y b como función de p.

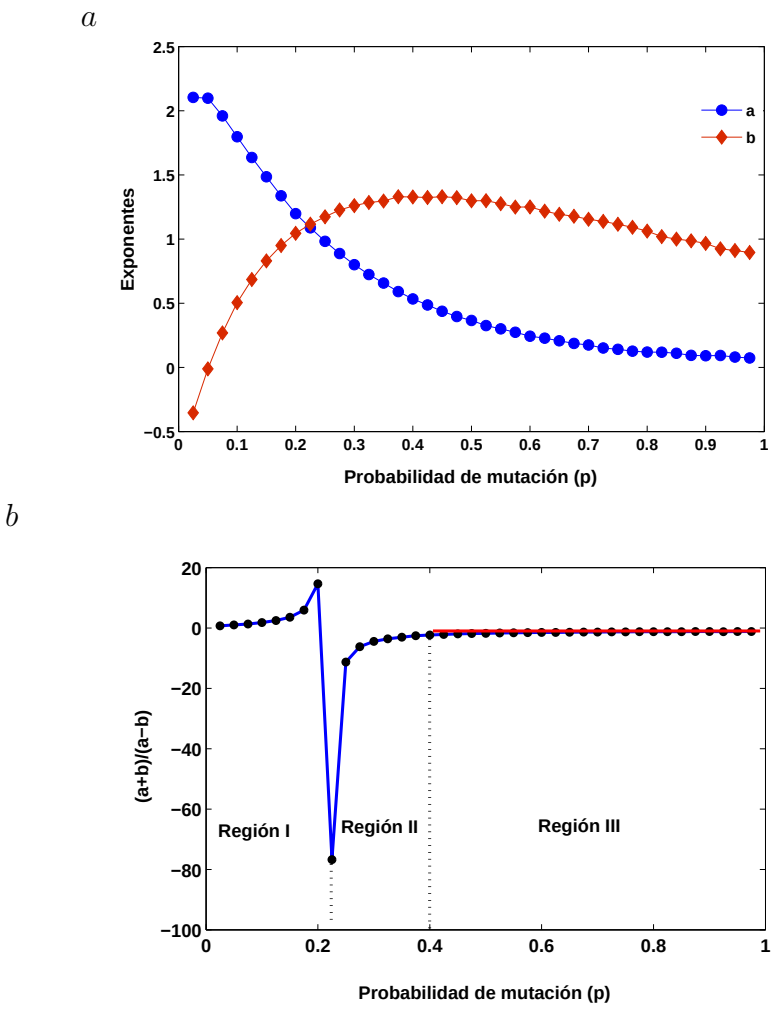


Figura 2.3: a) Dependencia de la probabilidad de mutación p con respecto al promedio de los parámetros de ajuste (a, b) determinados con 10 distribuciones orden-frecuencia de octámeros generados por realizaciones independientes del algoritmo de expansión-modificación (ecuación 2.2), con $p = q$. b) Los mismos datos graficados $\frac{a+b}{a-b}$ *vs* p, el mínimo separa la región I de la región II descrita en el texto y la línea sólida en rojo es una guía visual para la identificación de la región III que inicia alrededor de $p = 0.400$

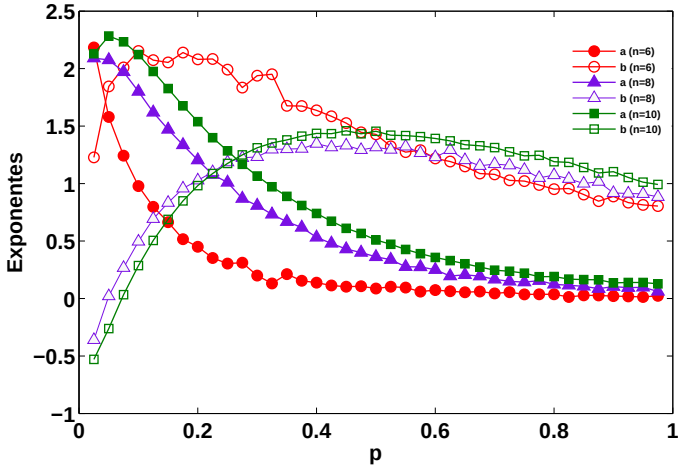


Figura 2.4: Variación de la FBC con respecto a los parámetros del ajuste (a, b) como función de p. Se muestran diferentes valores de los *n-ómeros*

.

En la figura 2.5 se muestra la entropía de Shannon para el caso de *octámeros* $S = -\sum_{i=1}^{2^n} f_i \ln f_i$, con f_i la frecuencia de aparición del octámero i-ésimo. Como se puede observar, ésta es una función monótona creciente con p que alcanza su máximo valor $S_{max} = \ln(2^8) \approx 5.54$ el cual es el valor de la entropía para una distribución uniforme. Por ello es posible conjeturar que los valores altos de a comparados con b están relacionados con la **permanencia**; mientras que los valores altos relativos de b están influenciados por el **cambio**. En las dos siguientes secciones precisaré esta interpretación.

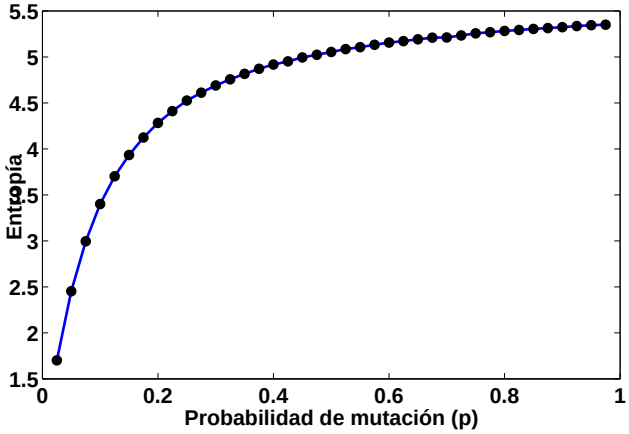


Figura 2.5: Entropía de Shannon, para *octámeros* generados por la ecuación, como función de la probabilidad de modificación p.

2.2.2. Parámetro a y correlaciones de largo alcance

Las correlaciones proporcionan un medio para la comprensión de los roles de a y b. En 1992 Wentian Li se enfocó en el comportamiento de correlaciones espaciales sobre secuencias generadas por el algoritmo de expansión-modificación [67]. Mediante el cálculo del espectro de potencia mostró la existencia de correlaciones de largo alcance –caracterizado por la existencia de una ley de potencia– para valores pequeños de p, cuando p aumenta lo que empieza a dominar es la expansión, esto rompe las correlaciones de largo alcance y deja de existir la ley de potencia. Aquí extenderemos este tratamiento para los *n-ómeros*, también observamos un comportamiento de ley de potencias en el espectro de potencias con pendiente negativa. (ver la fig.2.6).

En este caso el espectro de potencias, se calcula sobre los octámeros representados en su forma decimal. Es decir interpretamos la correlación de encontrar dos octámeros a una distancia d arbitraria cuando los octámeros se expresan en su representación decimal.

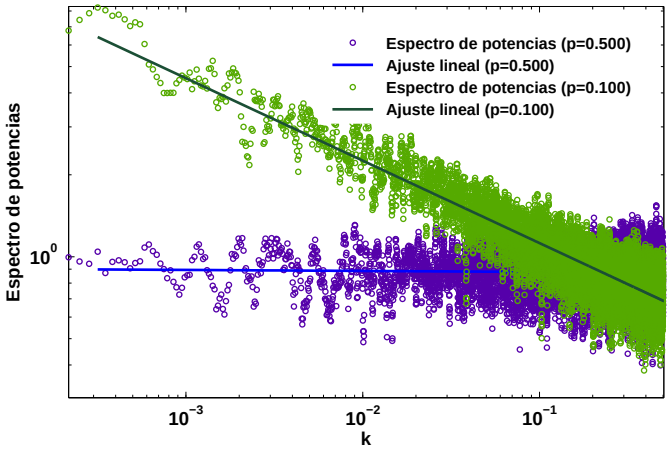


Figura 2.6: Espectros de potencia de *octámeros* generados por la el algoritmo de expansión-modificación para $p = 0.100$ y $p = 0.500$.

Cuando p se incrementa, la pendiente tiende a cero y las amplitudes de las fluctuaciones también crecen. La fig.2.7 muestra la función de correlación de los *octámeros* como función de p, los valores promedio de las pendientes sobre 10 realizaciones. Se muestra un régimen con tendencia lineal seguido de un régimen intermedio que culmina con una región de pendiente cero, característica del ruido blanco, indicativo del comportamiento no correlacionado. La línea recta en la figura es una ayuda visual para indicar el comportamiento lineal que culmina en $p_l = 0.225$. Notemos que este valor de p coincide con la probabilidad de transición p_t mostrada en la fig.2.3. Después de este punto los cambios en el desorden inducido por el incremento de p, modifican el decaimiento de las correlaciones de largo alcance. La pérdida de la tendencia lineal señala la transición orden-desorden reflejada por los valores de (a, b). Los parámetros a y b son sensibles a estos fenómenos y son una forma de determinar el punto de transición. Con esta evidencia podemos afirmar que a registra el peso de la presencia de correlaciones de largo alcance y b la importancia de los factores aleatorios. Esta observación no es particular del caso de $n = 8$ también lo hemos verificado para $n = 6, 10, 12$.

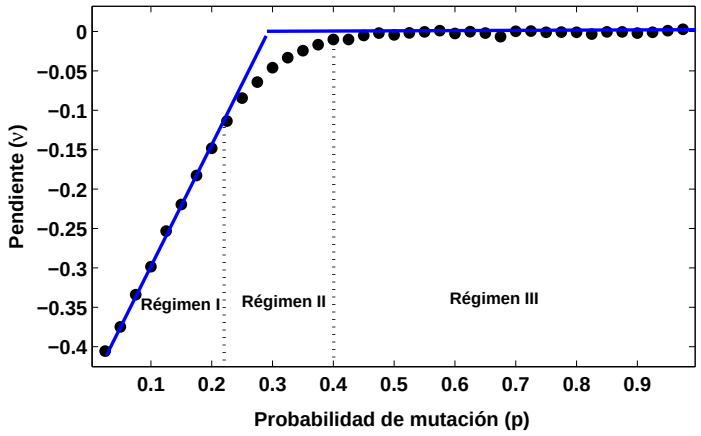


Figura 2.7: Pendientes promedios (ν) sobre 10 realizaciones del espectro de potencias como función de la probabilidad de modificación. La línea es una guía visual para la identificación de dos diferentes regímenes cuando la pendiente se incrementa con p. El régimen lineal culmina en $p_l = 0.225$ el cual coincide con la transición en el punto p_t mostrado en la Fig. 2.3a. La línea horizontal $\nu = 0$ exhibe el colapso de las correlaciones de largo alcance y el origen del régimen III el cual es cercano al inicio de la tercera región mostrada en la Fig.2.3b

Adicionalmente podemos afirmar que la transición y su relación con a se establece al considerar el decaimiento de a como función de p. La fig. 2.8 muestra al parámetro a como función de p junto con un ajuste exponencial para el caso $n = 8$. El valor característico de relajación p_r del ajuste es 0.245 que cae dentro del rango $p_t = 0.225 \pm 0.025$. Para $n = 10$ tenemos $p_r = 0.29$, $p_t = 0.264 \pm 0.025$ y para $n = 12$, $p_r = 0.423$, $p_t = 0.431 \pm 0.025$[3]. Esto significa que el parámetro a puede ser considerado un indicador en la determinación del valor en el cual se produce la de transición p_t, ahora mediante una "*longitud característica p*".

[3]La genericidad de este resultado consiste en que el valor característico de relajación da una indicación de donde se observa el cruce de a con b

La explicación a este hecho puede verse de la siguiente manera. La probabilidad de obtener una secuencia de tamaño l puede ser aproximada por

$$P(l) = p^2(1-p)^l = p^2 exp^{l\ln(1-p)} = p^2 exp(-l/lo)$$

.

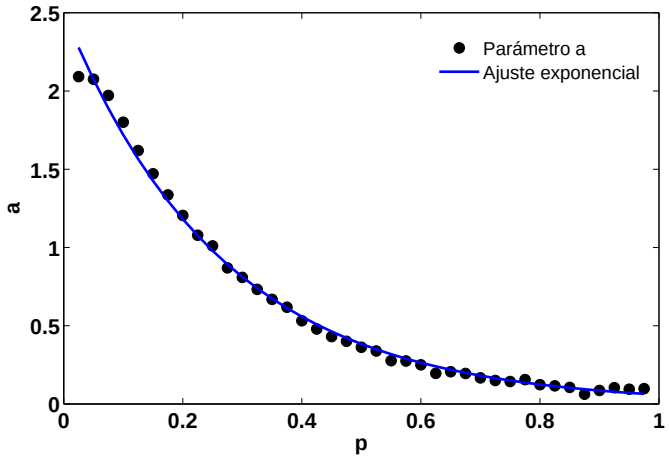


Figura 2.8: Ajuste exponencial del parámetro a de la FDB como función de la probabilidad de modificación p para *octámeros* generados por expansión-modificación con $R^2 = 0.99$

2.2.3. Párameteros (a, b) y eventos inusuales

Para $n = 8$, el número total de posibles *n-ómeros* es $2^8 = 256$, después de alcanzar una secuencia grande de 0 y 1 (típicamente tomamos secuencias de 1 X 10^7 n-omeros, en los cuales los valores de a, b coinciden aún si tomamos secuencias mayores a éstas, esto indica que la probabilidad de encontrar una secuencia ya no cambia) mediante la aplicación del algoritmo de expansion-modificación existe siempre la posibilidad de no encontrar un subconjunto de *n-ómeros*, es decir que su frecuencia de aparición sea 0. A éstos los llamaremos –naturalmente– *n-ómeros ausentes*. La ausencia de estas secuencias es un resultado poco probable que puede ser considerado como un evento raro. En la fig.2.9a, mostramos el número promedio de valores *n-ómeros ausentes* con respecto al parámetro de ajuste a, tomados sobre 10 realizaciones. Notemos

la línea recta del ajuste de mínimos cuadrados no lineales en la gráfica en semi-log para valores del promedio $a \geq 0.53$. Este crecimiento exponencial no está presente para valores menores que 0.53 y ocurre en $p = 0.4$ el cual se origina en la región III como se muestra en la fig.2.3b. En la figura 2.9c se muestra la gráfica análoga a la anterior, pero en términos de b. Se observa que con el incremento exponencial inicia con valores de la entropía de Shannon cercanos al máximo, es decir con $p \approx 1$ y este crecimiento alcanza su máximo valor para $p = 0.450$ (ver figura 2.3a). La línea recta en la gráfica semi-log es nuevamente un indicativo de un incremento exponencial de los *n-ómeros ausentes* con b.

En la figura 2.7 se observa que para valores inferiores a $p = 0.45$ las correlaciones largas se incrementan, esta dependencia con b es válida en la ausencia de correlaciones d. Después de una pequeña meseta alrededor del máximo valor de b con p disminuyendo, a pesar de que b decrece los eventos raros continúan incrementándose. En esta etapa el crecimiento exponencial con a mencionado previamente parece acelerarse.
El incremento de *n-ómeros ausentes* debido a a y b es de diferente naturaleza: para valores grandes de a el dominio de pocos *n-ómeros* produce una ocurrencia improbable de la mayor parte de los restantes. Para el caso de b, si empezamos de las condiciones más desordenadas, la homogeneidad característica de valores de entropía grandes es gradualmente interrumpida con el aumento de b permitiendo el aumento de la heterogeneidad y un incremento del número de eventos raros.

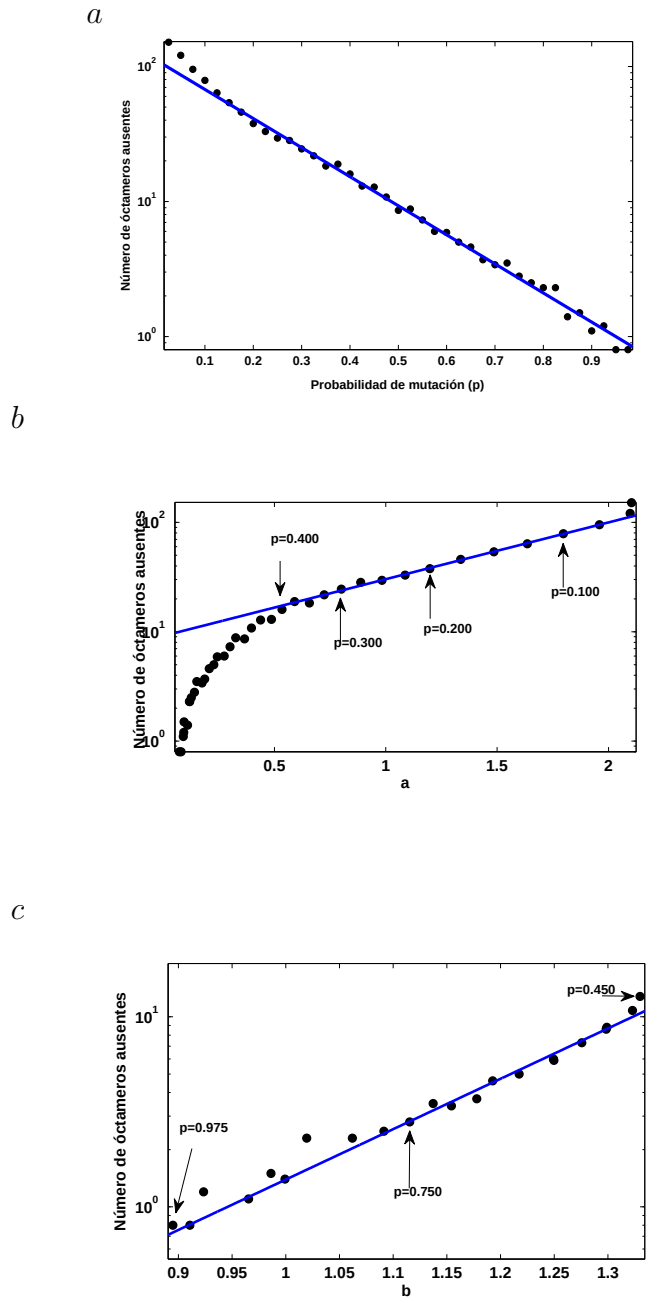


Figura 2.9: *a*) Dependencia del número *n-ómeros ausentes* como función de la probabilidad de mutación p sobre 10 realizaciones. *b*) Número de *n-ómeros ausentes* como función del valor promedio del parámetro a sobre 10 realizaciones. La línea recta es un ajuste no lineal usando el método de Levenberg-Marquardt con $R^2 = 0.99$, determinado en el rango de $a = 0.53$ a 2.01, correspondiente a valores de p entre 0.4 y 0.05. *b*) Variación del número de *n-ómeros ausentes* como fucnión del promedio del parámetro b correspondientes a los valores de $p \in (0.45, 0.975)$. La línea recta es un ajuste no lineal con $R^2 = 0.99$. En todos los casos se trata de 1x10^6 octámeros.

2.3. Variaciones del algoritmo de expansión-modificación

2.3.1. Retraso

Una variante del algoritmo de expansión-modificación consiste en permitir la posibilidad de mutar sólo después de un número d de pasos . Dentro de la analogía con respecto a las secuencias genéticas, esta implementación es más realista. Consideramos retrasos de $d = 3$ y 10, para los casos *6-ómeros* y *8-ómeros*. La fig.2.10 muestra que a, b la variación con respecto a la probabilidad de mutación p para $n = 8$ y $d = 3$. Notemos que la probabilidad de transición p_t cambia para valores más grandes con respecto a uno por $d = 0$ mostrado en la figura 2.3. Esto no es sorprendente puesto que la supresión de las modificaciones se pueden interpretar en términos de una menor probabilidad de modificación efectiva. En la ausencia de retrasos, el valor de p_t para este probabilidad efectiva de no retraso corresponde a un p_t con retraso. Notemos que las características cualitativas de la figura 2.3 se preservan con el retraso como se muestra en la figura 2.10.

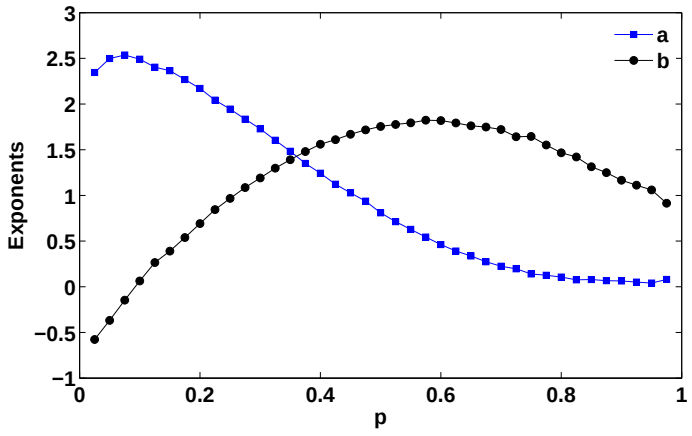


Figura 2.10: Parámetros de la FBC (a, b) como función de la probabilidad de modificación p para datos generados por la ecuación 2 con un retraso de $d = 3$

2.3.2. Caso $p \neq q$

Una gran variedad de comportamientos dinámicos pueden ser contemplados en los modelos de expansión-modificación cambiando las características deterministas y estocásticas en la evolución del mecanismo generatriz, es decir cambiando las reglas con las cuales los elementos binarios evolucionan y la probabilidad con la cual se llevan a cabo. Aquí, como ilustración, mostramos en la figura 2.11 la riqueza dinámica producida una vez que abandonamos la restricción $p = q$. La gráfica tridimensional exhibe la variación de los parámetros de ajuste (a, b) con cambios en las probabilidades p y q como se muestra en la ecuación 2.3

Debido a que p y q son independientes las relaciones entre los parámetros a y b con las características orden-desorden es menos directa y se vuelve dependiente del caso tratado. [4].

[4]Por ejemplo para un algoritmo diferente con ramificación constante:

$$
\begin{aligned}
1 &\rightarrow \begin{cases} 00 & \{p\} \\ 11 & \{1-p\} \end{cases} \\
0 &\rightarrow \begin{cases} 10 & \{q\} \\ 00 & \{1-q\} \end{cases}
\end{aligned}
$$

Se ha mostrado que para el caso $p = 2q$ para n-ómeros las relaciones funcionales entre p y q producen correlaciones de largo alcance para todos los valores de probabilidades [76, 77]

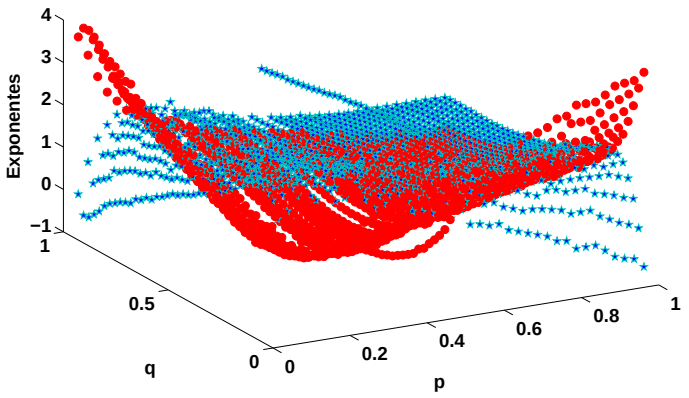


Figura 2.11: Gráfica de valores de los exponentes a y b como función de p y q, a está representada por circulos rojos y b por las estrellas azules.

2.4. Conexión con procesos multiplicativos.

Los procesos de expansión-modificación son mecanismos de ramificación que generan secuencias simbólicas en los que cada realización, después de N pasos, se alcanza con probabilidad dada por el producto de los factores que participan en la ruta de la ramificación. La rama está determinada por las probabilidades a priori de la implementación de las R posibles reglas definidas por el algoritmo esto es ejemplificado para $R = 4$ con el algoritmo de la ecuación 2.2 en donde las probabilidades de los diferentes resultados se muestran para 2 iniciando con una semilla 0 a $t = 0$

$$
\begin{array}{rc}
t = 1 & \\
00 & 1 - q \\
1 & q \\
t = 2 & \\
0000 & (1 - q)^3 \\
001 & (1 - q)^2 q \\
100 & q(1 - q)^2
\end{array}
$$

$$\begin{array}{cc} 11 & (1-q)q^2 \\ 11 & q(1-p) \\ 0 & qp \end{array}$$

cuando nos fijamos en un estado dado de la evolución de las secuencias la configuración de un *nómero* particular proviene, como primera aproximación, de los elementos previos que generan la configuración mediante el menor número de pasos. De aquí que, la probabilidad de obtener un *nómero* con el menor número de pasos es una buena aproximación a la probabilidad de encontrar el *n-ómero* en la secuencia final, una vez que las condiciones estacionarias se han alcanzado. Nuestro principal propósito para presentar esta perspectiva multiplicativa es que hace contacto con los cálculos algebraicos [25, 53] con las distribuciones orden-frecuencia tipo FBC que son generadas por procesos multiplicativos.

Es importante destacar la naturaleza cualitativa de los resultados aquí presentados, ya que lo que se pretende es la exposición y caracterización de comportamientos estadísticos, la complementaridad de descripciones diferentes y el vínculo con la dinámica de los procesos involucrados, todas en relación con los parámetros del ajuste (a, b).

Hemos identificado regiones con comportamiento colectivo claramente distinto, sin embargo, los valores de transición y las fronteras entre ellos deben ser vistos como estimaciones y no como determinaciones precisas. Es también importante mencionar que, a pesar de que concentramos nuestra atención en el caso de *octámeros*, hemos verificado que las principales características cualitativas que hemos obtenido son válidas para *n-ómeros* con $n = 6, 10, 12$. Hemos probado que éstas se preservan cuando incrementamos el tamaño de la secuencia generada por los procesos de expansión-modificación desde $240,000$ monómeros a $150,000$ y $360,000$. Este es un resultado esperado dado dado que las distribuciones normalizadas para diferentes tamaños N son muy similares y tienden a converger a una distribución única cuando N crece.

Un comentario general es que a pesar de que la dinámica en conflicto podría no ser una condición necesaria para las distribuciones beta orden-frecuencia, ésta proporciona condiciones que favorecen su ocurrencia. Hemos corroborado que eso también es el caso para familias de mapeos no lineales con dinámicas en conflicto mediante su dinámica simbólica que se presentamos en el capítulo siguiente.

Capítulo 3

Modelos deterministas

"Todo en la naturaleza es el resultado de leyes fijas" **Charles Darwin**
"Todo está determinado, el principio así como el fin, por fuerzas sobre las cuales no tenemos control. Todo está determinado tanto para el insecto como para la estrella. Los seres humanos, vegetales o el polvo cósmico, todos danzamos una misteriosa tonada, tocada en la distancia por un misterioso flautista."**Albert Einstein.**
"You need chaos in your soul to give birth to a dancing star". **Friedrich Nietzsche**

En este capítulo mostraremos un modelo caótico-determinista para la obtención de dinámicas simbólicas, en particular, trataremos con familias de mapeos caóticos unimodales. Con éstas exploramos, mediante la variación de un parámetro, las características topológicas y dinámicas del mapeo, estudiamos mediante la dinámica simbólica la aparición de la FBC y relacionamos los exponentes de ajuste a y b al parámetro que define la familia.

Se utiliza el formalismo termodinámico de los sistemas dinámicos, es decir, empleamos la estructura matemática de la física estadística para el estudio de microestados, que son, en términos dinámicos, las particiones del espacio fase de los mapeos, para calcular la función de partición y así encontrar las funciones termodinámicas de interés: en nuestro caso la(s) energía(s) libre(s). Mostramos que se encuentran transiciones de fase de primer y segundo orden (asociados a singularidades y/o discontinuidades de esta energía libre o sus derivadas) que están relacionadas con los cambios relativos de los parámetros de la FBC.

3.1. Mapeos unimodales caóticos

Los mapeos no-lineales en una dimensión, además de su importancia *per se*, han sido utilizados- por su simplicidad- como un laboratorio para el estudio de los distintos fenómenos que aparecen en los sistemas dinámicos caóticos, entre éstos se encuentran: la intermitencia, las rutas al caos, los exponentes de Lyapunov, las ventanas periódicas,etc. Dos de los representantes canónicos de estos mapeos son el logístico y el mapeo tienda.

Definición 1. *Un mapeo $f(x)$ es unimodal si existe una relación $x_{n+1} = f(x_n)$, con $f(x) : [a, b] \to [a, b]$, $f(a) = f(b) = 0$, y existe un único $x^* \in (a, b)$ tal que $f(x)$ es creciente para $0 \leq x < x^*$, y decreciente para $x^* \leq x \leq 1$*

En este trabajo estudiamos no sólo mapeos unimodales sino familias de éstos. La caracterización de una familia se determina mediante un parámetro adicional que selecciona un elemento dentro de un conjunto más amplio. En este capítulo estudiamos, particularmente, familias de mapeos que tienen dos regiones bien definidas por sus dinámicas opuestas,es decir, que existan regiones caracterizadas con el tiempo típico de permanencia de iteraciones: queremos una región laminar (permanencia) y otra expansiva (cambio)[78].[1]

Con la conjetura establecida en el capítulo anterior sobre los significados de a y b, queremos determinar analíticamente si un modelo permite corroborar y calcular propiedades de estos valores cuyo significado prefiguramos en el capítulo 2.

A continuación, presentamos las familias de mapeos unimodales que tratamos en este trabajo. Éstas tienen la característica de que el parámetro que determina el elemento de la familia modifica también sus propiedades de concavidad, exponentes de Liapunov (λ)[2], derivada schwarziana[3], y el tamaño de sus regiones laminar y caótica.

3.1.1. Familia logística generalizada

Una de las propuestas más interesantes y fructíferas abordadas en este trabajo consiste en el estudio de la siguiente familia de mapeos:

$$x_{n+1} = \lambda(1 - |1 - 2x_n|^\nu) = f(x; \lambda, \nu) \tag{3.1}$$

[1] Usamos la dicotomía laminar-caótico en analogía con dinámica de fluídos.

[2] Para mapeos unimodales en una dimensión $f(x)$, $\lambda = \frac{1}{n}\Pi_{i=0}^{n-1} \ln |f'(x_i)|$

[3] $(Sf)(z) = \left(\frac{f''(x)}{f'(x)}\right)' - \frac{1}{2}\left(\frac{f''(x)}{f(x)}\right)^2$

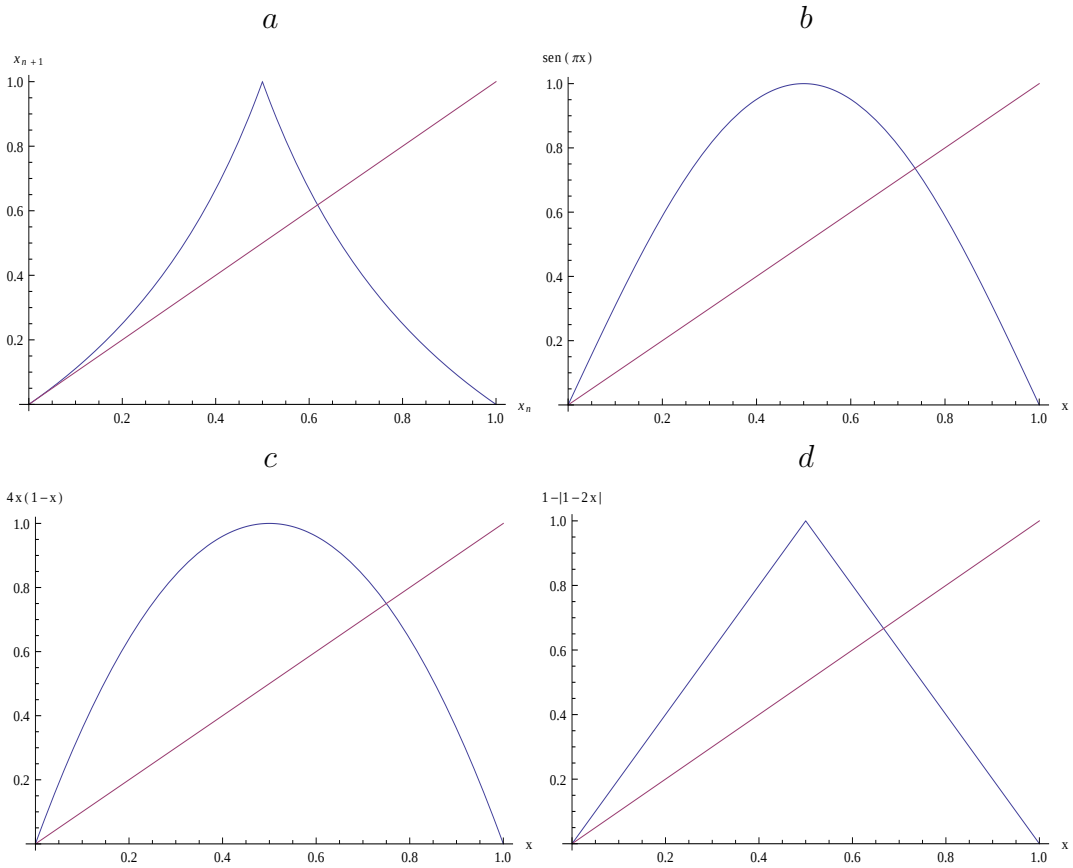


Figura 3.1: Ejemplos de mapeos unimodales conocidos. Se grafica también la función identidad para referencia. **a.** Mapeo de Farey. **b.** Mapeo seno $x_{n+1} = \sin(\pi x_n)$.**c.** Mapeo logístico $x_{n+1} = 4x_n(1 - x_n)$.**d.** Mapeo tienda $x_{n+1} = 1 - |1 - 2x_n|$.

con $0 \leq x \leq 1$, $0 < \lambda \leq 1$ y $\frac{1}{2} \leq \nu < \infty$.

El parámetro ν caracteriza a la familia – da el orden del máximo– , y λ determina la caoticidad del mapeo ya que controla la duplicación de período como ruta al caos. Con $\lambda = 1$ todos los mapeos de la familia se encuentran en el régimen caótico (ver figura 3.2), por lo que, en lo sucesivo supondremos este valor para todo elemento de la familia.

Una de las ventajas de esta familia de mapeos consiste en que para algunos valores particulares de ν se obtienen mapeos conocidos (ver figura 3.4A),en particular con $\nu = 2$ se tiene el mapeo logístico, por ello a la familia de mapeos definida por la ecuación 3.1 la denominaremos logística generalizada:

- ($\nu = \frac{1}{2}$) Mapeo tangente a la recta identidad.

$$x_{n+1} = 1 - \sqrt{|1 - 2x_n|}$$

- ($\nu = 1$) Mapeo tienda.

$$x_{x+1} = 1 - |1 - 2x_n|$$

- ($\nu = 2$) Mapeo logístico.

$$x_{n+1} = 1 - (1 - 2x_n)^2 = 4x_n(1 - x_n)$$

.

Se observa que conforme ν aumenta las propiedades de los mapeos cambian, con $\nu = 1/2$ el mapeo es tangente a la recta identidad en el origen, es decir que $f'(x = 0) = 1$, esta propiedad es una característica de los mapeos intermitentes con lo que la región laminar predomina sobre la región caótica; para $\nu = 1$, el mapeo tienda marca un límite en la concavidad para $\nu > 1$, por lo que los mapeos son convexos. La derivada schwarziana se vuelve negativa para $\nu > 1$ (ruta de duplicación de periodo hacia el caos).

En nuestro estudio de n-omeros mediante mapeos, lo que se toma en cuenta es la frecuencia relativa de secuencias de símbolos consecutivos de tamaño finito, que en el caso de un número grande de realizaciones, se puede equiparar con la probabilidad de que dado un punto arbitrario $x_0 \in [0, 1]$ éste genere,

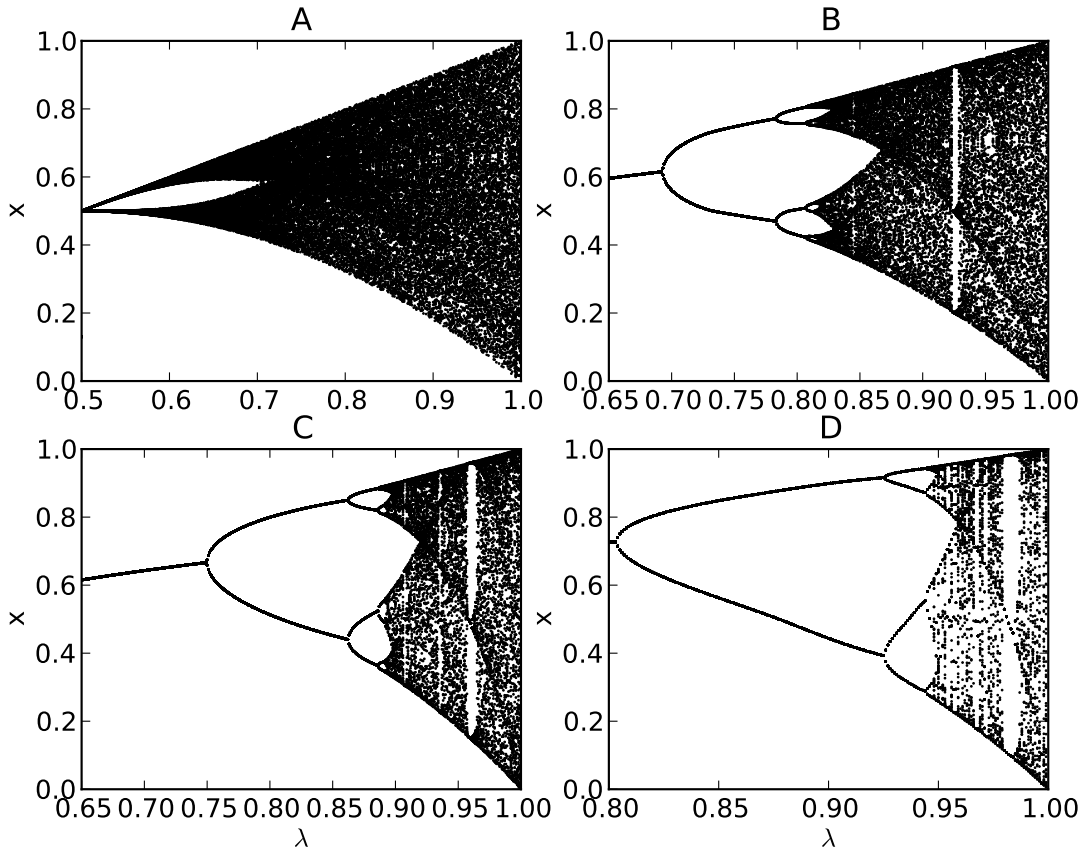


Figura 3.2: Diagrama de bifurcaciones de la familia propuesta en la ecuación 3.1. **A** con $\nu = 1$ Mapeo tienda, **B** $\nu = 1.5$, **C** con $\nu = 2$ mapeo logístico, **D** $\nu = 3$.

mediante iteraciones sucesivas, dicha secuencia. Es por ello, que nos enfocaremos al estudio de medidas probabilísticas asociadas a secuencias simbólicas, con este motivo tabajaremos con el operador de Perron-Frobenius, las densidades invariantes, las medidas probabilísticas, los cilindros y particiones de mapeos unimodales caóticos.

3.1.2. Operador de Perron-Frobenius y densidad invariante

Encontramos el operador de Perron-Frobenius (OPF) para la familia logística. Recordemos que el OPF es un operador sobre el espacio de densidades y permite explorar las distribuciones de puntos bajo la acción de un mapeo, que contrasta con el estudio de puntos en el espacio fase (trayectorias) que constituye el estudio tradicional de los sistemas dinámicos discretos.

El OPF también es conocido como operador de transferencia (como operador de Ruelle y como operador de Ruelle-Perron-Frobenius) se utiliza principalmente para estudiar sistemas dinámicos, mecánica estadística y el caos cuántico [79].

Definición 2. *Sea un mapeo $M(x)$ y una función de densidad $\rho(x)$ ($\rho_n(x)$ es la n-ésima aplicación del OPF sobre $\rho_0(x)$), entonces el operador de Perron-Frobenius $\hat{\mathbf{P}}$ (OPF, de aquí en adelante) se define de la siguiente forma:*

El OPF se aplica sobre densidades, entenderemos este término como funciones $\rho(x) : [0,1] \to [0,1]$ de tal forma que $\int_0^1 \rho(x)dx = 1$

$$\hat{\mathbf{P}}\rho(x) \equiv \frac{d}{dx}\int_{M^{-1}[a,x]} \rho(u)du \tag{3.2}$$
$$\rho_{n+1}(x) \equiv \frac{d}{dx}\int_{M^{-1}[a,x]} \rho_n(u)du$$

con $M^{-1}[a,x] = \{s|M(s) \in [a,x]\}$ la imagen inversa del intervalo $[a,x]$.

Las definiciones previas muestran ambas perspectivas del OPF: es decir, visto como un operador $\hat{\mathbf{P}}$ que actúa sobre un espacio de funciones de distribución (FDP); o como una ecuación de recurrencia. Desde estos puntos de vista, la densidad invariante $\rho^*(x)$ es respectivamente la eigenfunción con eigenvalor $\gamma = 1$ del operador $\hat{\mathbf{P}}$, es decir $\hat{\mathbf{P}}\rho^*(x) = \rho^*(x)$; o la densidad invariante es un punto fijo de la ecuación de recurrencia para las densidades[4].

[4]El operador de Perron-Frobenius $\hat{\mathbf{P}}$ tiene entre otras las siguientes propiedades:

Si el mapeo es diferenciable e invertible a pedazos la ecuación 3.3 puede escribirse equivalentemente como

$$\hat{\mathbf{P}}\rho(x) = \sum_{x=M^{-1}(u)} \frac{\rho(u)}{|M'(u)|} \tag{3.3}$$

El conjunto de operadores PF para la familia logística es el siguiente. Determinemos primero la imagen inversa de la familia de mapeos definida por la ecuación 3.1.

$$x' = 1 - |1 - 2x|^{\nu} \tag{3.4}$$

En el intervalo $1/2 \leq \nu < \infty$ y $0 \leq x \leq 1/2$. La solución para x' es

$$x = \frac{1}{2}(1 - (1 - x')^{\frac{1}{\nu}}) \tag{3.5}$$

con lo que, el conjunto buscado consiste en los intervalos siguientes: (se aprovecha la simetría de los mapeos, es decir para $1/2 \leq x \leq 1$, el mapeo se obtiene sustituyendo $x \to 1 - x$, con $0 \leq x \leq \frac{1}{2}$)

$$f^{-1}(x;\nu)([0,x]) = [0, \frac{1}{2}(1 - (1-x)^{\frac{1}{\nu}})] \bigcup [\frac{1}{2}(1 + (1-x)^{\frac{1}{\nu}}), 1] \tag{3.6}$$

Si sustituimos en la ecuación 3.3

$$\hat{\mathbf{P}}_{\nu}\rho(x) = \frac{d}{dx}\int_0^{\frac{1}{2}(1-(1-x)^{\frac{1}{\nu}})} \rho(u)du + \frac{d}{dx}\int_{\frac{1}{2}(1+(1-x)^{\frac{1}{\nu}})}^1 \rho(u)du \tag{3.7}$$

La familia de OPFs determinada por el parámetro ν queda explícitamente definida de la siguiente forma:

$$\hat{\mathbf{P}}_{\nu}\rho(x) = \frac{1}{2\nu}(1-x)^{\frac{1}{\nu}-1}\{\rho(\frac{1}{2} - \frac{1}{2}(1-x)^{\frac{1}{\nu}}) + \rho(\frac{1}{2} + \frac{1}{2}(1-x)^{\frac{1}{\nu}})\} \tag{3.8}$$

Las $\rho^*(x)$ invariantes para esta familia de OPFs son sólo conocidas para un conjunto muy particular de valores del parámetro $\nu = \frac{1}{2}, 1, 2$ [80].

1. $\hat{\mathbf{P}}f \geq 0 \; \forall f \geq 0$
2. $||\hat{\mathbf{P}}f|| \leq ||f||$, $||f||$ denota la L^1 norma.
3. $||\hat{\mathbf{P}}^n f_1 - \hat{\mathbf{P}}^n f_2|| \leq ||\hat{\mathbf{P}}^{n-1} f_1 - \hat{\mathbf{P}}^{n-1} f_2||$, propiedad de contracción bajo la acción del operador.

- $\nu = \frac{1}{2}$ la densidad invariante es

$$\rho^*(x) = 2(1-x)$$

- Para $\nu = 1$

$$\rho^*(x) = 1$$

- Para $\nu = 2$

$$\rho^*(x) = \frac{1}{\pi\sqrt{x(1-x)}}$$

Una cuestión interesante es saber si existe una familia de densidades invariantes asociada a la familia logística generalizada, es decir, que sean soluciones de las ecuaciones 3.8 que, naturalmente, incluya los casos arriba señalados[5].

En este trabajo proponemos una familia de densidades como solución a la ecuación 3.8.

3.1.2.1. Propuesta de densidad invariante.

Proponemos la siguiente familia de funciones como *ansatz* a las densidades invariantes en el intervalo $\nu < 1 < \infty$.

$$\rho(x) = \frac{\Gamma(\frac{2}{\nu})}{\Gamma(\frac{1}{\nu})^2} x^{\frac{1}{\nu}-1}(1-x)^{\frac{1}{\nu}-1} \tag{3.9}$$

con $\Gamma(x)$ la función gama defina por

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$$

y

$$\Gamma(n) = (n-1)!$$

para n un número natural. La constante

$$\frac{\Gamma(\frac{2}{\nu})}{\Gamma(\frac{1}{\nu})^2}$$

[5]Encontrar densidades invariantes es complicado, de hecho se saben las formas de las densidades invariantes de sólo algunos mapeos de importancia.

normaliza la expresión 3.9. Para el caso de $\nu = 1$ da la expresión correcta –incluida la constante de normalización– de la densidad invariante, es decir : $\rho(x) = 1$, lo mismo sucede para $\nu = 2$ con lo que se tiene: $\rho(x) = \frac{1}{\pi\sqrt{x(1-x)}}$.

Una forma un tanto redundante, pero ilustrativa del efecto del OPF sobre las densidades consiste en sustituir las expresiones de la densidad invariante para $\nu = 0.5, 1, 2$ y mostrar que son funciones propias de $\hat{\mathbf{P}}$

Con $\nu = \frac{1}{2}$ la densidad invariante es $\rho(x) = 2(1-x)$

$$\hat{\mathbf{P}}_{\frac{1}{2}} 2(1-x) = (1-x)\{(1+(1-x)^2) + (1-(1-x)^2)\} = 2(1-x) \quad (3.10)$$

Para $\nu = 1$, (mapeo tienda), el punto fijo de este operador es la densidad invariante $\rho(x) = 1$ que es función propia del OPF para la familia logística, como se muestra en reemplazando esta densidad en el OPF:

$$\hat{\mathbf{P}}_1 1 = \frac{1}{2}(1-x)^{1-1}(1+1) = 1 \quad (3.11)$$

y con $\nu = 2$ para $\rho(x) = \frac{1}{\pi\sqrt{x(1-x)}}$.

$$\begin{aligned} \hat{\mathbf{P}}_2 \frac{1}{\pi\sqrt{x(1-x)}} &= \tfrac{1}{4}(1-x)^{\frac{1}{2}-1}(\rho(\tfrac{1}{2} - \tfrac{1}{2}(1-x)^{\frac{1}{2}}) + \rho(\tfrac{1}{2} + \tfrac{1}{2}(1-x)^{\frac{1}{2}})) \\ &= \tfrac{1}{4}(1-x)^{\frac{1}{2}}(\tfrac{2}{\pi\sqrt{x}} + \tfrac{2}{\pi\sqrt{x}}) = \tfrac{1}{\pi\sqrt{x(1-x)}} \end{aligned} \quad (3.12)$$

La propuesta de densidad invariante aquí descrita, debe ser una función propia del OPF para todos los valores de ν permitidos, es decir $\frac{1}{2} \leq \nu < \infty$.

Como se muestra en la siguiente ecuación 3.13 esta propuesta no es correcta

$$\hat{\mathbf{P}}_\nu \rho(x) = \frac{(x(1-x))^{\frac{1}{\nu}-1}\Gamma\left(\frac{2}{\nu}\right)}{\Gamma\left(\frac{1}{\nu}\right)^2} \left[\frac{2\left(1-(1-x)^{2/\nu}\right)^{\frac{1}{\nu}-1}(1-x)^{\frac{1}{\nu}-1}\Gamma\left(\frac{1}{2}+\frac{1}{\nu}\right)}{\sqrt{\pi}\nu\Gamma\left(\frac{1}{\nu}\right)} \right] \neq \rho(x) \quad (3.13)$$

En el anexo A se aplica iterativamente el OPF para obtener las densidades resultantes, a partir de la densidad inicial como el *ansatz* de la ecuación 3.9.

Sin embargo, esta familia de densidades, parece ser una buena aproximación a la densidad invariante para el intervalo $1 \leq \nu \leq 2$, como se muestra

en la figura 3.3. En esta gráfica se exhibe la existencia de las densidades invariantes asociadas a la familia logística, ya que los histogramas convergen; además se grafican el *ansatz* 3.9 para efecto comparativo, como se observa nuestra propuesta parece ser un buen candidato a la densidad invariante de la familia logística.

Este hecho nos permitirá trabajar con la ecuación 3.9 para aproximar las densidades invariantes de la familia logística .

3.1.3. Dinámica simbólica

A pesar de que los mapeos son en el espacio continuos, no así en el tiempo, es posible expresarlos de manera discreta sin perder información en el proceso de convertirlos en iteraciones sobre conjuntos de celdas.

Es por ello que, con el objetivo de estudiarlos utlizaremos algunas herramientas de discretización y con ello poder comparar modelos distintos entre sí. La dinámica simbólica es por si misma motivo de interesantes investigaciones [81, 82], sin embargo, en este trabajo la usamos como herramienta y como marco común de distintos modelos cuya expresión en dinámicas simbólicas queremos comparar.

En el caso de mapeos unimodales es posible establecer una dinámica simbólica asociada mediante el proceso usual, es decir: se divide el espacio fase en N particiones ϕ_k, $k = 1, 2 \ldots, N$ y con $\phi_j \bigcap \phi_i = \phi, \forall\, i, j \in 1, 2, \ldots, N$ asignándoles un símbolo a cada una y se consideran secuencias de iteraciones de los mapeos, expresadas ahora como secuencia de símbolos y no de trayectorias.

3.1.4. Particiones

Consideremos un sistema dinámico en un espacio fase unidimensional, las particiones se definen como el conjunto $\{A_i\}$ tales que todos los elementos son disjuntos y la unión de ellos constituya el espacio fase total X.

$$\bigcup_{i=1}^{K} A_i = X \tag{3.14}$$

$$A_i \cap A_j = \emptyset \tag{3.15}$$

$\forall i, j$

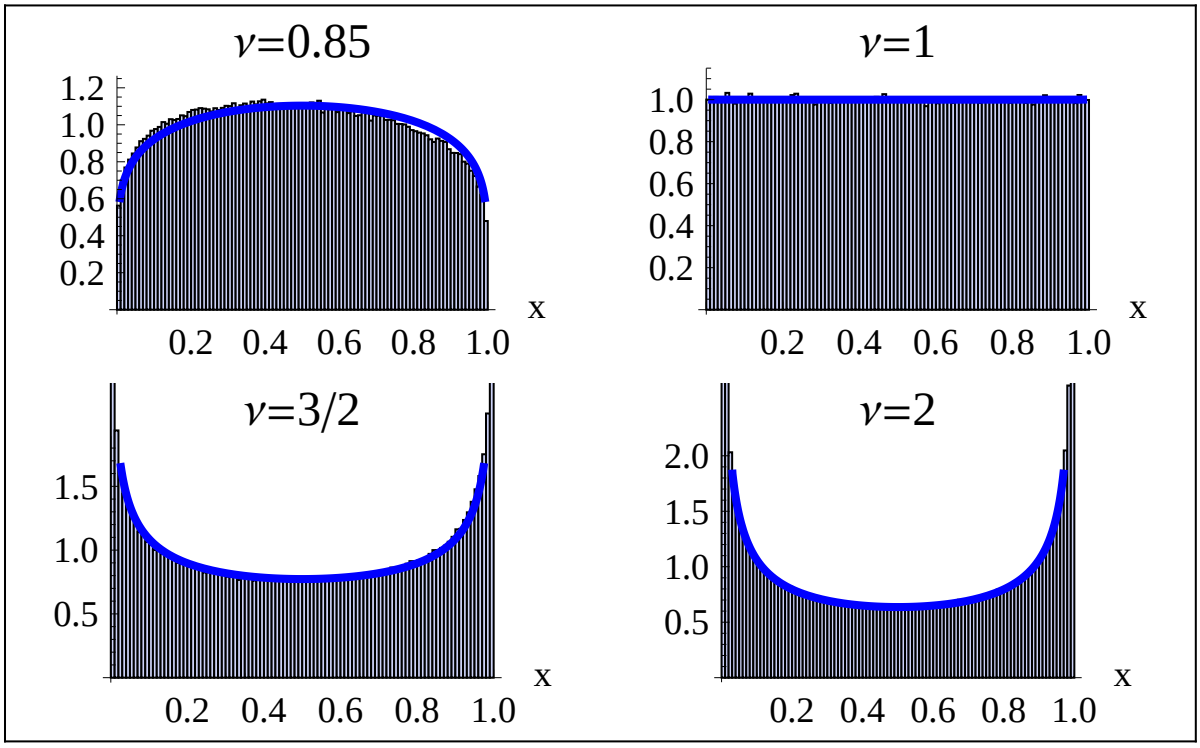


Figura 3.3: Histogramas que muestran un comparativo entre el *ansatz* de la ecuación 3.9 y los histogramas de las iteraciones de los mapeos con valores de los parámetros señalados en la figura algunos de los cuales ($\nu = 0.85, 1.5$) no se conoce la expresión correcta de su densidad invariante. Los histogramas están calculados para 10^6 iteraciones con un transitorio de $100,000$ y 1000 particiones iguales del intervalo $[0, 1]$.

Consideremos una partición del espacio fase $\mathcal{X}$ de tamaño K , etiquetemos entonces cada celda consecutivamente $i = 1, 2, 3, \ldots K$. La trayectoria de $x_0 \in X$ en dinámica simbólica consiste en la secuencia de símbolos o etiquetas establecidas para cada celda del espacio fase.

3.1.4.1. Partición generatriz

Una partición fundamental para el estudio de los dinámica simbólica es la partición generatriz debido a sus muy particulares propiedades:

Definición 3. *Consideremos una partición $\mathcal{G}$, de tamaño R, se dice que $\mathcal{G}$, es una partición generatriz de un mapeo $M(x)$, si la secuencia simbólica infinita generada $S = s_0, s_1, s_2, s_3, \ldots$ está determinada de manera unívoca por un único valor inicial $x_0 \in \mathcal{X}$.*

No es fácil saber si dado un mapeo $M(x)$ arbitrario existe o no una partición generatriz; además tampoco existe una forma general de obtener una partición de este tipo. En general depende del mapeo, si no se conoce una partición entonces se pueden hacer aproximaciones, considerando, por ejemplo, celdas del mismo tamaño, que en el límite cuando el tamaño de las celdas tiende a cero, puede ser considerada una buena aproximación ($R \approx \epsilon^{-1}$ a los cilindros, con ϵ el tamaño típico de una celda). Sin embargo, para una clase muy particular de mapeos, encontrar una partición generatriz es muy sencillo: los mapeos unimodales definen una partición generatriz a partir de dos celdas *izquierda* y *derecha*, delimitadas por la posición del máximo. Así, para el mapeo logístico, la partición generatriz es $\mathcal{G} = [0, x_{max}] \cup [x_{max}, 1]$, para este trabajo definiremos de manera genérica, dos símbolos para estas celdas $S = \{0, 1\}$, con ello, mediante un proceso iterativo, se generan secuencias simbólicas de $0, 1$ a partir de condiciones iniciales $x_0 \in [0, 1]$.Es decir, tienen la propiedad de que $\forall\ s_0 s_1 s_2 \ldots, s_n$ con $n \to \infty$ existe un único x_0 que genera esa secuencia. La dinámica simbólica para el mapeo logístico ha sido ampliamente estudiada: se han estudiado órbitas periódicas de distintos tamaños, la estabilidad de las órbitas, la secuencia *kneading*, entre otras [83, 84, 85, 86].

Siempre que sea posible, ya sea vía el OPF u algún otro método, es muy valioso encontrar la densidad invariante asociada a un mapeo, éste permite establecer algunas propiedades interesantes, por ejemplo, se puede definir una medida para conjuntos del espacio fase de la manera usual:

Definición 4. *Sean $\rho(x)$ una función de densidad invariante de un mapeo*

$M(x)$ y $A \subseteq X$ entonces la medida invariante $m(A)$ de A es:

$$m(A) = \int_A \rho(x)dx \tag{3.16}$$

La medida m satisface algunas propiedades particularmente útiles para los propósitos de este trabajo:

para una partición del $\mathcal{X}$ con $A_i \subseteq \mathcal{X}$,

$$m(A_i \cup A_j) = m(A_i) + m(A_j) \tag{3.17}$$

$$m(\mathcal{X}) = 1 \tag{3.18}$$

3.1.5. Cilindros

La relación entre la frecuencia de n-ómeros y la medida de los conjuntos definida en la sección anterior se clarifica, si podemos asociar estos conjuntos A_i con las condiciones iniciales que generan, bajo la iteración de un mapeo, las secuencias de 0 y 1.

Definición 5. *Sea $\{A_i\}$, $i = 1, 2, \ldots N$ una partición del espacio fase $\mathcal{X}$, un cilindro $J[s_0, s_1, \ldots, s_{n-1}]$ de un sistema dinámico $M(u)$, se define como el conjunto*
$J[s_0, s_1, \ldots, s_{n-1}] = \{x \in \mathcal{X} | (x_0 = M^{(0)}(x_0) \rightarrow s_0) \wedge (x_1 = M^{(1)}(x_0) \rightarrow s_1) \wedge \ldots \wedge (x_{n-1} = M^{(n)}(x_0) \rightarrow s_{n-1})\}$

con $s_0, s_1, \ldots, s_{n-1}$, símbolos de una partición de X.

En términos llanos, los cilindros son el conjunto de condiciones iniciales que, bajo la acción de un mapeo, dan lugar a una secuencia simbólica dada de tamaño n. Por ejemplo, si bajo la acción de un mapeo no es posible acceder a alguna secuencia simbólica (secuencia prohibida) entonces:

$$J[s_0, s_1, \ldots, s_{n-1}] = \phi$$

además

$$\bigcup_{s_0, s_1, \ldots, s_{n-1}} J[s_0, s_1, \ldots, s_{n-1}] = \mathcal{X}$$

De esta manera se puede calcular dada la medida definida por la ecuación 3.16 de un cilindro arbitrario

$$m(J[s_0, s_1, \ldots, s_{n-1}]) = \int_{J[s_0, s_1, \ldots, s_{n-1}]} \rho(x)dx \tag{3.19}$$

La ecuación anterior puede entenderse como el "peso relativo" de la secuencia simbólica $s_0, s_1, \ldots, s_{n-1}$ para un mapeo dada, por supuesto, que dados dos mapeos esto valores son en general distintos, debido a la propiedad de aditividad y normalización esta medida es considerada como una medida probabilística [87], por lo que esta medida establece un indicativo de la probabilidad (frecuencia relativa) de una secuencia arbitraria bajo la iteración de un mapeo.

3.2. Dinámica simbólica y n-ómeros

En esta tesis utilizamos los mapeos unimodales expresados en dinámica simbólica para estudiar la frecuencia de aparición de secuencias simbólicas de distintos tamaños mediante particiones generatrices. En particular, el estudio de las frecuencias de elementos consecutivos de símbolos que llamaremos *n-ómeros (dímeros, trimeros, tetrámeros, etc)* ordenados descendentemente de acuerdo a su frecuencia para ajustarlos a una FBC. Es decir, consideramos la partición generatriz para generar secuencias de símbolos

$$S = s_0 s_1 \ldots s_n \ldots$$

con

$$s_i = 1, 0$$

en donde los símbolos están determinados por la regla usual:

$$s_j \quad = \quad \begin{cases} 0 & f^{(j)}(x_0) < \frac{1}{2} \\ 1 & f^{(j)}(x_0) \geq \frac{1}{2} \end{cases} \tag{3.20}$$

con $x_0 \in (0, 1)$ un punto inicial arbitrario y $M^{(j)}(x_0)$ la j-ésima iteración del mapeo unimodal $M(x)$.

Para un mapeo unimodal fijo nuestro objetivo consiste en estudiar secuencias simbólicas obtenidas mediante iteraciones con el procedimiento previamente explicado. Por ejemplo, para el caso de 8 elementos consecutivos consideramos la estadística de octámeros en secuencias S, es decir:

$$S = \underbrace{00101011}_{43}\underbrace{00110011}_{51}\underbrace{10110101}_{181}\ldots^{6}$$

[6]Debajo de cada conjunto de 8 símbolos se escribe su correspondiente expresión en base 10.

En el caso de la familia logística, consideramos una condición inicial al azar x_0 y obtenemos una secuencia de 8 X 10^6 símbolos y calculamos la frecuencia de aparición de octámeros, que son en total $2^8 = 256$ posibles, el proceso se repite con una condición inicial distinta, después de 10 realizaciones promediamos las frecuencias. Graficamos estos datos en orden-frecuencia y ajustamos una FBC.

Posteriormente variamos el parámetro que define al mapeo para explorar el efecto del cambio de regiones laminares y caóticas y su reflejo en la determinación de los parámetros del ajuste de la FBC.

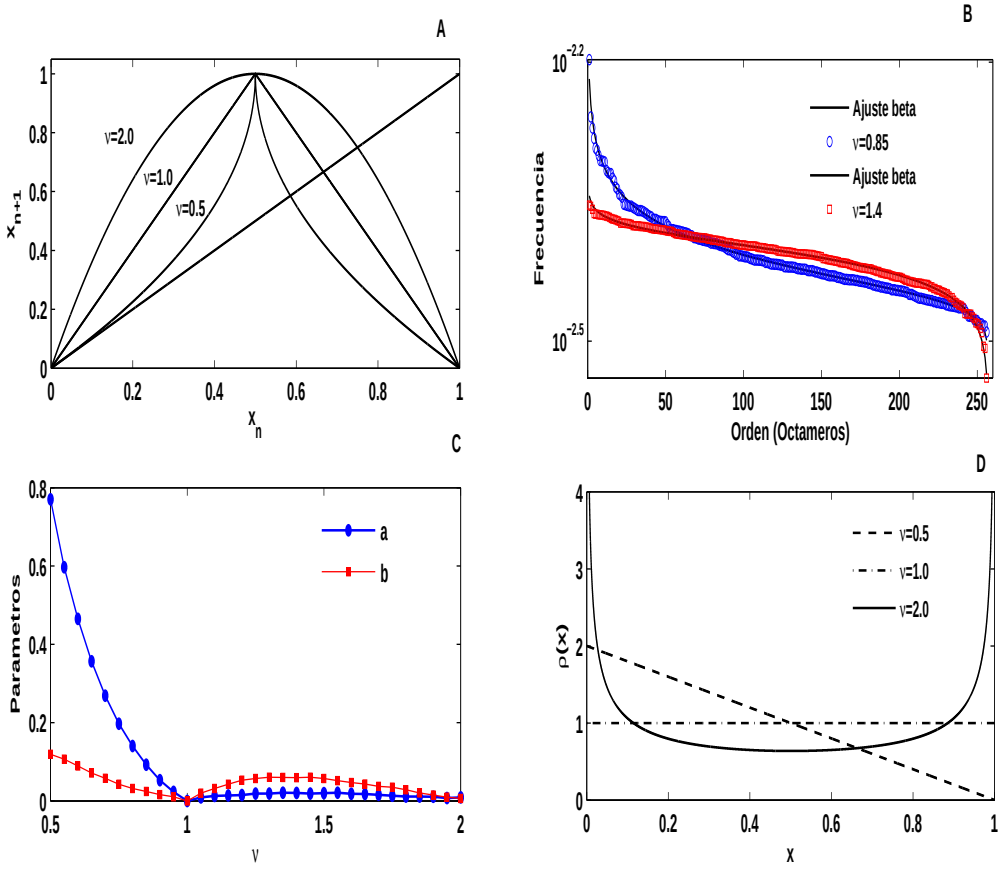


Figura 3.4: ***A***. La familia logística generalizada para distintos valores del parámetro $\nu = 0.5, 1, 2$ correspondientes a los mapeos tangente, tienda, y logístico respectivamente. La transición intermitencia-caos se da con el aumento del valor del parámetro ν. ***B***. Ajuste a la función beta para los valores del parámetro señalados en la figura, se muestran casos extremos en donde $a > b$ con $\nu < 1$ y $a < b$ con $\nu > 1$. ***C***. Parámetros de ajuste *a,b vs.* ν ver ecuación 3.1 con $\lambda = 1$. ***D***. Densidades invariantes para $\nu = 0.5$ (intermitencia), $\nu = 1$ $\rho(x) = 1$ (mapeo tienda) inicio de la duplicación como período como ruta al caos, y $\nu = 2$ (mapeo logístico) $\rho(x) = \frac{1}{\pi\sqrt{x(1-x)}}$.

De esta forma para el caso de *octámeros* se observa la gráfica orden-frecuencia en las figuras 3.4B para el caso de la familia logística. Al igual

que en el caso de los modelos de expansión-modificación observamos que la variación de la dinámica se refleja en los valores relativos de a, b:

- con $\frac{1}{2} < \nu < 1$, dominan las regiones laminares , $a \neq 0$ y $b \approx 0$ y $a > b$, se observa la posible presencia de correlaciones largas en esta zona y la región de intermitencia gobierna en estos valores de ν, además del predominio de algunas pocas secuencias, en donde abundan los 0.

- $\nu = 1$, es decir en el mapeo tienda, marca el límite del dominio de a sobre b, en este punto $a = b = 0$, como debería esperarse debido a que la densidad invariante es la función uniforme ($\rho(x) = 1$), todos los octámeros tienen la misma probabilidad de aparecer que tiende a $1/2^8$.

- $\nu > 1$ partir de este valor la concavidad de los mapeos cambia y el régimen expansivo domina sobre la región laminar, y además $a < b$, para una región, sin embargo con $\nu = 2$ (mapeo logístico) los valores a y b vuelven a coincidir debido a que el mapeo tienda y el logístico son equivalentes topológicos, a pesar de que las densidades difieren, pero al calcular los cilindros y las medidas de éstos, coinciden para valores grandes de número de iteraciones, con las probabilidades (frecuencias) de los n-ómeros.

Queremos encontrar otras familias de mapeos que compartan estas características para intentar mostrar la existencia de un comportamiento genérico de los exponentes de ajuste a, b.

3.2.1. Familia ϵ

Proponemos una variación interesante de la familia logística generalizada: la familia épsilon, que es una combinación lineal del mapeo tienda y logístico.

$$x_{n+1} = 1 - \epsilon|1 - 2x_n| + (\epsilon - 1)(1 - 2x_n)^2 = f(x_n) \tag{3.21}$$

No sorprende que el comportamiento sea similar al observado con la familia logística como se muestra en la figura 3.5.

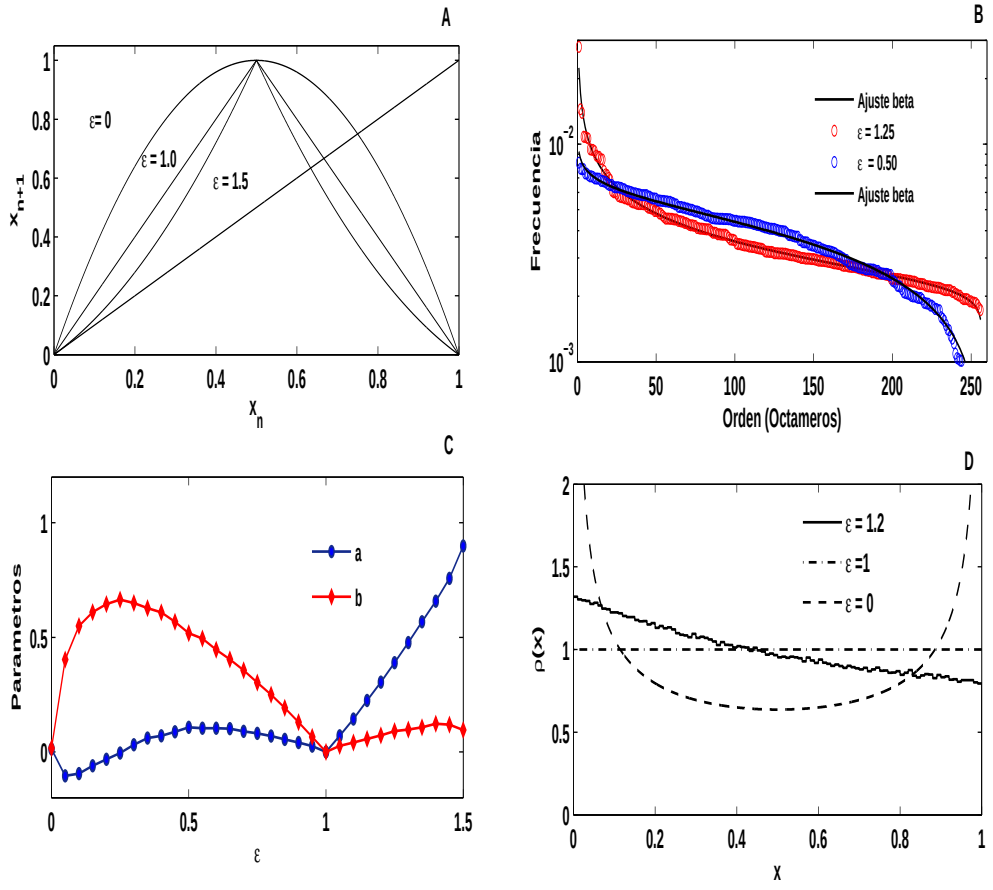


Figura 3.5: ***A***. La familia ϵ definida en la ecuación 3.21 para distintos valores del parámetro $\epsilon = 3/2, 1, 0$ correspondientes a los mapeos tangente, tienda, y logístico respectivamente. La transición intermitencia-caos se da con el aumento del valor del parámetro ϵ. ***B***. Ajuste a la función beta para los valores del parámetro señalados en la figura, se muestran casos extremos en los cuales $a > b$ con $\epsilon > 1$ y $a < b$ con $\epsilon < 1$. ***C***. Parámetros de ajuste a, b *vs.* ϵ. La inversión de los valores de a, b se da en $\epsilon = 1$ (mapeo tienda). ***D***. Densidades invariantes para $\epsilon = 1.2$ (intermitencia, en este caso se muestra un histograma ya se desconoce la expresión exacta de $\rho(x)$) , $\epsilon = 1$, $\rho(x) = 1$ (mapeo tienda) inicio de la duplicación como período como ruta al caos, y $\epsilon = 0$ (mapeo logístico), $\rho(x) = \frac{1}{\pi\sqrt{x(1-x)}}$.

3.3. Divergencias, intermitencia-caoticidad y betas

Ahora probaremos una conjetura acerca de la naturaleza de la transición asociada al pasar de un régimen con $a > b$ a otro en donde se inveirte esta relación $b < a$. En el capítulo anterior, después de nuestro estudio numérico establecimos la hipótesis de que se trata de una transición "orden-desorden". En esta caso estableceremos una manera de formalizar y tratar de demostrar la existencia de un transición, para ello utilizamos el formalismo termodinámico de los sistemas dinámicos. Ver anexo B.

3.3.1. Perron-Frobenius inverso

Queremos probar la siguiente conjetura: las divergencias de las densidades invariantes son la causa de la transición, debido a que cambian la forma de escalamiento por regiones del soporte de la densidad invariante. por ello es conveniente trabajar de entrada con familias de densidades invariantes y entonces determinar que mapeos unimodales caóticos generan. Este problema es una versión restringida del problema inverso de Perron-Frobenius, que ha sido abordado de manera satisfactoria en las siguientes referencias [88, 89, 90].

Los detalles de la implementación pueden consultarse en las referencias arriba mencionadas, por el momento estamos interesados en ligar propiedades de las divergencias en las densidades con transiciones de fase y cambios en los valores a, b, así como las propiedades de intermitencia y caos como un elemento genérico que determina el significado de los valores de ajuste de los parámetros a, b .

Densidades con singularidades en un extremo

Consideraremos densidades con divergencias en alguno de los extremos de su soporte, es decir $x = 0$ y $x = 1$, buscamos mapeos cuya densidad invariante sean de la forma $\rho_1(x) = A_1 x^{\alpha_1 - 1}$ o $\rho_2(x) = A_2(1-x)^{\alpha_2 - 1}$.

Deseamos determinar cuál es el mapeo unimodal caótico y simétrico[7] que se obtiene con cada una de las densidades propuestas.

Para $\rho_1(x)$, $\rho_2(x)$ la ecuación de Perron-Frobenius (ecuación 3.3) es la siguiente, es decir la solución es un mapeo $y(x)$:

$$\frac{dy}{dx} = \frac{x^{\alpha_1} + (1-x)^{\alpha_1}}{y^{\alpha_1}} \tag{3.22}$$

[7]Esto se pide para garantizar la presencia de una partición generatriz

$$\frac{dy}{dx} = \frac{x^{\alpha_2} + (1-x)^{\alpha_2}}{(1-y)^{\alpha_2}} \tag{3.23}$$

Estas ecuaciones quedan como:

$$y^{\alpha_1} dy = x^{\alpha_1} + (1-x)^{\alpha_1} dx \tag{3.24}$$
$$(1-y)^{\alpha_2} dy = x^{\alpha_2} + (1-x)^{\alpha_2} dx \tag{3.25}$$

Después de una integración directa, obtenemos los siguientes mapeos:

$$y = f_1(x_n) = x_{n+1} = (1 - |x_n^r - (1-x_n)^r|)^{\frac{1}{r}} \tag{3.26}$$
$$y = f_2(x_n) = x_{n+1} = 1 - |x_n^r - (1-x_n)^r|^{\frac{1}{r}} \tag{3.27}$$

Utilizamos el parámetro $r = \alpha_i + 1$,con $i = 1, 2$, para ser consistentes con el resto del texto, éste permite explorar distintos mapeos con distintas propiedades dinámicas.

Las formas de uno de estos mapeos se muestran en las figuras 3.6.

En la figura se observa, en primer lugar, el dominio de los valores a sobre b, es decir $a > b$, con $1 < r < 2$ en este corresponde a un régimen laminar. Conforme el parámetro r disminuye se alcanza un valor común $a = b = 0$, a partir de este valor se invierte la relación entre a y b, es decir conforme la región caótica aumenta, $b > a$ para $r < 1$. Estos resultados coinciden con los mostrados en la familia logística, la familia épsilon; así como con los mostrados en el capítulo anterior con el modelos estocástico de expansión-modificación.

Estos mapeos en los límites para $r \to \infty$ se puede pegar tanto como se desee, a la recta identidad esto permite explorar la preeminencia de una región laminar sobre una región caótica y su relación con los valores de ajuste de la FBC.

3.4. Formalismo termodinámico de los sistemas dinámicos

Finalmente discutiremos una formalización de los resultados previos mediante el formalismo termodinámico de los sistemas dinámicos (FTSD).

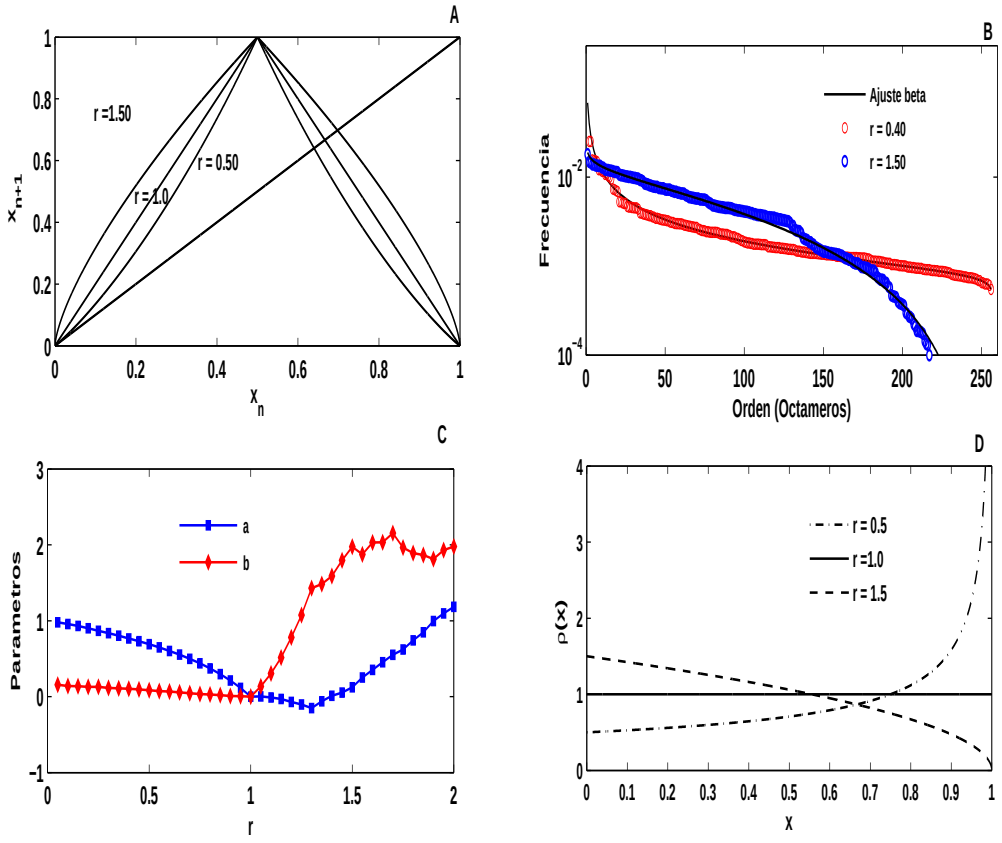


Figura 3.6: ***A***. La familia de mapeos definida en la ecuación 3.27 para distintos valores del parámetro $r = 3/2, 1, 0.5$. La transición intermitencia-caos se da con el aumento del valor del parámetro r. ***B***. Ajuste a la función beta para los valores del parámetro señalados en la figura, se muestran casos extremos en los cuales $a > b$ con $r < 1$ y $a < b$ con $r > 1$. ***C***. Parámetros de ajuste a, b *vs.* r. La inversión de los valores de a, b se da en $r = 1$ (mapeo tienda). ***D***. Densidades invariantes para $r = 0.5$ (intermitencia), $r = 1$ $\rho(x) = 1$ (mapeo tienda), y $r = 1.5$.

El estudio de los sistemas dinámicos caóticos mediante las herramientas de de la física estadística es un campo abordado inicialmente por Sinai, Ruelle y Bowen a principios de los años setenta del siglo pasado [91, 92, 93, 94] Un pequeño resumen con los principales resultados de esta teoría se puede encontrar en el anexo B.

3.4.1. Familia logística

Consideremos la familia logística entonces tomamos una partición del espacio fase $\mathbf{X}$, y teniendo como válida que la densidad invariante está dada aproximadamente por el *ansatz* 3.9, tenemos que: si dividimos el espacio fase en celdas de tamaño ϵ , (es decir el total de ellas es $n = 1/\epsilon$), la probabilidad P_i de encontrar la secuencia simbólica generada por la celda i-ésima está determinada por la siguiente expresión:

$$P_i = A_\nu \int_{(i-1)\epsilon}^{i\epsilon} x^{\frac{1}{\nu}-1}(1-x)^{\frac{1}{\nu}-1}dx \tag{3.28}$$

$i = 1, 2, \ldots, \frac{1}{\epsilon}$, y $A_\nu = \frac{\Gamma\left(\frac{2}{\nu}\right)}{\Gamma\left(\frac{1}{\nu}\right)^2}$ es la constante de normalización de cada función de densidad.

Estas integrales se pueden resolver de manera exacta mediante la función beta incompleta regularizada, $B_z(\alpha,\beta) \equiv \frac{B(z;a,b)}{B(a,b)} \equiv A_\nu \int_0^z t^{\alpha-1}(1-t)^{\beta-1}dt$

$$P_i = A_\nu(B_{i\epsilon}(\frac{1}{\nu},\frac{1}{\nu}) - B_{(i-1)\epsilon}(\frac{1}{\nu},\frac{1}{\nu})) \tag{3.29}$$

En particular, los dos extremos escalan de manera distinta al resto:

$$P_1 = B_\epsilon(\frac{1}{\nu},\frac{1}{\nu}) \approx \epsilon^{\frac{1}{\nu}} \tag{3.30}$$

$$P_{\frac{1}{\epsilon}} = 1 - B_{1-\frac{1}{\epsilon}}(\frac{1}{\nu},\frac{1}{\nu}) \approx (1-\epsilon)^{\frac{1}{\nu}} \tag{3.31}$$

Consideramos ahora la siguiente suma:

$$\sum_{i=1}^{\frac{1}{\epsilon}} P_i^\beta = (A_\nu)^\beta \sum_{i=1}^{\frac{1}{\epsilon}} (B_{i\epsilon}(\frac{1}{\nu},\frac{1}{\nu}) - B_{(i-1)\epsilon}(\frac{1}{\nu},\frac{1}{\nu}))^\beta \tag{3.32}$$

La probabilidad, es decir, el valor de la integral, tiene dos términos típicos: los extremos van como $\epsilon^{\frac{1}{\nu}-1}$ y el resto, como ϵ , es decir la longitud típica de una celda.

Utilizamos la distribución escort [8] $\mathcal{P}$, para que a partir de la distribución previa encontrar otra que va "pesando" las probabilidades escaneando su peso relativo, mediante un parámetro arbitrario β

$$\mathcal{P}_i \equiv \frac{P_i^\beta}{\sum\limits_{i=1}^{2^n} P_i^\beta} \tag{3.33}$$

Si atribuimos la frecuencia de aparición a una "energía de interacción" de tal forma que las secuencias con menor energía se ven favorecidas con respecto a aquellas con mayor energía. Por lo que proponemos una función "hamiltoniana" $H(s_1 s_2 \dots s_n)$ que refleje la observación anterior,

$$H(s_1 s_2 \dots s_n) = -\ln P(s_1 s_2 \dots sn) = -\ln \int_{cilindro_i} \rho(x) dx \tag{3.34}$$

de tal forma que la distribución escort queda de la siguiente manera

$$\mathcal{P}_i = \frac{P_i^\beta}{\sum\limits_{i=1}^{2^n} P_i^\beta} = \frac{exp(-\beta H(s_1 s_2 \dots s_n))}{\sum\limits_{\{s_1 s_2 \dots s_n\}} exp(-\beta H(s_1 s_2 \dots s_n)} \tag{3.35}$$

con lo que la distribución escort es análoga a la función de partición canónica.

La función de partición asociada estas cantidades está relacionada con la dimensión generalizada de Rényi $D(\beta)$ (ver ecuación B.11):

$$\sum_{i=1}^{\frac{1}{\epsilon}} P_i^\beta \approx (A_\nu)^\beta (2\epsilon^{\frac{\beta}{\nu}} + \sum_{i=2}^{\frac{1}{\epsilon}-1} (B_{i\epsilon}(\frac{1}{\nu}, \frac{1}{\nu}) - B_{(i-1)\epsilon}(\frac{1}{\nu}, \frac{1}{\nu}))^\beta) \tag{3.36}$$

$$\sum_{i=1}^{\frac{1}{\epsilon}} P_i^\beta \approx (A_\nu)^\beta (2\epsilon^{\frac{\beta}{\nu}} + (\frac{1}{\epsilon} - 2)\epsilon^\beta) \approx (\epsilon^{\frac{\beta}{\nu}} + \epsilon^{\beta-1}) \approx \epsilon^{(\beta-1)D(\beta)} \tag{3.37}$$

$$(\beta - 1)D(\beta) = min \left\{ \frac{\beta}{\nu}, \beta - 1 \right\} \tag{3.38}$$

$$\tag{3.39}$$

En donde β es un parámetro arbitrario que juega el papel de la "temperatura" (ver ecuación B.7).

[8] Utilizaremos el término en inglés ya que incluso en los pocos textos en español esta distribución adopta el término en inglés.

Entonces la energía libre de Helmholtz por unidad de volumen $\varphi(\beta)$ (ver tabla B.1) está dada por la siguiente expresión:

$$\varphi(\beta) = \begin{cases} 1 - \frac{1}{\beta} & \beta < \beta_c \\ \frac{1}{\nu} & \beta \geq \beta_c \end{cases} \tag{3.40}$$

El punto crítico es, el punto en el cual se intercambia el dominio de un conjunto de términos sobre otro en la contribución a la energía libre $\varphi(\beta)$.

$$\beta_c = \frac{\nu}{\nu - 1} \tag{3.41}$$

Si graficamos la energía libre de Helmholtz para cada valor de ν podemos observar que es lo que sucede en β_c (Ver Figura 3.7a). De esta manera podemos observar que existen transiciones de fase de primer orden que están asociadas a la primera derivada de la energía libre en el punto crítico β_c, sin embargo en $\nu \leq 1$, éstas divergencias dejan de existir, por lo que el mapeo tienda señala un punto límite entre la existencia o no de transiciones de fase, con esto hemos establecido la singularidad de este punto. En la figura 3.7 se muestra la superficie de energía libre para la familia logística. Es importante resaltar que el cálculo anterior resulta clave la divergencia de la densidad invariante para que el comportamiento de los extremos determine el dominio los elementos en la suma que constituye la función de partición. A pesar de no contar con una forma exacta para la densidad invariante, basta sólo con saber que existen divergencias para que los resultados aquí presentados sigan siendo válidos.

El significado "físico" de la transición de fase es el siguiente: la distribución escort nos permite escanear , mediante un parámetro arbitrario (β), la influencia de distintas regiones del espacio fase $\mathcal{X}$. En la transición de fase la contribución dominante de la energía libre súbitamente cambia a una región distinta del espacio fase.

El equivalente de fases "desordenada-condensada "puede interpretarse de la siguiente forma: la fase "desordenada" corresponde al caso en el cual los elementos del espacio fase contribuyen de manera (casi)equivalente y la fase "condensada" consiste en el hecho de que los puntos cercanos a los extremos son los términos dominantes.

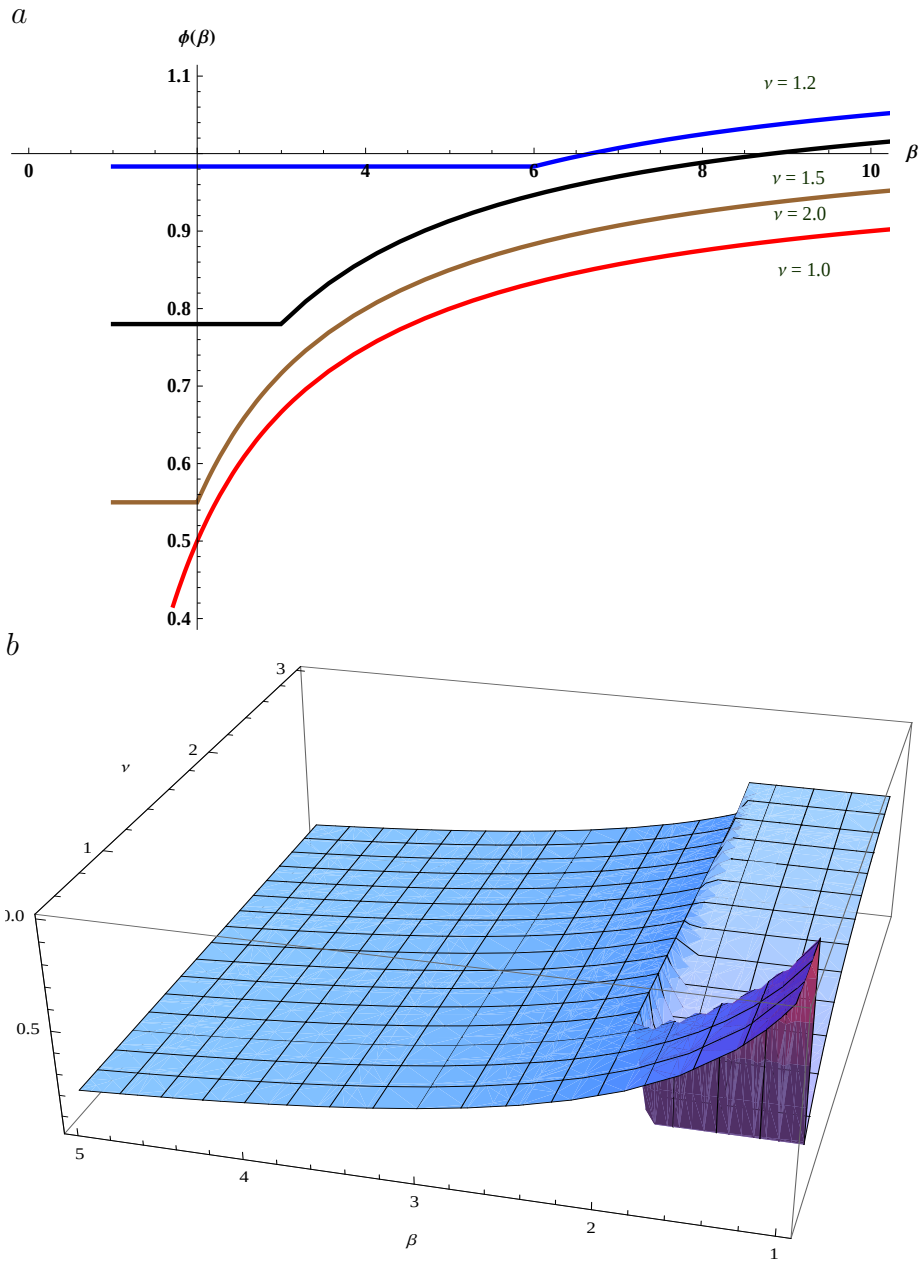


Figura 3.7: **a.** Energía libre de Helmholtz por unidad de volumen $\phi(\beta)$ para distintos mapeos pertenecientes a la familia logística. Se ha sumado una cantidad constante a cada función para una mejor apreciación de las formas funcionales. Las funciones son continuas pero en el valor β_c la primera derivada no existe indicando una transición de fase de primer orden **b.** La figura muestra la superficie generada al considerar la familia logística $\phi(\beta, \nu)$, se observa con mayor nitídez el límite marcado en el valor $\nu = 1$ en el comportamiento de la energía libre.

3.4.1.1. Familia épsilon

A pesar de no conocer la densidad invariante es posible observar en 3.5 que existen un fenómeno análogo al de la familia logística. Se observa un intercambio de valores relativos entre los parámetros que coincide con la variación de los valores de los parámetros de ajuste a, b de la FBD.

Nuevamente el punto límite es el mapeo tienda, a excepción de este y del caso límite en donde es tangente a la identidad en el origen, el resto de los mapeos son no hiperbólicos, por lo que el comportamiento distinto entre ellos se debe a una variación en el peso de la función de distribución de probabilidad que se refleja en las medidas de los intervalos del sistema. Se observa también el efecto de las divergencias en las densidades como límite entre las dos regímenes (Ver figura 3.5D).

3.4.1.2. Divergencias en un extremo

Los mapeos presentados generados a partir de densidades invariantes con divergencias en uno de los extremos del intervalo (ecuaciones 3.27) son análogo a la familia logística, la única diferencia reside en el hecho de que en lugar de tener dos puntos en los cuales la probabilidad escala de manera distinta, se trata sólo de uno, por lo que la contribución en la expresión 3.28, es la misma en el límite termodinámico con lo que las expresiones y gráficas asociadas son las mismas que las del caso previo de la beta simétrica.

Este resultado, de hecho, refuerza, la observación hecha sobre la naturaleza del límite de $a > b$ con la (in)existencia de transiciones de fase. En este caso, partimos de la densidad invariante (con lo que no hacemos suposiciones sobre las divergencias) se obtienen los mapeos y de éstos se generan las dinámicas simbólicas y la obtención de las FBC.

Hemos demostrado que los parámetros de ajuste de la distribución beta $\{a, b\}$ están relacionados con los siguientes fenómenos:

1. a está censando el orden del sistema mediante estructuras jerárquicas : (leyes de potencia, intermitencia, sesgo en las distribuciones de probabilidad, etc), en cambio, el parámetro b refleja el caos del sistema expresado en términos de mayor desorden (entropías máximas, exponentes de Liapunov máximos,etc.).

2. La caracterización anterior permite identificar dos regiones de dinámicas distintas con $a > b$ intermitencia, en términos del FT : una energía

externa que favorece ciertos "microestados" sobre otros, por otra parte, con $a < b$.

3. El punto en el que se intercambian estos comportamientos, es decir $a = b$, se marca con transiciones de fase de primer orden reflejadas en la energía libre de Helmholtz $\varphi(\beta)$ como función de un parámetro externo β (que juega el papel de la temperatura).

Capítulo 4

Evolución molecular: un modelo biológico

En este capítulo se presenta el modelo de evolución molecular propuesto por Eigen y Schuster [95, 96] para la evolución de un conjunto de polímeros poseedores de dos características particulares: pueden replicarse y mutar puntualmente. Este modelo fue propuesto para simular moléculas auto-replicantes en un ambiente pre-biótico, formulados como un mecanismo de evolución de polímeros simples que funge como hipótesis del origen de la vida .

Mostraremos que la presencia de estos los dos procesos anteriores en este modelo de evolución molecular generan en la distribución del número de especies ordenados conforme a su frecuencia dan lugar a la FBC, observaremos también, que el hecho conocido como la catástrofe del error, modifica esta distribución indicando un cambio similar a los ya reportados, es decir, que con $a = b$ se observa una transición, en este caso, formalmente, una transición de fase en un modelo equivalente de espines.

4.1. El modelo

La ecuación propuesta 4.1 por Eigen y posteriormente Schuster en los años 70 del siglo pasado en los artículos seminales [96, 97, 98], ilustra el mecanismo de evolución de secuencias de símbolos de L elementos, auto-replicantes y con una tasa de mutación puntual.

4.1.1. Supuestos

Como todo modelo tiene supuestos que lo limitan y por lo tanto lo definen, en el caso del de cuasi-especies éstos son las siguientes:

1. Existe una secuencia con mayor adecuación (*fitness*), que puede medirse como el éxito reproductivo.
2. Las tasas de mutación son constantes en el tiempo y genotipos.
3. Los mutantes muestran adecuaciones invariantes a pesar del número de mutaciones.
4. Las moléculas se degradan a tasas constantes.
5. No hay recombinación genética.

El modelo básico de cuasi-especies describe la dinámica de poblaciones de secuencias macromoleculares mantenidas por replicación. Las mutaciones suceden durante el proceso de copia, pueden modificar la tasa de replicación de las nuevas secuencias que aparecen, permitiendo entonces a la selección darwiniana operar.

La situación antes descrita se expresa mediante el siguiente sistema de ecuaciones

$$\frac{dx_i}{dt} = (A_i Q_{ii} - D_i)x_i(t) + \sum_{j \neq i} A_j Q_{ij} x_j(t) - \Phi(t) x_i(t) \tag{4.1}$$

con x_i la frecuencia relativa de la i-ésima especie o secuencia de símbolos, A_i es el *fitness* (éxito reproductivo) de la especie i, D_i es la tasa de degradación de la especie i, la tasa de replicación efectiva de la propia especie es Q_{ii}, en tanto que, Q_{ij} es la tasa de mutación de la especie j en la especie i (es decir mide la probabilidad de que mediante mutación la especie j se convierta en la especie i), Φ es un flujo que se impone para que la población se mantenga constante, por ello es llamada la condición de población constante, en particular, si se quiere que $\sum_{i=1}^{s} x_i = 1$, el flujo Φ está determinado por

$$\Phi = \sum_{i=1}^{s} (A_i - D_i)x_i(t) = \sum_{i=1}^{s} E_i x_i(t) = \langle E(t) \rangle$$

E_i es la productividad de la especie i y $\langle E(t)\rangle$ es la productividad promedio, notemos que este término hace que el sistema de ecuaciones 4.1 se convierta en no lineal (de hecho es un sistema cuadrático en las variables de interés $x_i(t)$).

Debido a que $\sum_{j=1}^{s} Q_{ij} = 1$ se tiene que:

$$\sum_{j\neq i}^{s} A_j Q_{ij} x_j(t) = \sum_{i=1}^{s} A_i(1 - Q_{ii})x_i(t) \tag{4.2}$$

El sistema 4.1 entonces se puede reducir a

$$\frac{dx_i}{dt} = \sum_{i=1}^{s} W_{ji}x_j(t) - \langle E(t)\rangle x_i(t) \tag{4.3}$$

$W_{ii} = A_i q_{ii} - D_i$ *valores selectivos* y los $W_{ij} = A_j Q_{ij}$ para $i \neq j$ se llaman *valores de mutación*

En forma matricial las ecuaciones anteriores se pueden escribir de la siguiente manera

$$\frac{d\mathbf{X}}{dt} = (\mathbf{W} - \langle E(t)\rangle)\mathbf{X} \tag{4.4}$$

Es posible encontrar soluciones analíticas del sistema. Si $\mathbf{O}$ es una matriz tal que que $\mathbf{OWO^{-1}} = \lambda\mathbf{I}$ y $\mathbf{U} = \mathbf{OX}$ o de manera equivalente se tiene que

$$\frac{du_i}{dt} = (\lambda_i - \langle E(t)\rangle)u_i(t) \tag{4.5}$$

$$\langle E(t)\rangle = \sum_{i=1}^{s} \lambda_i u_i \tag{4.6}$$

Si la mutación de las secuencias es puntual, entonces la matriz de mutación Q_{ij} está dada por

$$Q_{ij} = A_i u^{h_{ij}}(1 - u)^{L-h_{ij}} \tag{4.7}$$

En dónde L es la longitud de las secuencias, h_{ij} es la distancia de Hamming entre las secuencias i y j, y u es la tasa de mutación.

4.1.2. Cuasi-especies

Con frecuencia, la teoría de evolución molecular es llamada la teoría de cuasi-especies. De hecho, la existencia de cuasi-especies moleculares junto con la catástrofe del error (sección 4.2) caracterizan a esta teoría. El término especie tiene significados distintos en biología y en química, los autores de la TCE, químicos de formación, se refieren a moléculas casi idénticas con el término especie y no tiene relación alguna con especies biológicas. Así una cuasi-especie es un "enjambre" de especies genéticamente relacionadas (secuencias cuya distancia de Hamming es pequeña) que constituye un punto de equilibrio estable del sistema de ecuaciones diferenciales no lineales definidos por el sistema 4.1

La solución analítica u_{sol} del sistema 4.1 es una combinación lineal de las especies $u_{sol} = \sum_{i=1}^{r} a_i x_i$, de esta manera hay una especie x_j con mayor sobrevivencia (contribución en el término solución) y la segunda en frecuencia x_k es "genéticamente similar" con ésta, y así sucesivamente, por lo que la solución del problema 4.1 genera no una única secuencia que sobrevive a este proceso, sino a un ensamble de secuencias genéticamente emparentadas, vecinas en el paisaje adaptativo, es esta colección de moléculas cuasi-idénticas (cuasi-especies), el principal, novedoso y empíricamente útil, resultado de esta teoría.

Hoy en día existe evidencia experimental de que ciertos virus de RNA (VIH, polio, influenza, entre otros), se comportan según la teoría de cuasi-especies [99, 100, 101, 102, 103]: es decir para organismos con tasas de mutación elevadas (que le sirven para adaptarse a ambientes hostiles) los virus que sobreviven son un subconjunto que difieren, en su secuencia de RNA, en pocos lugares, esta colección de secuencias se llama las cuasi-especies de virus de RNA. La teoría de cuasi-especies ofrece un marco conceptual para la evolución de virus de RNA y se han convertido en un paradigma virológico [104, 105, 106, 107, 108, 109, 110, 111].

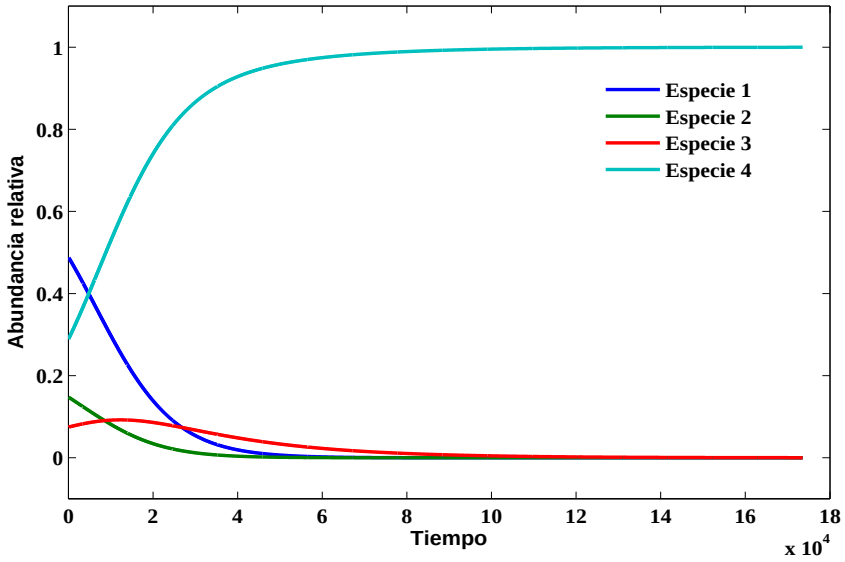


Figura 4.1: Solución del problema de evolución molecular sin mutación (es decir la matriz de mutación sólo con elementos en la diagonal). Existe selección, es decir sobrevive la secuencia con mayor adecuación, pero no hay evolución. Se ilustra la solución para 4 especies

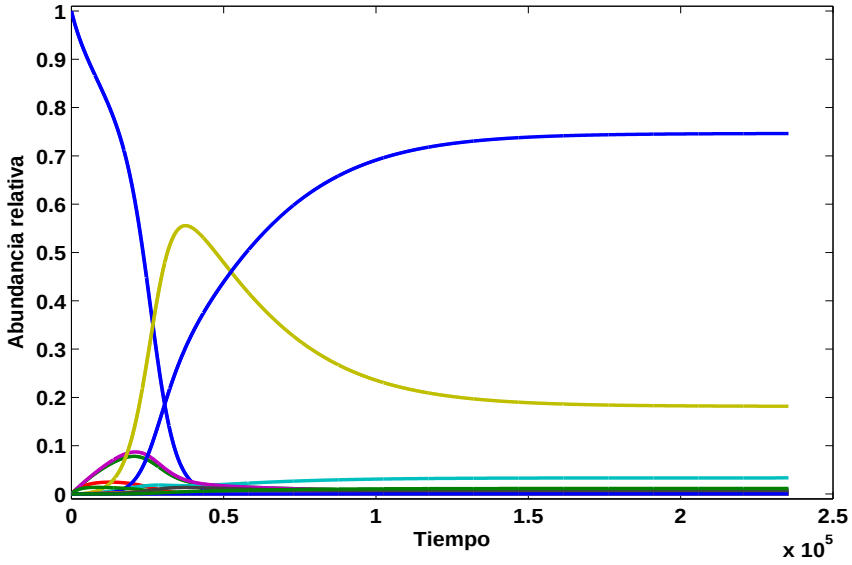


Figura 4.2: Solución del problema de evolución molecular con cuasi-especies. Sobreviven no el de mayor adecuación, sino un conjunto relacionado "genéticamente" entre si. En este caso se trata de 16 especies.

4.2. La catástrofe del error

Las ecuaciones de cuasi-especies 4.1 describen la trayectoria de una población en el espacio de todas las posibles secuencias de tamaño fijo finito. Las cuasi-especies "censan" el paisaje adaptativo[1], e intentan escalar los máximos locales (es decir áquellas secuencias con mayor éxito reproductivo) y buscan permanecer en estos máximos (locales o globales).

Si la tasa de mutación es muy grande, la capacidad de la cuasi-especie de escalar y permanecer en la cima de un máximo se reduce debido a la alta variabilidad. Debido a esto, existe un umbral de mutación que es compatible con la adaptación (entendida como la capacidad de la cuasi-especie de encontrar máximos de adaptación y permanecer en ellos), así que, si el valor de mutación excede este valor umbral, entonces, no será posible la adaptación.

[1]El paisaje adpatativo es un espacio $N^n \times R$, en donde a cada secuencia simbólica se le asigna una valor real positivo ($f : N^n \to R$) que corresponde a su adecuación, es decir su éxito reproductivo.

Supongamos válida, esta premisa, llamada paisaje adaptativo de un sólo pico (*single-peak fitness landscape*), cuando la tasa de mutación es baja, la solución en el equilibrio describe una cuasiespecie centrada en este único máximo. Las secuencias lejanas al pico tiene frecuencias muy bajas, la cuasiespecie, está entonces, adaptada a este máximo de adecuación. Cuando la tasa de mutación aumenta la distribución de cuasi-especies se ensancha, existe entonces una tasa crítica u_c más allá de la cual la cuasi-especie deja de sentir el efecto del pico, y no queda localizada alrededor de él. Es decir, no existe adaptación.

A este hecho se le conoce como catástrofe del error (*error catastrophe*). Este fenómeno es muy familiar en el ámbito de la física, y es posible hablar, formalmente, de una transición de fase.

4.2.1. Transición de fase

Es posible mapear el modelo de cuasi-especies en un modelo de Ising bidimensional con lo que se puede establecer una correspondencia formal entre las transiciones de fase y la catástrofe del error.

Con el siguiente cambio de variables se transforma un sistema al otro, con los valores de "espín"

$$s = 1, 0$$

en los siguientes referencias se trata este problema como una cadena de Ising cuántica [112, 113] en los cuales se puede establecer la existencia de transición de fase mediante la siguiente transformación de coordenadas:

$$y_i(t) = x_i(t) e^{\int_0^t dt' \sum_{j=1}^{N} f_j x_j(t')}$$

El análogo de la temperatura T del sistema se relaciona con la tasa de mutación u de las secuencias de la siguiente manera [114, 115, 116, 117]:

$$\frac{1}{T} = \left| \log \frac{1-u}{u} \right|$$

Con esta transformación es posible equiparar ambos formalismos y entonces demostrar que la catástrofe del error es una transición de fase [118]. La fase ergódica corresponde a la ausencia de la secuencia maestra (cuasi-especie) y entonces el sistema se difumina sobre todo el espacio de secuencias, mientras que en, la fase no ergódica existen siempre las cuasi-especies.

4.2.2. Matriz de transferencia

La matriz **W** definida en la ecuación 4.7 es equivalente a la matriz de transferencia de un modelo de espines en 2 dimensiones En un trabajo en proceso estamos buscando la relación entre esta matriz de transferencia para el sistema de casi-especies y el método del operador de transferencia para mapeos unimodales [119].

4.3. Distribución de especies y FBC

Exhibimos de manera panorámica la teoría de cuasi-especies. La relación entre el modelo de cuasi-especies, los modelos de expansión-modificación y las familias de mapeos unimodales. Éstos comparten la presencia de dos procesos con dinámicas opuestas entre si que se ven modulados por un parámetro (mutación, parámetro de la familia de mapeos y tasa de mutación) que permiten modular la dominancia de un procesos sobre el otro. En estos casos la presencia de un proceso, que genéricamente podemos llamar permanencia, establece correlaciones largas en las secuencias simbólicas y el otro proceso, que podemos englobar en el concepto de cambio, rompe las correlaciones.

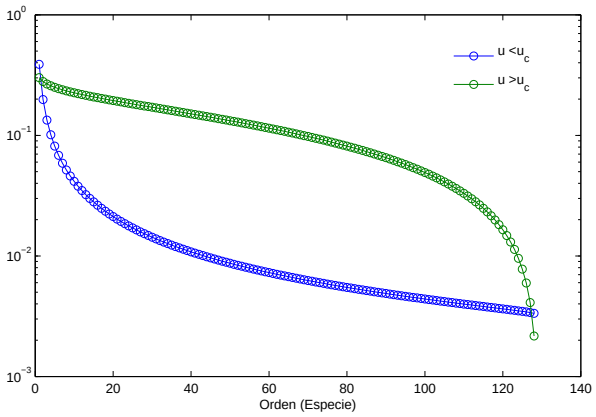


Figura 4.3: Ajuste de la FBC para la frecuencia relativa de especies como resultado de resolver el sistema de ecuaciones diferenciales del modelo de cuasiespecies para un total de 128 de ellas. Son dos realizaciones: una con probabilidad de mutación menor a u_c en donde se oberva una ley de potencia con claridad, y con $u > u_c$ en la que se observa la aparición de un cambio en los valores de a y b, en este caso $a < b$.

Este capítulo no pretende ahondar en la muy interesante teoría de cuasi-especies sino en explorar otro sistema con dinámicas en conflicto, y estudiarlo en términos de orden-frecuencia para aportar más evidencia a la hipótesis de este trabajo.

Análogamente a los dos capítulos previos podemos ahora resolver numéricamente la ecuación de cuasi-especies para N especies moleculares y encontrar la frecuencia de aparición agruparla de acuerdo a esta frecuencia de forma descendente y graficarla para valores de la tasa mutación previa a la catástrofe del error y después de ella. Los resultados se muestran en la siguiente gráfica sintetizan los resultados que podrían esperarse de acuerdo a lo visto previamente. Es decir, cuando se tiene orden asociado a invariabilidad de escala , obtenemos $a > b$ en la FBC; después de una transición de fase, los valores se invierten y tenemos $b > a$. Con lo que los valores de la FBC están reflejando la dinámica de los procesos involucrados y pueden inclusive utilizarse para caracterizar un fenómeno a partir de sus distribución de frecuencias. Resumiendo de manera sucinta: hemos observado 3 procesos distintos entre sí expresado de forma simbólica reflejan en su expresión de un observable, en esta caso orden-frecuencia distribuciones FBC y además el

comportamiento de los exponentes de ajuste se comporta de manera genérica. Es decir con la permanencia $a > b$ y con el cambio $b > a$.

Capítulo 5

Otros enfoques y conclusiones

"En una ocasión Richter me dijo: "los médicos no deberían decir "lo he curado", sino "no se me ha muerto". Del mismo modo en la física se podría decir "he proporcionado causas a las que finalmente no se les puede señalar lo absurdo" en vez de "lo he explicado".
Aforismos
Georg Christoph Lichtenberg

5.1. Otros enfoques

El problema de elucidar el origen de la ubicuidad de la FBC y el posible significado de los ajustes de los exponentes (a, b) ha sido abordado recientemente mediante dos enfoques complementarios. El primero de ellos a través de la operación resta de variables estocásticas (de forma análoga al operador de convolución) [30], en ese trabajo se observa que mediante la aplicación sucesiva de un operador se obtienen FDPs tipo beta. Se observa que esta la aplicación sucesiva de esta operación tiende a converger en el espacio de FDPs hacia funciones tipo beta, con lo que se sugiere algún límite asociado a esta operación.

En segundo lugar también se ha tratado este problema, a través de un proceso estocástico de nacimiento y muerte para redes (Cocho G., Rodríguez R. F., Martínez-Mekler G. Birth and Death Master Equation for the Evolution of Complex Networks), en este abordaje se utiliza una ecuación maestra cuyas soluciones están dadas por PDFs beta.

En ambos casos, las soluciones encontradas son FDPs (tipo-)beta. Éstas, a su vez, se pueden expresar en orden-frecuencia vía la inversión de la función

de probabilidad acumulada, sin embargo, las formas funcionales son muy distintas a la FBC, sólo en en el caso de leyes de potencia ambas formas coinciden, sin embargo los exponentes son distintos ($\alpha \rightarrow \frac{1}{\alpha+1}$) [20].

Por otra parte, adicionalmente hemos encontrado que en procesos en donde su utiliza el ajuste de datos vía orden-frecuencia(tamaño) podemos aproximar FDP (χ^2, χ, y otras distribuciones), mediante la inversión de la función de distribución acumulada y una expansión en serie de potencias de las colas de la distribución para relacionarlas con los parámetros (a, b).

Además es posible encontrar una fórmula general para la FDP a partir de la FBC para algunos casos particulares de los parámetros de ajuste (Ver anexo C), este último resultado permite el cálculo de a, b mediante la función de verosimilitud (MLE) (Ateneodo Celia, Martinez-Mekler Gustavo y Alvarez-Martinez Roberto, artículo en proceso).

En ese trabajo se considera que la ubicuidad de la FBC proviene de la expresión de funciones de distribución de probabilidad en su versión FAP. Con las funciones de distribuciones cuando éstas son unimodales, con ello se pueden relacionar las colas de las FDPs con los exponentes de ajuste de la FBC.

Finalmente, de acuerdo a los resultados aquí presentados, podemos señalar las siguientes conclusiones:

5.2. Conclusiones

La FBC ha mostrado una enorme correspondencia con un diverso conjunto de datos cuando se grafican en orden-frecuencia u orden-tamaño. Los datos provienen de una gran variedad de fuentes de diversos orígenes, sin embargo, en este trabajo abordamos el estudio de secuencias simbólicas generadas por dos procesos en competencia entre si, en estos casos, la presencia de la FBC se debe a la similitud de las dinámicas involucradas; de esta manera es posible asignarle significados a los parámetros de ajuste de la FBC (a, b) en términos de las características dinámicas del proceso.

Los diferentes ajustes propuestos pueden compararse mediante métodos de información en los que se muestra que la FBC, para un amplio conjunto de datos, es superior a un conjunto importante de otras funciones usadas en la literatura científica para ajustar datos en orden-frecuencia(tamaño).

Dada la ubicuidad de las distribuciones orden-frecuencia (orden-tamaño) en un conjunto amplio de fenómenos de diversa naturaleza, estamos enfrenta-

dos con dos principales tareas: entender cuál es el significado de los parámetros de ajuste y entender cuál es origen de tal ubicuidad.

Debido a la extensa variedad de ejemplos encontrados hasta ahora en donde la FBC es un buen modelo, es poco probable establecer, a pesar de los esfuerzos en esta dirección, un mecanismo general o meta-mecanismo que sea común a todos estos ejemplos. La hipótesis central de este trabajo consiste en que fenómenos que comparten una dinámica similar, a pesar de ser de distintos dominios, puedan ser entendidos y comparados debido a sus semejanzas; entonces es posible entender y asignar significados generales a los parámetros de ajuste de la FBC.

En esta tesis nos dedicamos a estudiar las dinámicas en conflicto, donde es posible identificar dos sub-procesos claramente discernibles. Uno de ellos caracteriza genéricamente cambio y el otro, permanencia. Estos procesos están detrás de las transiciones asociadas a los valores relativos de los parámetros de ajuste de la FBC.

En los sistemas de expansión-modificación existen dos procesos opuestos que genéricamente se pueden caracterizar de la siguiente manera: uno de ellos, la expansión, preserva las correlaciones de largo alcance entre los bloques de símbolos y el otro, modificación, (o mejor dicho, inserción) corta estás correlaciones. Este método está controlado por el parámetro de mutación p. Se pueden identificar regiones bien definidas con comportamientos asociados a correlaciones de largo alcance y correlaciones cortas conforme crece p, los resultados son independientes del número de bloques considerados. Existe además un límite en en el cual los comportamientos se invierten $a = b$. La entropía aumenta hasta alcanzar su límite teórico máximo para una secuencias binaria de longitud N, es decir $\ln 2^N$.

Con nuestro enfoque hemos estudiado algunos aspectos de la dinámicas en conflicto en sistemas finitos determinados por los algoritmos de expansión-modificación ramificada. En estos procesos hemos encontrado una transición del tipo orden-desorden en términos de la interpretación del significado de los parámetros de ajuste de la beta: el parámetro a es favorecido por la invariabilidad de escala y registra la presencia de correlaciones de largo alcance, mientras que el parámetro b es sensible al desorden y está relacionada, en la ausencia de correlaciones, a la aparición de eventos improbables y al desarrollo de heterogeneidad. Puesto que en los modelos de expansión-modificación el conflicto entre permanencia y cambio puede ser cuantificado por las probabilidades p y q, éstas proporcionan mecanismos para la variación y la comprensión de los parámetros del ajuste, el cual ayuda a identificar las carac-

terísticas relevantes y proporcionar indicios sobre la fenomenología observada en expresiones sociales, antropológicas, finanzas, biológicas y artísticas, por mencionar algunas, en donde aparece la FBC.

Adicionalmente hemos evidenciado las interconexiones entre varias características singulares de la transición, tales relaciones entre cambios en la dependencia funcional de la invariabilidad de escala del espectro de potencias con el desorden, de longitudes características para relajaciones del parámetro, localización de los puntos de transición, la dependencia exponencial del número de eventos improbables con los parámetros del ajuste y otros criterios para el análisis del comportamiento complejo entre el orden y desorden.

En el caso de familias de mapeos unimodales que mediante la variación de un parámetro atraviesan una plétora de estados entre ellos regímenes intermitencia hacia regímenes caóticos con lo que se modifican las secuencias más favorecidas debido al cambio del parámetro que caracteriza al sistema:

1. Favorecen estados con menor "energía de interacción".

2. Conforme se modifica el parámetro se atraviesan una transición de fase, en dónde otros estados llevan el peso de la distribución escort.

3. Este comportamiento se puede generalizar a mapeos cuya densidad invariante diverge en un número finito de puntos del soporte.

La dinámica entre estados "laminares" y caóticos mediada por un parámetro permite explorar el significado de los parámetros a, b.Exploramos mediante el FTSD los "microestados" asociados a los n-omeros y observamos un conjunto de transiciones de fase de primer orden, que aparecen cuando la densidad invariante tiene divergencias.

El mecanismo de evolución molecular presentado en el capítulo 4, muestra un sistema de ecuaciones diferenciales no lineales para modelar un conjunto de polímeros con dos procesos involucrados: reproducción y mutación. Como solución se obtiene un enjambre de soluciones similares entre sí como punto fijo estable del sistema (cuasi-especies). Si la tasa de mutación u del sistema aumenta se atraviesa una u_c que formalmente es un punto crítico asociado a una transición de fase de segundo orden (la catástrofe del error).

Expresadas las frecuencias de aparición de cada especie podemos ordenarlos para ajustarlos a una FBC, con lo que obtenemos que:

1. Cuando la tasa de mutación es inferior a la crítica u_c se tiene que $a > b$

2. Para $b > a$ se obtienen soluciones en donde la información se pierde y cualquier secuencia del espacio de secuencias binarias es igualmente probable, con lo que se pasa de una región ordenada a un desorden total reflejado en términos de la entropía de Shannon de los bloques de símbolos binarios se alcanza un máximo.

5.2.1. Dinámica simbólica

A pesar de que la mayor parte de los problemas en física se formulan de manera natural en los números reales, es decir son continuos, conviene adoptar un marco común en el cual representar distintos sistemas de manera simbólica. En general la discretización de un sistema puede entenderse, en muchas ocasiones, como una forma de simplificar un problema que permite su abordaje teórico y/o computacional, por ejemplo, autómatas celulares (el espacio y tiempo son continuos), mapeos (el espacio continuo y el tiempo es discreto), redes de mapeos acoplados (espacio y tiempo discretos, no así las variables). Esto permite estudiarlos en sistemas más simples pero que, de preferencia, conserven las características fundamentales de los sistemas continuos, ya sea que su expresión natural sea en el discreto, como los modelos de expansión-modificación) o que se estudien como discretos , como la dinámica simbólica de los mapeos unimodales, la dinámica simbólica permite contextualizarlos para poder compararlos a partir de las dinámicas que están detrás de ellos [120].

Estudiar las medidas de complejidad de estos sistemas vía su expresión en dinámica simbólica es relativamente más simple, podemos establecer que cuando el proceso de permanencia domina sobre el de cambio, se tienen estructuras jerarquícas, correlaciones de largo alcance, y leyes de potencia, cuando el proceso se invierte entonces la estructura (información) se pierde y se alcanza un punto de extremo de desorden absoluto, en donde se observa una carencia de cualquier jerarquía. De esta manera a está determinando el orden y b censa el caos.

Apéndice A

Anexo Perron-Frobenius

Cálculo de la densidad invariante como punto fijo del OPF.

$$\rho_0(x) = \frac{\Gamma\left(\frac{2}{\nu}\right)(1-x)^{\frac{1}{\nu}-1}x^{\frac{1}{\nu}-1}}{\Gamma\left(\frac{1}{\nu}\right)^2} \tag{A.1}$$

Las dos primeras iteraciones de esta propuesta son las siguientes expresiones:

$$\rho_1(x) = \frac{\Gamma\left(\frac{2}{\nu}\right)\left(-\frac{\Gamma\left(\frac{2}{\nu}\right)((1-x)x)^{\frac{1}{\nu}-2}\left(\Gamma\left(\frac{2}{\nu}\right)((1-x)x)^{1/\nu}+(x-1)x\Gamma\left(\frac{1}{\nu}\right)^2\right)}{\Gamma\left(\frac{1}{\nu}\right)^4}\right)^{\frac{1}{\nu}-1}}{\Gamma\left(\frac{1}{\nu}\right)^2} \tag{A.2}$$

$$\rho_2(x) = \frac{1}{\Gamma\left(\frac{1}{\nu}\right)^2}\Gamma\left(\frac{2}{\nu}\right)$$
$$\left(1 \Big/ \Gamma\left(\frac{1}{\nu}\right)^2 \Gamma\left(\frac{2}{\nu}\right)\left(-\frac{\Gamma\left(\frac{2}{\nu}\right)((1-x)x)^{\frac{1}{\nu}-2}\left(\Gamma\left(\frac{2}{\nu}\right)((1-x)x)^{1/\nu}+(x-1)x\Gamma\left(\frac{1}{\nu}\right)^2\right)}{\Gamma\left(\frac{1}{\nu}\right)^4}\right)^{\frac{1}{\nu}-1}\right.$$
$$\left.\left(1-\frac{\Gamma\left(\frac{2}{\nu}\right)\left(-\frac{\Gamma\left(\frac{2}{\nu}\right)((1-x)x)^{\frac{1}{\nu}-2}\left(\Gamma\left(\frac{2}{\nu}\right)((1-x)x)^{1/\nu}+(x-1)x\Gamma\left(\frac{1}{\nu}\right)^2\right)}{\Gamma\left(\frac{1}{\nu}\right)^4}\right)^{\frac{1}{\nu}-1}}{\Gamma\left(\frac{1}{\nu}\right)^2}\right)\right)^{\frac{1}{\nu}-1} \tag{A.3}$$

Apéndice B

Formalismo termodinámico

El formalismo termodinámico es un conjunto de métodos que se inspiran en la física estadística para ser aplicados en los sistemas dinámicos. Los trabajos pioneros en esta área fueron desarrollados por Sinaí [121], Ruelle [122, 123], Bowen [124] en los años 70 del siglo pasado.

Las dos principales objetivos por alcanzar mediante el abordaje de los sistemas dinámicos vía la termodinámica estadística son los siguientes:

1. La caracterización cuantitativa del movimiento caótico mediante procedimientos de la termodinámica.

2. Análisis termodinámico de los conjuntos multifractales.

Para sistemas no-lineales independientemente de su dimensionalidad- el caos es un fenómeno genérico. Por lo que la predicción de una trayectoria, a tiempos grandes, sólo puede ser obtenida de manera empírica. Esto está asociado a la sensibilidad a las condiciones iniciales, por lo que las trayectorias sólo pueden ser descritas por métodos estadísticos.

Una de las formas de caracterizar estadísticamente estas trayectorias consisten en encontrar distribuciones de probabilidad de las iteraciones.

1. La densidad diverge en un punto o en una infinidad de puntos.

2. El atractor tiene estructura fractal.

3. Espectros de Liapunov

La necesidad de utilizar el formalismo termodinámico no surge de una alta dimensionalidad del espacio fase, sino de la complejidad inherente de los sistemas no lineales, de esta forma podemos caracterizar lo conceptos de los sistemas dinámicos caóticos con la dimensión fractal, la entropía con conceptos termodinámicos tales como la entropía, la temperatura, la presión y la energía libre.

La pregunta fundamental es ¿cómo extender estas analogías y cómo es que resultan ser útiles para el estudio de los sistemas dinámicos?

El concepto vinculante entre estás dos campos son las distribuciones *escort*. Dada una distribución de probabilidades p_i, entonces se puede construir, a partir de ésta, una nueva distribución P_i con un parámetro real arbitrario β de la siguiente manera

$$P_i = \frac{p_i^\beta}{\sum_{j=1}^r p_j^\beta} \tag{B.1}$$

Las distribuciones escort tienes las mismas propiedades que las distribuciones canónicas generalizadas. Éstas son las herramientas para el escaneo de las FDP y con ellas podemos formalizar las analogías con la termodinámica.

B.1. Función de partición canónica

El análogo a los "microestados" en los sistemas dinámicos es el peso de los cilindros $c_n(S)$, de una secuencia simbólica $S = s_1 s_2 \dots s_n$ de longitud n [121], con lo que la función hamiltoniana del sistema es:

$$H(s_1 s_2 \dots s_n) = -\ln p(s_1 s_2 \dots s_n) = \int_{c_n(S)} \rho(x) dx \tag{B.2}$$

La ecuación anterior B.2 reproduce lo que se espera, es decir para secuencias poco probables la "energía" es grande mientras que para secuencias muy probables la energía es pequeña. El siguiente desarrollo muestra la relación entre la distribución escort y la función de partición.

$$P(s_1 s_2 \dots s_n) = \frac{(p_(s_1 s_2 \dots s_n))^\beta}{\sum_{\{s_1 s_2 \dots s_n\}} (p(s_1 s_2 \dots s_n))^\beta} = \frac{\exp(-\beta H(s_1 s_2 \dots s_n))}{\sum_{\{s_1 s_2 \dots s_n\}} \exp(-\beta H(s_1 s_2 \dots s_n))} \tag{B.3}$$

la suma de la ecuación anterior se efectúa sobre todas las realizaciones de secuencias de longitud n. La temperatura está entonces relacionada con el parámetro β de la manera usual $\beta = \frac{1}{K_B T}$.

Así, la energía libre de Helmholtz se puede escribir en estos términos como:

$$F(\beta) = -\frac{1}{\beta}\log Z(\beta) \tag{B.4}$$

en donde $Z(\beta)$ es la función de partición

$$Z(\beta) = \sum_{\{s_1 s_2 \ldots s_n\}} \exp(-\beta H(s_1 s_2 \ldots s_n)) \tag{B.5}$$

La dimensión generalizada de Renyi ($D(\beta)$) [125, 126, 127, 128] y la información de Renyi (I_β) están definidas de la siguiente manera :

$$I_\beta(p) \equiv \frac{1}{\beta - 1}\ln\sum_{i=1}^{r}(p_i)^\beta \tag{B.6}$$

$$D(\beta) \equiv -\lim_{\epsilon\to 0}\frac{1}{\epsilon(\beta-1)}\ln\sum_{i=1}^{r}(p_i)^\beta \tag{B.7}$$

por lo que están relacionadas mediante la siguiente ecuación

$$\frac{I_\beta}{V} = \frac{1}{(\beta-1)V}\ln\sum_{i=1}^{r}(p_i)^\beta \tag{B.8}$$

$$D(\beta) = -\lim_{V\to\infty}\frac{I_\beta}{V} \tag{B.9}$$

con $V = -\ln\epsilon$, con ϵ el tamaño de celda del espacio fase. Se toma el límite con $\epsilon \to 0$ de tal forma que $V \to \infty$ de forma análoga al límite termodinámico.

La relación entre la función de partición $Z(\beta)$ y la información de Rényi $I_\beta(p)$

$$I_\beta(p) = \frac{1}{\beta-1}\ln Z(\beta) \tag{B.10}$$

establece que en el límite $\epsilon \to 0$ (límite termodinámico) la función de partición escala como:

$$Z(\beta) \sim \epsilon^{(\beta-1)D(\beta)} \tag{B.11}$$

La dimensión generalizada de Rényi está asociada para $\beta = 1$ a la información de Shannon, por lo que es llamada la *dimensión de información*, $D(2)$ se llama dimensión de correlación [129].

Con la expresión de la función de partición se puede entonces calcular las variables termodinámicas usuales asociadas a la función de partición canónica y sus análogas en el FTSD, los resultados se pueden observar en esta tabla

Función de partición canónica	
$H(s_1 s_2 \ldots, s_n)$	$-\ln p(s_1 s_2, \ldots, s_n) = -\int_{c(s_1 s_2, \ldots, s_n)} \rho(x) dx$
$\mathcal{V}$	$-\ln \epsilon$
kT	β^{-1}
$\mathcal{F}(\mathcal{V},T) = -kT \ln Z(\beta)$	$-\frac{\beta-1}{\beta} D(\beta) \ln \epsilon = (1 - kT) D(\frac{1}{kT}) \mathcal{V}$
$\phi(\beta) = \frac{\mathcal{F}}{\mathcal{V}}$	$\frac{\beta-1}{\beta} D(\beta)$

Se pueden relacionar otro tipo de cantidades "macroscópicas" como promedios termodinámicos, asociados a otras funciones de partición, tales como las entropías de Rényi y los exponentes de Liapunov. En las referencias antes citadas se encuentra más información.

Apéndice C

De la FBC a la FDP

En este anexo mostramos la forma de obtener a partir de la FBC la FDP correspondiente a través de la función de distribición de probabilidad acumulada.

Sea la FDP con $r = 1, 2, \ldots N$, y la variable que denota la frecuencia o el tamaño de alguna propiedad, la FBC es:

$$y = f(r) = A\frac{(N + 1 - r)^b}{r^a} \tag{C.1}$$

nuestro objetivo consiste en invertir está expresión:

$$r = f^{-1}(y) \tag{C.2}$$

y después normalizar mediante el número total de elementos (N), entonces esta es la función de distribución acumulada:

$$FDA(y) = \frac{r(y)}{N} = \frac{f^{-1}(y)}{N} \tag{C.3}$$

La derivada de la ecuación anterior es la FDP:

$$p(y) = FDP(y) = \frac{d}{dy}FDA(y) = \frac{1}{N}\frac{d}{dy}f^{-1}(y) = \frac{1}{N}\frac{1}{dy/dr} \tag{C.4}$$

(con, $p(y)dy = dr/N$)

C.1. FBC y FDP

En el caso de la FBC se puede obtener la FDP de manera cerrada para algunos casos particulares de los parámetros.

C.1.1. $a = b = \alpha$

$$y = f(r) = A\frac{(N+1-r)^{\alpha}}{r^{\alpha}} \tag{C.5}$$

$$(\frac{y}{A})^{1/\alpha} = \frac{N+1-r}{r} \tag{C.6}$$

$$(\frac{y}{A})^{1/\alpha} = \frac{N}{r} - \frac{1}{r} - 1 \tag{C.7}$$

$$(\frac{y}{A})^{1/\alpha} + 1 = \frac{1}{r}(N+1) \tag{C.8}$$

$$\frac{(\frac{y}{A})^{1/\alpha} + 1}{N+1} = \frac{1}{r} \tag{C.9}$$

$$r = \frac{N+1}{(\frac{y}{A})^{1/\alpha} + 1} \tag{C.10}$$

$$p(y) = \frac{1}{N}\frac{d}{dy}r(y) \tag{C.11}$$

$$p(y) = \frac{N+1}{N}\frac{d}{dy}((\frac{y}{A})^{1/\alpha} + 1)^{-1} \tag{C.12}$$

$$p(y) = -\frac{N+1}{AN\alpha}\frac{(\frac{y}{A})^{\frac{1}{\alpha}-1}}{((\frac{y}{A})^{\frac{1}{\alpha}} + 1)^2} \tag{C.13}$$

También se puede resolver para los casos $a = 1, b = 2$, $a = 2, b = 1$, $a = 3, b = 2$,$a = 2, b = 3$.

C.2. Función de verosimilitud (*likelihood*) y MLE

Si tenemos datos $\{y_i\}$ $(i = 1, 2, \ldots N)$, entonces la función de verosimilitud *likelihood* $L(a, b|y)$ está definida como [130, 131]:

$$L(a, b|y) = \Pi_{i=1}^{N} p(a, b; y_i) \tag{C.14}$$

La log-verosimilitud $\mathcal{L}$ está dada por el logaritmo de L

$$\mathcal{L}(a,b|y) = \sum_{i=1}^{N} \ln p(a,b;y_i) \tag{C.15}$$

con $p(a,b;y_i)$ la FDP con parámetros a, b.

para determinar los parámetros estimados $\hat{a}$ y $\hat{b}$ de la distribución para un conjunto de datos y_i, se resuelven las siguientes ecuaciones que hacen que la log-verosimilitud sea un extremo, aeste método se le conoce como *Maximum Likelihod Estimation (MLE)*:

$$\frac{\partial}{\partial a}\mathcal{L}(a,b|y) = 0 \tag{C.16}$$

$$\frac{\partial}{\partial b}\mathcal{L}(a,b|y) = 0 \tag{C.17}$$

C.3. Determinación de los coeficientes vía MLE

$$\mathcal{L}(a,b|y) = \sum_{i=1}^{M} \ln p(a,b;y_i) \tag{C.18}$$

Se tiene el estimado del parámetro a $\hat{a}$

$$\frac{\partial}{\partial a}\mathcal{L}(a,b|y) = 0 \tag{C.19}$$

Se tiene el estimado de b $\hat{b}$

$$\frac{\partial}{\partial b}\mathcal{L}(a,b|y) = 0 \tag{C.20}$$

$$p(a,b;y) = \frac{N+1}{AN\alpha} \frac{(\frac{y}{A})^{\frac{1}{\alpha}-1}}{((\frac{y}{A})^{\frac{1}{\alpha}}+1)^2} \tag{C.21}$$

Para este caso, la función $\mathcal{L}$ está dada por

$$K \sum_{i=1}^{M} \ln \frac{(\frac{y_i}{A})^{\frac{1}{\alpha}-1}}{((\frac{y_i}{A})^{\frac{1}{\alpha}}+1)^2} \tag{C.22}$$

Con la determinación de la función

$$\mathcal{L}(a,b|y) = K\sum_{i=1}^{M}(\frac{1}{\alpha}-1)\ln(\frac{y_i}{A}) - 2\ln((\frac{y_i}{A})^{\frac{1}{\alpha}}+1) \tag{C.23}$$

con $K = \log\frac{N+1}{An\alpha}$

$$\frac{\partial\mathcal{L}}{\partial\alpha} = K\sum_{i=1}^{M}(-\frac{1}{\alpha^2})\log(\frac{y_i}{A}) - 2\frac{\frac{1}{\alpha}(\frac{y_i}{A})^{\frac{1}{\alpha}-1}}{(\frac{y_i}{A})^{\frac{1}{\alpha}}+1} = 0 \tag{C.24}$$

Estas ecuaciones se resuelven numéricamente para encontrar la estimación vía MLE.

Bibliografía

[1] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *ACM SIGCOMM Computer Communication Review*, volume 29, pages 251–262. ACM, 1999.

[2] L.A. Adamic, R.M. Lukose, A.R. Puniyani, and B.A. Huberman. Search in power-law networks. *Physical review E*, 64(4):046135, 2001.

[3] F. Lucchin and S. Matarrese. Power-law inflation. *Physical Review D*, 32(6):1316, 1985.

[4] L.A. Adamic and B.A. Huberman. Power-law distribution of the world wide web. *Science*, 287(5461):2115, 2000.

[5] X. Gabaix, P. Gopikrishnan, V. Plerou, and H.E. Stanley. A theory of power-law distributions in financial market fluctuations. *Nature*, 423(6937):267–270, 2003.

[6] F. Auerbach. Das gesetz der Bevölkerungskonzentration. *Petermanns Geographische Mitteilungen*, 59(13):73–76, 1913.

[7] A.J. Lotka and DC) Washington Academy of Sciences (Washington. The frequency distribution of scientific productivity. Washington Academy of Sciences, 1926.

[8] H.W. Singer. The "Courbe des Populations." A Parallel to Pareto's Law. *The Economic Journal*, 46(182):254–263, 1936.

[9] V. Pareto. Cours d\'economie politique. 1896.

[10] R.Gibrat. Les Inegalities Economiques; Applications: aux inegalities des richesses, a la concentration des entreprises, aux populations des

villes, aux statistiques des familles, etc., d une loi nouvelle, la loi de l effet proportionnel. *Librarie du Recueil Sirey,Paris*, 1931.

[11] G.K. Zipf. Human behavior and the principle of least effort. 1949.

[12] D.J. Watts and S.H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684):440–442, 1998.

[13] A.L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.

[14] R. Albert and A.L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[15] B. Zhang and S. Horvath. A general framework for weighted gene coexpression network analysis. *Statistical applications in genetics and molecular biology*, 4(1):1128, 2005.

[16] MEJ Newman. Power laws, Pareto distributions and Zipf's law. *Arxiv preprint cond-mat/0412004*, 2004.

[17] G.U. Yule. A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, pages 21–87, 1925.

[18] I.S. Gradshteyn and I.M. Ryzhik. Table of integrals. *Series, and Products (Academic, New York, 1980)*, (1.421), 1980.

[19] Per Bak, Chao Tang, and Kurt Wiesenfeld. Self-organized criticality: An explanation of the 1/f noise. *Phys. Rev. Lett.*, 59(4):381–384, Jul 1987.

[20] A. Clauset, C.R. Shalizi, and MEJ Newman. Power-law distributions in empirical data. *arXiv*, 706, 2007.

[21] S. Redner. How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 4(2):131–134, 1998.

[22] R. Mansilla, E. Köppen, G. Cocho, and P. Miramontes. On the behavior of journal impact factor rank-order distribution. *Journal of Informetrics*, 1(2):155–160, 2007.

[23] G Martinez-Mekler, R Alvarez-Martinez, M Beltran del Rio, R Mansilla, P Miramontes, and G. Cocho. Universality of Rank-Ordering Distributions in the Arts and Sciences. *PLOS ONE*, 4(3):e4971, 2009.

[24] I.I. Popescu. On a Zipf's law extension to impact factors. *Glottometrics*, 6:83–93, 2003.

[25] M. Beltran del Río, G. Cocho, and GG Naumis. Universality in the tail of musical note rank distribution. *Physica A: Statistical Mechanics and its Applications*, 387(22):5552–5560, 2008.

[26] W. Li. Fitting chinese syllable-to-character mapping spectrum by the beta rank function. *Physica A: Statistical Mechanics and its Applications*, 2011.

[27] W. Li, P. Miramontes, and G. Cocho. Fitting Ranked Linguistic Data with Two-Parameter Functions. *Entropy*, 12(7):1743–1764, 2010.

[28] W. Li and P. Miramontes. Fitting ranked english and spanish letter frequency distribution in us and mexican presidential speeches. *Arxiv preprint arXiv:1103.2950*, 2011.

[29] D.M. Bates and D.G. Watts. *Nonlinear regression analysis and its applications*, volume 2. Wiley Online Library, 1988.

[30] M. Beltran del Rio, G. Cocho, and R. Mansilla. General model of subtraction of stochastic variables. Attractor and stability analysis. *Physica A: Statistical Mechanics and its Applications*, 2010.

[31] W.J. Reed and B.D. Hughes. From gene families and genera to incomes and internet file sizes: Why power laws are so common in nature. *Physical Review E*, 66(6):067103, 2002.

[32] K. Sneppen, P. Bak, H. Flyvbjerg, and M.H. Jensen. Evolution as a self-organized critical phenomenon. *Proceedings of the National Academy of Sciences*, 92(11):5209, 1995.

[33] MEJ Newman and RG Palmer. Models of extinction. *A Review.(adaporg/9908002)*, 1999.

[34] H.A. Simon. On a class of skew distribution functions. *Biometrika*, 42(3/4):425–440, 1955.

[35] D.S. Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306, 1976.

[36] P.L. Krapivsky, S. Redner, and F. Leyvraz. Connectivity of growing random networks. *Physical review letters*, 85(21):4629–4632, 2000.

[37] R.K. Merton. The matthew effect in science. *Science*, 159(3810):56–63, 1968.

[38] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin. Structure of growing networks with preferential linking. *Physical Review Letters*, 85(21):4633, 2000.

[39] P. Bak and P. Bak. *How nature works: the science of self-organized criticality*, volume 212. Copernicus New York, 1996.

[40] P. Bak and C. Tang. Earthquakes as a self-organized critical phenomenon. *Journal of Geophysical Research*, 94(B11):15635–15, 1989.

[41] Z. Olami, H.J.S. Feder, and K. Christensen. Self-organized criticality in a continuous, nonconservative cellular automaton modeling earthquakes. *Physical Review Letters*, 68(8):1244–1247, 1992.

[42] E.T. Lu and R.J. Hamilton. Avalanches and the distribution of solar flares. *The Astrophysical Journal*, 380:L89–L92, 1991.

[43] P. Bak and K. Sneppen. Punctuated equilibrium and criticality in a simple model of evolution. *Physical Review Letters*, 71(24):4083–4086, 1993.

[44] H. Hong, M. Ha, and H. Park. Finite-size scaling in complex networks. *Physical review letters*, 98(25):258701, 2007.

[45] R. Ferrer-i Cancho and B. Elvevåg. Random texts do not exhibit the real zipf's law-like rank distribution. *Plos One*, 5(3):e9411, 2010.

[46] L.A.N. Amaral, A. Scala, M. Barthélémy, and H.E. Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences*, 97(21):11149, 2000.

[47] D. Sornette. *Critical phenomena in natural sciences: chaos, fractals, selforganization, and disorder: concepts and tools*. Springer Verlag, 2004.

[48] J. Aitchison and J.A. Brown. *The lognormal distribution*. Cambridge Univ. Pr., 1969.

[49] E. Limpert, W.A. Stahel, and M. Abbt. Log-normal distributions across the sciences: keys and clues. *Bioscience*, 51(5):341–352, 2001.

[50] D. Sornette. Multiplicative processes and power laws. *Physical Review E*, 57(4):4811–4813, 1998.

[51] D. Sornette and R. Cont. Convergent multiplicative processes repelled from zero: power laws and truncated power laws. *J. Phys. I France*, 7:431–444, 1997.

[52] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet mathematics*, 1(2):226–251, 2004.

[53] GG Naumis and G. Cocho. Tail universalities in rank distributions as an algebraic problem: The beta-like function. *Physica A: Statistical Mechanics and its Applications*, 387(1):84–96, 2008.

[54] H. Jeong, SP Mason, A.L. Barabási, and ZN Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

[55] A. Dragulescu and V.M. Yakovenko. Evidence for the exponential distribution of income in the usa. *The European Physical Journal B-Condensed Matter and Complex Systems*, 20(4):585–589, 2001.

[56] J.M. Montoya, S.L. Pimm, and R.V. Solé. Ecological networks and their fragility. *Nature*, 442(7100):259–264, 2006.

[57] J. Laherrere and D. Sornette. Stretched exponential distributions in nature and economy:"fat tails"with characteristic scales. *The European Physical Journal B*, 2(4):525–539, 1998.

[58] JC Phillips. Stretched exponential relaxation in molecular and electronic glasses. *Reports on Progress in Physics*, 59:1133–1207, 1996.

[59] GG Naumis and G. Cocho. The tails of rank-size distributions due to multiplicative processes: from power laws to stretched exponentials and beta-like functions. *New Journal of Physics*, 9:286, 2007.

[60] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

[61] G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[62] H. Akaike. On entropy maximization principle. *Application of statistics*, 1977.

[63] A.D.R. McQuarrie and C.L. Tsai. *Regression and time series model selection*. World Scientific Pub Co Inc, 1998.

[64] W. Li. Spatial $1/f$ Spectra in Open Dynamical Systems. *Europhys. Lett*, 10(5):395–400, 1989.

[65] R. Alvarez-Martinez, G. Martinez-Mekler, and G. Cocho. Order-disorder transition in conflicting dynamics leading to rank-frequency generalized beta distributions. *Physica A: Statistical Mechanics and its Applications*, 390(1):120–130, 2011.

[66] M. Nowak. *Evolutionary Dynamics: Exploring the Equations of Life*. Cambridge, MA: Harvard University Press, 2006.

[67] W. Li and K. Kaneko. Long-range correlation and partial 1/f α spectrum in a noncoding DNA sequence. *Europhys. Lett*, 17(7):655–660, 1992.

[68] CK Peng, SV Buldyrev, AL Goldberger, S. Havlin, F. Sciortino, M. Simons, HE Stanley, et al. Long-range correlations in nucleotide sequences. *Nature*, 356(6365):168–170, 1992.

[69] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626, 1968.

[70] M. Kimura. *The neutral theory of molecular evolution*. Cambridge Univ Pr, 1985.

[71] M. Kimura. DNA and the neutral theory. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 312(1154):343, 1986.

[72] L.H. Tang. Random-tiling quasicrystal in three dimensions. *Physical review letters*, 64(20):2390–2393, 1990.

[73] C.L. Henley. Random tiling models. *Quasicrystals: the state of the art*, 16:459–560, 1991.

[74] L.J. Shaw, V. Elser, and C.L. Henley. Long-range order in a three-dimensional random-tiling quasicrystal. *Physical Review B*, 43(4):3423–3433, 1991.

[75] M. Widom, DP Deng, and CL Henley. Transfer-matrix analysis of a two-dimensional quasicrystal. *Physical review letters*, 63(3):310–313, 1989.

[76] A. Czirok, R.N. Mantegna, S. Havlin, and H.E. Stanley. Correlations in binary sequences and a generalized Zipf analysis. *Physical Review E.*, 52(1):446–452, 1995.

[77] A. Czirók, H.E. Stanley, and T. Vicsek. Possible origin of power-law behavior in n-tuple Zipf analysis. *Physical Review E*, 53(6):6371–6375, 1996.

[78] T. Prellberg and J. Slawny. Maps of intervals with indifferent fixed points: thermodynamic formalism and phase transitions. *Journal of statistical physics*, 66(1):503–514, 1992.

[79] M.C. Mackey. *Time's arrow: the origins of thermodynamic behavior*. Dover Pubns, 2003.

[80] S.H. Strogatz. *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering*. Westview Pr, 1994.

[81] B. Kitchens. *Symbolic dynamics: one-sided, two-sided, and countable state Markov shifts*. Springer Verlag, 1998.

[82] D.A. Lind and B. Marcus. *An introduction to symbolic dynamics and coding*. Cambridge Univ Pr, 1995.

[83] P. Collet and J.P. Eckmann. *Iterated maps on the interval as dynamical systems*, volume 1. Birkhauser, 1997.

[84] B. Derrida, A. Grevois, and Y. Pomeau. Universal metric properties of bifurcations of endomorphisms. *Journal of Physics A: Mathematical and General*, 12:269, 1979.

[85] M. Metropolis, HL Stein, and PR Stein. J. combinatorial theory. *Aj*, 15:25, 1973.

[86] N. Metropolis, ML Stein, and PR Stein. On finite limit sets for transformations on the unit interval. *Journal of Combinatorial Theory, Series A*, 15(1):25–44, 1973.

[87] A. Lasota and M.C. Mackey. *Chaos, fractals, and noise: stochastic aspects of dynamics*, volume 97. Springer, 1994.

[88] D. Pingel, P. Schmelcher, and FK Diakonos. General theory and examples of the inverse Frobenius-Perron problem. *Arxiv preprint chao-dyn/9801028*, 1998.

[89] D. Pingel, P. Schmelcher, and FK Diakonos. Theory and examples of the inverse Frobenius–Perron problem for complete chaotic maps. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 9:357, 1999.

[90] FK Diakonos, D. Pingel, and P. Schmelcher. A stochastic approach to the construction of one-dimensional chaotic maps with prescribed statistical properties. *Physics Letters A*, 264(2-3):162–170, 1999.

[91] R. Bowen. Markov partitions for Axiom A diffeomorphisms. *American Journal of Mathematics*, pages 725–747, 1970.

[92] R. Bowen and J.R. Chazottes. *Equilibrium states and the ergodic theory of Anosov diffeomorphisms*. Springer Verlag, 2008.

[93] D. Ruelle and G. Gallavotti. *Thermodynamic formalism*. Addison-Wesley Reading, Massachusetts, 1978.

[94] Y.G. Sinai. Gibbs measures in ergodic theory. *Russian Mathematical Surveys*, 27:21, 1972.

[95] M. Eigen. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58(10):465–523, 1971.

[96] M. Eigen and P. Schuster. A principle of natural self-organization. *Naturwissenschaften*, 64(11):541–565, 1977.

[97] M. Eigen, W. Gardiner, P. Schuster, and R. Winkler-Oswatitsch. The origin of genetic information. *Scientific American*, 244:88–92, 1981.

[98] M. Eigen, J. McCaskill, and P. Schuster. Molecular quasi-species. *The Journal of Physical Chemistry*, 92(24):6881–6891, 1988.

[99] G.W. Hoffmann. Co-selection in immune network theory and in aids pathogenesis. *Immunology and cell biology*, 72(4):338–346, 1994.

[100] E. Domingo and JJ Holland. Rna virus mutations and fitness for survival. *Annual Reviews in Microbiology*, 51(1):151–178, 1997.

[101] M.A. Nowak, R.M. May, and R.M. Anderson. The evolutionary dynamics of HIV-1 quasispecies and the development of immunodeficiency disease. *Aids*, 4(11):1095–1103, 1990.

[102] E. Domingo, C.K. Biebricher, M. Eigen, and J.J. Holland. *Quasispecies and RNA virus evolution: principles and consequences*. Landes Bioscience Austin (Texas), 2001.

[103] C.L. Burch and L. Chao. Evolvability of an rna virus is determined by its mutational neighbourhood. *Nature*, 406(6796):625–628, 2000.

[104] EA Duarte, IS Novella, SC Weaver, E. Domingo, S. Wain-Hobson, DK Clarke, A. Moya, SF Elena, JC De La Torre, and JJ Holland. Rna virus quasispecies: significance for viral disease and epidemiology. *Infectious agents and disease*, 3(4):201, 1994.

[105] JC De La Torre and JJ Holland. Rna virus quasispecies populations can suppress vastly superior mutant progeny. *Journal of virology*, 64(12):6278, 1990.

[106] E. Domingo, E. Martínez-Salas, F. Sobrino, J.C. de la Torre, A. Portela, J. Ortín, C. López-Galindez, P. Pérez-Breña, N. Villanueva, R. Nájera, et al. The quasispecies (extremely heterogeneous) nature of viral rna genome populations: biological relevance–a review. *Gene*, 40(1):1–8, 1985.

[107] E. Domingo, E. Baranowski, C.M. Ruiz-Jarabo, A.M. Martín-Hernández, J.C. Sáiz, and C. Escarmís. Quasispecies structure and persistence of rna viruses. *Emerging Infectious Diseases*, 4(4):521, 1998.

[108] EA Duarte, IS Novella, SC Weaver, E. Domingo, S. Wain-Hobson, DK Clarke, A. Moya, SF Elena, JC De La Torre, and JJ Holland. Rna virus quasispecies: significance for viral disease and epidemiology. *Infectious agents and disease*, 3(4):201, 1994.

[109] M. Eigen, C.K. Biebricher, et al. Sequence space and quasispecies distribution. *RNA genetics*, 3:211–245, 1988.

[110] JJ Holland, JC De La Torre, and DA Steinhauer. RNA virus populations as quasispecies. *Current topics in microbiology and immunology*, 176:1, 1992.

[111] W.L. Schneider and M.J. Roossinck. Genetic diversity in rna virus quasispecies is controlled by host-virus interactions. *Journal of virology*, 75(14):6566, 2001.

[112] I. Leuthäusser. Statistical mechanics of eigen's evolution model. *Journal of statistical physics*, 48(1):343–360, 1987.

[113] P. Tarazona. Error thresholds for molecular quasispecies as phase transitions: From simple landscapes to spin-glass models. *Physical Review A*, 45(8):6038, 1992.

[114] E. Baake, M. Baake, and H. Wagner. Ising quantum chain is equivalent to a model of biological evolution. *Physical review letters*, 78(3):559–562, 1997.

[115] D.B. Saakian, C.K. Hu, and H. Khachatryan. Solvable biological evolution models with general fitness functions and multiple mutations in parallel mutation-selection scheme. *Physical Review E*, 70(4):041908, 2004.

[116] D.B. Saakian and C.K. Hu. Solvable biological evolution model with a parallel mutation-selection scheme. *Phys Rev E*, 69:046121, 2004.

[117] D. Saakian and C.K. Hu. Eigen model as a quantum spin chain: Exact dynamics. *Arxiv preprint cond-mat/0402212*, 2004.

[118] B. Luque. An introduction to physical theory of molecular evolution. *Central European Journal of Physics*, 1(3):516–555, 2003.

[119] C. Beck and F. Schlögl. *Thermodynamics of chaotic systems: an introduction*, volume 4. Cambridge Univ Pr, 1995.

[120] R. Badii and A. Politi. *Complexity: hierarchical structures and scaling in physics*, volume 6. Cambridge Univ Pr, 1999.

[121] Y.G. Sinai. Construction of dynamics in one-dimensional systems of statistical mechanics. *Theoretical and Mathematical Physics*, 11(2):487–494, 1972.

[122] J.P. Eckmann and D. Ruelle. Ergodic theory of chaos and strange attractors. *Reviews of modern physics*, 57(3):617, 1985.

[123] D. Ruelle. An inequality for the entropy of differentiable maps. *Bulletin of the Brazilian Mathematical Society*, 9(1):83–87, 1978.

[124] R. Bowen. Equilibrium states and the ergodic theory of axiom a diffeomorphisms (lecture notes in mathematics, 470), 1975.

[125] A. Renyi. *Foundations of probability*, volume 9. Holden-Day San Francisco, 1970.

[126] B.B. Mandelbrot. Intermittent turbulence in self-similar cascades- divergence of high moments and dimension of the carrier. *Journal of Fluid Mechanics*, 62(2):331–358, 1974.

[127] HGE Hentschel and I. Procaccia. The infinite number of generalized dimensions of fractals and strange attractors. *Physica D: Nonlinear Phenomena*, 8(3):435–444, 1983.

[128] P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena*, 9(1-2):189–208, 1983.

[129] C. Beck. Upper and lower bounds on the renyi dimensions and the uniformity of multifractals. *Physica D: Nonlinear Phenomena*, 41(1):67–78, 1990.

[130] R.A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368, 1922.

[131] A.W.F. Edwards. *Likelihood*. Cambridge Univ Pr, 1984.