UNIVERSIDAD NACIONAL AUTONOMA
DE MEXICO

POSGRADO EN CIENCIAS FÍSICAS

"UBICUIDAD DE LA DISTRIBUCIÓN BETA DE DOS PARÁMETROS. FENOMENOLOGÍA Y MODELOS"

T E S I S

QUE PARA OBTENER EL GRADO DE:

DOCTOR EN CIENCIAS (FÍSICA)

PRESENTA:

Manuel Beltrán del Río García.

DIRECTOR DE TESIS: DR. Germinal Cocho Gil.

MIEMBRO DE COMITÉ TUTORAL: DR. Denis P. Boyer.
MIEMBRO DE COMITÉ TUTORAL: DR. Gustavo Martínez-Mekler.

pcf

posgrado en ciencias físicas
u n a m

MÉXICO, D.F. 2010

Ubicuidad de la Distribución Beta de Dos Parámetros. Fenomenología y Modelos.

Manuel Beltrán del Río García

11 de agosto de 2010

Resumen

El presente trabajo expone nuestra investigación sobre la ubicuidad de distribuciones estadísticas tipo "rango-tamaño" o "rango-frecuencia" [1], que pueden ser bien imitadas (o representadas) por la distribución Beta de dos parámetros (DBDP)

$$s = N\frac{(R-r)^{\beta}}{r^{\alpha}}. \tag{1}$$

La ubicuidad de la distribución beta fue estudiada mediante dos tipos fundamentalmente diferentes de trabajo:

- La creación y análisis de un corpus fenomenológico de distribuciones; el desarrollo de técnicas y criterios, computacionales y analíticos, para establecer su validez como ejemplos en el ámbito de la presente investigación; y la fabricación heurística de generadores de distribuciones de este tipo.
- La búsqueda de una última razón a través de la cuál se explique, de la manera más global y sencilla posible, la presencia de la regularidad bajo estudio. Para dicho fin se desarrollaron modelos generales.

Los bloques principales que componen la tesis son:

- Una introducción sobre la historia del estudio de regularidades estadísticas. Descripción de nuestra circunstancia, intereses y motivaciones originales.
- Un capítulo sobre el planteamiento del problema y metodología.
- Un capítulo sobre el corpus de datos obtenidos, procedimientos y criterios de "aprobación" y análisis.
- Un capítulo sobre los modelos propuestos, sus características y aplicabilidad.
- Una comparación entre los modelos, con una discusión de su carácter global y Conclusiones.

Outline

The present work surveys our research work about the ubiquity of rank-size and rank-frequency distributions [1] that are well fitted by the two parameter beta distribution (TPBD)

$$s = N\frac{(R-r)^{\beta}}{r^{\alpha}}. \tag{2}$$

Our work was done through two fundamentally different kinds of work:

- The creation and analysis of a phenomenologic corpus of distributions. The development of numerical and analytic techniques and criteria to establish their validity as examples and study matter.
- The search for a mechanism that would explain the wide occurrence of the TPBD, as generic and inclusive as possible.

This work's main sections are:

- A brief introduction to the history of the research of statistical regularities such as ours.
- A chapter stating our particular problem and the methodology involved.
- A chapter that summaries the set of obtained or created data.
- A chapter about the developed models that produce a TPBD
- Concluding remarks with a comparison between the models and their relationship to some of the empirical findings.

Índice general

Capítulo 1

INTRODUCCIÓN

1.1. Antecedentes

Del medir y ponderar, cimientos del quehacer científico, deriva inevitablemente una condensación de información: el objeto de estudio, con todas sus peculiaridades y circunstancias, es reducido a un conjunto abstracto de propiedades presumible e idealmente características. Y es solo así, claramente si no deseamos invocar alguna suerte de facultad de precepción metafísica, que el inquisidor natural puede sistematizar su estudio. Una vez llevada a cabo esta reducción es posible, y es el propósito de este trabajo mostrar en efecto sucede, que de dos fenómenos visiblemente disímbolos se puede extraer consistentemente el mismo tipo de patrón de medidas. Cuando así sucede ha demostrado ser prudente pensar *a priori* que existe una razón en común: intuitivamente, la casualidad es "costosa en términos de probabilidad"[1] , y dado que hemos partido de la tesis de que las fuentes de ambos conjuntos de información parecen ajenos después de un estudio superficial, es pertinente pensar que se ha encontrado una generalidad subyacente o que se ha caído en el ridículo de haber desarrollado una técnica para obtener sistemáticamente el mismo resultado, sin importar el origen de los datos. Es necesario pues dotarse de un conjunto de contraejemplos o de diferentes categorías de resultados encontrados para desembarazarse de esta última posibilidad. De todos los casos históricos, quizás los más pertinentes sean el del Teorema del Límite Central, y el de las leyes de potencia.

1.1.1. Teorema del Límite Central

El Teorema del Límite Central, sin duda el más célebre caso de una descripción detallada de un fenómeno de ubicuidad, es un ejemplo de la situación atrás descrita.

La enunciación estándar del Teorema del Límite Central es :

[1] Siguiendo el principio de parsimonia de Occam interpretado por Russell [3].

La suma, normalizada por $1/\sqrt{N}$, de N variables aleatorias identicamente distribuidas, de promedio cero y con varianza finita σ^2, es una variable aleatoria con una distribución de densidad de probabilidad que converge en el sentido de medida a una Gaussiana con varianza σ^2 en el límite $N \to \infty$ [1].

El Teorema del Límite Central, y variantes cercanas, engloban un vasto grupo de fenómenos bajo la misma estampa: la distribución de probabilidad común que resulta de un muestreo suficientemente extenso de ciertas proyecciones. Cabe mencionar que en este caso, lo que antes considerabamos como la consecuencia de una "compresión" o pérdida de información inherente a la medición, es ahora el producto de dos procesos diferentes de condensación o abstracción. En el caso del Teorema del Límite Central, el muestreo estadístico que reduce la información se hace sobre un objeto cuya naturaleza ha sido ya enmascarada, *i.e.*, la multitud de dimensiones que caracterizan su circunstancia han sido fundidas, de acuerdo a una receta muy particular dictada por el teorema, en un fenómeno subordinado aunque todavía complejo.

Las condiciones que impone el Teorema del Límite Central a esta primera reducción se pueden relajar a:

- Que la amalgama de resultados de la interacción de los numerosos factores se pueda matematizar como la convolución de las distribuciones de probabilidad de cada uno de éstos, *i.e.*, que cada resultado sea el producto decoherente de eventos independientes, y:
- Que la dispersión de éstos sea del mismo orden de magnitud.

En virtud de la primera imposición se logra que la probabilidad de cada resultado se pueda determinar por el conjunto de probabilidades que cada una de sus condiciones favorables, sin tener que espeificar el orden, temporal o jerárquico, de éstas. La segunda constricción asegura que el rastrear el origen de un resultado en particular nos lleve no a un solo juego particular de circunstancias, sino a uno conjunto múltiple de ellas. Estas condiciones son suficientemente laxas, o al menos suficientemente "naturales" , para explicar la notoria ubicuidad de las distribuciones de Gauss y Lévy.

1.1.2. Leyes de potencia

Las leyes de potencia aparecen como distribuciones en incontables procesos de muy diversas areas [4, 5, 6, 8]. Esta ubicuidad ha sido tema de estudio y debate durante más de un siglo. Tenemos como ejemplos célebres a la ley de Zipf, y a la ley de Pareto [4, 7]. La primera es una aplicación de una ley de potencia a la distribución rango-frecuencia de todas las palabras diferentes que se encuentran en un texto dado (se puede generalizar también a conjuntos comunes de dos o más palabras, procedimiento aplicado comunmente en lenguajes tonales como el Mandarín) ver Figs. 1.1 y 1.2. La segunda es también una ley de potencias que se aplicó originalmente a las distribuciones de ingresos monetarios en familias

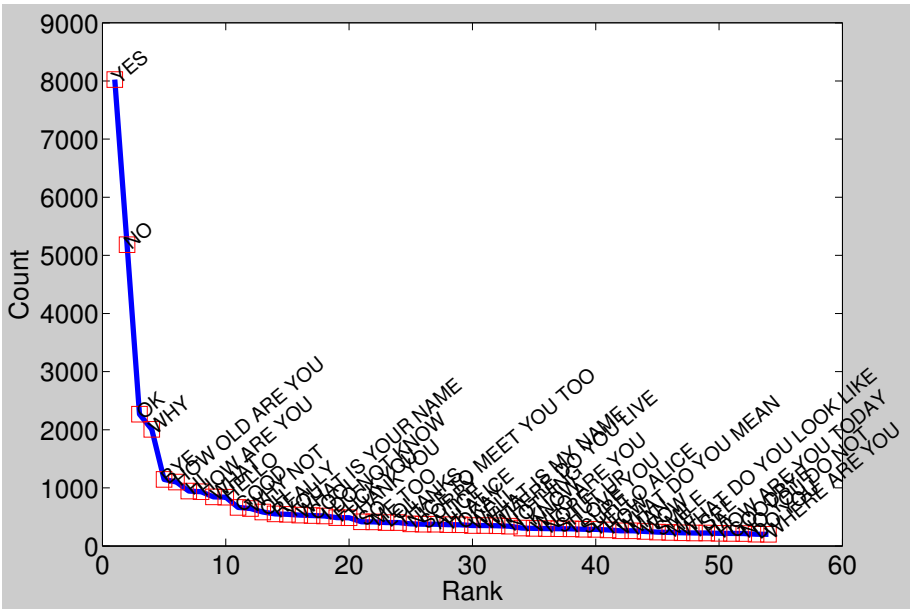


Figura 1.1: Distribución rango–frecuencia de las palabras y frases cortas diferentes en una muestra de lenguaje. Esta muestra fue tomada no de un texto, como suele hacercse al ilustrar la ley de Zipf, sino de un generador de lenguaje a base de inteligencia artificial [17].

[4]. Otros ajustes de potencias a distribuciones se han encontrado en las fluctuaciones de los valores bursátiles [15], en las distribuciones de familias de genes y proteínas [9], en las distribuciones de equilibrio de sistemas compuestos [10], en las distribuciones de "motivos" en "imágenes naturales" [11], en procesos multiplicativos [12] y en redes complejas de muy diversas naturalezas [13], por citar solo algunos.

Se ha recurrido a una gran cantidad de modelos diferentes para generar leyes de potencia. Existen modelos dinámicos, aditivos, multiplicativos, de fragmentación, de optimización, *etc*, y en muchos de éstos casos se han cometido abusos al adjudicar la forma de ley de potencia a el modelo creado. La misma ley de Zipf es un ejemplo de ésto. El primer modelo utilizado para explicar esta regularidad fue criticado fuertemente y un famoso debate surgió entre Mandelbrot y Simon [14], en favor y en contra respectivamente, de un modelo de optimización de los recursos de la gramática. Otra fuente común de crítica es la dificultad que puede presentarse para discriminar una ley de potencia de otras posibles, en una distribución empírica dada [15]. Las distribuciones log–normales, por ejemplo, se pueden confundir con leyes de potencia si se considera solamente un dominio suficientemente pequeño, ya que en una representación log-log una distribución log-normal se asemeja a una función recta [8]. Inclusive existen

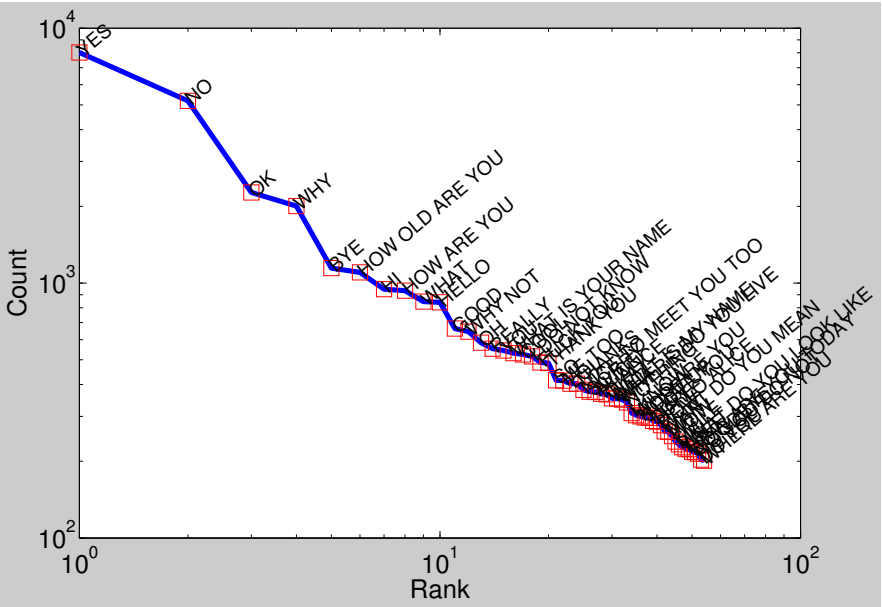


Figura 1.2: Misma colección de datos que en la figura 1.1 [17], en escala log-log para evidenciar la recta que denota una ley de potencia.

maneras de aproximar leyes de potencia mediante exponenciales, que se usan para aligerar la carga de cómputo al hacer cálculos numéricos [16]. Estas dificultades deben ser cuidadosamente descartadas después de haber adjudicado una ley de potencia a algún fenómeno en particular, que es en realidad otra dificultad importante y común, que suele depender principalmente de un corpus de datos suficientemente grande como para hacer estadística significativa, y de qué manera se realiza y califica un ajuste [18].

Problemas en ajustes con leyes de potencia Muchas veces cuando se quiere ajustar una colección de datos, se tienen regiones de la muestra que pertenecen a dinámicas diferentes, especialmente en el caso de una muestra de un sistema físico. Cuando ésto sucede, se puede optar por usar en cada una de estas regiones un ajuste diferente, correspondiente a la dinámica que generó esa región en particular. Discriminar entre este y el caso en que en realidad toda la muestra fue generada por el mismo proceso, no es en general una tarea trivial. Un caso típico de una muestra mixta se suele dar en la forma de una colección de datos que en escala log-log presenta una sección lineal intermedia a dos regímenes diferentes [7]. En dicha sección media es posible hacer un ajuste de ley de potencia si existe una razón *a priori* para descartar a los datos circundantes como parte significativa de la muestra, ya sea porque se piense que son producidas por una dinámica diferente a la que se piensa modelar o porque simplemente no tienen un peso significativo (ver fig. 1.4). Entonces es difícil adjudicar una

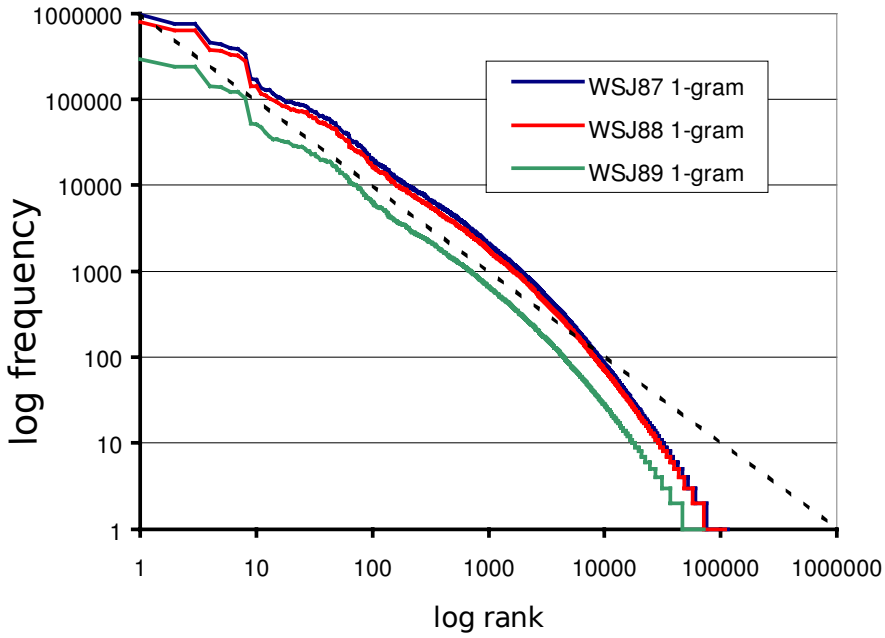
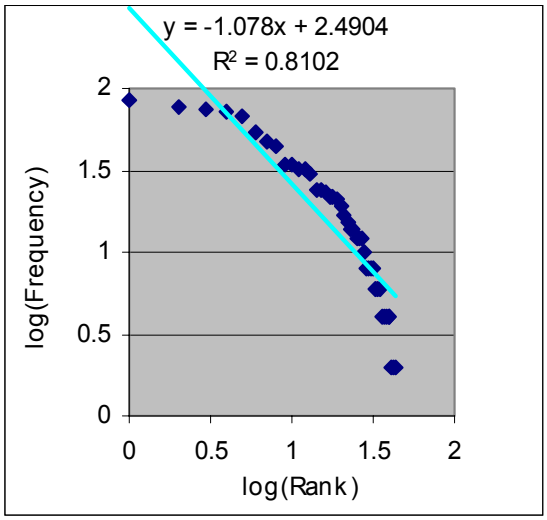


Figura 1.3: Desviación de la ley de Zipf en un texto grande (texto del Wall Street Journal, tomada de [7]).

ley de potencia a una muestra que no sea estrictamente lineal en escala log–log. Ejemplo de esto es el fenómeno descrito en [7], en donde las distribuciones de palabras empiezan a desviarse de la ley de Zipf cuando el tamaño de la muestra pasa cierto límite fig 1.3.

Del hecho de que en una muestra continua, en escala log-log, una sección suficientemente pequeña es recta, se deriva la necesidad de asegurar que el dominio de la distribución en donde se usa una ley de potencia sea satisfactoriamente grande.

Vemos entonces que de la sencillez misma de las leyes de potencia se derivan los problemas para identificarlas y validarlas. La poca flexibilidad de forma que poseen hace que los criterios de aplicación sean estrictos. La ubicuidad de distribuciones en forma de potencias con colas ha motivado la búsqueda de correcciones a las leyes de potencia para obtener ajustes más satisfactorios, para los cuales no sea necesario seccionar el dominio [24].



Pitch distribution for Bach's Orchestral Suite No.3
in D '2. Air on the G String', BWV.1068.

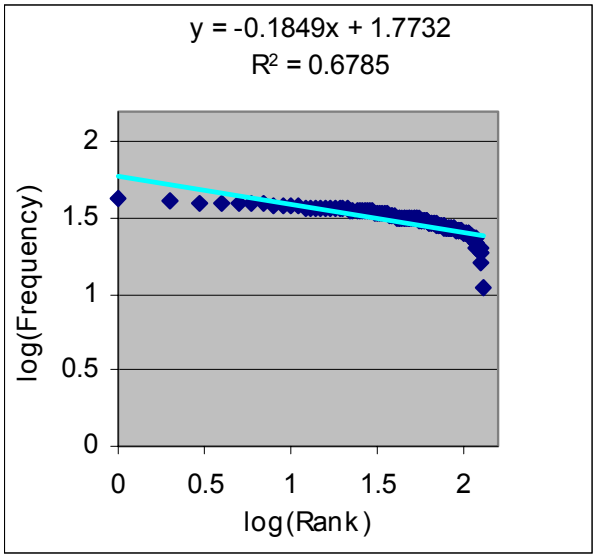


Figura 1.4: Ejemplos de una distribución mixta y un ajuste de potencias para la sección media entre dos regímenes diferentes. [19] y [20]

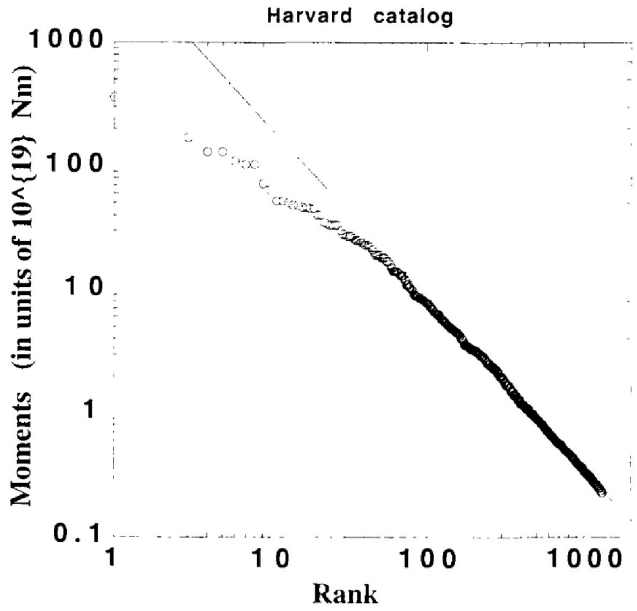


Figura 1.5: Distribución rango-tamaño de terremotos registrados según su magnitud (en doble logaritmo), tomado de [1].

1.1.3. Rango–tamaño y rango–frecuencia

Rango-tamaño.

La distribución rango tamaño es simplemente la función $t(r)$ que resulta de ordenar los datos de cierta colección $\{t_1 = t_{\text{máx}} \geq t_2 \geq t_3 \geq \ldots\}$, normalmente de mayor a menor, mediante una nueva variable llamada "rango". Asignamos pues el rango $r = 1$ al mayor de los valores, $t_{\text{máx}}$, *i.e.*, $t(1) = t_{\text{máx}}$, $r = 2$ al segundo, *etc.* La variable r también puede ser tomada como contínua, para propósitos teóricos. A manera de ejemplo, podemos ver en la figura 1.5 la distribución rango-tamaño de los terremotos registrados según su magnitud [1].

Si en vez de una colección de datos lo que se tiene es la distribución de probabilidad $\rho(t)$ de los valores que puede adoptar la variable t, entonces la distribución rango–tamaño se puede obtener de la inversión de la función $r(t)$, que a su vez se puede obtener como consecuencia de lo siguiente:

$$r(t) \propto \int_t^{t_{max}} \rho(\xi)\mathbf{d}\xi. \tag{1.1}$$

Dicho de otra manera, el rango de un valor es proporcional a la probabilidad de encontrar un valor mayor cualquiera.

Rango–frecuencia

La distribución rango-frecuencia es un histograma $f(r)$, ordenado de mayor a menor, *i.e.*, $f(1)$ es el número de veces (o la frecuencia relativa, si así se desea normalizar) con la que aparece el valor más frecuente en la muestra, $f(2)$ aquella del valor que le sigue y así sucesivamente. Como ejemplo, en la figura 1.1 en la sección anterior, tenemos la distribución rango-frecuencia de las palabras que conforman un texto en particular.

1.2. Distribución beta de dos parámetros

Definimos la distribución beta de dos parámetros (desde ahora DBDP) como:

$$s = N\frac{(R-r)^{\beta}}{r^{\alpha}}. \tag{1.2}$$

Aquí s y r son las variables de "tamaño" y "rango" respectivamente, R es el número de rangos distintos [2], α y β son los parámetros de las potencias, generalmente números positivos. N es un factor de normalización que puede ser interpretado de distintas maneras. Por ejemplo, si deseamos representar una distribución discreta rango–tamaño, entonces haríamos:

$$N = S_{\text{máx}}\frac{1}{(R-1)^{\beta}}, \tag{1.3}$$

con $S_{\text{máx}}$ el valor ("tamaño") más grande de la muestra, para instantáneamente asignarle a este último el rango $r = 1$.
Si la Ec.1.2 se trata de una ditribución rango-frecuencia continua, entonces la variable s representa la densidad de probabilidad de la "r-ava"muestra más abundante. En dicho caso la normalización queda:

$$N^{-1} = R^{-\alpha+\beta-1}\mathbf{B}(-\alpha+1, \beta+1), \tag{1.4}$$

en donde $\mathbf{B}(z, w)$, la función beta [21], se define como

$$\mathbf{B}(z, w) \equiv \int_0^1 t^{z-1}(1-t)^{w-1}\mathbf{d}t, \tag{1.5}$$

y entonces la distribución beta queda normalizada como una distribución de probabilidad.

Las fiugras 1.6 y 1.7 muestran ejemplos de la DBDP con valores de los parámetros dentro del intervalo que típicamente utilizamos. De dichas figuras también se puede inferir que, *a grosso modo*, los parámetros α y β controlan la curvatura de las "colas" de la distribución, izquierda y derecha respectivamente, con realmente poca influencia en el extremo opuesto.

[2] En problemas discretos se suele usar $R-1$ como el número total de rangos.

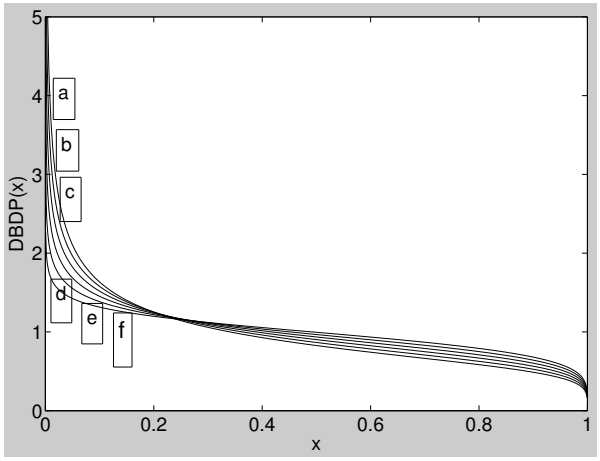


Figura 1.6: Dependencia de la distribución beta del parámetro α, los valores para cada curva son: (a): $\alpha = 0.35$,(b): $\alpha = 0.3$,(c): $\alpha = 0.25$,(d): $\alpha = 0.2$,(e): $\alpha = 0.15$,(f): $\alpha = 0.1$ y para todas $\beta = 0.2$

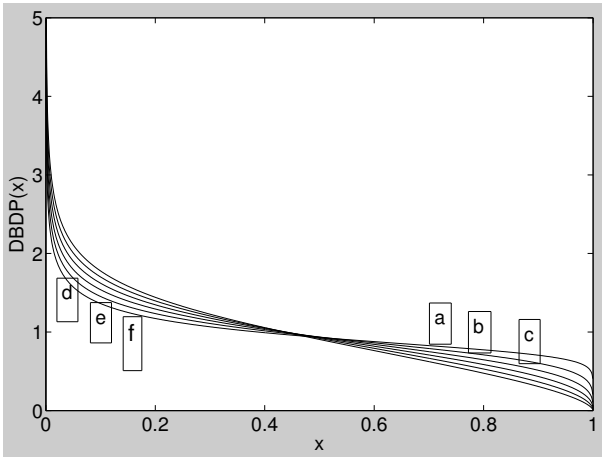


Figura 1.7: Dependencia de la distribución Beta del parámetro β, los valores para cada curva son: (a): $\beta = 0.1$,(b): $\beta = 0.2$,(c): $\beta = 0.3$,(d): $\beta = 0.4$,(e): $\beta = 0.5$,(f): $\beta = 0.6$ y para todas $\alpha = 0.2$

1.3. Problema

Una vez dispuesto el glosario técnico en la sub-sección anterior, podemos proceder a plantear el problema principal. Su origen recae en el hallazgo fortuito, totalmente heurístico y a manos del Dr. Germinal Cocho, de que la distribución rango-frecuencia de las ternas de bases nitrogenadas (codones) en la cadena de ADN de procariontes, es, bajo gran precisión numérica, una DBDP (Ver sección 2.2). Inmediatamente después se demostró que no solo las ternas que corresponden al marco de lectura de transcripción siguen esta ley, los marcos desplazados e inclusive la distribución de grupos de cuatro o más de bases específicas también lo hacen [49]. A partir de este último descubrimiento empezó la sospecha de que detrás de la DBDP hay un mecanismo general. Poco después se corroboró la presencia de la DBDP utilizando un modelo que imita, mediante cadenas binarias, la composición estadística de las secuencias nucleótidos del ADN: el modelo de expansión–modificación [22]. En paralelo también se utilizaron secuencias simbólicas generadas con mapeos caóticos (Sección 3.1.2 y [49]), cuyas distribuciones rango–frecuencia fueran DBDP, para tratar de asociar elementos de la dinámica caótica, como intermitencia, a las propiedades del modelo de expansión modificación y descubrir una alternativa más general a este último para generar la DBDP.

1.3.1. Ubicuidad de la DBDP

Durante el trabajo inicial con mapeos caóticos y motivados por la sospecha de que la DBDP pertenece a una clase muy general de distribuciones "naturales", se encontró poco a poco gran cantidad de distribuciones de muy diversos orígenes que siguen este patrón estadístico, desde sistemas sociales como la distribución de poblaciones en ciertas regiones [49], hasta en contextos propios de la física como la distribución de degeneraciones por nivel de sistemas discretos de bosones (Sección 3.2), en modelos de ecología de poblaciones [46], la distribución rango-frecuencia de notas en composiciones musicales (Artículos [24] y [25], reproducidos en la sección 2.1), solo por nombrar algunas. La sección 2 es un recuento detallado del corpus de ejemplos que hemos encontrado.

1.3.2. Búsqueda de un modelo genérico

Dada la gran colección creciente de fenomenología, y en vista de la discusión planteada en la sección 1.1, la búsqueda de un modelo genérico, en vez de buscar explicaciones *ad hoc* para cada uno de los diferentes sistemas, parece ser la estrategia más apropiada para abordar el problema.

Antes de haber derivado explicaciones generales, *i.e.* de carácter abstracto, se estudiaron, en la medida de lo posible, las causas individuales que hacen que la forma de las distribuciones se ajuste satisfactoriamente a una DBDP, y solo entonces, a partir de las similitudes entre ellas, se buscaron pistas para confeccionar un modelo lo más sencillo y general posible. Es así que, a manera de ejemplos, para explicar el caso de las distribuciones en comunidades ecológicas

se recurriera a ecuaciones estocásticas, al modelo de expansión–modificación para aquel de las secuencias genéticas, a modelos de fragmentación o deposición irreversible para entender la aparición de la DBDP en el caso de motivos en fachadas arquitectónicas o en pinturas, *etc.*

1.3.3. Artículos y siguientes capítulos

Después de comentar los artículos [25] y [24] (Reproducidos en el apéndice A) en el capítulo 2.1, en el capítulo 2 exponemos la colección de tipos y casos de distribuciones que hemos encontrado se ajustan satisfactoriamente con una DBDP y la calidad de dicha representación en cada ámbito. Finalmente, en el capítulo 3, con la presentación de los modelos y conjeturas desarrollados para explicar el fenómeno de la ubicuidad de la DBDP, que es la parte medular del trabajo.

Capítulo 2

FENOMENOLOGÍA

2.1. Distribución rango–frecuencia de notas en música armónica

El caso de la distribución rango-frecuencia de las notas en composiciones musicales demostró ser de particular ayuda al total del trabajo: Para obtener las distribuciones pertinentes se utilizaron archivos codificados MIDI (*Musical Instrument Digital Interface*). Dado que los archivos MIDI son de fácil acceso, pequeños y de dominio público, fue posible obtener muestras de cerca de 2000 piezas, dando un respaldo estadístico considerable. Por eso los avances más rápidos se lograron en este ámbito y en aquél de las secuencias genéticas, de las cuales existe también una colección vastísima. Los dos artículos que presentamos, tratan de esta sección de nuestro trabajo. Aquella que usó como fuentes de distribuciones archivos MIDI y de a qué particularidades se le atribuye a estos el desplegar DBDP.

En el artículo [24] se publicaron los primeros resultados sobre DBDP y música. Después de una breve introducción a la DBDP análoga a la de ésta tesis se muestra la precisión con la cuál se ajusta la DBDP a las distribuciones rango-frecuencia de las notas musicales en una pieza dada, con particular énfasis en la diversidad histórica y musical de las piezas utilizadas. La primera parte de la discusión se enfoca a destacar como en la gran mayoría de las distribuciones de notas existe una estructura con dos curvaturas que claramente no puede ser descrita por una ley de potencia, como se pretendió frecuentemente [14, 19, 20], y cómo en cambio la forma sigmoidea de la DBDP resulta ideal para describir estas distribuciones.

La segunda sección contiene una gráfica del espacio de parámetros en la cuál se incluye un punto por cada par de parámetros correspondiente al ajuste de cada una de las casi 2000 piezas, que se utilizó en el intento de atribuir a los parámetros de los ajustes algún sentido estadístico o musical. Finalmente, se explica como dicha relación entre pieza y ajuste se encontró después de notar que en el espacio de parámetros las piezas se encontraban distribuidas de acuerdo con

ciertas características armónicas básicas, lo que permitió darnos cuenta de cómo se podían separar las distribuciones rango-frecuencia de notas en la convolución de varias distribuciones independientes, estableciendo así una pista importante para investigar el origen de la DBDP en música.

El segundo artículo concerniente a DBDP en el contexto de música ([25]) contiene también un recuento introductorio acerca de la DBDP, el formato MIDI, y las distribuciones rango-frecuencia. Las contribuciones originales de este segundo artículo son la profundización sobre el significado estructural de las distribuciones independientes que componen una distribución rango-frecuencia de notas en música y principalmente la comparación que hacemos entre éstas y las distribuciones parciales de arbustos y árboles que se presentan en [38], sugiriendo así que quizás la DBDP es el resultado general de sumas de distribuciones independientes en un límite lejano al de grandes números, una suma de solo unas cuantas variables estocásticas.

2.2. Genética

La distribución rango-frecuencia de ternas de bases nitrogenadas en secuencias genéticas fue el primer encuentro con la DBDP por parte de nuestro equipo. El trabajo original era una propuesta para el origen del alfabeto ternario de la síntesis protéica [39]. Durante el trabajo para dicha investigación, se concluyó que la DBDP era una excelente representación para las distribuciones rango tamaño de tripletes y otros conjuntos de "n-ómeros", pero solamente para las secuencias genéticas de procariontes o para regiones específicas de las secuencias de eucariontes. Esta última característica resultó ser de gran importancia, dado que la principal diferencia (en términos estadísticos) entre la constitución de las cadenas genéticas de procariontes y aquellas de eucariontes es la existencia en estas últimas de grandes secciones de "código espúreo", constituidas por numerosas repeticiones de una misma secuencia corta (generalmente menor a 300 bases nitrogenadas [44]).

La figura 2.1 muestra la diferencia general entre las distribuciones rango–tamaño del contenido de codones en cadenas de procariontes y cadenas completas (incluyendo secciones no-codificadoras) de eucariontes, representados aquí por la bacteria *E. Coli* y por *H. Sapiens* respectivamente. La figura incluye el ajuste de DBDP para la muestra de procarionte. En el caso de las cadenas eucariontes es notoria una subestructura en la composición de los codones menos representados, es decir, en la región derecha de la distribución rango–frecuencia, y se hace evidente que una función simple de dos curvaturas como la DBDP no puede adecuarse bien a este tipo de distribuciones. Lo primero que sugerimos fue que la diferencia quizás proviene porque en el caso de eucariontes se incluyen las distribuciones diferentes que corresponden a las regiones codificadores y no codificadoras, pero no se llegó a caracterizar de manera suficientemente clara la diferencia entre las composiciones de uno y otro tipos de secuencias. Aunque es

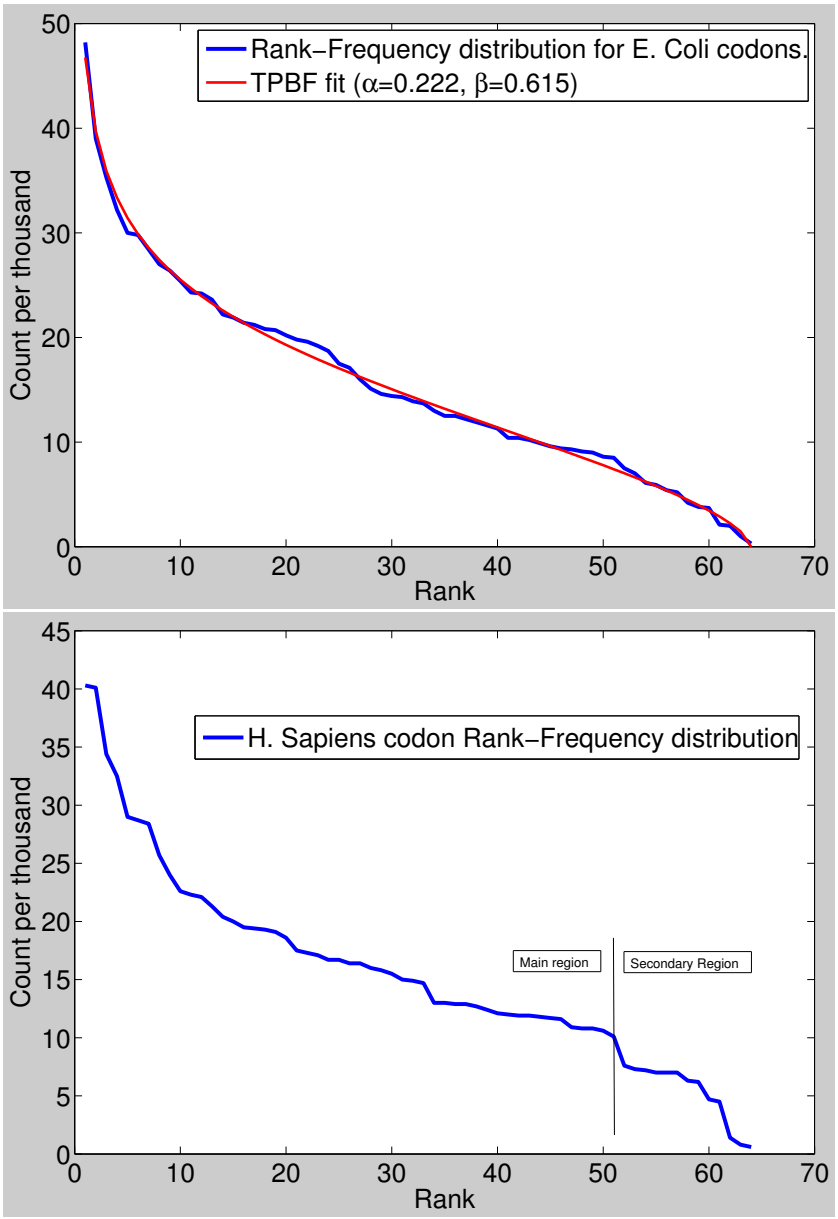


Figura 2.1: Distribución Rango-Tamaño del contenido de codones en *E.Choli* y *H. Sapiens* respectivamente.

sabido que las secciones no codificadoras se constituyen principalmente, como ya lo habiamos mencionado, de numerosas repeticiones de un patrón relativamente pequeño, no encontramos en estos componentes, en primera inspección, alguna diferencia notoria de la cual sospechar responsable de las discrepancias entre las distribuciones de eucariontes y procariontes [44]. Creemos, sin embargo, que descubrir la causa de esta diferencia no necesitará mucho más esfuerzo que una simple inspección superficial, y que además similitudes recientemente encontradas con el tipo de histogramas que se encuentran al estudiar otro tipo de secuencias simbólicas, como la música o el lenguaje escrito (figuras .4 y 5 del segundo artículo en el apéndice A), facilitarían un rápido recuento de posibles explicaciones.

2.3. Distribución rango-tamaño de cobertura de vegetación

Una de las muestras más satisfactorias que tenemos es aquella que encontramos en [38], en donde se muestran los datos sobre la presencia (en área) de las diferentes especies de árboles y arbustos que crecieron en un determinado lote en el estado de Illinois, EE.UU.A. La motivación original residía en los avances que a la fecha se habían hecho en el ámbito de la DBDP en un célebre trabajo sobre teoría neutralista de distribución de especies por S.Hubbel [46].

De dichos datos se extrajeron la distribuciones rango-tamaño de arbustos, matas y árboles, y también, de todas las especies juntas (Ver última figura del segundo artículo en el apéndice A.

Si bien las distribuciones individuales para cada tipo de planta no son plenamente satisfactorias en cuanto a ser representadas por la DBDP, la distribución que resulta de incluir las tres lo es cabalmente. Este hecho alimentó la sospecha de que quizás la DBDP es una distribución estable bajo algún esquema sencillo de "adición", en un sentido análogo al de la distribución gaussiana bajo convolución. Sobre esta sospecha se elaboró un extenso trabajo que presentamos en la sección 3.4, y otro reciente usando los trabajos previos en distribuciones de notas en composiciones armónicas [25].

2.4. Factor de impacto en revistas científicas

En la referencia [42], se expone un ejemplo de distribución en excelente acuerdo con la DBDP, la distribución rango-tamaño del "factor de impacto" (*Journal Impact Factor*) de las revistas de publicación científica.

A diferencia de los anteriores ejemplos, éste surge en un ámbito social, sin reglas claras de crecimiento, aunque a veces asociado con modelos de la dinámica de nodos de redes.

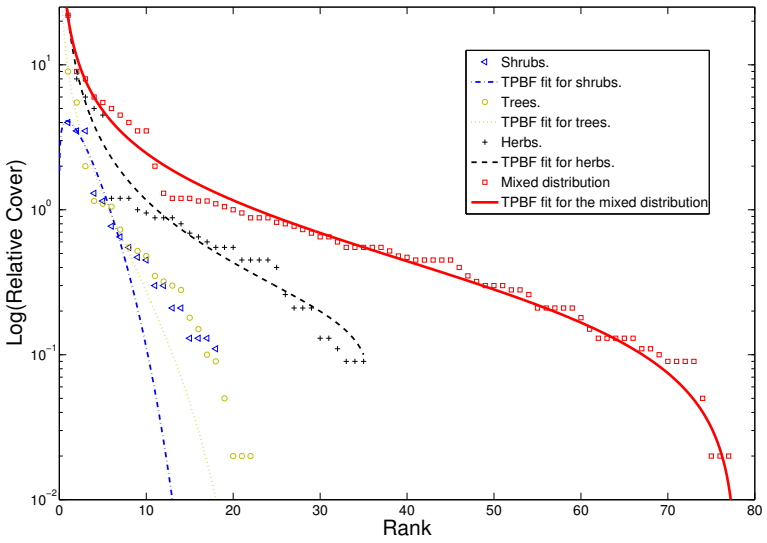


Figura 2.2: Distribución rango-tamaño de especies arbóreas en un lote abandonado en el estado de Illinois, EEUUA.

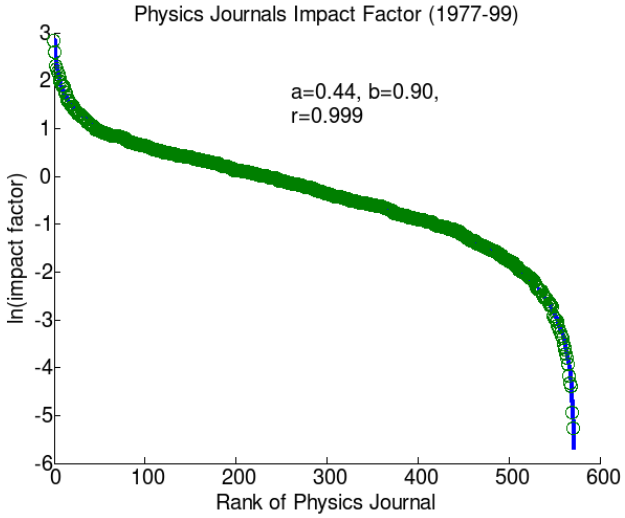


Figura 2.3: Distribución rango-tamaño del factor de impacto de revistas científicas [42].

2.5. Artículo comprensivo de fenomenologia

En [49] contamos con un compendio publicado acerca del conjunto de ejemplos que nuestro grupo ha encontrado. En el apéndice B reproducimos dicho artículo.

Capítulo 3

MODELOS GENERADORES DE DBDP

En este capítulo exponemos varios de los modelos que hemos desarrollado para explicar la aparición de la DBDP.

Los siguientes modelos son muestra de algunos intentos de explicar la ubicuidad de la DBDP que hemos desarrollado. Su conjunto representa la mayor parte del trabajo en torno al presente tema. La sección 3.1 es una breve exposición de los primeros modelos que fueron desarrollados en torno a la DBDP. Todos los modelos presentados en esta sección fueron desarrollados principalmente por el Dr. Germinal Cocho y su grupo. Salvo por contribuciones menores y simulaciones numéricas implementadas para el modelo descrito en la sub-sección 3.1.2, el autor del presente trabajo no tiene mayores colaboraciones en el desarrollo de los modelos descritos en 3.1. La sección 3.2 es un ejercicio combinatorio, típico de mecánica cuántica, a través del cual se generan de manera sencilla funciones que en cierto límite coinciden con la DBDP. En la sección 3.4 se expone el modelo estocástico de resta de variables que es una modificación sencilla del aspecto dinámico del Teorema Del Límite Central y parte medular del conjunto de trabajos de la presente.

3.1. Primeros modelos

3.1.1. Expansión–Modificación

El modelo de expansión modificación (MEM [22]) es un sistema dinámico simbólico que genera secuencias binarias mediante mecanismos que imitan las maneras más comunes de mutaciones en una secuencia genética. La manera de generar secuencias del modelo es la siguiente:

1. **Se da una semilla arbitraria.** De manera arbitraria se da una secuencia binaria pequeña, de dos o tres dígitos.

2. **Empieza la iteración.** Cada elemento de la secuencia es expuesto al siguiente proceso: con una probabilidad p el elemento cambia al valor opuesto ($1 \to 0$ ó $0 \to 1$).En caso de que no se cumpla el evento de probabilidad p (con una probabilidad $1-p$) el elemento no cambia y se añade un elemento idéntico en el siguiente espacio de la secuencia.

3. **Se agranda la secuencia.** Repitiendo el procedimiento se agranda la secuencia, para lograr una distribución rango-tamaño de grupos de n elemntos consecutivos es suficiente una longitud de $n1000$.

Por ejemplo, escogemos que la probabilidad de cambio sea $p = 0.6$ y empezamos con la semilla

$$1101.$$

Para cada elemento se escoge un número aleatorio $\epsilon \in [0, 1]$.

$$\overbrace{1}^{.4753}\ \overbrace{1}^{.9541}\ \overbrace{0}^{.2168}\ \overbrace{1}^{.8743}$$

Si ϵ para un elemento resulta ser $\epsilon < .6$ entonces cambiamos el valor del elemento, de otra manera se inserta un elemento igual justo en el espacio siguiente:

$$\overbrace{1}^{.4753}\ \overbrace{1}^{.9541}\ \overbrace{0}^{.2168}\ \overbrace{1}^{.8743}$$

$$\to$$

$$\overbrace{0}\ \overbrace{11}\ \overbrace{1}\ \overbrace{11}$$

para enctonces acabar con la secuencia:

$$011111$$

lo cual se vuelve a volver a iterar hasta lograr una secuencia larga.

La distriución rango-frecuencia de una de estas secuencias, sin importar del valor del parámetro p o el número n para el tamaño de los grupos a contar, es ajustada con éxito [1] por una DBDP.

La conección entre las secuencias genéticas y el modelo de espansión-modificación se da por construcción [22]. Existe evidencia clara de que los cambios en las secuencias genéticas se dan a nivel de pares de bases o de grupos numerosos de ellos pero no particularmente en codones. El hecho de que en primera aproximación el número de aminoácidos que aparecen en una secuencia suficientemente grande está relacionado directamente con el número de degeneraciones en el código que transcriben ese aminoácido. En otras palabras, si un aminoácido en

[1] El valor del parámetro r^2 para los ajustes rara vez difería de 1 por más de una centésima.

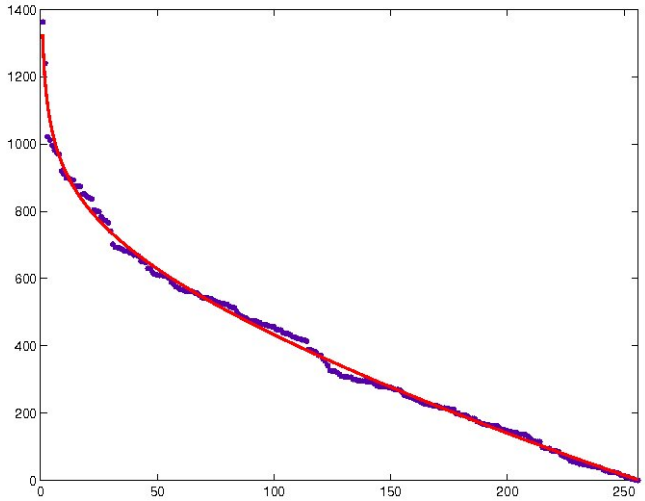


Figura 3.1: Distribución rango-tamaño de una secuencia generada por modelo de expansión-modificación para una probabilidad de cambio $p = .9$ y grupos de ocho elementos. El valor del parámetro r^2 para el ajuste (línea contínua en la gráfica) es de 0.996 con $a = 0.622$ y $b = 1.010$.

particular es codificado por cuatro codones diferentes, entonces es más probable que aparezca que un aminoácido que corresponde solamente dos o tres codones diferentes.

Dada esta independencia de las mutaciones y los marcos de lectura, podemos inferir que un sistema que imite los mecanismos de mutación más comunes, aunque sea una secuencia en donde no existan asimetrías como las del marco de lectura natural, tendería a una distribución del mismo tipo que las secuencias genéticas reales. Éste es el caso del modelo de expansión–modificación.

Un modelo que solo incluya uno de los dos factores de cambio, expansión o modificación, produciría distribuciones uniformes o gaussianas. Por ejemplo, consideramos una secuencia binaria aleatoria, arbitrariamente larga, y la sometemos a un proceso que incluya solamente modificación de símbolos individuales con probabilidad p, entonces el resultado es una secuencia con una distribución idéntica a la de la secuencia original, una distribución uniforme. Este hecho se hace evidente cuando pensamos de la siguiente manera: si partimos de una secuencia binaria aleatoria suficientemente grande, obtendremos distribuciones rango–tamaño uniformes. Si ahora invertimos los valores individuales ($1 \rightarrow 0$ ó $0 \rightarrow 1$) de todos los elementos obtendremos, por supuesto, una distribución idéntica a la original. Si en vez de cambiar todos los símbolos solo los modificamos con probabilidad p. Entonces tenemos, después del proceso, dos subconjuntos de nuestra secuencia aleatoria original, uno que quedó igual y otro

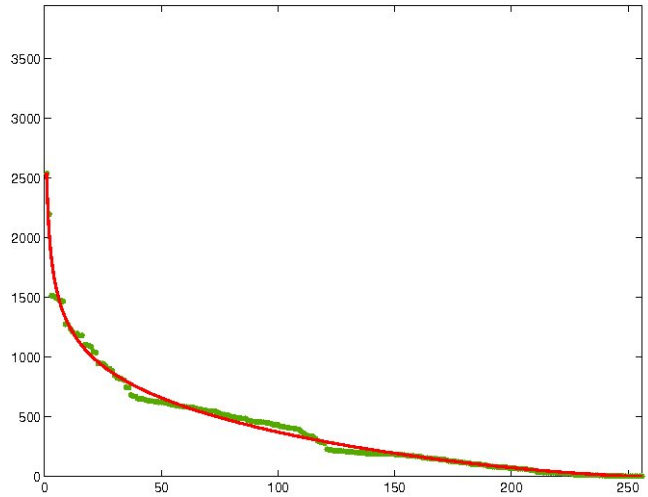


Figura 3.2: Distribución rango-tamaño de una secuencia generada por modelo de expansión–modificación para una probabilidad de cambio $p = .5$ y grupos de ocho elementos. El valor del parámetro r^2 para el ajuste (línea contínua en la gráfica) es de 0.989 con $a = 0.732$ y $b = 2.376$.

cuyos elementos cambiarion por su valor opuesto, para una secuencia lo suficientemente grande[2] y un valor de p suficientemente alejado de 1, el primer conjunto tendría una distribución parecida a la original, y el segundo tendría la distribucíon de una serie aleatoria cuyos elementos en su totalidad fueron cambiados de valor, es decir, como vimos arriba, tendría también una distribución homogénea.

Los exponentes a y b han sido asociados a partes opuestas de la dinámica del MEM. El primer parámetro está asociado a la inercia del sistema y el último, b, está ligado a la tendencia que tuvo el sistema a cambiar, iteración con iteración, la secuencia original. Ver [49] para una discusión más extensa. Vemos entonces que la aparición de una distribución Beta en el modelo expansión–modificación (figuras 3.1,3.2 y 3.3) es consecuencia del conflicto entre los dos tipos de cambio [23].

[2]Si el conjunto que no cambió sigue siendo considerablemente grande entonces es una muestra representativa de todo el conjunto, dada la aleatoriedad. El tamaño de dicho subconjunto depende no solo de el tamaño original de la muestra, sino también del valor de p, que de alguna manera representa las proporciones entre la serie original y su subconjunto que no cambió.

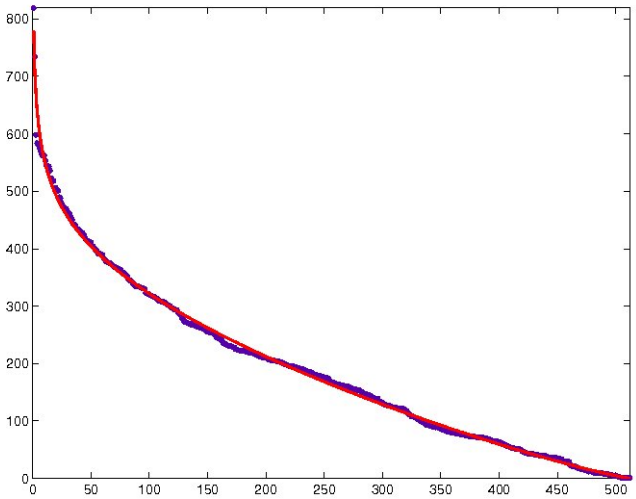


Figura 3.3: Distribución rango–tamaño de una secuencia generada por modelo de expansión–modificación para una probabilidad de cambio $p = .9$ y grupos de nueve elementos. El valor del parámetro r^2 para el ajuste (línea continua en la gráfica) es de 0.997.

3.1.2. Mapeos unimodales.

Una vez establecida la relación entre el MEM y la DBDP, se extendió la analogía a las secuencias simbólicas generadas a través del bien conocido mapeo logístico,

$$x_{n+1} = \mu x_n(1 - x_n). \tag{3.1}$$

Dichas secuencias se generan asignando a cada uno de los iterados $x_n \in [0, 1]$ uno de dos símbolos, dependiendo de si $x_i \in [0, \frac{1}{2})$ o no. Dependiendo del valor del parámetro μ, las secuencias de grupos de n pueden ser, desde una secuencia de un solo símbolo, en el caso de valores de μ en donde el mapeo no se comporte de manera caótica, hasta una combinación caótica de los 2^n símbolos diferentes, si el valor de μ corresponde a alguna ventana periódica. Si calculamos la distribución rango-tamaño de secuencias con un valor de μ en la región no caótica obtendremos datos solo para unos cuantos símbolos, puesto que la dinámica adopta rápidamente orbitas que recorren sistemáticamente unos cuantos puntos, pero si acercamos el valor de μ al valor de ergodicidad total ($\mu \approx 4$), entonces en la distribución rango–frecuencia empiezan a aparecer todos los símbolos posibles, y consistentemente tienen una misma forma funcional de DBDP.

3.1.3. Ecuación maestra

Modelo neutralista de Hubbel. En un trabajo previo del Dr.Germinal Cocho, inspirado en el famoso modelo de ecología neutralista de Hubbel [46], él demostró que mediante la elección adecuada de ciertos coeficientes, y tomando cierto límite, podemos obtener una DBDP de la solución estacionaria de la siguiente ecuación maestra determinada por el operador diferencial **M**:

$$\frac{\partial P(k,t)}{\partial t} = \mathbf{M}(P(k,t)) \equiv d(k+1)P(k+1,t)+b(k-1)P(k-1,t)-(b+d)P(k,t). \tag{3.2}$$

En el trabajo original de Hubbel la ecuación de evolución de la probabilidad es [46]:

$$\frac{\partial N(k,t)}{\partial t} = aN(k,t) - cN(k,t)F(\sigma(t)) + \mathbf{M}(N(k,t)), \tag{3.3}$$

en donde $N(k,t)$ es la probabilidad de que una especie tenga k individuos al tiempo t, $\sigma(t)$ es el número total de individuos de todas las epecies, y $d(k)$ y $b(k)$, que derivan sus nombres de los términos *death* y *birth*, fueron definidas originalmente como amplitudes de transición en $+1$ y -1 respectivamente del número de individuos de la especie k. Podemos transformar esta última ecuación en la ecuación 3.2 si hacemos el cambio a la variable normalizada $n(k,t) \equiv \frac{N(k,t)}{\sigma}$. Entonces obtenemos dos ecuaciones:

$$\frac{\partial n}{\partial t} = \mathbf{M}(n(k,t)) \tag{3.4}$$

y:

$$\frac{d\sigma}{dt} = (a\sigma - c\sigma F(\sigma)). \tag{3.5}$$

En lo siguiente, solo consideraremos la ec. (3.4) y su solución estacionaria, que es de la forma [48]:

$$n_{\mathbf{est}}(k) = n_0 \prod_{i=1}^{k} \frac{b(i-1)}{d(i)}, \tag{3.6}$$

en donde n_0 es una constante que se obtiene de normalización.

Coeficientes. ¿Qué coeficientes hacen que la solución (3.6) sea una DBDP? Para empezar podemos considerar el siguiente par de amplitudes de transición:

$$d(k) = \lambda_-(C_2 + k)(N_2 - k) \tag{3.7}$$

y

$$b(k) = \lambda_+(C_1 + k)(N_1 - k), \tag{3.8}$$

,en donde todas las constantes son positivas.

Si introducimos la forma explícita de estas amplitudes en el desarrollo de la solución estacionaria (ec.3.6) obtendremos:

$$n_{est}(k) = \left(\frac{\lambda_+}{\lambda_-}\right)^k \frac{C_2!}{(C_1-1)!}\frac{N_1!}{(N_2-1)!}\frac{(k+(C_1-1))!}{(k+C_2)!}\frac{(N_2-k-1)!}{(N_1-k)!} \tag{3.9}$$

Que en el límite:

$$\frac{N_2-N_1}{(N_2+N_1)/2} \equiv \frac{\Delta N}{\overline{N}} \approx 0 \tag{3.10}$$

y

$$k \gg \overline{C} \equiv (C_2+C_1)/2 \tag{3.11}$$

vemos que queda:

$$n_{est}(k) \simeq A e^{-\log\left(\frac{\lambda_+}{\lambda_-}\right)k} k^{C_1-1-C_2}(\overline{N}-k)^{\Delta N}, \tag{3.12}$$

en donde A es un término de normalización.

Como se puede apreciar, esta última expresión para la solución estacionaria tiene la forma que buscábamos, siempre y cuando acordemos fijar $\lambda_+ = \lambda_-$. Si los coeficientes b y d con las formas que acabamos de presentar dan, después de los límites (3.10) y (3.11), una distribución beta, quizás eventualmente se pueda demostrar que una clase más general de coeficientes cumplen el cometido. En primera aproximación cualquier par de funciones expandibles, con un único máximo, podrían preservar la forma de productos de factoriales en la expresión (3.9), y con ello mantener la forma de productos de potencias característica de la distribución beta. De demostrarse, se generalizaría aún más la clase de contextos que generan una distribución beta de primera clase.

3.1.4. Ecuación Fokker–Planck y distribución beta

La ecuación maestra que presentamos es una realización markoviana de una ecuación contínua, que estamos interesados en resolver, para tener una idea de como evolucionan los componentes individuales, cuyas "trayectorias" en promedio están distribuidas de acuerdo a nuestra representación.

Ecuación Fokker–Planck. La primera parte de esta sección es la conversión de la ecuación 3.4 a un régimen continuo, en donde las reglas de difusión sean válidas en cualquier escala. El resultado de dicho límite es una ecuación de Fokker–Planck.

Para obtener el equivalente continuo de la ec.(3.2), podemos considerar un límite en diferencias del típo:

$$\frac{\partial^2 f}{\partial x^2} = \lim_{\epsilon^2 \to 0} \frac{f(x+\epsilon)+f(x-\epsilon)-2f(x)}{\epsilon^2} \tag{3.13}$$

y:

$$\frac{\partial f}{\partial x} = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon}. \tag{3.14}$$

Usando esos límites, podemos reorganizar la ecuación maestra y sumar nuevos términos (en aras de la brevedad adoptamos la notación: $f_+ \equiv f(k+1)$, $f_- \equiv f(k-1)$ y $f \equiv f(k)$):

$$\frac{\partial P}{\partial t} = d_+ P_+ + b_- P_- - (bP + dP) \tag{3.15}$$

$$= d_+P_+ + b_-P_- - bP - dP + (b_+P_+ - b_+P_+) + (d_-P_- - d_-P_-) - (dP - dP) - (bP - bP) \tag{3.16}$$

$$= \overbrace{d_-P_- + d_+P_+ - 2dP} + \overbrace{b_-P_- + b_+P_+ - 2bP} - \overbrace{(b_+P_+ - bP)} - \overbrace{(d_-P_- - dP)} \tag{3.17}$$

$$\approx \frac{\partial^2 (bP/2)}{\partial k^2} + \frac{\partial^2 (dP/2)}{\partial k^2} - \frac{\partial bP}{\partial k} + \frac{\partial bP}{\partial k} \tag{3.18}$$

para finalmente obtener

$$\frac{\partial P}{\partial t} = -\frac{\partial (b-d)P}{\partial k} + \frac{1}{2}\frac{\partial^2 (b+d)P}{\partial k^2}. \tag{3.19}$$

Una manera más elegante de obtener este resultado [47] es mediante la expansión de la ecuación Chapman–Kolmogorov. Si denotamos a la amplitud de transición de un estado k a uno k' como $\langle k|w|k'\rangle$, entonces la ecuación Chapman–Kolmogorov queda:

$$\frac{\partial P}{\partial t} = -P(k,t) \int \langle k|w|k'\rangle dk' + \int P(k',t)\langle k'|w|k\rangle dk'. \tag{3.20}$$

Para expandir la ecuación hacemos un cambio de variable $k' = k + r$ y definimos $w(k,r) \equiv \langle k|w|k'\rangle$. Al utilizar la expansión del generador infinitesimal de desplazamientos:

$$P(k-r,t) = e^{-r\frac{\partial}{\partial k}} P(k,t) = \sum_{n=0}^{\infty} \frac{(-r)^n}{n!} \frac{\partial^n}{\partial k^n} \{P(k,t)\}, \tag{3.21}$$

la ecuación (3.20) cambia a

$$\frac{\partial P}{\partial t} = \int -P(k,t)w(k,r)dr + \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \frac{\partial^n}{\partial k^n} \int r^n P(k,t)w(k,r)dr \tag{3.22}$$

$$= \sum_{n=1}^{\infty} \frac{(-1)^n}{n!} \frac{\partial^n}{\partial k^n} \int r^n P(k,t)w(k,r)dr. \tag{3.23}$$

En nuestro caso (discreto) tenemos que

$$w(k,r) = b(k)\delta(r-1) + d(k)\delta(r+1) \tag{3.24}$$

como nuestra expresión es markoviana[3], obtenemos

$$\frac{\partial P}{\partial t} = \sum_{n=1}^{2} \frac{(-1)^n}{n!} \frac{\partial^n}{\partial k^n} \int r^n P(k,t) w(k,r) dr \tag{3.25}$$

$$= -\frac{\partial (b-d)P}{\partial k} + \frac{1}{2}\frac{\partial^2 (b+d)P}{\partial k^2}, \tag{3.26}$$

recuperado así el resultado del método por diferencias.

Solución La solución estacionaria de la ec. 3.19 se obtiene de manera inmediata si $B \equiv b+d$ y $A \equiv b-d$:

$$\begin{aligned}
\frac{\partial P}{\partial t} &= 0 \Rightarrow \\
(AP) &= \frac{1}{2}\frac{\partial (BP)}{\partial k} \\
\frac{2A - (\partial_k B)}{B} &= \frac{1}{P}\frac{\partial P}{\partial k} \\
\ln(P(k)) &= 2\int_0^k \frac{2(A)' - (\partial_k B')}{B'} dk'
\end{aligned}$$

$$P(k) = C \exp 2 \int_0^k \frac{2(A)' - (\partial_k B')}{B'} dk' \tag{3.27}$$

Una manera alternativa en la que suele encontrarse en la literatura [45] es

$$P(k) = \frac{C}{B} \exp\left(2\int_0^k \frac{A}{B} dk' \right). \tag{3.28}$$

Si los coeficientes b y d cumplen $N_2 > -C_1$ o $N_1 > -C_2$ y $\lambda_+ = \lambda_-$, entonces la expresión $b+d$ tiene dos raíces reales y la integral dentro de la exponencial tiene una solución sencilla:

$$\int \frac{A}{B} dk' = \int \frac{(k'+a)}{(b-k')(k'+c)} dk' = \frac{(a-b)\ln(d+k)}{b+c} - \frac{(a-c)\ln(c-k)}{b+c}. \tag{3.29}$$

Después de exponenciar, la solución (3.27) queda:

[3] El límite de ecuación maestra a Fokker–Planck exige que la expansión se corte después del segundo término, a manera de preservar $\langle l \rangle (\equiv \frac{\Delta l}{\tau})$ y $\frac{\langle l^2 \rangle - \langle l \rangle^2}{\tau} (\propto D)$.

$$P(k) = C \exp\left[2\left(\frac{(a-b)\ln(d+k)}{b+c} - \frac{(a-c)\ln(c-k)}{b+c}\right)\right] \tag{3.30}$$

$$P(k) = C'(k+c_1)^{a'}(c_2-k)^{b'} \tag{3.31}$$

Identificando los coeficientes, vemos que $c_1 \approx \overline{C}$, y por el límtie (3.11) obtenemos la misma forma de la solución de la ecuación maestra con las condiciones (3.10):

$$P(k) = C'(k)^{a'}(c_2-k)^{b'}; \tag{3.32}$$

$$P(k) = C' k^{\Delta C - 1}(\overline{N}-k)^{\Delta N}. \tag{3.33}$$

Hemos mostrado que las restricciones (3.10) y (3.11), y el límite al contínuo de la ecuación maestra son equivalentes.

3.1.5. Ecuación Langevin

No unicidad En el marco de Itō[4], el paso de una ecuación de Langevin generalizada a una ecuación de Fokker-Planck se realiza de manera inmediata[5]:

$$\dot{y} = A(y) + B(y)\eta(t) \tag{3.34}$$

$$\frac{\partial P(y,t)}{\partial t} = -\frac{\partial(A(y)P)}{\partial y} + \frac{1}{2}\frac{\partial^2(B^2(y)P)}{\partial y^2}, \tag{3.35}$$

en donde $\eta(t)$ es ruido con correlación tipo $\delta(t-t')$.

Haciendo lo opuesto, podríamos partir de la ecuación Fokker-Planck (3.35), y descomponemos la modulación del ruido blanco en una suma de N_{max} términos arbitrarios:

$$B(y)^2 = \sum_i^{N_{max}} G_i^2(y) \tag{3.36}$$

Notamos que además de la ec.(3.34), la siguiente ecuación también satisface la relación Langevin $\leftrightarrow$ Fokker–Planck:

$$\dot{y} = A(y) + \sum_i^{N_{max}} G_i(y)\eta_i(t) \tag{3.37}$$

si las variables estocásticas $\eta_i(t)$ son independientes.

[4]La interpretación de Itō del paso a la ecuación de evolución de la densidad de probabilidad resulta de considerar la amplitud del ruido y el valor de la fuerza actuante (fricción, campo, *etc*) como aquellos que tenía el caminante *antes* de realizar el salto discreto [1].

[5]Mediante una integración en diferencias a primer orden.

Interpretación. Para una primera interpretación podemos quedarnos con el caso más sencillo de las ecuaciones de Langevin que son compatibles con nuestra ec. Fokker-Planck:

$$\dot{k} = A(k) + \sqrt{B(k)}\eta(t). \tag{3.38}$$

Sabemos por la condición impuesta a los coeficientes b y d ($\lambda_+ = \lambda_-$), que $A(k) = b - d$ es una función lineal en k. Entonces el término $A(k)$ puede ser interpretado como una fricción estándar.
El término estocástico está modulado por un factor con un máximo único, y para los valores $\overline{N} < k < -\overline{C}$ la difusión se convierte en una difusión anómala que regresa a los caminantes a la región complementaria $\overline{N} > k > -\overline{C}$. Entonces de acuerdo con el desarrollo que culminó con a ec.(3.33), la anterior ecuación estocástica debe tener como solución estacionaria para la densidad de probabilidad una DBDP.

3.2. Sistema binario de bosones

En esta sección presentamos un modelo generador que es perfectamente análogo al cálculo de la degeneración de los microestados en función de la energía de un sistema cuántico de dos niveles de energía y N bosones.

3.2.1. Motivación

Una primera modificación al proceso general del que depende el Teorema del Límite Central, es la inclusión de correlaciones "débiles" entre las variables estocásticas, de manera que dejen de ser totalmente independientes.

En un trabajo previo [32] se mostró que a través de ciertos procesos multiplicativos es posible modelar la DBDP. El componente principal de dicho proceso es considerar como "uno solo" ciertos resultados en los cálculos combinatorios, sin importar de cuantas maneras se puedan reordenar sus componentes, considerándolos indistinguibles.

El papel de esta indistinguibilidad sugiere la inclusión de ciertos elementos de estadística cuántica, específicamente en las consideraciones de degeneración de bosones. En la estadística cuántica se puede considerar cierta interacción efectiva de atracción entre bosones y repulsión entre fermiones. A éstas llamadas "interacciones estadísticas" se les puede pensar responsables de las antes mencionadas correlaciones entre variables estocásticas de un proceso aditivo.

Existen además, en dos dimensiones, los llamados "anyones", que obedecen estadísticas "intermediarias" entre fermiones y bosonoes. En dichas estadísticas existe un parámetro p, que en el caso de fermiones corresponde a uno y en el de bosones a cero, y que puede tomar cualquier otro valor intermedio. Es posible mostrar que un sistema de anyones con un parámetro $p \neq 0, 1$ corresponde a un caso de bosones con una interacción cuántica-estadística adicional.

Haldane [26] definió las estadísticas "no enteras" o "fractional exclusion statistics", solo en referencia a dimensiones espaciales, dado que la teoría de anyones solo es válida en dos dimensiones. Haldane propuso la siguiente fórmula combinatoria para el número de estados (g) para N partículas idénticas en ν niveles diferentes:

$$g = \frac{[\nu + N - 1 - p(N-1)]!}{N!(\nu - 1 - p(N-1))!}. \tag{3.39}$$

De nuevo, para el caso particular de $p = 1$ y $p = 0$, la ec.3.39 reproduce las bien conocidas fórmulas para fermiones y bosones, *i.e*:

$$g = \frac{\nu!}{N!(\nu - N))!}, \tag{3.40}$$

y:

$$g = \frac{[\nu + N - 1]!}{N!(\nu - 1)!} \tag{3.41}$$

respectivamente.

Nótese que la ec.(3.39) puede reescribirse como:

$$g = \frac{[\nu' + N - 1]!}{N!(\nu' - 1)!}, \tag{3.42}$$

en donde $\nu' = \nu - p(N-1)$ es un "número efectivo de niveles", al que puede dársele cualquier valor real, incluyendo números negativos. Sin embargo, para ciertos valores de ν', tanto $(\nu' - 1)!$ como la ec.(3.42) pueden tener valores negativos.

3.2.2. El modelo

Consideremos un sistema cuántico de N bosones y ν niveles de energía. Para simplificar los cálculos partiremos a estos niveles en dos grupos con energías ϵ_1 y ϵ_2, con $\epsilon_2 < \epsilon_1$, cada uno con degeneración o sub-niveles ν_1 y ν_2, tales que $\nu_1 + \nu_2 = \nu$. Los N bosones han de ser acomodados en estos niveles, n_1 en el "nivel superior" y los otros $n_2 \equiv N - n_1$ en el nivel inferior.

En el caso particular $\nu_2 = 1$, el número de microestados con energía $E = n_1\epsilon_1 + n_2\epsilon_2$ está dado exactamente por la siguiente ecuación:

$$g(E) = g(n_1) = \frac{(n_1 + \nu_1 - 1)!}{n_1!(\nu_1 - 1)!}. \tag{3.43}$$

Para obtener la distribución del número de estado ordenados por su energía podemos utilizar la siguiente expresión:

$$r(s) = \sum_{n_1 = s}^{N} g(n_1), \tag{3.44}$$

en la cuál r es el rango de la frecuencia, s es una variable proporcional a la energía y $g(s)$ es la degeneración de microestados con una energía correspondiente a s. El resultado de la suma es de forma analítica para los casos especiales $\nu_1 = 1$ o $\nu_2 = 1$,

$$\begin{aligned} r &= \sum_{n_1=0}^{N} \frac{(n_1+\nu_1-1)!}{n_1!(\nu_1-1)!} - \sum_{n_1=0}^{s-1} \frac{(n_1+\nu_1-1)!}{n_1!(\nu_1-1)!} \\ &= \frac{(N+\nu_1)!}{N!\nu_1!} - \frac{(s-1+\nu_1)!}{(s-1)!\nu_1!}, \end{aligned}$$

en el caso de éste último y:

$$\begin{aligned} r = \sum_{n_1=s}^{N} \frac{(N-n_1+\nu_2-1)!}{(N-n_2)!(\nu_2-1)!} &= \sum_{j=0}^{N-s} \frac{(j+\nu_2-1)}{j!(\nu_2-1)} \\ &= \frac{(N-s+\nu_2)!}{(N-s)!\nu_2!} \end{aligned}$$

para el caso $\nu_1 = 1$. Tomando el límite contínuo de la variable s y reemplazando los factoriales pertinentes por funciones de Stirling, obtenemos para $\nu_2 = 1$ que

$$(N^{\nu_1} - \nu_1! r)^{\frac{1}{\nu_1}} = s, \tag{3.45}$$

y para $\nu_1 = 1$ que

$$N - (\nu_2! r)^{\frac{1}{\nu_2}} = s. \tag{3.46}$$

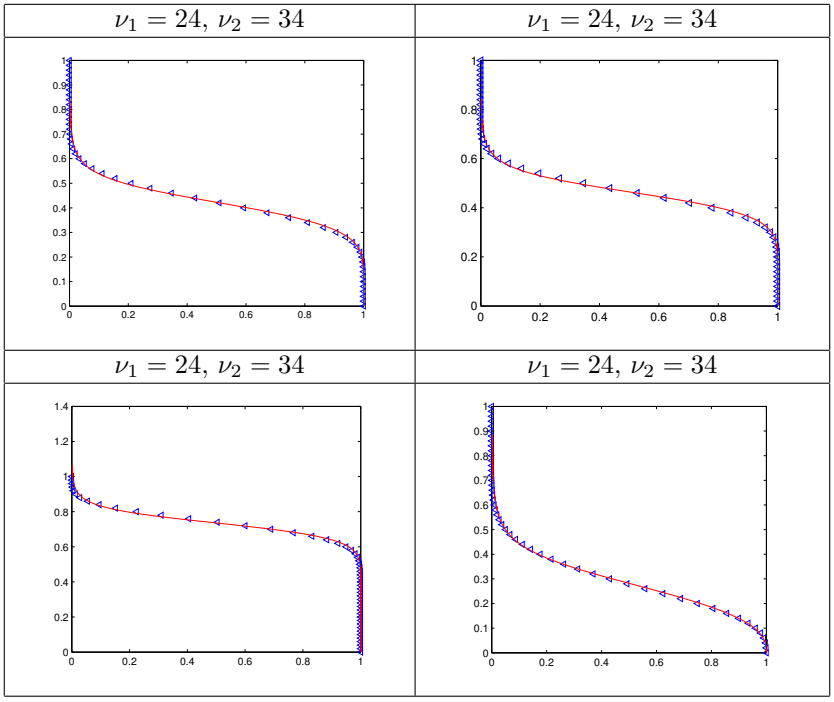
Si el sistema estuviera compuesto de partículas de Haldane entonces el número efectivo ν_i podría tomar valores arbitrarios, dando libertad numérica al valor del exponente $\frac{1}{\nu_i}$.

En un caso general, con números de ocupación n_1 y n_2 arbitrarios (con $N = n_1 + n_2$), el número de microestados está determinado por:

$$g(E) = g(n_1) = \frac{(n_1+\nu_1-1)!}{n_1!(\nu_1-1)!} \frac{(n_2+\nu_2-1)!}{n_2!(\nu_2-1)!} \tag{3.47}$$

$$= \frac{(n_1+\nu_1-1)!}{n_1!(\nu_1-1)!} \frac{(N-n_1+\nu_2-1)!}{(N-n_1)!(\nu_2-1)!} \tag{3.48}$$

Hay que notar que la forma funcional que describe la distribución rango–tamaño de los microestados en donde la variable ordenada es el parámetro s que es proporcional a la energía, en el caso $\nu_2 = 1$, es un caso particular de la DBDP; de otra manera dicha función solo tiene un parecido numérico al caso complementario $\beta = 0$ de la DBDP. El equivalente de las ecs.3.45 y 3.46 para un conjunto arbitrario de los parámetros del sistema ν_1 y ν_2 debe ser una función que interpole entre los dos casos anteriores, tanto como la DBDP es una función interpolante en las funciones potenciales $\alpha = 0$ y $\beta = 0$. Para verificar esta similitud es necesario recurrir a métodos numéricos.



Cuadro 3.1: Muestras generadas con el modelo de niveles de bosones y sus ajustes.

3.2.3. Resultados comparativos

Mediante métodos numéricos generamos múltiples muestras de la distribución rango–tamaño de los microestados de nuestro modelo, y nos dispusimos a ajustar cada una con la forma de la ec.(1.2). Los valores ν_1 y ν_2 fueron tomados de $\nu_1, \nu_2 \in [1, 60]$. Algunas muestras tomadas al azar se muestran en la figura 3.1.

El acuerdo entre el ajuste y las distribuciones generadas es claro tanto visualmente como por el coeficiente de bondad r^2: el promedio de dicho coeficiente es $\hat{r^2} = .998$ para las $60 \times 60 = 3600$ muestras tomadas, con una desviación estándar $\sqrt{\sigma^2} = 0.0007$. La Fig.3.4 muetra el coeficiente r^2 como función de los parámetros ν_1 y ν_2. El valor alto de dicho parámetro cerca de los ejes es testimonio del hecho de que la distribución obtenida de valores cualesquiera de ν_1 y ν_2 es una interpolación de los casos particulares $\nu_1 = 1$ y $\nu_2 = 1$, en donde se mostró analiticamente que el ajuste es apropiado.

La dependencia de los parámetros $\alpha(\nu_1, \nu_2)$ y $\beta(\nu_1, \nu_2)$ de la ec.1.2 se muestra en las Figs.3.5 y 3.6.

La gráfica de los parámetros $\alpha(\nu_1, \nu_2)$ y $\beta(\nu_1, \nu_2)$ está de acuerdo con el hecho que cuando $\nu_2 = 1$, el valor correspondiente de α es cero.

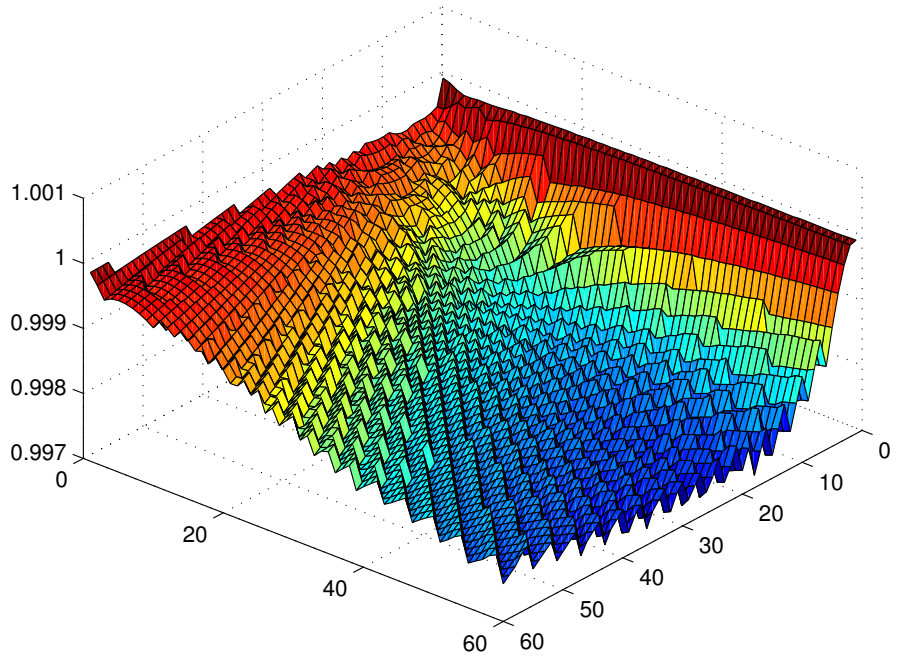


Figura 3.4: Parámetro r^2 del ajuste del modelo para pares ν_1, ν_2.

El modelo propuesto genera una distribución rango-tamaño que asemeja numericamente de manera extraordinaria a la DBDP, en todo su espacio de configuración, y en algunos caso se ha mostrado tiene su forma exacta. El modelo se ha tenido de manera sencilla, en su forma de solo dos grupos de niveles, pero como veremos en la sección 3.4, dividir los "factores" responsables en categorías "favorable" y "desfavorable" podría también ser un primer paso para separarnos sutilmente de las condiciones del teorema del límite central, para conservar su generalidad pero cambiar totalmente el tipo de atractor que prescribe.

3.3. Deposición secuencial polidispersa aleatoria

Justo como algunos modelos sencillos reproducen bien el comportamiento de leyes de potencia en secuencias simbólicas, hemos descubierto que cierto modelo de deposición balística produce ditribuciones que son bien ajustadas por una DBDP. Este modelo es una realización numérica del problema de deposición secuencial polidispersa aleatoria (o RSA por sus siglas en inglés), en el

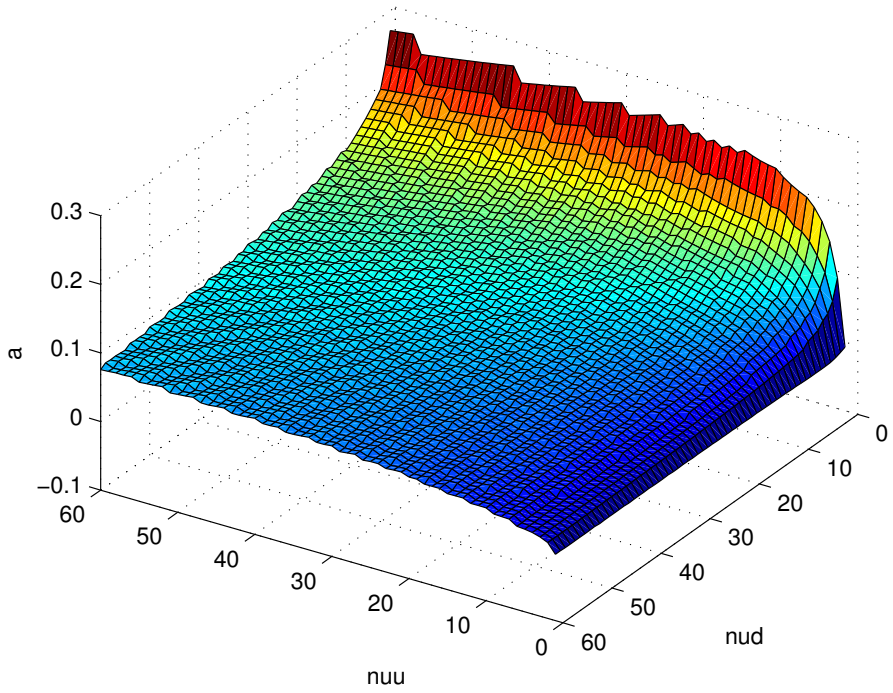


Figura 3.5: $\alpha(\nu_1, \nu_2)$

cuál cuerpos inhomogéneos son seleccionados al azar, siguiendo una receta probabilística, y fijados irreversible y aleatoriamente a una superficie, evitando la superposición. La RSA se ha tratado extensamente [40] y se ha aplicado en una gran variedad de campos [41].

3.3.1. Algoritmo

Nuestro algoritmo particular es:

1. Consideramos un área unitaria cuadrada.
2. Con probabilidad uniforme escogemos un punto en dicha superficie, este punto será el centro de un círculo.
3. Generamos el radio del círculo de la distribución de probabilidad

$$\rho(r) = 2(\gamma + 1)(2r)^{\gamma}, \tag{3.49}$$

con un corte en:
$r > 1/2$.

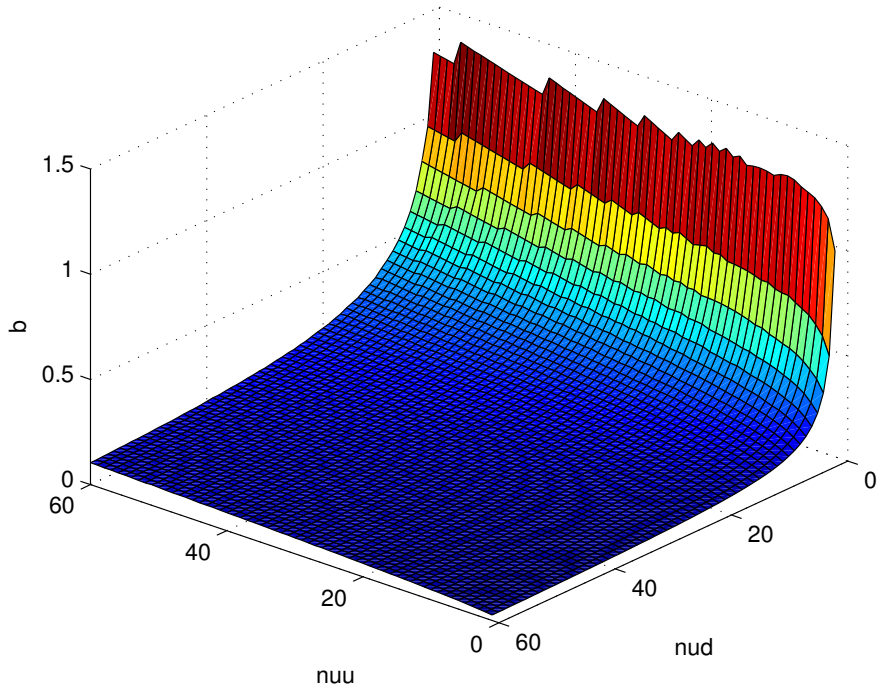


Figura 3.6: $\beta(\nu_1, \nu_2)$

4. Si el círculo generado resulta estár superpuesto sobre algún otro, o sobre el borde de la superficie, entonces se anula y se repite el proceso. De otra manera se queda sin posibilidad de reacomodo o cambio de tamaño.

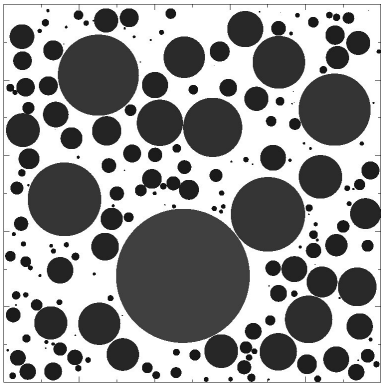
5. Se repite el proceso hasta alcanzar un número predeterminado de inserciones exitosas.

6. Se obtiene la distribución rango-tamaño de los radios de los círculos exitosamente generados.

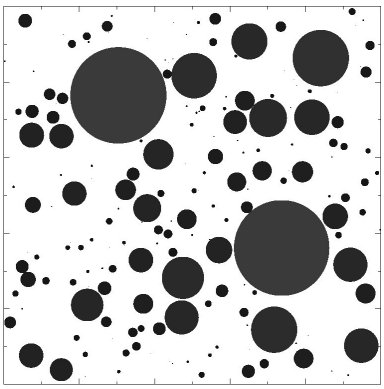
7. Se repite todo el ciclo con diferentes parámetros. Los valores del número de inserciones exitosas se tomaron del intervalo $[25, 2000]$, y el valor del exponente γ de la distribución de probabilidad de los radios se acotó con el intervalo $(-1, 4]$.

3.3.2. Resultados numéricos

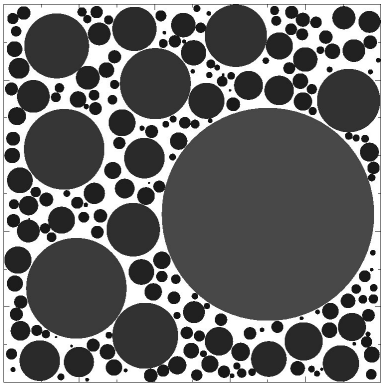
Además del antecedente de los motivos pictóricos, recurrimos a este modelo tomando en cuenta el siguiente argumento intuitivo: Los primeros círculos



$\gamma = 0$



$\gamma < 0$



$\gamma > 0$

Figura 3.7: Resultados típicos para realizaciones con distintos valores del parámetro γ.

añadidos serán poco suceptibles a ser rechazados, dado que el área de adsorción está vacía. Entonces tendremos, especialmente para $\gamma > 0$, que favorece radios grandes, un comportamiento de potencia para los radios grandes. Una vez que se haya logrado cierta saturación, las subsiguientes inserciones tendrán que ser invariantes de escala, puesto que el efecto de la frontera del área original se diluirá ante la creciente dificultad para encontrar espacio entre los obstáculos ya dispuestos. Entonces también la cola de la distribución tiene que tener forma de potencia, justo como la DBDP. Cabe agregar que este modelo solo ilustra el tipo de restricciones que llevan a distribuciones tipo DBDP, y que no hay intención alguna en imitar aspectos no estructurales propios de las muestras en arte gráfico que analizamos, así como los modelos de lenguajes aleatorios [14] no imitan características gramaticales o estéticas, aunque siguen la ley de Zipf.

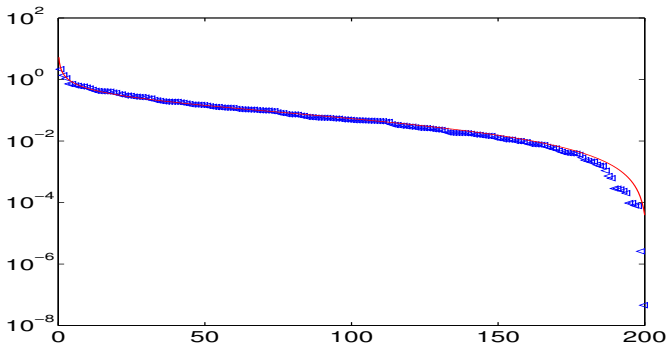
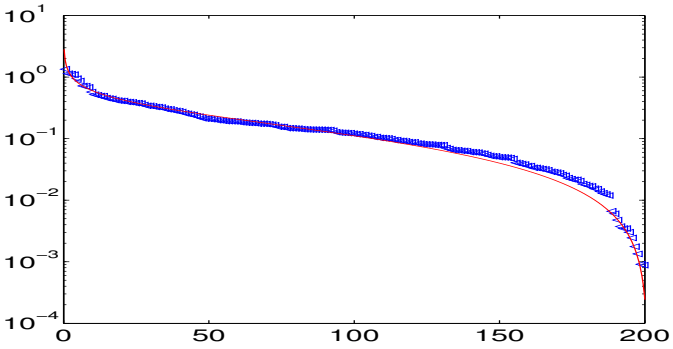
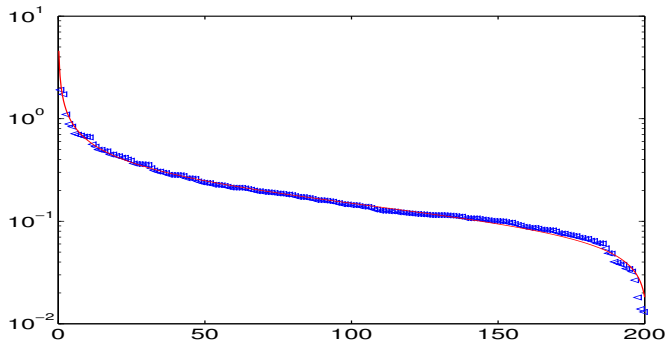
El siguiente paso fue corroborar la hipótesis de que las distribuciones rango-tamaño de los radios tendrían una buena representación de DBDP, así que se prosiguió por hacer ajustes de la colección de distribuciones. Para cada una de ellas se hicieron dos ajustes, uno en la totalidad de la distribución, y otro eliminando el primer 10 % de los radios añadidos. Este segundo lote de ajustes se hizo para ver el comportamiento de una hipotética distribución sin efecto de bordes.
A través de las gráficas de la tabla 3.2 se concluyó que las aproximaciones numéricas de la DBDP eran adecuadas. Hay que notar que existe cierta subestructura en la región de rangos altos; esta subestructura se debe a que las muestras realizadas no corresponden al límite de número infinito de círculos. Variando el número de estos, se encontró que una vez que la muestra se satura por encima de un cierto valor del diámetro d', es decir, es imposible introducir un círculo cuyo diámetro cumpla $d > d'$, la subestructura mencionada aparece solamente para los rangos r tales que $d(r) < d'$.

3.3.3. Resultados analíticos

En vista del éxito de las simulaciones numéricas, nos aventuramos a encontrar resultados analíticos que asociaran de manera directa la DBDP a ciertos procesos de deposición irreversible. En vista de las complicaciones que requieren los cálculos sobre las realizaciones numéricas en dos dimensiones, empezamos nuestras indagaciones analíticas en el mismo modelo pero en una sola dimensión. La meta es obtener alguna expresión cerrada para la distribución de probabilidad de una longitud en particular, después de n inserciones. A partir de ésta se podría determinar fácilmente la distribución rango–tamaño y comparar las expresiones con la DBDP.

Aún mediante la reducción a una dimensión y hasta la fecha, solo hemos logrado algunos resultados preliminares:

Supongamos que tenemos originalmente un intervalo de longitud c, y que

Exp.	Ejemplo de ajuste: Distribución completa.
$\gamma < 0$	
$\gamma = 0$	
$\gamma < 0$	

Cuadro 3.2: Muestras y ajustes de las distribuciones rango–tamaño generadas a través del modelo de Deposición Aleatoria.

generamos un segmento de longitud l, mediante la distribución de probabilidad

$$\rho(l) = \frac{(\gamma+1)}{c^{\gamma+1}} l^{\gamma}, \tag{3.50}$$

con $\gamma > -1$ y $l \in [0, c]$. Ahora deseamos colocar este segmento dentro del intervalo inicial escogiendo al azar la posición de su centro dentro del intervalo. Si el segmento no cabe se repite todo el procedimiento. Si la inserción es exitosa el segmento se queda y ahora repetimos el procedimiento en las porciones laterales libres que quedan del intervalo inicial. (La probabilidad de que el segmento se haya insertado de manera que no quede espacio de uno, o inclusive ambos lados, es identicamente cero.), hasta insertar un número predeterminado de "piezas".

Dado este algoritmo tenemos los siguientes resultados:

- La probabilidad de que en el intervalo $[\frac{-c}{2}, \frac{c}{2}]$, se inserte un fragmento con centro en x.
  Dicha probabilidad es igual integral de la probabilidad de las longitudes de todos los intervalos que puedan "caber" allí, es entonces:

  $$\rho(x) = N(L/2 - |x|)^{\gamma+1}), \tag{3.51}$$

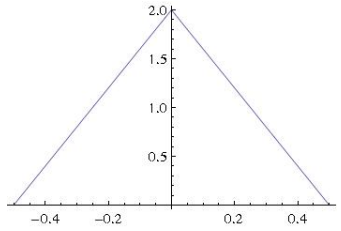
  en donde N es el factor de normalización. Ejemplos de ésta función para diferentes valores del exponente de la distribución de probabilidad de los intervalos se muestran en la Fig.3.8.

- Probabilidad de inserción de un segmento de longitud l.
  Dado un intervalo de longitud c, la probabilidad de que la siguiente inserción mida l está dada por la probabilidad conjunta de escoger dicha longitud (eq.3.50) y de escoger un centro apropiado. Como se trata de eventos independientes, la distribución de probabilidad es:

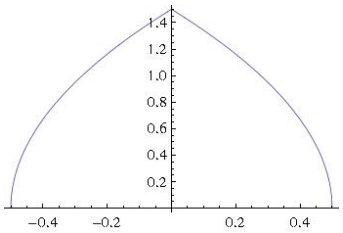
  $$\rho(l|c) = \frac{l^{\gamma}(1 - \frac{l}{c})}{c^{\gamma+2} B(\gamma+1, 2)}. \tag{3.52}$$

  Para normalizar dicha distribución hemos recurrido a la función Beta completa, $B(a, b)$ ec.(1.5). En la fig. 3.9 mostramos los tres diferentes regímenes según el signo o nulidad del exponente γ.

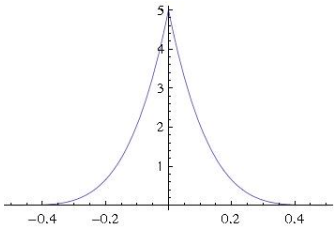
- Probabilidad de insertar una pieza de tamaño l, en un segmento de longitud c dividido en n espacios arbitrarios.
  Nuestro resultado más avanzado hasta ahora describe la situación del algoritmo en la cuál, después haber hecho un número determinado de inserciones, el espacio original se encuentra dividido en n intervalos separados por piezas ya dispuestas. Cabe notar que esta distribución de probabilidad funciona cuando las longitudes de los segmentos libres que quedan se toman al azar, y que en realidad, dichos segmentos tienden a tener longitudes similares, pues es siempre más probable un intento exitoso de inserción en el mayor de los intervalos libres. La probabilidad se obtiene después de



$\gamma = 0$

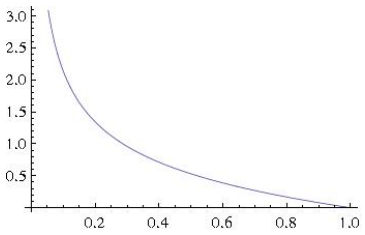


$\gamma < 0$

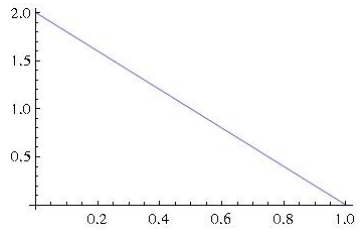


$\gamma > 0$

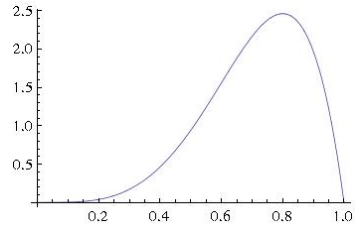
Figura 3.8: Distribución de probabilidad $\rho(x) = N(L/2-|x|)^{\gamma+1}$ del centro de una inserción para distintos valores del exponente γ ($L = 1$).



$\gamma < 0$

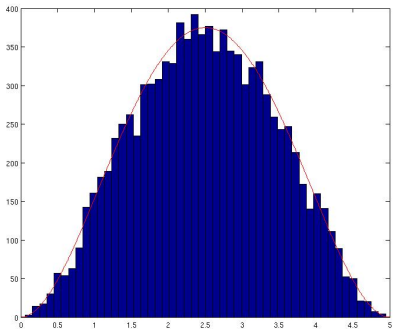


$\gamma = 0$

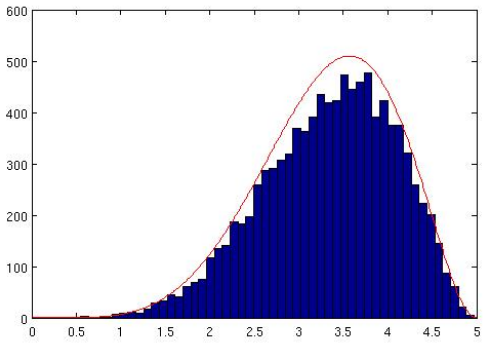


$\gamma > 0$

Figura 3.9: Distribución de probabilidad $\rho(l|c) = \frac{l^{\gamma}(1-\frac{l}{c})}{c^{\gamma+2}B(\gamma+1,2)}$ de la longitud de un segmento insertado exitosamente en un intervalo de longitud c $(c = 1)$.



$\gamma = 2$,$n = 2$



$\gamma = 7$,$n = 2$

Figura 3.10: Histograma de la corroboración numérica de la distribución de probabilidad de la longitud de un segmento insertado exitosamente en un intervalo de longitud c con n separaciones arbitrarias, ec. (3.53).

hacer una integral sobre el espacio de configuraciones posibles de las divisiones del intervalo. La distribución de probabilidad tiene, todavía como las demás, una forma relativamente sencilla:

$$\rho_n(l|c) = \frac{l^\gamma (1 - \frac{l}{c})^n}{c^{\gamma+n+1} B(\gamma + 1, n + 1)}. \tag{3.53}$$

Aunque los resultados analíticos distan de ser contundentes para el propósito que fueron dilucidados, es una señal fortuita que todos los resultados parciales que obtuvimos, en la forma de distribuciones, son de hecho del tipo DBDP. Para futuras excursiones en este modelo hay que tomar en cuenta que, por ejemplo, si no exigimios una cota mínima a la distribución de las longitudes

ec.(3.50), es posible que la distribución rango-tamaño de las longitudes de un proceso tomado *ad infinitum*, no tenga un soporte acotado, como es el caso de la DBDP.

3.4. Resta de variables estocásticas

El modelo de resta de variables estocásticas es posiblemente el más general y aplicable de todos los que hemos mencionado aquí. A través de éste mostramos que las variables que dependen de la resta sistemática de variables estocásticas independientes, y que siguen además cierta restricción de positividad, tienen por densidades de probabilidad a funciones que son bien representadas numericamente por la DBDP, aunque se conocen resultados analíticos, algunos de los cuales no coinciden con la forma exacta de la DBDP. Como el modelo se trata esencialmente de una variación al Teorema del Límite Central, es útil presentarlos en paralelo.

3.4.1. Suma y resta de variables estocásticas

Si la variable Y depende de manera aditiva de X_1 y X_2, variables independientes, caracterizadas por las distribuciones de probabilidad P_1 y P_2 respectivamente, entonces la función de probabilidad de $Y \equiv X_1 + X_2$ se obtiene de manera inmediata mediante la convolución de P_1 y P_2:

$$P(Y) = (P_2 \oplus P_1)(Y) \equiv N \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbf{d}\xi_1 \mathbf{d}\xi_2 P_1(\xi_1) P_2(\xi_2) \delta(\xi_2 + \xi_1 - Y), \quad (3.54)$$

en donde N es un factor de normalización. Dado que la convolución es un producto asociativo y conmutativo, la función de distribución de probabilidad de la variable $Y \equiv \sum_1^M X_i$ definida como la suma de un número M de variables estocásticas independientes $\{Xi\}_\beta$, está dada por la convolución múltiple:

$$P(Y) = P_M \oplus P_{M-1} \oplus \cdots \oplus P_2 \oplus P_1. \quad (3.55)$$

Se puede mostrar fácilmente, como además está enunciado en el Teorema del Límite Central, que bajo una renormalización adecuada de la variable Y, y ciertas condiciones sobre los cumulantes de las variables X_i, el límite $M \to \infty$ procura la forma de una función gaussiana para la distribución $P(Y)$ ([1], sección **2.3**).

Se suele recurrir a la imagen intuitiva de un caminante aleatorio que realiza M saltos cuyas longitudes fueron caracterizadas probabilísticamente por las variables X_i.

3.4.2. Correlación de integrandos positivos

Si estamos interesados, por el contrario, en una variable que dependa de la resta de dos variables estocásticas independientes, podemos seguir esencialmente el mismo formalizmo enunciado en el Teorema del Límite Central, reemplazando simplemente la operación de convolución por la correlación entre dos funciones: Sean P_1 y P_2 distribuciones de probabilidad "bien comportadas" en el intervalo $[0, 1]$. La correlación cruzada de ellas se define como:

$$P_2 \star P_1(y) \quad \equiv \quad N \int_0^1 \int_0^1 \mathbf{d}\xi_1 \mathbf{d}\xi_2 P_2(\xi_2) P_1(\xi_1) \delta(\xi_2 - \xi_1 - y),$$

en donde N se introdujo, como de costumbre, como un factor de normalización.

Las hipótesis del Teorema del Límite Central pueden todavía ser válidas aunque estemos considerando la resta de variables estocásticas. Es evidente que podemos definir nuevas densidades de probabilidad de manera que el resultado de restar las originales sea equivalente a sumar nuestras nuevas variables, forzando a la distribución resultante a ser de tipo gaussiano o de Levy. Añadimos entonces otra restricción: en cada operación de correlación se han de evitar los resultados negativos de la substracción, introduciendo así una "barrera absorbente". Si hemos de considerar solamente el dominio positivo de esta nueva función, podemos reescribirla de la siguiente manera:

$$P_2 \star P_1|_{y>0}(y) = P_2 \ominus P_1(y) = N' \int_y^1 \mathbf{d}\xi P_2(\xi) P_1(\xi - y), \tag{3.56}$$

en cuyo caso el factor de normalización tendría que ser cambiado también:

$$(N')^{-1} = \int_0^1 \mathbf{d}y \left(\int_y^1 \mathbf{d}\xi P_2(\xi) P_1(\xi - y) \right). \tag{3.57}$$

Al tomar $y > 0$ y escoger la normalización adecuada, hemos logrado que la distribución de probabilidad que resulta de este producto pertenezca a la misma clase de funciones de donde hemos tomado sus factores P_1 y P_2, *i.e.*, el producto definido de esta manera es cerrado.

Una vez habiendo definido nuestro nuevo producto a traves de la eq. (3.56) y notando que no es ni conmutativo ni asociativo, nos podemos preguntar si las sucesiones "izquierdas" de la forma:

$$P = P_M \ominus (P_{M-1} \ominus \cdots \ominus (P_2 \ominus P_1)) \cdots), \tag{3.58}$$

o las "derechas":

$$P = (((P_1 \ominus P_2) \ominus P_3) \ominus \cdots \ominus P_{M-1}) \ominus P_M, \tag{3.59}$$

tienen un atractor funcional como aquellas en la Eq.3.55, cuales son sus propiedades de estabilidad y, en particular, si la DBDP es una buena representación numérica de dichas funciones.

Como un detalle final a la presentación de nuestro problema, hemos de decir que existe también para las series recién definidas, como en el caso del Teorema del Límite Central, un análogo de caminante aleatorio. Las series derechas de la forma de la ec. (3.59) tienen un análogo de caminante aleatorio sencillo. Intuitivamente corresponden a una población de caminantes concentrados en el intervalo $(0, 1)$ y que tienen la capacidad de realizar saltos con longitudes

negativas, con una barrera absorbente en el origen. El límite de iterar dicho sistema *ad infinitum*, si en cada paso renormalizamos la población de caminantes, es una distribución de Dirac centrada en el origen, y es por eso que en lo siguiente no necesitaremos tratar series derechas. Las series izquierdas tienen un comportamiento mucho más complicado inclusive que aquellas que se utilizan para ilustrar difusión estándar y el Teorema del Límite Central. El análogo de caminata aleatoria de las series izquierdas se puede explicar como un algoritmo:

- Consideramos una población de caminantes distribuidos de acuerdo a la densidad P_0 en el intervalo $(0, 1)$.
- Los caminantes realizan un salto de longitud $-1 < l < 0$, con probabilidad $P_1(l)$.
- Se eliminan del sistema todos aquellos que después de haber realizado el salto sigan teniendo una posición positiva.
- La nueva densidad de caminantes P' es entonces una función con soporte en $(-1, 0)$.
- Invertimos la densidad de probabilidad P' para crear la función $P(\xi) = P'(-\xi)$.
- La función P, con soporte en $(0, 1)$, es el resultado de la primera iteración y se somete al proceso completo una vez más tomando el lugar de P_0 en el primer paso.

El algoritmo no es tan transparente como aquél de las caminatas brownianas, pero aún así puede asistir intuitivamente al entendimiento del modelo.

3.4.3. Estabilidad de la DBDP bajo correlación

Nuestro primer paso fue preguntarnos si la DBDP es estable bajo nuestro nuevo producto de correlación. Antes de empezar es necesario introducir la notación: $\{a, b\}_B$, para referirnos económicamente a una DBDP con dichos parámetros, *i.e.*:

$$f(x) = \{a, b\}_B \equiv N\frac{(d-x)^b}{x^a}, . \tag{3.60}$$

El problema arriba enunciado se reduce entonces a saber si existen α y β parámetros tales que la ecuación:

$$\{\alpha, \beta\}_B \approx \{a_1, b_1\}_B \ominus \{a_2, b_2\}_B \tag{3.61}$$

se puede resolver a nivel numérico, para a_1, b_1, a_2, b_2 dados. En lo siguiente utilizaremos los parámetros griegos α_i y β_i para referirnos a aquellos obtenidos mediante ajustes, y latinos a_i, b_i para denotar los parámetros de las DBDP usadas en los cálculos de los productos con $\ominus$.

Empezamos la exploración numérica escogiendo un conjunto sensato de dominios para los parámetros a_1, b_1, a_2 y b_2. Para garantizar la normalización, todos los parámetros a_i deben cumplir $a_i < 1$, y dado que estamos interesados primordialmente en funciones monótonas decrecientes, porque nos interesa representar distribuciones ordenadas, podemos restringir aún más los parámetros con $0 \leq a_i$ y $0 \leq b_i$. Finalmente, dado que necesitamos un rango finito para los parámetros, tomamos $b_i < 3$ de manera arbitraria, y como $\{a, b\}_B$ no tiene puntos críticos para a fija y $b > 3$, podemos suponer que no perdemos, al hacer esto, ninguna particularidad del modelo.

Una vez fijados los intervalos $0 \leq a < 1$ y $0 \leq b < 3$, los dividimos en 10 y 23 partes iguales, respectivamente, para obtener un conjunto discreto de valores para utilizar durante el cómputo:

$$\begin{aligned} a &\in \{0, 0.09, 0.18, \ldots, 0.81\} \\ b &\in \{0, \frac{3}{12}, 2\frac{3}{12}, \ldots, 3 - \frac{3}{12}\}. \end{aligned}$$

Evaluamos numéricamente todas las combinaciones posibles de $\{a_1, b_1\}_B \ominus \{a_2, b_2\}_B$ utilizando los conjuntos anteriores, para lograr un total de $10 \times 10 \times 12 \times 12 = 14,400$ funciones, cada una de las cuales fue ajustada por una DBDP, para confirmar la validez de la ec. (3.61).

Primero inspeccionamos visualmente ajustes escogidos al azar, dado el gran número de ellos, y solo entonces, habiendo confirmado que un valor alto del parámetro de bondad r^2 estaba en general asociado a un buen ajuste, se procedió a hacer estudios estadísticos tomando en cuenta todas las muestras. En la Fig.3.11 mostramos varios ajustes de la DBDP tomados al azar.

Para la totalidad de los 14,400 ajustes evaluados, obtuvimos para r^2 un promedio de $\overline{r^2} = 0.9984$, con desviación $\sigma = 0.0027$. De todas las muestras, solo el 1.03 % obtuvo un parámetro en el intervalo $r^2 < 0.99$. En la Fig.3.12 mostramos el histograma del valor de r^2 tomando solo este último 1.03 %.

La inspección de ese conjunto reveló que todos sus elementos eran de la forma $\{a_1, b_1\}_B \ominus \{a_2, 0\}_B$, con $a_2 \leq 0.18$, y de éstos la mitad eran además de la forma $\{a_1, b_1\}_B \ominus \{0, 0\}_B$.

3.4.4. Límite por iteración

Puesto que no podemos, salvo para un conjunto muy particular de casos, resolver el límite $M \to \infty$ para las series con la forma de la ec. (3.58), recurrimos a la evaluación numérica de las series izquierdas de la forma:

$$P = \overbrace{K \ominus (K \ominus \cdots \ominus (K \ominus P_0)) \ldots)}^{M}, \tag{3.62}$$

para obtener, después de M iteraciones, y para cada elección arbitraria de $K(y)$ y una "condición inicial" P_0, una representación numérica del atractor P.

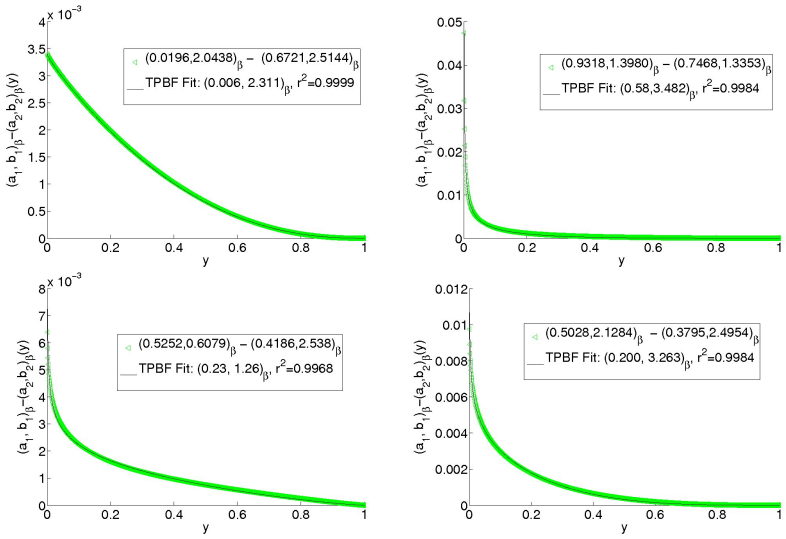


Figura 3.11: Cuatro muestras tomadas al azar. De izquierda a derecha, y de arriba hacia abajo: $(0.9318, 1.3980)_B \ominus (0.7468, 1.3353)_B$, $(0.5252, 0.6079)_B \ominus (0.4186, 2.538)_B$, $(0.0196, 2.0438)_B \ominus (0.6721, 2.5144)_B$, $(0.5028, 2.1284)_B \ominus (0.3795, 2.4954)_B$

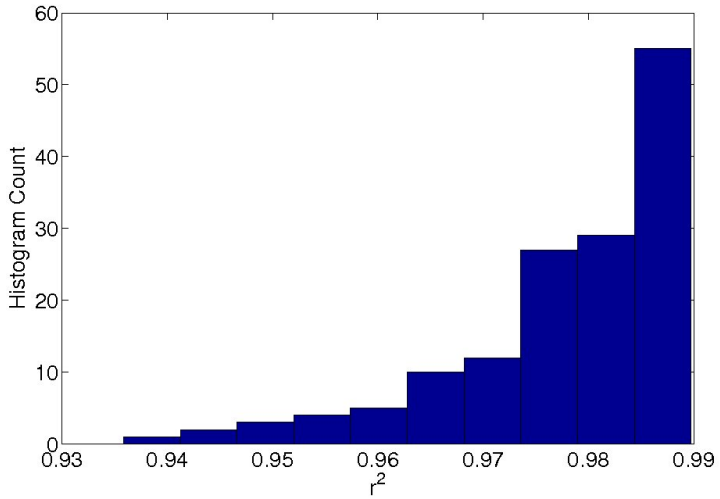


Figura 3.12: Histograma de los valores del parámetro r^2 para el más bajo 1.03 % del conjunto de los ajustes de las evaluaciones numéricas de $\{a_1, b_1\}_B \ominus \{a_2, b_2\}_B$.

$K(x)$	α	β	r^2
$e^{-2\sqrt{x}}$	0.064	1.033	0.997
e^{-2x}	0.041	1.317	0.997
$\frac{e^{-x}}{\sqrt{x}}$	0.167	1.180	0.993
$\sqrt{x(1-x)}$	-0.148	1.184	0.990
$x(1-x)$	-0.219	1.456	0.984
$(x(1-x))^2$	-0.337	1.879	0.975
$(x(1-x))^{-0.75}$	0.056	0.246	0.997
$1-0.2x$	-0.050	0.873	0.997
$1-0.6x$	-0.017	1.010	0.999
$1-0.9\sqrt{x}$	-0.009	1.389	0.999
1	-0.066	0.848	0.995
$(1-0.25\sin(15x))(1-x)$	-0.053	1.754	0.995
$(1-0.5\sin(25x))(1-x)$	-0.046	1.766	0.992
$\frac{(1-0.25\sin(25x))}{\sqrt{x}}$	-0.053	1.754	0.995
$\frac{(1-0.5\sin(25x))}{\sqrt{x}}$	0.086	0.840	0.996
$e^{-0.5(x-0.5)^2}$	-0.077	0.891	0.994

Cuadro 3.3: Parámetros de ajuste y bondad para el atractor de las series izquierdas tipo (3.58) para diferentes funciones K.

Para darnos una primera idea del tipo de atractores que tiene la ec. (3.62) tiene, empezamos por escoger $K(y)$ de la forma $\{a,b\}_B$ y tomamos la "condición inicial" P_0 como la distribución uniforme sobre el intervalo $[0,1]$.

Será demostrado más tarde que el atractor es único, y que entonces la elección de condición inicial es irrelevante para el problema de encontrar los atractores. Para asegurar la normalización tomamos valores del parámetro a, de las funciones $K(y) = \{a,b\}_B$, de el conjunto $(-\infty, 1)$.

La elección de K como una función tipo $K(y) = \{a,b\}_B$ es justificable aunque parezca arbitraria: dado que se demostró que las DBDP son estables bajo el proceso, escoger funciones K no corrompe la generalidad del estudio. De cualquier manera se hicieron pruebas para funciones K de muchos tipos. La tabla 3.3 muestra los resultados de ajustar una DBDP en el límite númerico $M \to \infty$ de series del tipo (3.62) para diferentes funciones K y $P_0 = K$. La lista no es exhaustiva ciertamente, pero contiene muchos tipos diferentes de comportamientos funcionales; funciones periódicas, exponenciales, potencias, funciones racionales y otras, para mostrar la versatilidad del atractor DBDP.

3.4.5. Resultados de la iteración

El esquema básico de exploración es el siguiente:

- Se escoge una función del conjunto de nuestras funciones de prueba $\{a, b\}_B$, *i.e.*,
$$K(y) = N\frac{(d-x)^b}{x^a}$$
- Una secuencia de funciones $\{P_i\}$ se genera mediante iteración, en donde:
$$\begin{aligned} P_1 &= K \ominus P_0 \equiv N' \int_y^1 \mathbf{d}\xi K(\xi) P_0(\xi - y) \\ P_2 &= K \ominus P_1 \\ &\vdots \end{aligned}$$
- Para cada una de estas funciones P_i, se realiza un ajuste con la DBDP, obteniendo así un par de valores α_i, β_i del ajuste.
- Si en efecto la DBDP resulta ser una buena representación para las P_i, entonces se traza la trayectoria de los valores α_i, β_i obtenidos de los ajustes, en el espacio $\alpha \times \beta$.

La Fig.3.13 muestra un ejemplo del conjunto de distribuciones P_i que resultan de varios procesos de iteración con una función $K(y)$ de la forma $\{a, 0\}_B$. También muestra el ajuste con la DBDP que se hizo sobre la función que resultó de iterar veinte veces (P_{20}). La velocidad de convergencia mostrada en este ejemplo es característica del resto de las funciones muestreadas. Todas las funciones se superponen prácticamente después de 4 iteraciones. La trayectoria en el espacio (α, β) del mismo conjunto de funciones se muestra en la Fig.3.14.

Primero búscamos los atractores de funciones K del tipo $K = \{a, 0\}_B$, con $0 \leq a < 1$. Usamos 40 funciones $K = \{a, 0\}_B$ con:

$$a \in \{0, 1/40, 2/40, \ldots 39/40)\}, \tag{3.63}$$

cada una de las cuáles se iteró 20 veces. las trayectorias en espacio $\alpha \times \beta$ resultantes se muestran en la Fig.3.15.

El valor promedio de r^2 para estos ajustes es $\overline{r^2} = 0.99$. El mismo procedimiento se repitió para funciones $K(y)$ de la forma $\{0, b\}_B$ y $\{a, a\}_B$, de nuevo considerando $0 \leq a < 1$ y $0 \leq b < 3$, en particular se usaron los valores:

$$\begin{aligned} a &\in \{0, \frac{1}{40}, \frac{2}{40}, \cdots, \frac{39}{40}\}, \\ b &\in \{0, 3\frac{1}{40}, 3\frac{2}{40}, \cdots, 3\frac{39}{40}\}. \end{aligned}$$

Los parámetros de los ajustes para estos iterados se muestran en las Figs.3.16 y 3.17.

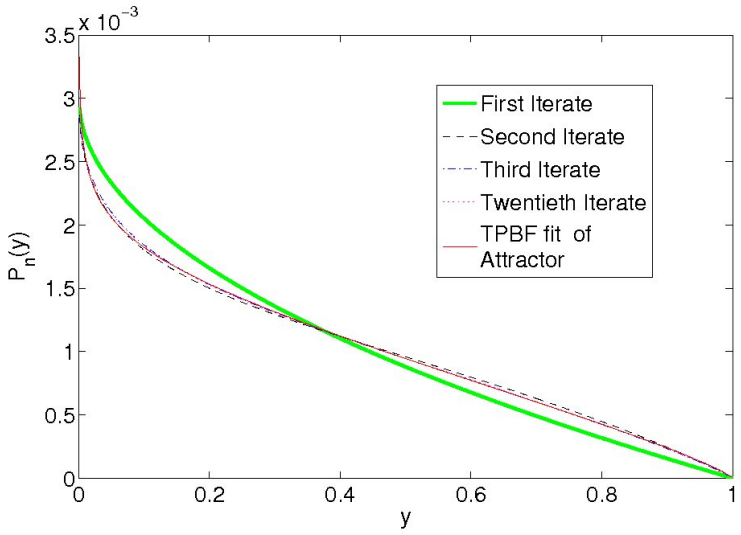


Figura 3.13: Distribuciones obtenidas después de una, dos, tres y veinte iteraciones y el ajuste con DBDP de esta última.

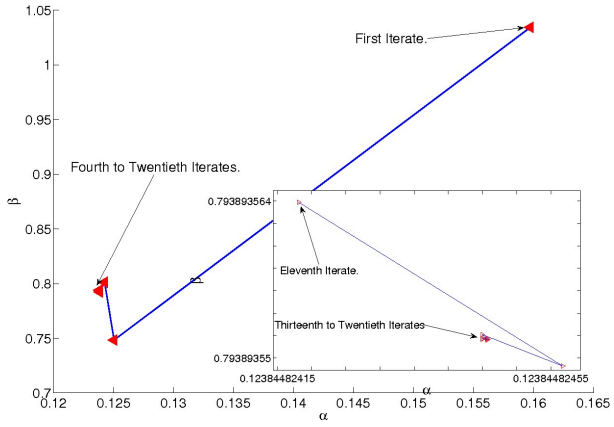


Figura 3.14: Trayectoria en el espacio (α, β) correspondiente a las primeras veinte iteraciones de $K = \{\frac{1}{2}, 0\}_B$. El *Inset* es un acercamiento que muestra la trayectoria después del onceavo iterado. Nótese que bastan unas cuantas iteraciones para acercar todos los parámetros subsecuentes a un rango menor a la incertidumbre ≈ 0.001.

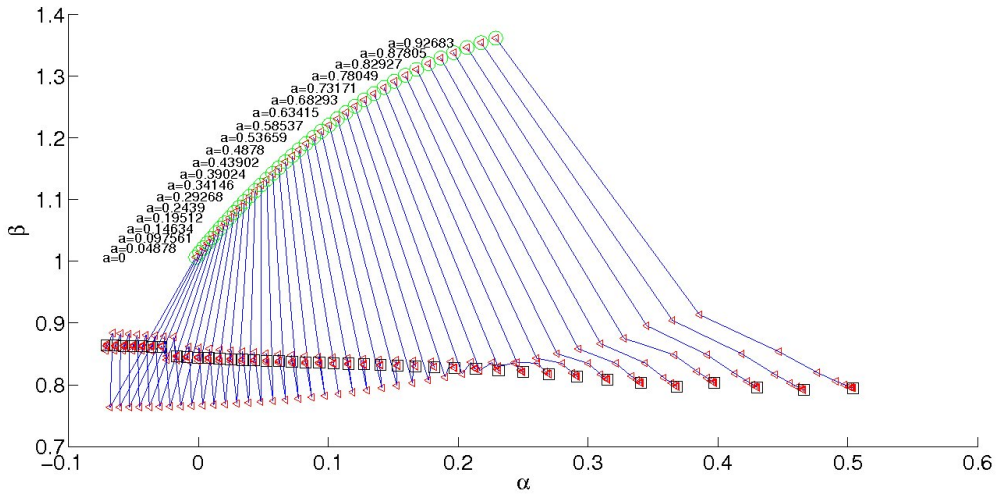


Figura 3.15: Conjunto de las trayectorias que consisten de veinte iteraciones de $\{a, 0\}_B$. *Círculos* y *cuadrados* indican los parámetros del primer y del último iterado, respectivamente. El valor de a se muestra solo sobre las trayectorias nones.

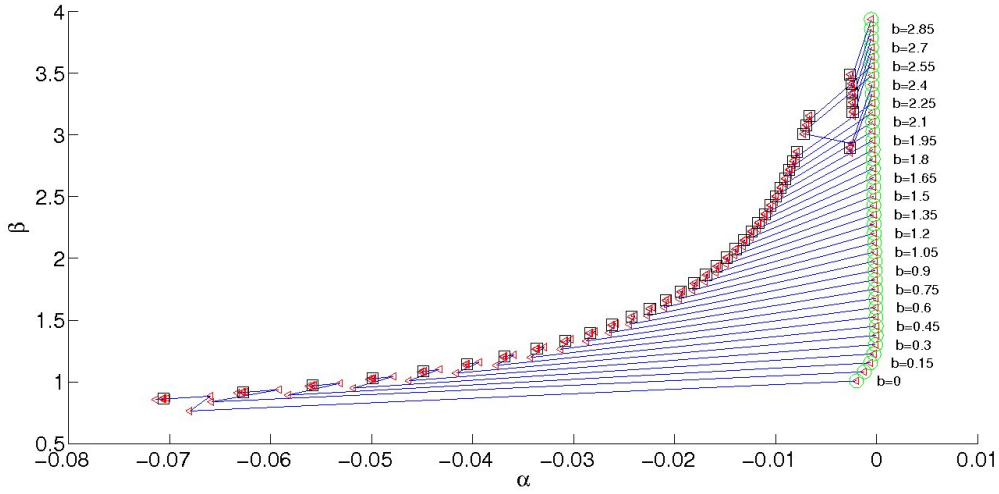


Figura 3.16: Conjunto de las trayectorias que consisten de veinte iteraciones de $\{0, b\}_B$. *Círculos* y *cuadrados* indican los parámetros del primer y del último iterado, respectivamente. El valor de b se muestra solo sobre las trayectorias nones.

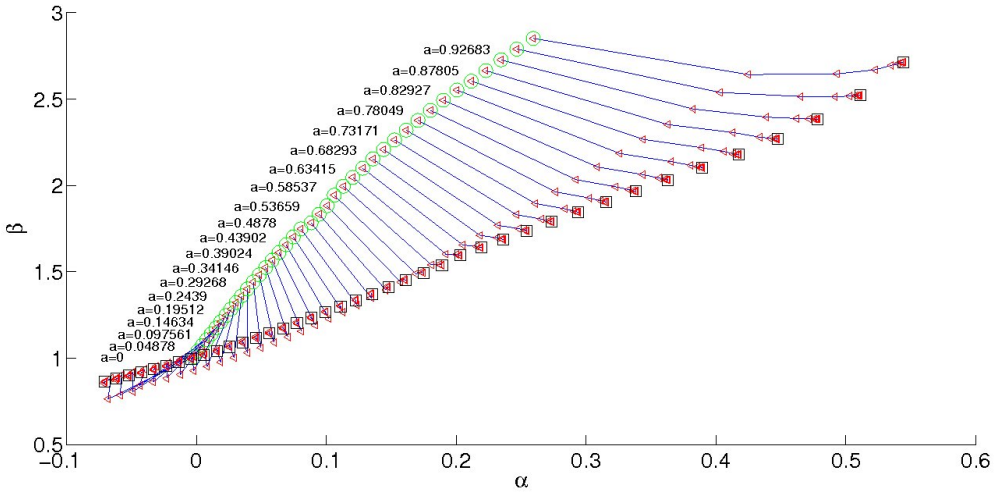


Figura 3.17: Conjunto de las trayectorias que consisten de veinte iteraciones de $\{a, a\}_B$. *Círculos* y *cuadrados* indican los parámetros del primer y del último iterado, respectivamente. El valor de a se muestra solo sobre las trayectorias nones.

La forma de los atractores que corresponden a algunas funciones $K = \{0, b\}_B$ y sus ajustes con la DBDP se muestran en la Fig.3.18. Vemos que para esta forma en particular de K, los atractores tienen ajustes cuya precisión mejora cuando el parámetro b aumenta, puesto que para pequeños valores de b aparece en el extremo izquierdo del ajuste una pendiente "espúrea", como lo atestiguan los valores negativos de α en la Fig. 3.16, que hacen que la DBDP sea no-monótona. Para valores más grandes del parámetro b en las funciones K, el atractor cambia de su forma convexa a una cóncava, y entonces el ajuste ya no presenta esta curvatura y se vuelve una mejor representación.

3.4.6. Aproximación por polinomios

Para el problema particular de encontrar P_2:

$$P_3 = P_2 \ominus P_1, \tag{3.64}$$

dados P_3 y P_1, podemos recurrir a una expansión polinomial de orden n de la función P_2 propuesta por el Dr. Germinal Cocho:

$$P_1(x) \approx \mathbf{P}_{1,n}(x) \equiv \sum_{i=0}^{n} \frac{a_i}{i!} x^i, \tag{3.65}$$

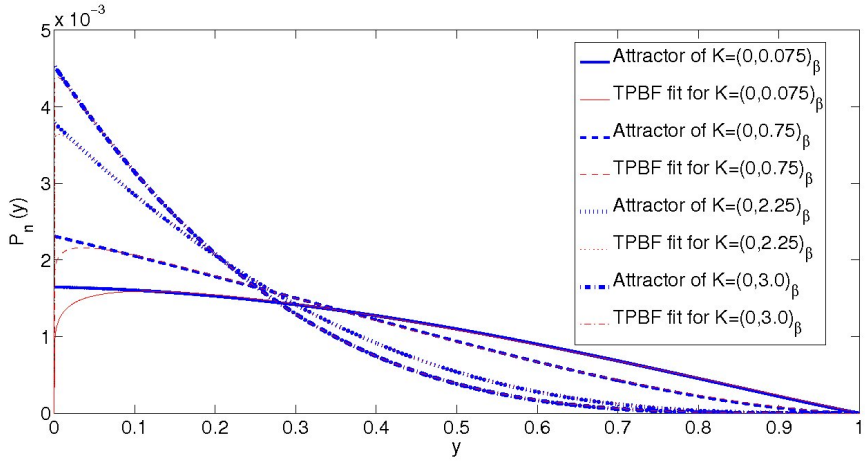


Figura 3.18: Atractores y sus ajustes para las funciones $K = \{0, 0.075\}_B$, $\{0, 0.75\}_B$, $\{0, 2.25\}_B$ y $\{0, 3\}_B$.

de tal suerte que después de $n+1$ derivadas sucesivas obtendremos:

$$\begin{aligned} P_3 &\approx \int_y^1 \mathbf{P}_{1,n}(\xi - y) P_2(\xi) \mathbf{d}\chi \\ \frac{\mathbf{d}P_3}{\mathbf{d}y} &\approx \sum_{i=0}^{n} \frac{a_i}{i!}(1-y)^i P_2(1) - a_0 P_2(y) - \int_y^1 \sum_{i=1}^{n} \frac{a_i}{i!}(1-y)^{i-1} P_2(\xi) \mathbf{d}\xi \\ &\vdots \\ \frac{\mathbf{d}^{(n+1)}P_3}{\mathbf{d}y^{n+1}} &\approx \mathbf{D}(P_2), \end{aligned} \tag{3.66}$$

en donde $\mathbf{D}(P_2)$ es un operador diferencial que depende de los coeficientes de expansión a_i y transforma a la Ec. (3.66) en una ecuación diferencial lineal de orden n, con coeficientes constantes.

3.4.7. Resumen

Después de subrayar los rasgos generales del modelo en el primer párrafo, en la sección 3.4.2 presentamos su origen y paralelo con el Teorema del Límite Central y las definiciones del producto principal detrás de la dinámica. En la sección 3.4.3 mostramos como la DBDP es estable bajo el producto $\ominus$, es decir, si $P_1 = \{a_1, b_1\}_B$ y $P_2 = \{a_2, b_2\}_B$, entonces $\exists\ a, b$ tales que $P_1 \approx \{a, b\}_B$. En las siguientes secciones se analizaron la velocidad, precisión y generalidad de la convergencia y se propuso un método de expansión de la solución del producto de correlación.

Capítulo 4

ANÁLISIS ÚLTIMO Y CONCLUSIONES

4.1. Acerca del conjunto revisado de ejemplos.

- El *corpus* hasta ahora amasado de ejemplos es sustancioso, y es solamente porque conocemos también muchos casos en los que una distribución rango-tamaño o rango-frecuencia no presenta ni remotamente semejanza con la DBDP (Como la distribución rango-tamaño de las poblaciones por pais, la distribución rango-tamaño del consumo de petróleo por pais, etc...), que la dimensión de dicho conjunto es importante.
- Aunque el modelo de distribuciones como el límite de un proceso de correlación es sin duda el más genérico, éste fue desarrollado heurísticamente, no para imitar o explicar algún subconjunto en particular de nuestra colección de ejemplos. Quizás demuestre ser una tarea sencilla encontrar algun fenómeno natural, quizás social, en el cual el producto último de una serie de eventualidades esté dado por algún tipo de correlación de los factores.

4.2. Acerca de los Modelos.

El modelo de resta de variables estocásticas es ciertamente el modelo más genérico y prometedor que tenemos. La gama de aplicaciones, dada la simpleza de las hipótesis del proceso, se vislumbra amplia, en inclusive varios de los fenómenos que reportamos en las primeras secciones de este trabajo parecieran tener una afinidad básica al producto de correlación, en particular el modelo presentado en la sección 3.1.3 pareciera poderse representar como la suma probabilística de las distribuciones que modelan la aparición y desaparición de especies.

Acerca del producto de correlación podemos concluir que:

- Presentamos un modelo estocástico muy general, cuyo producto es el resultado de la resta estocástica de variables. En cada proceso la distribución debe ser renormalizada de una manera prescrita simple, no muy restrictiva.

- En la sección 3.4.4, mediante inspección numérica mostramos que la iteración secuencial del operador de correlación de funciones generales que fungen de distribuciones de probabilidad resulta en la rápida convergencia hacia un atractor.

- Las restricciones a dicho proceso son suficientemente fuertes para diferenciar la dinámica de aquella que surge de aplicar secuencialmente una transformación de convolución convencional, y por lo tanto no obtener atractores tipo gaussianos o de Levy, sino funciones que no tienen en general una forma analítica sencilla.

- En las secciones 3.4.3 y 3.4.6 mostramos como éstas funciones son representadas con precisión numérica por la DBDP, mientras que ciertamente no poseen la misma expresión. Podemos mencionar el caso particular $(0,1)_\beta \ominus (0,1)_\beta = \frac{1}{3} - \frac{y}{2} + \frac{y^3}{6}$, que claramente no es de la forma $(a,b)_B$y sin embargo la función $(-0.015, 1.765)_B$ es una extraordinaria aproximación con $r^2 = 0.9998$.

- La sección 3.4.5 ejemplifica la manera en que los coeficientes se comportan al ser reiterados en el modelo. A primera vista no existe un patrón sencillo que determine los coeficientes a_∞ y b_∞ del atractor a partir de aquellos de los factores.

- Aunque el papel del producto por correlación tiene un cariz genérico, no tenemos todavía una única receta para adjudicar la DBDP a las distribuciones de todos los ejemplos que hemos revisado. De los modelos inspirados para cada uno de los ámbitos estudiados quizás aquél que mostramos en [25] sea el que se puede extender de manera más general al caso de secuencias genéticas y quizás a otros de secuencias simbólicas, aunque la manera de plantearlo no sea de ninguna manera operativa, sino meramente descriptiva.

4.3. Trabajo futuro

- Existen todavía muchas ramas prometedoras y sin explorar de este trabajo, especialemnte en el caso de las distribuciones de longitudes en el modelo de absorción irreversible no-homogénea, en donde se dan visos de resultados que unan alguna forma explícita para las distribuciones con la DBDP o almenos con algún comportamiento de doble cola de potencia con región intermedia.

- El producto de correlación es perfectamente equivalente a un caminante aleatorio sobre el cual no se ha hecho trabajo alguno. Si bien la definición de su dinámica no es cómoda, es posible que una exploración con ese enfoque pueda abrir conecciones o inclusive aportar resultados inmediatos al estudio de la ubicuidad de la DBDP o al desarrollo de la versión modificada del Teorema del Límite Central que vimos aquí.

- En vista de cierta capacidad insospechada de la DBDP, recientemente observada, de representar distribuciones rango-tamaño tomadas de muestras generadas con distribuciones bien conocidas, cuyas formas funcionales se conocen y no corresponden a aquellas de la DBDP (Por ejemplo, en los ajustes de algunas soluciones analíticas del modelo de resta de variables estocásticas obtenemos resultados extraordinarios y sin embargo las formas funcionales son conocidas y diferentes), es quizás imperioso hacer por primera vez un análisis de la vastedad del espacio de distribuciones que tienen ajustes satisfactorios dada una cuota arbitraria del parámetro r^2 o quizás otros estimadores más pertinentes.

Apéndice A

ARTÍCULOS

Physica A 387 (2008) 5552–5560

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Universality in the tail of musical note rank distribution

M. Beltrán del Río [a], G. Cocho [a], G.G. Naumis [b,c,*,1]

[a] Departamento de Sistemas Complejos, Instituto de Fisica, Universidad Nacional Autónoma de México, Apdo, Postal 20-364, 01000, México D.F., Mexico

[b] Facultad de Ciencias, Universidad Autónoma del Estado de Morelos, Av. Universidad 1001, Cuernavaca, Morelos, Mexico

[c] Departamento de Física-Matematica, Universidad Iberoamericana, Prolongación Paseo de la Reforma 880, Colonia Lomas de Santa Fe, 01210, Distrito Federal, Mexico

ARTICLE INFO

Article history:
Received 25 January 2008
Received in revised form 12 May 2008
Available online 25 May 2008

PACS:
02.60.Ed
43.75.+a

Keywords:
Ranking distributions
Power law distribution
Multiplicative processes
Music

ABSTRACT

Although power laws have been used to fit rank distributions in many different contexts, they usually fail at the tails. Languages as sequences of symbols have been a popular subject for ranking distributions, and for this purpose, music can be treated as such. Here we show that more than 1800 musical compositions are very well fitted by the first kind two parameter beta distribution, which arises in the ranking of multiplicative stochastic processes. The parameters a and b are obtained for classical, jazz and rock music, revealing interesting features. Specially, we have obtained a clear trend in the values of the parameters for major and minor tonal modes. Finally, we discuss the distribution of notes for each octave and its connection with the ranking of the notes.

1. Introduction

With few exceptions music has been considered as the art which has devoted itself not to the reproduction of natural phenomena, but rather to the expression of the "artist's soul" through sounds. Music is the most "non-material" of the arts. However, musical compositions can be seen as messages written in an alphabet, with different sounds as letters. In most of the musical compositions there are at most 70 different notes (without taking into account the time length of the notes). Such analogy has been the source of many efforts to compare the information content of music with respect to other languages, using tools borrowed from statistics, statistical physics and even fractal geometry. In a text written using a natural language, the elements or "words" can be taken as the different letters, twenty six for English. Other relevant example appears in biology, where the DNA codes the genetic information. In DNA, the "words" are the 61 triplets (codons) without taking into account the STOP word codon. Both natural language texts and DNA sequences present power laws in the observed frequency of a word as a function of its rank (r), where the rank is just the ordinal position of a word, if all words are ordered according to their decreasing frequency. The most frequent word has rank 1, the next most frequent rank 2 and so on. The power law behavior of the ranking is known in languages as the Zipf law [2]. This law is also very common in different fields like in physics, biology, geography, etc. [2]. In physics one can cite the rank distribution of stick-slip events in sheared granular media [3], radionuclides half-life time and nuclides mass number [4]. Other complex systems share the same phenomenology, as networks [5], biological clocks [6] and metabolic networks [7].

* Corresponding author. Tel.: +52 555 5622 51 74; fax: +52 555 5622 50 08.
E-mail addresses: mbeltrandlerio@fisica.unam.mx (M. Beltrán del Río), cocho@fisica.unam.mx (G. Cocho), naumis@fisica.unam.mx (G.G. Naumis).

[1] On sabbatical leave from: Departamento de Física-Quimica, Instituto de Física, Universidad Nacional Autónoma de Mexico., Apdo. Postal 20-364, 01000, Mexico D.F., Mexico.

0378-4371/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2008.05.031

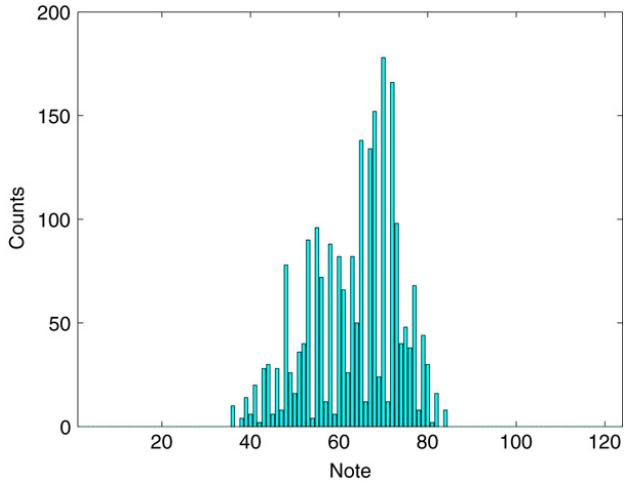


Fig. 1. Number of times that each note is used for Bach's prelude BWV 881 in F minor, taken from the Well Tempered Klavier II.

Many attempts have been made to fit musical compositions with Zipf, Simon and Yule power law formulae with reasonable results [8–11]. However, in most of the fittings presented so far, the power law is not satisfied in all regions, and usually strong deviations are observed at the tails of the ranking [8,9]. Although this is an important feature observed in many different contexts, this observation has not been fully addressed or recognized, even if these events are very important due to their rare appearance. Here we will show that such departures are well fitted by a modified power law, known as the first class two-parameter Beta distribution [1]. In a previous paper [12], we have proved that such law arises in many different systems, like in the ranking of human population, shear-slip events in granular media, in the genome of prokaryotes, in the impact factor of scientific journals [17], etc. Furthermore, we proved that such law arises when multinomial events are ranked in the limit of many random variables. The two-parameter Beta distribution is:

$$f(r) = \frac{K(R+1-r)^b}{r^a}, \tag{1}$$

where $f(r)$ is the frequency of the word r. The parameters a and b are fitted from the data, r is the rank and R is the maximal r. If $f(r)$ is normalized to 1, we have,

$$K \equiv 1/\sum_{r=1}^{N} \frac{K(R+1-r)^b}{r^a}.$$

For $R \gg 1$, K can be transformed into an integral that yields the Complete Beta Function (not to be confused with the Beta distribution) $K \approx \Gamma(b-a+2)/\Gamma(1-a)\Gamma(1+b)$ where $\Gamma(x)$ is the Gamma function. Notice that Eq. (1) has the virtue of reproducing a power law for intermediate ranges, which means that the Zipf law is valid in the body of the distribution, as is well known from many different studies. Of course that there are many other fitting functions that have been proposed [14–16]. Some of these functions use only one fit parameter, like the Lavalette law, and some others use two, like the Yule–Simon distribution [9]. As discussed in Ref. [12], Eq. (1) uses two parameters, and it is clear that in general it provides a better fit than one-parameter functions at the expense of more parameters. However, Eq. (1) has a theoretical derivation based on the ranking of multinomial events, in which many choices are available in a given step of a tree decision process.

In this article we will show that the tail of the rank-frequency distribution (RFD) for many different musical pieces also follows the same modified beta-like law. As a result, the tail of the RFD is almost universal for music. As we shall see, our results indicate some general trends in the parameters a and b, depending upon the type of music and the tonal scale used by the composer, providing a much more accurate description of music than the Zipf law.

2. Methods and results

The RFD of the musical compositions studied was obtained through their MIDI encoded versions (which is an industry-standard protocol that enables electronic musical instruments, computers and other equipment to communicate, control and synchronize with each other). Each MIDI file can be easily transformed into a standard text file that contain all the information of the composition as a list of all the events in the piece, such as "on" and "off" cues of the notes, speed and volume changes.

These text files are stripped of all other content than the cues of the notes. Each note corresponds to a given pitch. For example, the lowest MIDI note is coded as "0", and corresponds to five octaves below middle C or 8.176 Hz in common Western musical tuning. The highest note is five octaves above the G above middle C or 12,544 Hz, and is designated as MIDI note 127. Thus, the pitch in the MIDI musical alphabet has 128 letters. A typical resulting histogram of the notes is presented in Fig. 1, for Bach's prelude BWV 881 in F minor, taken from the Well Tempered Klavier II. Once the files were translated into

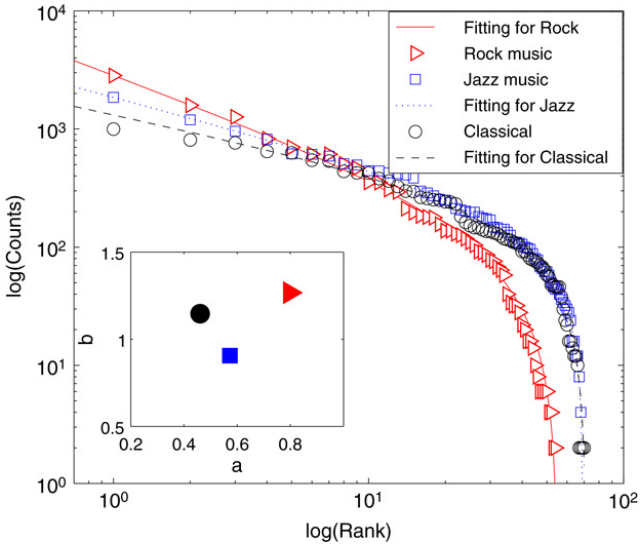


Fig. 2. A log–log plot of the frequency of notes against the rank. The classical piece is Beethoven's Quartet Op. 131 ($a = 0.461$, $b = 1.147$, $r^2 = 0.9968$), the jazz piece is "A good one" by Benny Goodman ($a = 0.573$, $b = 0.905$, $r^2 = 0.9993$) and for rock music, the song "Sweet child of mine" by Guns&Roses ($a = 0.798$, $b = 1.267$, $r^2 = 0.9992$).

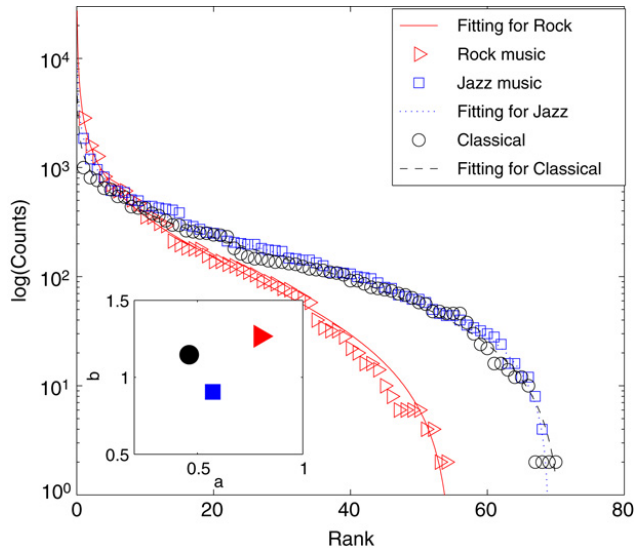


Fig. 3. A semilog plot of the frequency of notes against the rank for the same pieces of the previous figure.

numbers, the notes were ranked according to their frequency. A fit was then performed using Eq. (1), and the parameters a and b were obtained as a product of such fit. All the pieces were analyzed in a quick batch process. The number of pieces using in this study was 941 for classical music, 487 for Jazz and 422 for Rock music.

The spectrum (RFD) and fits of some pieces taken at random for each style analyzed (classical, jazz and rock music) are shown in Figs. 2 and 3. Usually, rank laws are analyzed in log–log plots to compare with a power law, as in Fig. 2. One can clearly see that the power law is only obtained for the most frequent notes, and a tail effect is seen for less frequent notes. In Fig. 3 we present the same set of data as in Fig. 2, but using a semilog plot, since for such kind of graphs the rank axis is not distorted as in the case of the log–log plot. Basically, one can clearly see how Eq. (1) produces an excellent fit, since not only the body of the RFD is well fitted, but also the tail. In fact, the average values of the correlation coefficient ($\mathcal{R}$) is for almost all of the cases larger than 0.99. In Table 1 we present the fitting results done in previous works for several different types of distributions. From the plot of these three particular cases (Beethoven, Benny Goodman and Guns&Roses), one can see that for example, the song by Guns and Roses contains much less harmonic content than the others, since basically notes with a high rank are used. The tail effect in Beethoven and Benny Goodman means that sometimes low rank notes are used, and thus there is more information content. This tail information is encoded in the corresponding values of the parameters

Table 1
Parameters a and b and goodness of fit $\mathcal{R}^2$ of the Beta representation for several symbolic sequences, obtained in previous works

Languages [19]	a	b	$\mathcal{R}^2$
Spanish	0.43	1.31	0.971
French	0.41	1.35	0.967
Latin	0.39	0.86	0.971
English	0.17	1.51	0.964
German	0.39	1.25	0.967
Finnish	0.09	1.41	0.981
Genetic sequences [20]			
Ch. Tracho	0.220	0.501	0.991
E. Coli	0.247	0.503	0.998
Homo Sapiens	0.164	0.365	0.989
Jannasch	0.370	1.243	0.978
Music			
Albeniz (Spanish Suite 5)	1.12	0.39	0.994
Bach (Double Concert BWV 1043)	0.27	1.88	0.988
Beethoven (Quartet Op. 131)	0.20	1.81	0.988
J. Satriani ("Rubina")	1.78	0.28	0.997
"Dizzy" Gillespie ("Manteca")	0.79	1.88	0.996

a and b. This can be seen by taking the derivative of $\ln f(r)$,

$$\frac{\mathrm{d}\ln f(r)}{\mathrm{d}r} = -\frac{b}{(R-r+1)} - \frac{a}{r}. \tag{2}$$

For $r \ll R$, corresponding to high rank notes, i.e., the left part of Fig. 2, we have that $(R-r+1) \approx R$ from where it follows that,

$$\frac{\mathrm{d}\ln f(r)}{\mathrm{d}r} \approx -\frac{b}{R(1-(r-1)/R)} - \frac{a}{r} \approx -\frac{a}{r}, \tag{3}$$

so the upward curvature of the tail is basically controlled by the size of parameter a. Furthermore, by taking the integral of Eq. (3) we get,

$$f(r) \approx r^{-a},$$

which allows us to recover the Zipf law. On the other hand, the lowest rank region $r \to R$, corresponding to the right of Fig. 2, can be approximated as,

$$\frac{\mathrm{d}\ln f(r)}{\mathrm{d}r} \approx -\frac{b}{(R-r)(1-1/(R-r))} - \frac{a}{R} \approx -\frac{b}{(R-r)}. \tag{4}$$

The previous approximation indicates that b controls the statistics of the low rank notes, and it follows a second power law with a cut-off at the maximum range R,

$$f(r) \approx (R-r)^b, \tag{5}$$

Fig. 4 shows a comparison between the fits on a piece by J. Brahms produced with a stretched exponential, a q-exponential [13], a parabolic fractal distribution [15], and the first class two-parameter Beta distribution. It is readily noted that only the TPBD presents a good fitting at both tails of the distribution, owing to its finite domain.

Once the fittings were performed, a pair of a and b parameters was obtained for each of the 1800 pieces. In Fig. 5 we present a plot of the values of b against the parameter a, considering that each piece has a Cartesian coordinate (a, b). Several features are worthwhile mentioning. The first is that we find a few points on the negative region of the a parameter axis. If the parameter a assumes a negative value, Eq. (1) is no longer monotonous and therefore it fails to accurately describe, by definition, a rank-size distribution, even if the fittings are still good. Nevertheless on most cases of negative a, the maximum of the curve described by the fitting is located at values of the rank less than 1, so that the fit is monotonous for admitted values of the rank and is therefore still acceptable. Secondly, rock music shows the smaller dispersion among the data. Jazz music seems to be the greatest dispersion among the values. In Figs. 6 and 7, we show the distribution of a and b for classical, jazz and rock music taken together. For all three cases a is small, and b is centered in a value close to 2. The average value $\langle a \rangle$ for classical music is lower than the average for rock and jazz music.

In terms of the frequency versus rank, that means a more intensive use of less frequent words in the classical music. This is also consistent with the fact that jazz and rock music are constructed around a main tonal center, while classical music presents many modulations, so the tonal centers vary with time. In the following table we include some of the obtained values in this work compared with others taken from the case of languages and genetic sequences:

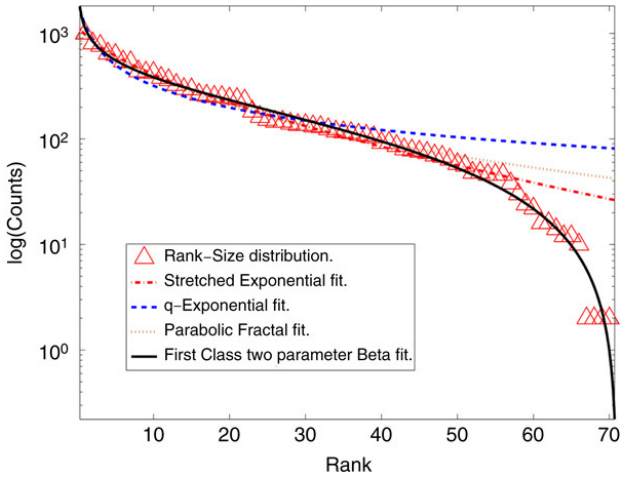


Fig. 4. Semilog plot of the rank-size distribution of Brahms' Symphony No. 4, 1st movement with superimposed fittings. the fittings to be compared are a stretched exponential, a q-exponential, a parabolic fractal distribution, and the first class two-parameter Beta distribution. Note the correct qualitative behavior of the TPBF in the rightmost region of the distribution.

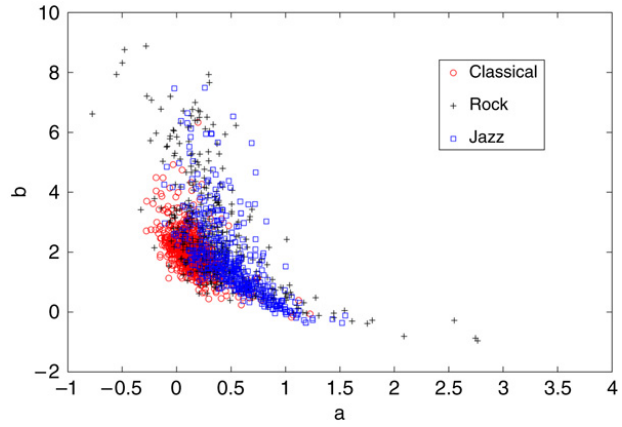


Fig. 5. Fitting parameters a and b plotted as coordinates for classical (circles), jazz (x), rock (cross) and rock without percussion (squares).

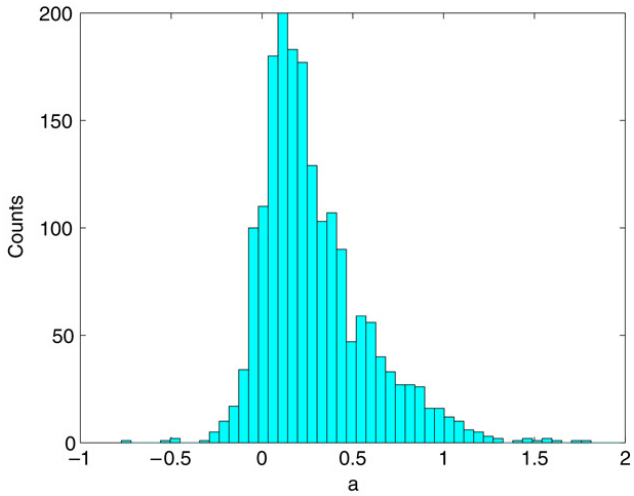


Fig. 6. Distribution of parameter a for the 1800 pieces analyzed.

From Figs. 5, 6 and 7 it is clear that in general $a < b$ for most of the pieces, although some exceptions are observed.

Furthermore, the fact that the beta-like function is observed in music means that behind the statistics of music, there is an underlying multiplicative process, as explained in Ref. [12] for many different systems. However, to gain insight about how this law arises for the particular case of music, let us consider with more detail the note statistics. The "musical code",

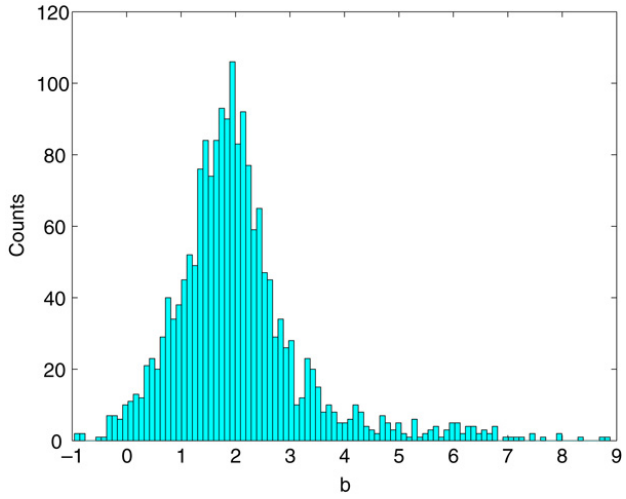


Fig. 7. Distribution of parameter b for the 1800 pieces analyzed.

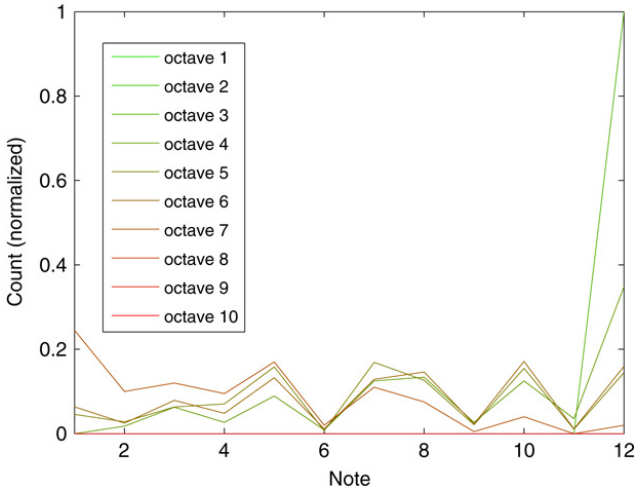


Fig. 8. Histogram per octave of the number of times that a certain note is used. The histogram is normalized to the frequency of each note in a certain octave. The number 1 note is C# and number 12 is C. Notice that F#, A and B are rarely used, as can be expected from harmonic arguments. The piece is Bach's prelude BWV 881 in F minor.

in which all the pieces studied were written in, was originally developed according to a set of conceptions about harmony that were later revised to form the so-called "well tempered scale". This set of notes is arranged in subsets called octaves. Each octave comprises twelve notes, which are given one twelve "names" (C, C#, D, Eb, E,...), each tag assigns the note to a definite role, regardless of the octave it belongs to according to the harmonic scheme of the composition is written in. So the "musical code" if one makes no distinction between two notes that have the same role (i.e. name) but have different sounds owing to different octaves. In such reduced scheme, note 1 corresponds to C#, note 2 to D and so on. Fig. 8 shows the resulting frequency histogram of the notes for Bach's prelude BWV 881 in F minor, divided by octaves. The histogram for each octave is normalized to one. Several features are observed. First one can see that in octave 1 only the note C is used, corresponding to the lowest note available. But the most important observation is that different octaves present only slightly different patterns. For example, notes F#, A and B are almost not used in all octaves, a fact that can be explained in a pure tonal basis, since for example F# is neither contained in the F minor scale nor in other neighbouring tonal centers. We have verified that many other pieces present a similar pattern.

Combining the previous observation with Figs. 1 and 8, it is possible to deduce that the ranking of notes has two different sources. The first is a general envelope given by the octave in which notes are played, since in Fig. 1 it is clear that the most used notes are at the central octaves. We have verified that most of the pieces have this property, although at the moment we can only speculate about this particular distribution (like for example, that the human voice coincides with that part of the spectrum, etc.). The second source of ranking has a harmonic nature, as shown in Fig. 8.

An interesting corroboration of the last assertion, that the total histogram of notes is the product of an envelope and an octave histogram, is the fact that a certain trend is observed for the values of a and b for major and minor modes, when we compare pieces that have the same instrumentation and musical structure. In Fig. 9 we present the a plot of the type

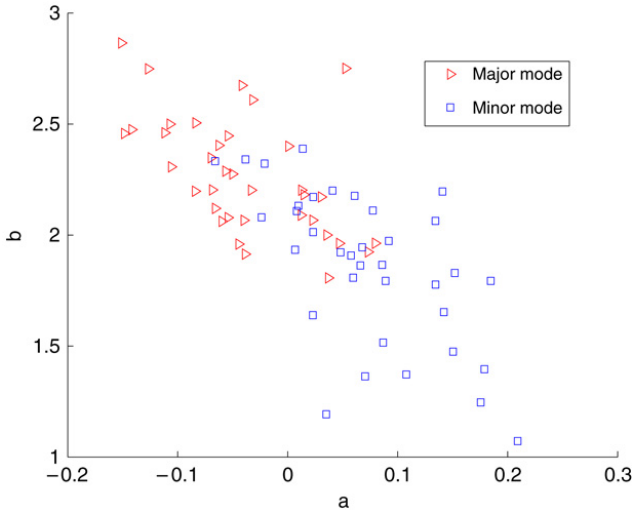


Fig. 9. A comparison of the a and b parameters for Bach's well tempered clavier distinguishing minor and major modes.

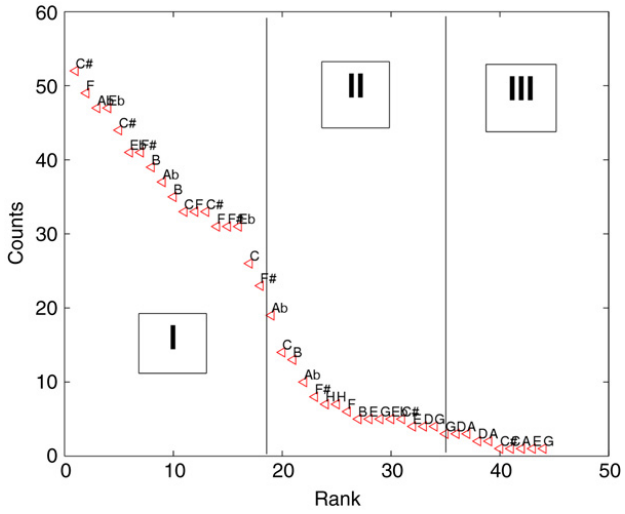


Fig. 10. The regions of the rank-size distribution of a fugue in $C\#$ Major, Region I Contains mostly notes that belong to the Tonal scale, region III is characterized by having mostly notes that do not belong to the scale and region II is a crossover section.

(a, b) for Bach's well tempered clavier. The triangles are major modes while the squares are minor modes. Note that there is a tendency of minor modes to have higher values of a, which leads to the idea that the values of (a, b) have something to do with the tonal structure of the pieces. And indeed, after some analysis it is noted that the ranked distribution of notes is composed of three sections. The first is made up exclusively of notes that belong to the tonal scale, the third is comprised only of notes that are not in the tonal scale, the middle zone is a mixture of the former two, see Fig. 10. These three sections are a consequence of the degeneration on notes by having both a favorable harmonic role (i.e. are in the tonal scale) and in central octave, just one of these characteristics or none. The subset of notes that are not included in the major scale has much less importance in major modes than the set of notes not used in the minor scale in minor modes, as shown by comparing the twelve-note rank-size distribution of minor and major modes shown in Figs. 11 and 12. Thus, in major modes the middle section of the full rank-size distribution is much smaller than that of the pieces written in minor mode, and the fittings tend in major modes to a smaller (even negative) a parameter to compensate for the bigger b that is needed to fit the narrower gap between the first and third sections.

3. Conclusions

We have shown that a Beta rank law can be used to improve upon the Zipf law, which has been widely used in music, even as a source to test music aesthetics [11]. Such law has its origins in multiplicative processes, like in decision trees [12]. In the particular case of music, it seems that ranking has two contributions, one is due to the position of the notes in the sound spectrum and the second is related with the harmonic nature of a piece. Such effect has a clear impact upon the parameters of the Beta rank law, since a clear distinction is observed for minor and major modes. These results suggest

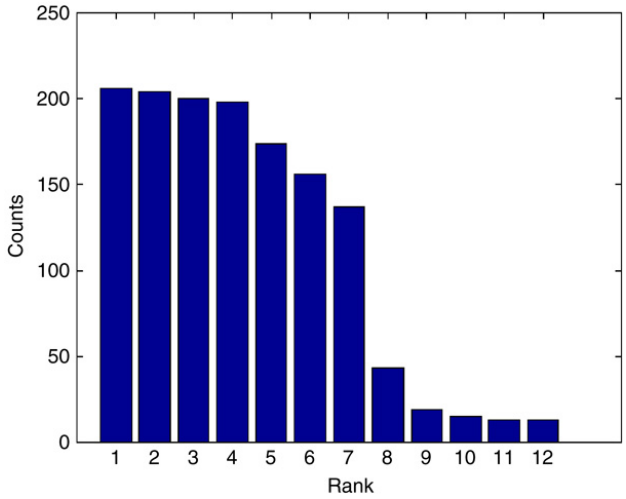


Fig. 11. Typical twelve-note rank-size distribution of a composition written in major mode. The sharp cut after the seventh note is responsible for the small intermediate region in the full rank-size distribution.

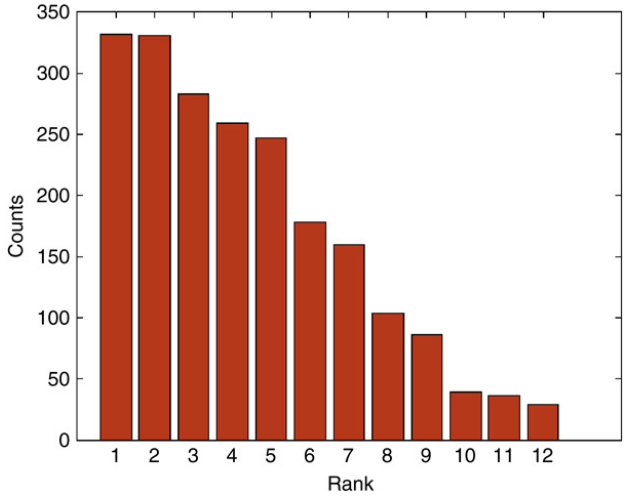


Fig. 12. Typical twelve-note rank-size distribution of a composition written in minor mode. Due to clear harmonic reasons, this distribution is smoother than that of major modes.

the presence of generic brain constraints associated to features of brain neural networks. Perhaps, important aspects of the dynamics of these neural networks could be modeled by the multiplication of number sequences [18]. In absence of external modulations, this behavior would be dominant, restricting the "free inspiration" of the musician. However, if the composers take into account these restrictions, they could follow or violate these restrictions, adding an extra dimension to music composition.

Acknowledgments

This work was supported by DGAPA UNAM project IN108502, and CONACyT 48783-F and 50368.

References

[1] J.B. McDonald, Some generalized functions for the size distribution of income, Econometrica (1984) 647–663.
[2] W. Li, Phys. Rev. E 43 (1991) 5240. see also: W. Li. http://www.nslij-genetics.org/wli/zipf (2003).
[3] M. Bretz, R. Zaretzki, S.B. Field, N. Mitarai, F. Nori, Europhysics Letters 74 (2006) 1116.
[4] G. Audi, O. Bersillon, J. Blachot, A.H. Wapstra, Nuclear Physics A624 (1997) 124.
[5] S. Fortunato, A. Flammini, F. Menczer, Physical Review Letters 96 (2006) 218701.
[6] A.C.C. Yang, S.S. Hseu, H.W Yien, A.L. Goldberger, C.K. Peng, Physical Review Letters 90 (2003) 108103.
[7] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, A.L. Barabasi, Nature 407 (2000) 651.
[8] M. Schroeder, Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise, W. H. Freeman, San Francisco, 1992.
[9] M.E.J. Newman, Contemporary Physics 46 (2005) 323–351.
[10] B. Manaris, P. Machado, C. McCauley, J. Romero, D. Krehbiel, Lecture Notes in Computer Science, Applications of Evolutionary Computing, in: LNCS, vol. 3449, Springer-Verlag, Berlin, 2005, pp. 498–507.

[11] B. Manaris, J. Romero, P. Machado, D. Krehbiel, T. Hirzel, W. Pharr, R.B. Davis, Computer Music Journal 29 (2005) 55.
[12] G.G. Naumis, G. Cocho, Physica A 387 (2008) 84.
[13] C. Tsallis, G. Bemski, R.S. Mendes, Physics Letters A 257 (1999) 9398.
[14] M. Montemurro, Physica A 300 (2001) 567578.
[15] J. Laherrere, D. Sornette, The European Physical Journal. B 2 (1998) 525.
[16] E.W. Montroll, M.F. Shlesinger, Journal of Statistical Physics 32 (1983) 209.
[17] R. Mansilla, G. Cocho, Complex Systems 12 (2000) 207.
[18] G.G. Naumis, G. Cocho, New Journal of Physics 9 (2007) 286.
[19] http://www.ultrasw.com/pawlowski/brendan/frequency.htm.
[20] http://www.kazusa.or.jp/codon/.

Rank-Size Distribution of Notes in Harmonic Music: Hierarchic Shuffling of Distributions

Manuel Beltrán del Río and Germinal Cocho

Departamento de Sistemas Complejos, Instituto de Fisica. Universidad Nacional Autónoma de México. Apdo. Postal 20-364, 01000, México D.F., Mexico
{mbeltrandelrio,cocho}@fisica.unam.mx

Abstract. We trace the rank size distribution of notes in harmonic music, which on previous works we suggested was much better represented by the Two-parameter, first class Beta distribution than the customary power law, to the ranked mixing of distributions dictated by the harmonic and instrumental nature of the piece. The same representation is shown to arise in other fields by the same type of ranked shuffling of distributions. We include the codon content of intergenic DNA sequences and the ranked distribution of sizes of trees in a determined area as examples. We show that the fittings proposed increase their accuracy with the number of distributions that are mixed and ranked.

Keywords: Ranking distributions, Power law distribution, Zipf law in Music.

1 Introduction

Since the publication of Zipf's law, Power Law and Rank-size distributions have been often tried in fields other than languages, Music among others. There have been several recent papers proposing different explanations and/or modifications to the pure Power Law [1,2,3,4]. In a previous work [5], we have shown that pure Power Laws are not a good representation of the rank size distribution of notes of a musical piece throughout its domain, and that the two-parameter, first class Beta Distribution (TPBF) [6]:

$$S(r) = \frac{N(R+1-r)^{\beta}}{r^{\alpha}}, \tag{1}$$

where $S(r)$ is the size against which the distribution is ranked, N a normalisation factor, R the number of total ranks and r the rank variable, can account for a better fit on both tails of the distribution. Refer to Figure 1 for some sample fittings, and to Figure 2 for a comparison of several types of representations.

When analysing the results of over 1500 fitted musical pieces in parameter space, a clear trend over major and minor modes was noted. To get a clearer picture of what made the parameters different for major and minor modes, the rank-size distribution of notes was broken into a distribution of octaves ranked

J. Zhou (Ed.): Complex 2009, Part II, LNICST 5, pp. 2222–2228, 2009.

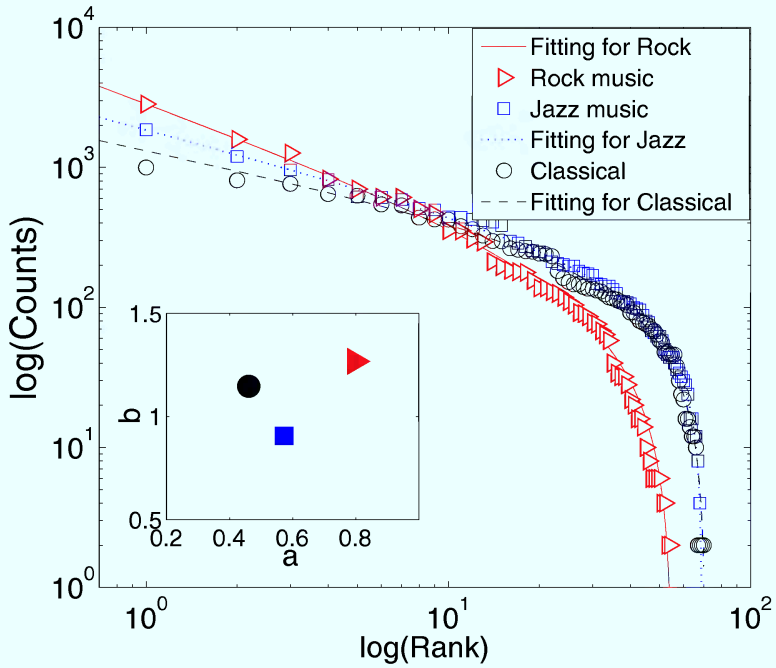


Fig. 1. A log-log plot of the frequency of notes against the rank. The classical piece is Beethoven´s Quartet No.14 Op. 131 ($\alpha = 0.461$, $\beta = 1.147$, $r^2 = 0.9968$), the jazz piece is "A good one" by Benny Goodman ($\alpha = 0.573$, $\beta = 0.905$, $r^2 = 0.9993$) and for rock music, the song "Sweet child of mine" by Guns&Roses ($\alpha = 0.798$, $\beta = 1.267$, $r^2 = 0.9992$).

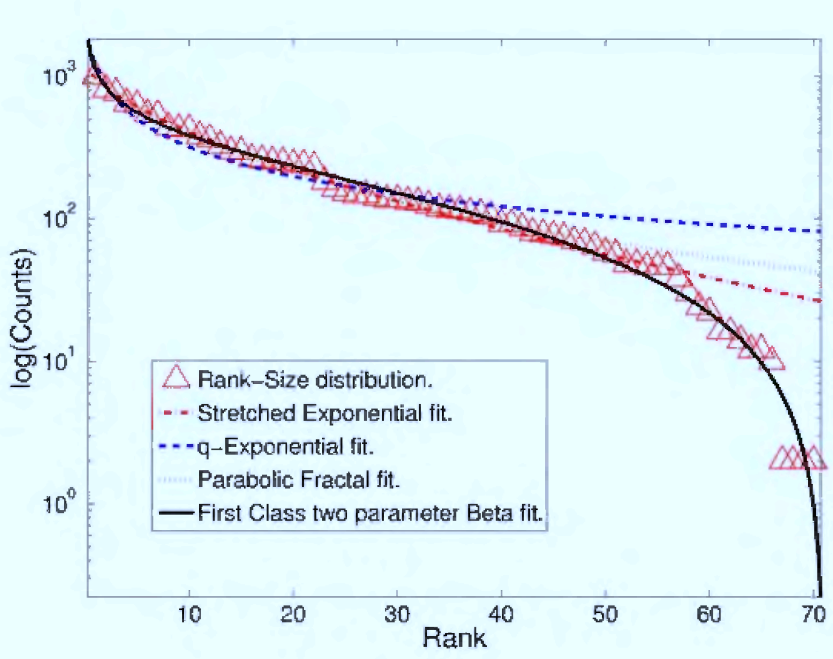


Fig. 2. Semilog plot of the rank-size distribution of Brahms' Symphony No.4, 1st movement with superimposed fittings. the fittings to be compared are a stretched exponential, a q-exponential, a parabolic fractal distribution, and the first class two parameter Beta distribution. Note the correct qualitative behaviour of the TPBF in the rightmost region of the distribution.

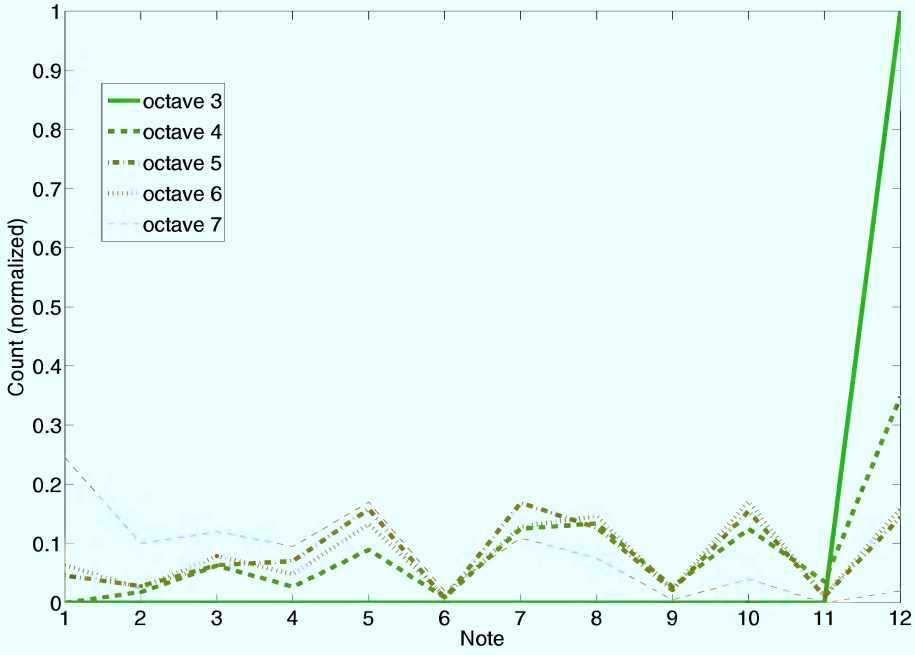


Fig. 3. Histogram per octave of the number of times that a certain note is used. The histogram is normalized to the frequency of each note in a certain octave. The number 1 note is C# and number 12 is C. Notice that F#, A and B are rarely used, as can be expected from harmonic arguments. The piece is J.S. Bach's prelude BWV 881 in F minor.

by the number of notes used within, and a distribution of the twelve notes of the scale, per octave, ranked by frequency of appearance. The second group of the above mentioned revealed that the use of a particular note was qualitatively the same in each octave, thus suggesting that the use of a particular note can be simulated stochastically according to its importance within the particular scale used, i.e. favouring, for example, the tonic, the dominant, etc... over notes that are not in the scale of the main key or those of the most common modulations. See Fig. 3. This subject is treated in [7] from the perspective of perception psychology.

2 TPBF and Ranked Shuffling of Distributions

Until now all work has been descriptive, showing the phenomenology and making emphasis on the ubiquity of the TPBF [8,9,10], which suggests an approach of great generality. In the following we show that it is suspiciously often the case that, under certain restrictions yet to be fully characterised, the mixing of distributions and their posterior ranking produces functional forms well represented by the TPBF, much in the same spirit as convolution leads to a Normal distribution, given certain restrictions and limits. The ubiquity of the findings

is the prime motivation to believe that the mentioned restrictions shouldn't be too strong as to make the systems of interest abound as they seem to.

The separation of the music rank-size histogram into distributions of the individual octaves, made it clear that the place a note adopted after being ranked was dependent on its function within the tonal scale on which the piece was set, and on the octave on which it was played. The final form of the distribution was thus shown to be a collection of regions that could be characterised by the different combinations of those two parameters that determine the "importance" of a note within the piece. For example, the leftmost region of the rank-frequency distribution, that which contains the notes with higher frequency, is characterised by containing notes placed typically in the middle octaves (the histogram of the octaves is usually unimodal and non monotonous) and belonging to the tonal scale, the intermediate region consisted predominantly on notes either played on border octaves and belonging to the scale or on middle octaves and not on the scale, etc... That music can be decomposed in that way also explains the major-minor modes trend in parameter space, since the 12 note histogram of each is different, given the inclusion of two extra notes in the minor mode scale. See Figs. 4 and 5

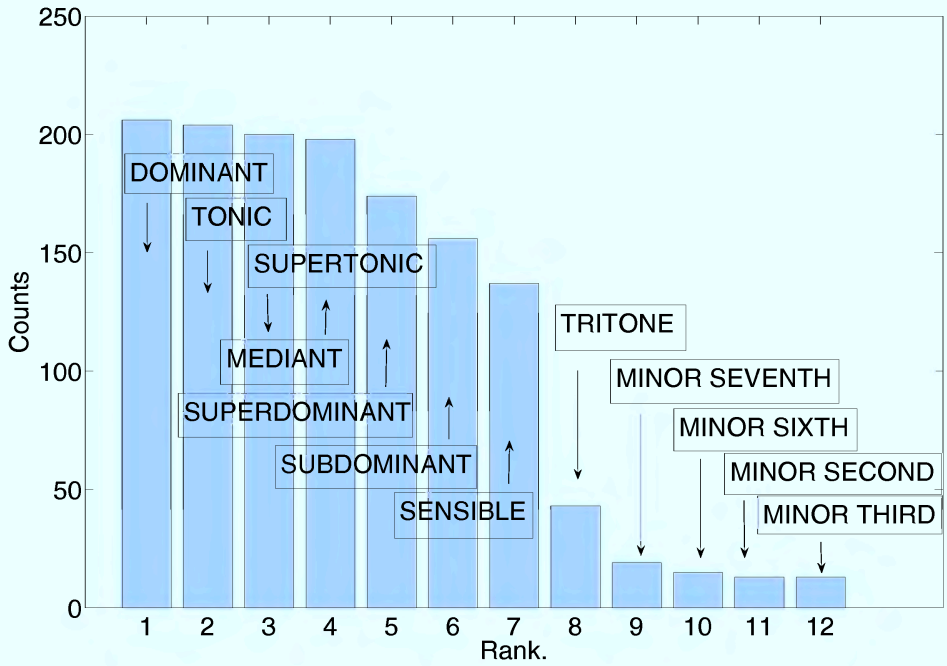


Fig. 4. Typical twelve-note rank-size distribution of a composition written in major mode. The sharp cut after the seventh rank marks the line between the notes on the scale and those used only as passing or grace notes.

It is this play between hierarchies what is consistently present in systems that show ranked distributions well fitted by the TPBF. Evidently, any distribution can be separated arbitrarily into smaller distributions in such way that the ranked shuffling of the constituents is the original distribution, so it is possible

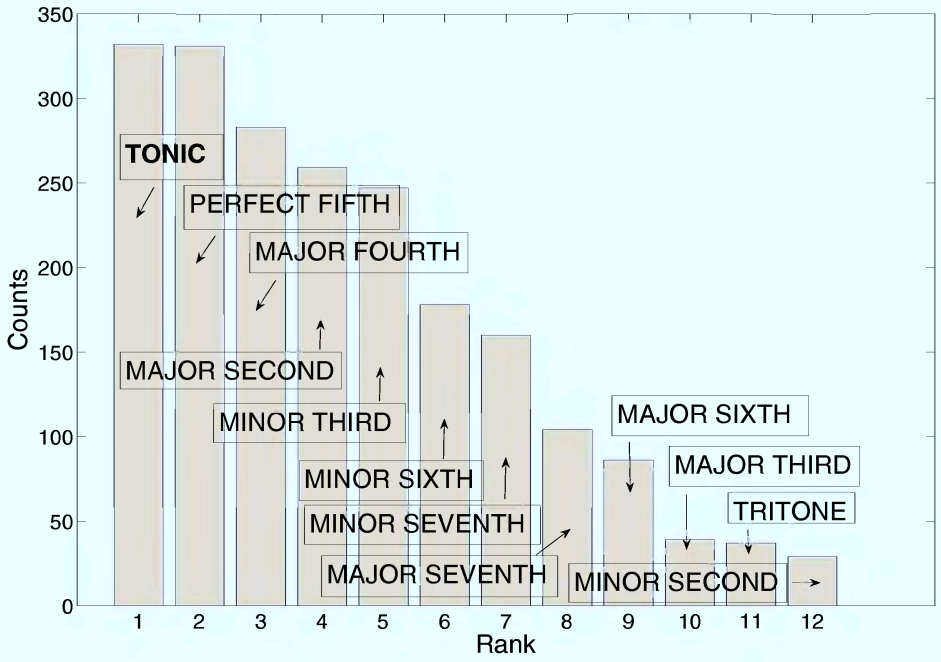


Fig. 5. Typical twelve-note rank-size distribution of a composition written in minor mode. This distribution has a less dramatic fall than that for major keys because minor tones have two more notes in the scale, depending on whether it is rising or falling.

that the restrictions needed on the ranked shuffling of distributions as to produce TPBF's act on the way the individual constituents of the distribution overlap.

2.1 Examples

To illustrate the ranked shuffling of distributions we present data taken from [11]. We obtained the distribution of trees, herbs and shrubs in a particular abandoned field in Illinois, USA, ranked according to the area of coverage. The fittings for each of these are not quite as satisfactory as that for the distribution of all areas together, that is, that for the ranked shuffle of the three distributions. Refer to Fig 6.

This progressive approach to a better TPBF fit with further shuffling also happens in the previous examples on Music. See Fig 7.

To end this section we mention, without much detail, that another important example we have worked with is the ranked distribution of triplets of nucleic bases in DNA. This last distributions reveals a structure similar to that of music, in that it can be separated in regions characterised by the degeneration of the parameters that determine the likeliness of appearance of a triplet. Whereas in Music this likeliness was determined by the octave and whether the note is in the tonal scale or not, in the distribution of triplets it is characterised by whether the first, second and third base are a "strong" or "weak" base, which in a loose sense equals to say that they belong or not to a privileged set, like, in music, the tonal scale.

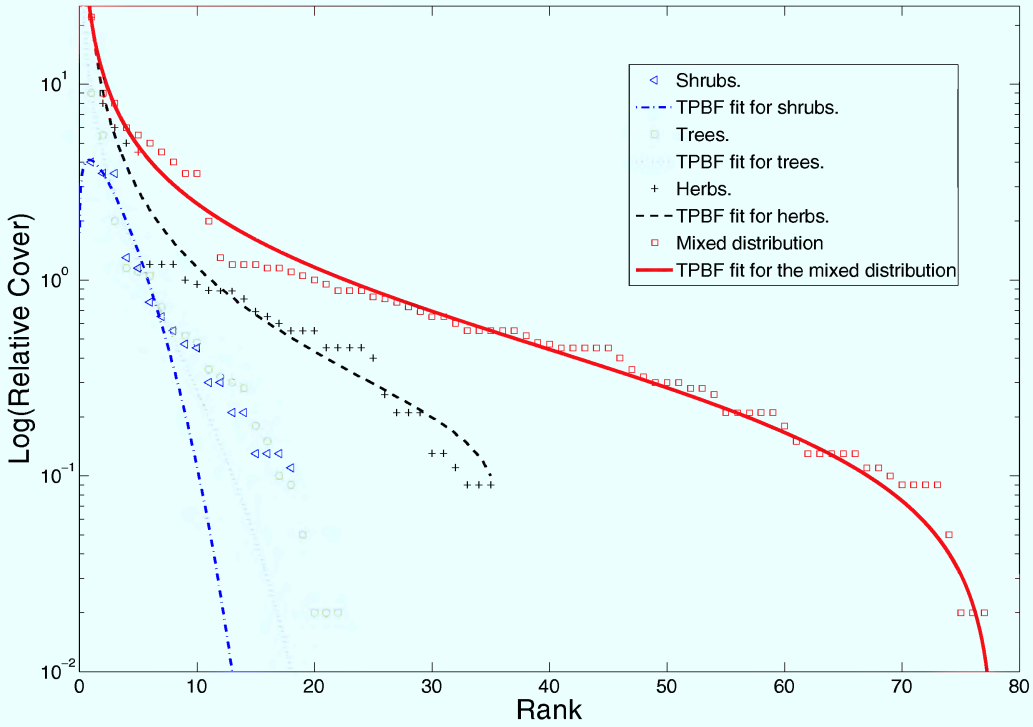


Fig. 6. Log Plot of the distribution of trees, herbs and shrubs, ranked according to relative cover (percentage of total cover) in an abandoned lot in Illinois, USA

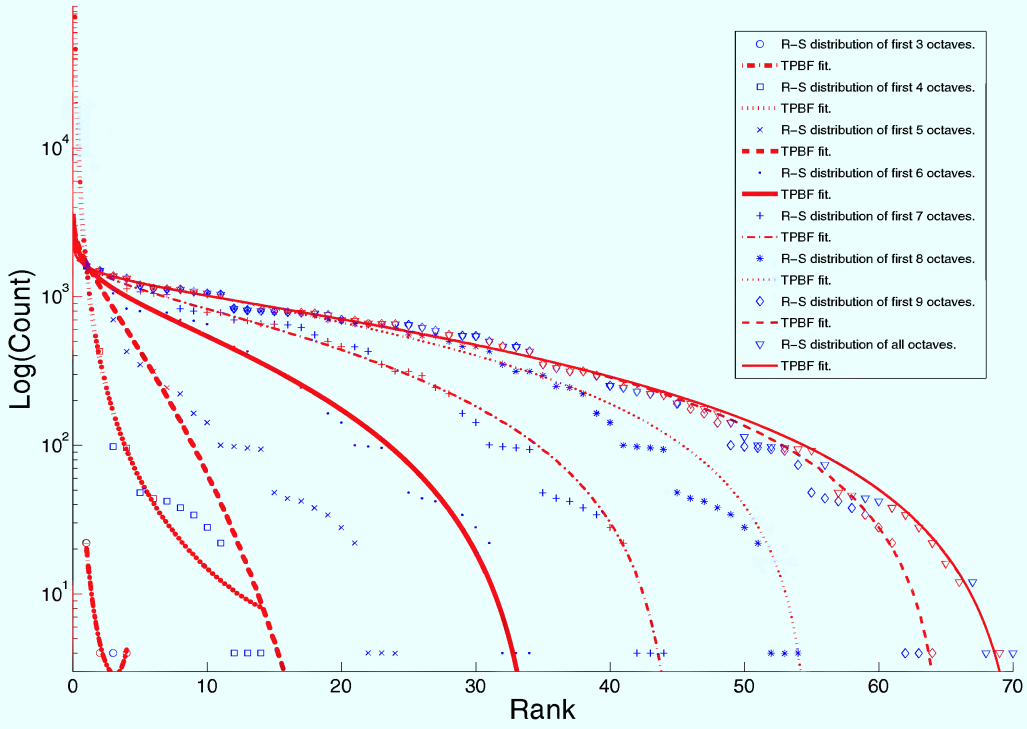


Fig. 7. Log Plot of the rank-frequency distributions and their TPBF fittings of an increasing number of octaves in Isaac Albeniz's *Sevilla* from *Suite Iberia*

3 Conclusions

The ranked distributions of notes in harmonic musical compositions are consistently well represented by the TPBF, and the parameters arising from the fittings have tendencies that become clear once one separates the whole into distributions of individual octaves, where the importance of the main key of the composition is revealed. Such separation leads to believe that the presence of the TPBF may arise from the ranked shuffling of distributions. Evidence from other fields, and from the general observation that goodness-of-fit parameters increase progressively on the number of distributions that are shuffled and ranked, strengthen this last idea. The necessary restrictions on the ranked shuffling to produce the TPBF can't be too strong since they need to account for the ubiquity of the findings.

References

1. Zanette, D.H.: Zipf's law and the creation of musical context. Musicae Scientiae 10, 3–18 (2006)
2. Zanette, D.H.: Playing by numbers. Nature 453(7198), 988–989 (2008)
3. Manaris, B., Machado, P., McCauley, C., Romero, J., Krehbiel, D.: Developing fitness functions for pleasant music: Zipf's law and interactive evolution systems. In: Rothlauf, F., Branke, J., Cagnoni, S., Corne, D.W., Drechsler, R., Jin, Y., Machado, P., Marchiori, E., Romero, J., Smith, G.D., Squillero, G. (eds.) EvoWorkshops 2005. LNCS, vol. 3449, pp. 498–507. Springer, Heidelberg (2005)
4. Manaris, B., Romero, J., Machado, P., Krehbiel, D., Hirzel, T., Pharr, W., Davis, R.B.: Computer Music Journal 29, 55 (2005)
5. Beltrán del Río, M., Cocho, G., Naumis, G.G.: Universality in the tail of musical notes rank distribution. Physica A 387(22), 5552–5560 (2008)
6. McDonald, J.B.: Some generalized functions for the size distribution of income. Econometrica, 52–53 (1984)
7. Krumhansl, C.L.: Cognitive Foundations of Musical Pitch. Oxford Psychology Series, vol. 17, pp. 29–31 (1990)
8. Mansilla, R., Cocho, G.: Complex Systems 12, 207 (2000)
9. Naumis, G.G., Cocho, G.: Physica A 387, 84 (2008)
10. Naumis, G.G., Cocho, G.: New J. Phys. 9, 286 (2007)
11. Bazzaz, F.A.: Plant Species Diversity in Old-Field Successional Ecosystems in Southern Illinois Ecology, vol. 56(2), pp. 485–488 (Early Spring 1975)

Apéndice B

COMPENDIO DE FENOMENOLOGÍA PUBLICADO EN PLoS

OPEN ACCESS Freely available online

PLoS one

Universality of Rank-Ordering Distributions in the Arts and Sciences

Gustavo Martínez-Mekler[1,2☯]*, Roberto Alvarez Martínez[2,3☯], Manuel Beltrán del Río[3], Ricardo Mansilla[4], Pedro Miramontes[5], Germinal Cocho[2,3☯]

1 Instituto de Ciencias Físicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, México, **2** Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Distrito Federal, México, **3** Instituto de Física, Universidad Nacional Autónoma de México, Distrito Federal, México, **4** Centro de Investigaciones Interdisciplinarias en Ciencias y Humanidades, Universidad Nacional Autónoma de México, Distrito Federal, México, **5** Facultad de Ciencias, Universidad Nacional Autónoma de México, Distrito Federal, México

Abstract

Searching for generic behaviors has been one of the driving forces leading to a deep understanding and classification of diverse phenomena. Usually a starting point is the development of a phenomenology based on observations. Such is the case for power law distributions encountered in a wealth of situations coming from physics, geophysics, biology, lexicography as well as social and financial networks. This finding is however restricted to a range of values outside of which finite size corrections are often invoked. Here we uncover a universal behavior of the way in which elements of a system are distributed according to their rank with respect to a given property, valid for the full range of values, regardless of whether or not a power law has previously been suggested. We propose a two parameter functional form for these rank-ordered distributions that gives excellent fits to an impressive amount of very diverse phenomena, coming from the arts, social and natural sciences. It is a discrete version of a generalized beta distribution, given by $f(r) = A(N+1-r)^b/r^a$, where r is the rank, N its maximum value, A the normalization constant and (a, b) two fitting exponents. Prompted by our genetic sequence observations we present a growth probabilistic model incorporating mutation-duplication features that generates data complying with this distribution. The competition between permanence and change appears to be a relevant, though not necessary feature. Additionally, our observations mainly of social phenomena suggest that a multifactorial quality resulting from the convergence of several heterogeneous underlying processes is an important feature. We also explore the significance of the distribution parameters and their classifying potential. The ubiquity of our findings suggests that there must be a fundamental underlying explanation, most probably of a statistical nature, such as an appropriate central limit theorem formulation.

Citation: Martínez-Mekler G, Martínez RA, del Río MB, Mansilla R, Miramontes P, et al. (2009) Universality of Rank-Ordering Distributions in the Arts and Sciences. PLoS ONE 4(3): e4791. doi:10.1371/journal.pone.0004791

Editor: Madalena Costa, Harvard University, United States of America

Received October 25, 2008; **Accepted** January 16, 2009; **Published** March 11, 2009

Funding: Support from CONACyT 47836-F, DGAPA-UNAM-IN115908 and DGAPA-UNAM-IN112407-3 grants are acknowledged. R. Alvarez thanks CONACyT for a scholarship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: mekler@fis.unam.mx

☯ These authors contributed equally to this work.

Introduction

During the past decade or so, a considerable amount of research has been devoted to power law behaviors, particularly with regard to complex networks [1,2]. However, when real data is analyzed, in most of the cases the power law trend holds only for an intermediate range of values; there is a power law breakdown in the distribution tails [3,4]. Both the breakdown point and the tail functional forms are of interest [5]. Several explanations have been provided for this phenomenon, such as finite size effects (e.g. insufficient data for good statistics) [6,7,8], network dilution, network growth constraints [3,7] and different underlying dynamical regimes, leading to power law corrections (sometimes referred to as scaling corrections) in the form of exponential, Gaussian, stretched exponential, gamma and various types of extreme value distributions [9,10]. In this work we focus on rank-ordered distributions, often related to cumulative distribution functions, which show the way in which a given property of a system is ordered decreasingly according to its importance (rank). Our main result is that a surprising amount of situations follow a two parameter distribution which incorporates the product of two power laws defined over the complete data set, one measured from "left to right" and the other from "right to left". The fit holds for the full range of values, tails included, with correlations that rival with, or generally improve on, power law correction schemes proposed in the literature.

In our work a functional universality is revealed for rank-ordered distributions, encompassing apparently unrelated phenomena coming from music, painting, ecology, urbanism, neuroscience, genetics and social networks, amongst others. In the following we develop a phenomenology based on a selection of the vast number of cases where we have encountered this functional form. Prompted by some of these observations we implement a conflicting dynamics model that generates this distribution and contributes to the identification of relevant underlying features of processes leading to it, as well as to a

characterization of its parameters. From our exploration we also detect that the convergence of multiple heterogeneous processes appears to be an important factor. Overall, our findings suggest that there must be a deep underlying explanation, possibly of a statistical nature.

Results

Phenomenology

Rank-ordered relations show how given property of a process decreases [1,2,7,8]. A well studied instance of this is the so called Zipf law [11] which originally referred to the frequency with which words are used in a specific language. Zipf showed that the logarithm of the frequencies with which words appear in the novel Ulysses by James Joyce, when plotted in decreasing order against the logarithm of their rank, fall on a straight line with slope -1, thus indicating a power law behavior. However, in general, this straight line behavior with negative slope holds only within an intermediate rank range [12,13]. Here we show that for this phenomenon of common occurrence, the power law corrections have themselves universal features, further more, a surprising amount of systems of very diverse nature which do not follow power laws at all, present a common statistical behavior expressed by a generic rank-ordered distribution function.

As a starting point we consider systems consisting of symbols arranged sequentially such as codons (nucleic acid triplets that code for amino acids) in genes or notes in musical scores. In Fig. 1A we show a log-log plot of the frequency with which the 61 possible codons (stop codons excluded) appear in the coding genetic sequences of the bacterium *Escherichia coli*, plotted in decreasing order from the most common to the least common one. Notice the power law like behavior in the intermediate range and the steeper finite size decay for the less frequent occurrences. If we plot the same data in a semi-log representation, together with the codons of the genes of *Nesisseria gonorrhea* and the worm *Caenorhabditis elegans*, we obtain the sigmoid type graphs shown in Fig. 1B. In this representation the full data range is given equal standing. The form in the semi-log graphs in the region to the left of the inflexion point is suggestive of a logarithmic decay, while the one to the right brings to mind a logarithmic behavior with the independent variable measured from right to left; we therefore test the pertinence of using a functional form incorporating the above mentioned features as a fit for the data, namely:

$$f(r) = A(N+1-r)^b \Big/ r^a,$$

where r is the rank value, N its maximum value, A a normalization constant and $(a,\ b)$ two fitting exponents. This expression is a discrete version of the continuous random variable generalized beta distribution and we shall refer to it from now on as DGBD [14,15]. The bold curves in Fig. 1 show that the functional form is a very good choice. The square of correlation coefficients, R^2, determined by a log-log multiple linear regression, lie between 0.98 and 0.99. We have obtained similar results, for tens of organisms covering archea, bacteria and eukaryotes, both for amino acid and codon distributions.

If we now look into the arts, we have that notes in musical scores provide another example of sequences of symbols where rank frequency DGBD are encountered. Fig. 2 shows compositions by Beethoven, Holst and the rock band Alice Cooper. Again correlation coefficients are very high, with R^2 above 0.98, notice that fit is very good for the whole range of values. The analysis of more than 1800 compositions shows that this type of behaviour is

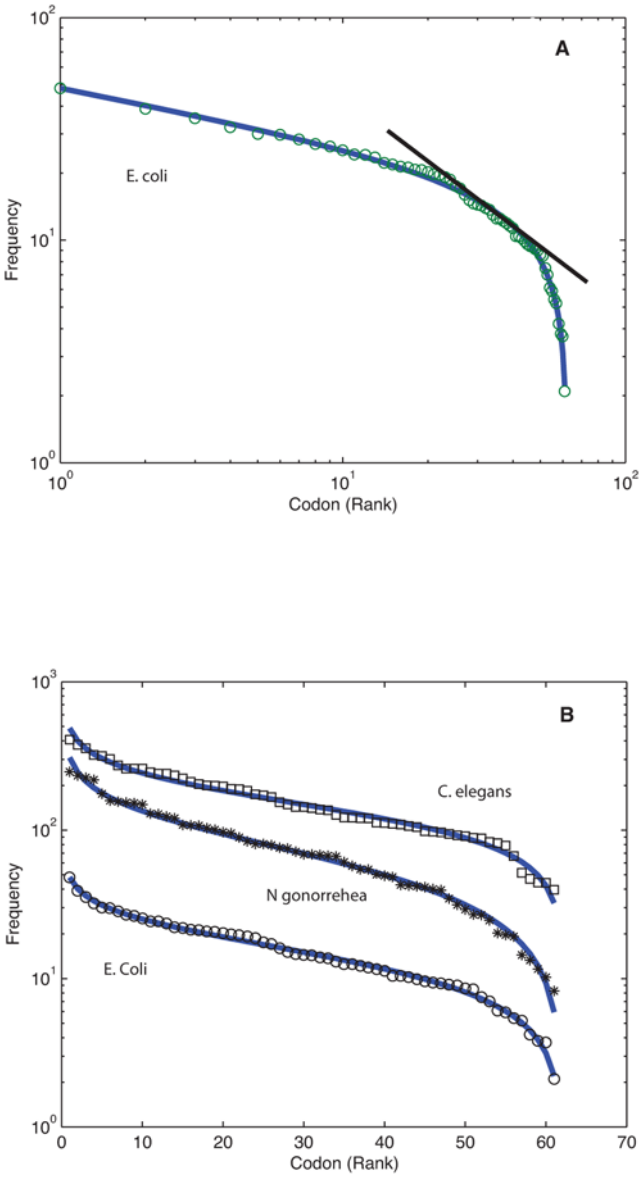


Figure 1. Frequency-rank in genetic sequences. (A) Log-log plot of the frequency, in descending order, with which the codons appear in the genome of E. coli. The bold line is the discrete generalized beta distribution (DGBD) fit with exponents and squared correlation coefficient $(a,b,R^2) = (0.25,\ 0.50,\ 0.99)$. The straight line is included as a guide to the eye of a power law behavior within a restricted range. **(B)** Semi-log plot of the frequency-ordered codons of the genomes of *C. elegans*, *N. gonorrehea* and *E. coli*. Solid lines are the fits with $(a,b,R^2) = (0.28,\ 0.38,\ 0.98)$, $(0.31,0.65,\ 0.99)$ corresponding to the first two, values for *E coli* are given in (A). Frequencies for *N. gonorrhea* have been multiplied by a factor of 5 and those of *C. elegans* by 10 in order to avoid overlaps.
doi:10.1371/journal.pone.0004791.g001

recurrent. Furthermore, fitting parameters (a,b) appear to be sensitive to whether the musical composition is in a minor or mayor scale [16].

Still in the arts, keeping in mind that the frequency of occurrence of a note is in some sense related to the "length" occupied in a given score, we determine the area occupied by specific geometric motifs in abstract painting, such as rectangles in canvases by Paul Klee and Piet Mondrian or circles in works of art by Kandinsky. We then order these determinations as rank-size

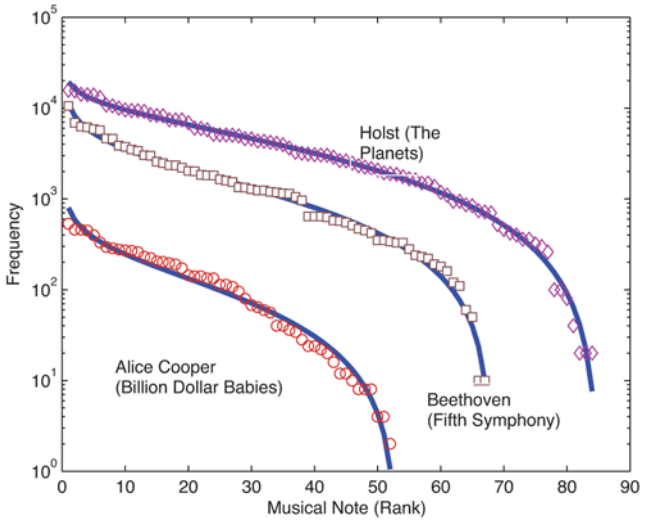


Figure 2. Frequency-rank distributions for musical scores. Plot of the occurrence of musical notes, ordered decreasingly, in the scores of Holst's "The Planets", Beethoven's first movement of the "Fifth Symphony" and Alice Cooper's "Billion Dollar Babies". Solid lines are DGDB fits with $(a,b,R^2) = (0.23, 1.54, 0.988)$, $(0.42, 1.25, 0.987)$, $(0.71, 1.06, 0.978)$.
doi:10.1371/journal.pone.0004791.g002

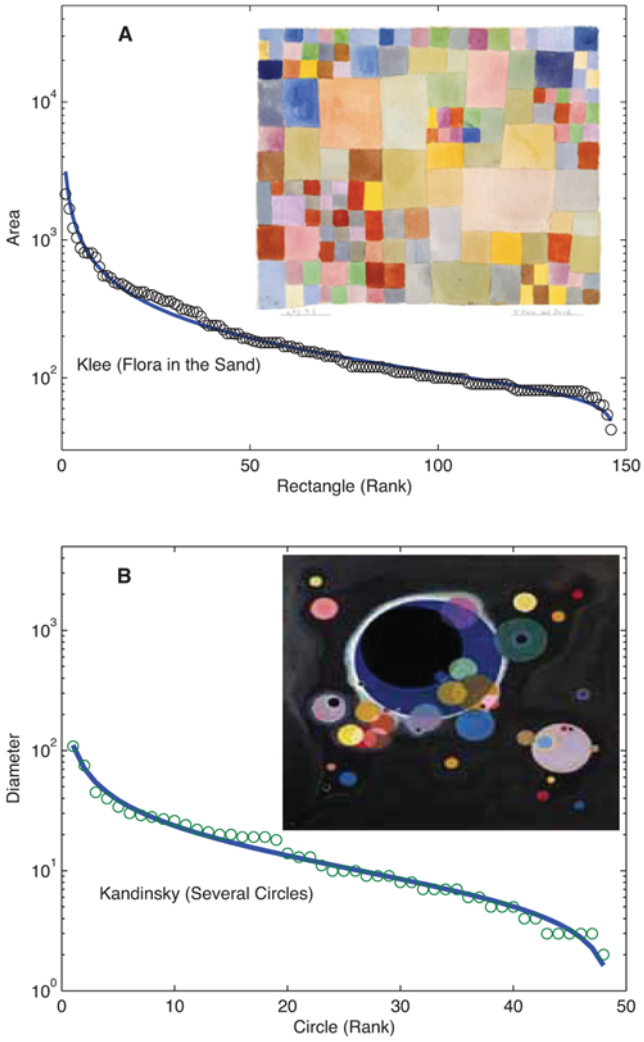


Figure 3. Size-ordered distributions in abstract paintings. (A) Plot of rectangle relative sizes in arbitrary units shown in decreasing order appearing in Klee's painting "Flora in the sand". Bold line is the DGBD fit with $(a,b,R^2) = (0.70, 0.14, 0.999$. **(B)** Plot of circle relative areas expressed in arbitrary units present in Kandinsky's "Several Circles" arranged in decreasing order, here the bold line fit has $(a,b,R^2) = (0.62, 0.32, 0.978)$.
doi:10.1371/journal.pone.0004791.g003

distributions and adjust DGB distributions. In Fig. 3 we show fits for Klee's "Flora on Sand" and Kandinsky's "Several Circles" respectively, again with R^2 values above 0.98.

An environmental case is shown in Fig. 4A for plant species diversity in old-field successional ecosystems. Here the rank ordered data refer to the relative cover values of the plant species encountered in 40 year old abandoned fields in Southern Illinois [17].

Fig. 4A is related to neurophysiology [18], it shows that rank ordered local field potential measurements in cat cerebral cortex during natural wake states follow very closely a DGBD, $(a,b,R^2) = (0.081,\ 0.239,\ 0.97)$. When slow wave sleep states (SWS) are considered the fit worsens while rapid-eye-movement (REM) periods resemble awake state results.

A rank ordered distribution related to society, is presented in Fig 5A for the world wide classification of universities according to their number of contributions to the journals Nature and Science between 2002 and 2006 [19]. Here the square of the correlation coefficient is 0.99. In Figs. 5 B,C we show fitting results with R^2 above 0.99 for two other examples of social bearing: the journal impact factor ranking [20] and population ordered municipalities of Spanish provinces, respectively [21].

As DGBD network examples we show the movie actor collaborative distribution [1] (see Fig. 6A) and the rank-size distribution of the out-bound links of the *E. coli* genetic regulatory network [22] (see Fig. 6B). In the former each node is an actor, and two actors are connected if they were cast in the same movie. Though this network has been extensively studied in the literature and good results for the connectivity probability have been found with alternative two parameter distribution functions [23], our DGBD fit reaches remarkable accuracy, reproducing qualitative features.

For comparative purposes in Table 1 we show the values of (a,b,R^2) for several representative examples of diverse nature, some of them taken from previous figures.

Model

The material presented so far is only a sample of the variety of situations where we have encountered a rank ordering statistical behavior following closely the DGBD. This poses the challenge of unraveling mechanisms or identifying characteristics that may contribute to some understanding of these findings [24]. Prompted by our analysis of genetic sequences, as a step in this direction we work with an expansion-modification dynamics introduced by Li [25,26], where two processes converge, one related to permanence the other to change. This model incorporates basic elements of a neutral evolution scheme in which the main mechanisms for change in sequences are duplications and point mutations. The simplest Boolean realization of this scheme is the following: *i)* consider a system with variables that can only take two values, say *0* and *1*; *ii)* initiate a process with either one of these values by applying with probability p the modification (point wise mutation) rule: *0* goes to *1*, or *1* goes to *0*, and with probability $1\text{-}p$ the expansion (duplication) rule: *0* goes to *00* or *1* goes to *11*, *iii)* generate a growing sequence of zeros and ones by a repeated application of the preceding algorithm. After a large number iterations of this algorithm, the statistical behavior of the ensuing

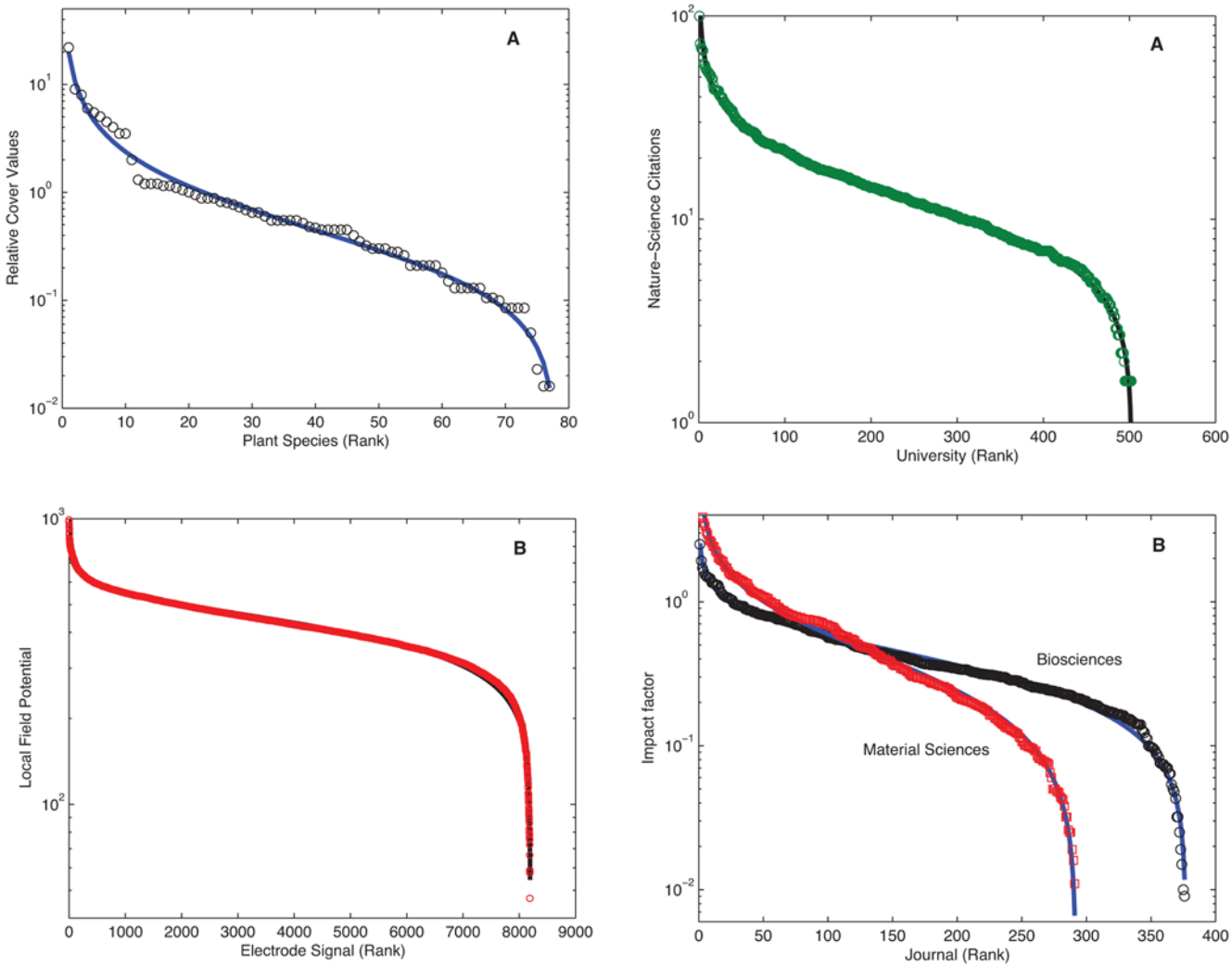


Figure 4. Rank-ordered distributions in biological systems. (A) Plot of the relative area occupied by different species in abandoned fields of Ilinois over a span of 40 years [17]. For this case $(a,b,R^2) = (0.88, 0.76, 0.98)$. **(B)** Local field potential measurements of cat cerebral cortex taken every 4 ms in an awake state, total of 8192 data points plotted in decreasing order [18] $(a,b,R^2) = (0.08,0.25,0.98)$.
doi:10.1371/journal.pone.0004791.g004

sequence can be tested by looking into the frequency-rank of *n-tuples* (non-overlapping groupings of n consecutive elements). Here we have implemented a slight variation of the algorithm described above which enhances expansion, namely *0* goes to *000* and 1 goes to *111*, both cases with probability *1-p*. This makes it somewhat more "realistic" in genetic terms. In practice we start with a 0 or 1 seed chosen with probability 0.5. After 128000 iterations the out coming sequence is treated as an initial condition and further iterated 10^6 times. The frequency with which non-overlapping sextuplets occur is then averaged over 10 realizations of this process. Fig. 7A shows this average frequency in decreasing order for two values of the modification probability p, as well as the corresponding DGBD. In Fig. 7B the values of the fitting parameters a and b are plotted against p. For p very small, $a>b$, point mutations are rare and expansion is favored, leading to extended intervals of zeros or ones; as p grows a and b eventually meet since a decreases and b increases. Above this threshold value p_{th}, $a<b$ and the higher likelihood of point mutations induces more disorder. From this perspective a is related to permanence and b to

Figure 5. Rank-ordered distributions in social phenomena. (A) Academic ranking of world Universities [19] based on the number of publications in Nature and Science,$(a,b,R^2) = (0.37,0.43,0.99)$. **(B)**. Bioscience and material science journals ordered by impact factor [20] $(a,b,R^2) = (0.59, 0.83, 0.99),(0.51,0.75,0.99)$ respectively. **(C)**. Population of the municipalities of the Spanish provinces of Zaragoza and Valladolid [21] $(a,b,R^2) = (0.95, 0.54, 0.99), (0.98,0.42,0.99)$ *respectively*.
doi:10.1371/journal.pone.0004791.g005

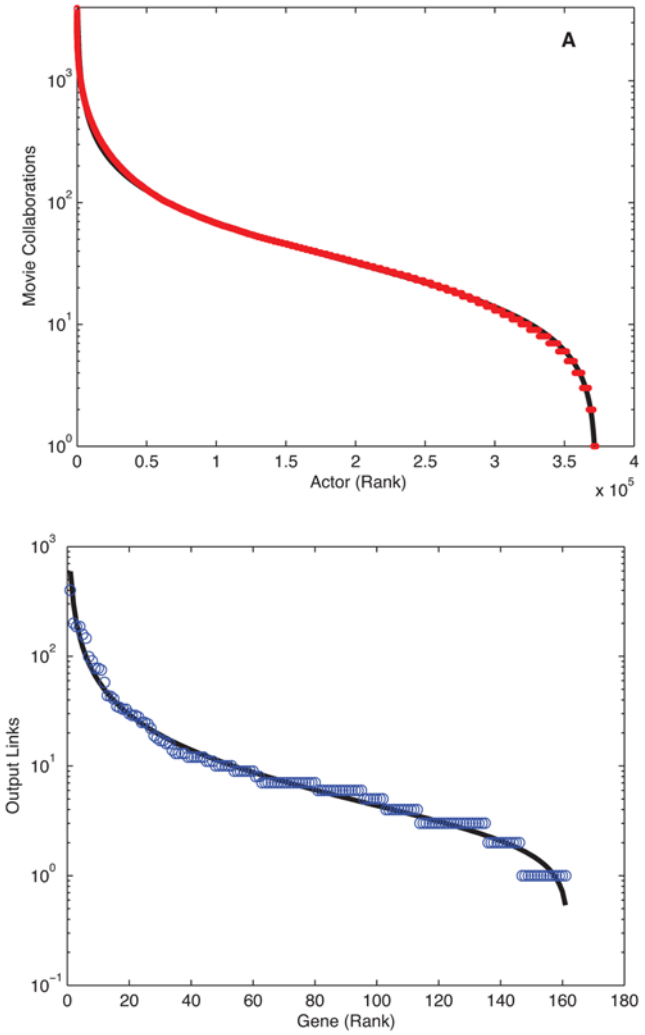


Figure 6. Rank-ordered distributions in networks. (A) Movie actor network based on the Internet Movie Database (c.f. http://www.nd.edu/~networks) containing 372,794 actors linked by movie collaborations $(a,b,R^2) = (0.71,0.61,0.99)$. **(B)** *E. coli* regulatory network nodes ordered by the number of output links based on the data of reference [22].
doi:10.1371/journal.pone.0004791.g006

change. Eventually, for values of p sufficiently large, intervals of alternating zeroes and ones start to dominate, reducing the degree of disorder and decreasing the value of b, which however continuous to be greater than a. Modifications of this model by introducing independent probabilities for mutation and modification, different expansion rates, as well as delays for mutation application, all produce sequences with good DGBD. Threshold values are sensitive to these changes and may even be absent.

This behavioral pattern is further reinforced by looking into families of deterministic discrete time evolution rules of continuous variables (mappings) where permanence relates to regular (laminar) behaviors and change appears from chaotic (turbulent) dynamics. For both the discrete models of the previous paragraph and these continuous models it can be shown that the point $a = b$ signals a disorder transition. In the first case this coincides with the end of scale invariant regions [25], in the second it marks the onset of maximum entropy.

Table 1. Fitting parameters a, b and correlation coefficient R^2 for diverse systems.

	a	*b*	R^2
Letters in English	0.18	1.31	0.97
Musical Notes in Haendel's Messiah	0.56	1.46	0.98
Area of Motifs in Malevich's Airplane Flying	1.1	0.57	0.98
Old-field Ecosystems	0.88	0.76	0.98
Local Field Potential in Cat Cerebral Cortex	0.08	0.24	0.97
Crashes of U.S. Stock Exchange	3.56	0.11	0.98
E.coli Genetic Regulatory Network	0.99	0.39	0.98
Movie Actors Network	0.71	0.61	0.99
Academic Ranking of World Universities	0.37	0.43	0.99
Biosciences Journal Impact Factor	0.59	0.83	0.99
Mexican State Population	0.44	0.68	0.99
Zaragoza Municipality Population	0.95	0.54	0.99
Valladolid Municipality Population	0.98	0.42	0.99
Chinese Province Population	0.14	0.98	0.99
Highway Distance from Guanajuato to Major Mexican Cities	1.52	3.87	0.99

Data sources are for: letters in the Concise Oxford Dictionary [29] (similar results hold for other 25 languages we have looked into), musical notes come from the musical score, relative area occupied by different species in abandoned fields of Illinois [17], journal impact factor in biosciences and material sciences journals [30], Mexican state population [31], Chinese population [32], Zaragoza and Valladolid municipality population, Mexican highways [31].
doi:10.1371/journal.pone.0004791.t001

Discussion

Overall we have encountered a universal behavior defined in terms of a functional relation for rank ordered distributions that holds accurately along the whole rank range for an impressive amount of phenomena of very diverse nature. It is not surprising that this expression goes beyond power laws since it is a two parameter relation that reduces to a power law when one of them is zero. Special interest arises when power laws require corrections due to finite size effects or other considerations. Under these circumstances they have often been modified by the inclusion of one or more additional parameters, e.g. Gaussian or exponential cut-offs. In most of the examples we have studied, though this type of correction often improves fits, our DGBD is quantitatively and above all qualitatively more satisfactory (see Fig. 8 for an example). Our main point is that, regardless of the presence of a power law, we have found a generic behavior previously not identified.

With regard to the meaning of the DGBD parameters, in some instances the exponent a can be related to behaviors generating power laws, as is the case of scale invariance in turbulence in the so called inertial range where energy is transferred between different scales at the same rate, while b seems to be associated with chaotic, disordered fluctuations, for example the dissipative range for turbulence [27]. The DGBD manages to encompass both types of regimes as well as their crossover. Further understanding of the exponents comes from our expansion-modification study where a conflicting dynamics leads to the DGBD. The expansion component which preserves a given trend is associated with a, on the other hand the modification part favors change and is related to b. Though we have shown that these conflicting permanence-change processes can produce DGBD, we are in no position to consider them as a requirement. On occasions we have perceived that parameters relations hold for certain instances, for

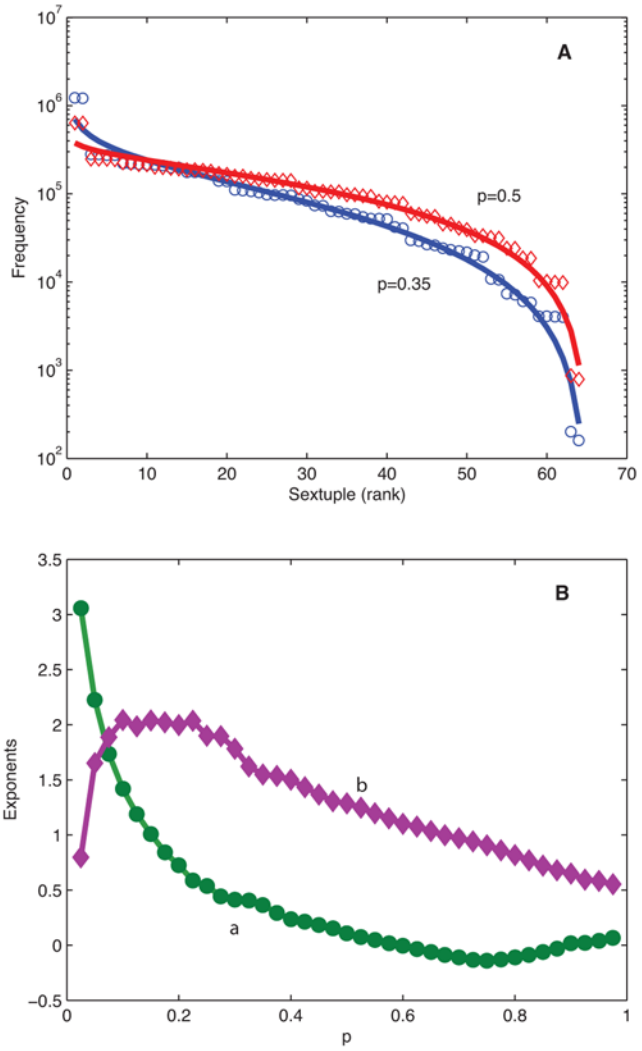


Figure 7. Frequency-rank distributions of sextuples generated by an expansion-modification algorithm. (A) Data is generated by the algorithm described in the text. Circles are determined with a modification probability $p = 0.35$, the corresponding solid line is the DGDB fit with $(a,b,R^2) = (0.36,1.55,0.96)$. For the rhomboids $p = 0.5$ and $(a,b,R^2) = (0.11,1.28,0.96)$. **(B)** shows the variation of the parameters (a,b) with probability p.
doi:10.1371/journal.pone.0004791.g007

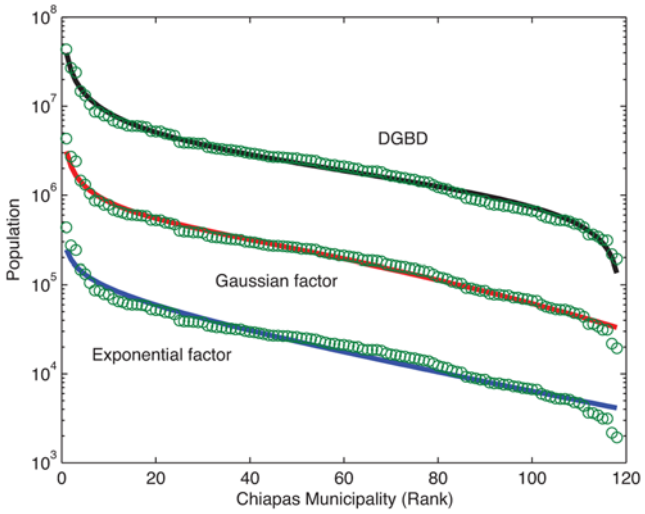


Figure 8. Two parameter fits for rank ordered data. The figure shows three fits for the population of the municipalities of the Mexican state of Chiapas [26] plotted in decreasing order. The bottom set of points corresponds to the original data, the other two sets have been obtained by successively multiplying by 10 in order to distinguish the behavior of the each fit. The top fit is the DGDB distribution, the middle one corresponds to a power law multiplied by a Gaussian factor and the bottom is a power law multiplied by an exponential factor. All fits have two adjustable parameters and produce good values for R^2, in the neighbourhood of 0.97. Notice however that the DGBD curve reproduces more successfully the overall form of the data, particularly at the two extremes.
doi:10.1371/journal.pone.0004791.g008

example the for the musical notes frequencies $a<b$ in general, while for network connectivity related situations $a>b$ is encountered more often. However, the role of exponents a and b as universality classifying parameters, as for example in critical phenomena [28], remains be investigated in further detail.

Our findings are most revealing when both parameters a and b are non-negligible and not too disparate. This usually happens for the social phenomena we have explored and which present some the most impressive fits. Based on these examples, it appears that DGBD fits are at their best when dealing with situations that result from the convergence of multiple heterogeneous processes. These are most probably weakly correlated, for example as a result of constrictions. Such considerations are in accordance with the old-field relative occupation studies previously mentioned [17] where data has been collected for various types of vegetation; we have found that the statistical behavior of each type considered separately follows less convincingly the DGBD than the integration of them shown in Fig 4A. From the above, it seems also worthwhile to analyze the role of constrictions in the art and music examples. Additionally, consideration of phenomena with processes operating at different scales, as well as multinomial multiplicative processes [24] seem promising for a better understanding of our observations. All in all, the ubiquity of our findings suggests that there ought to be a fundamental underlying explanation of a statistical nature, such as a central limit theorem extension or reformulation for the class of systems we have been encountering.

Acknowledgments

We thank A. Destexhe for providing his data on cat cerebral cortex neural activity, and M. Aldana for discussions and sharing his computing skills.

Author Contributions

Conceived and designed the experiments: GMM RAM MBdR RM PM GC. Performed the experiments: GMM RAM MBdR RM PM GC. Analyzed the data: GMM RAM MBdR RM PM GC. Contributed reagents/materials/analysis tools: GMM RAM MBdR RM PM GC. Wrote the paper: GMM RAM GC.

References

1. Barabasi A-L, Albert R (1999) Emergence of Scaling in Random Networks. Science 286: 509–512.
2. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393: 440–442.

3. Amaral LAN, Scala A, Barthélémy M, Stanley HE. Classes of small-world networks. PNAS 97: 11149–11152.
4. Newman MJ (2005) Power Laws, Pareto Distributions and Zipf's Law. Contemporary Physics 46: 323–351.
5. Sornette D (2003) Critical Markets Crashes. Physics Reports 378: 1–98.
6. Hong H, Ha M, Park H (2007) Finite-Size Scaling in Complex Networks. Phys Rev Lett 98: 258701 -1-4.
7. Watts DJ (1999) Small Worlds: The Dynamics of Networks Between Order and Randomness. Princeton, NJ: Princeton University Press. 264 p.
8. Albert R, Barabasi A-L (2002) Statistical mechanics of complex networks. Rev Mod Phys 74: 47–97.
9. Laherrere J, Sornette D (1998) Stretched exponential distribution in nature and economy: "fat tails" with characteristic scales. Eur Phys J B 2 2: 525–539.
10. Montroll EW, Shlesinger MF (1983) Maximum entropy formalism, fractals, scaling phenomena,and 1/f noise: A tale of tails. J Stat Phys 32: 209–230.
11. Zipf GK (1949) Human Behavior and the Principle of Least Effort. Cambridge, MA: Addison-Wesley Press. 573 p.
12. Quan HL, Sicilia-García EL, Ming J, Smith FJ (2000) Procceedings of the 17th International Conference on ComputerLinguistics, Montreal, 2002.
13. Ferrer i Cancho F, Sole R (2001) Two Regimes in the Frequency of Words and the Origins of Complex Lexicons: Zipf's Law Revisited, Quantitative Linguistics 8: 165–173.
14. Bury K (1999) Statistical Distributions in Engineering. Cambridge, UK: Cambridge University Press. 362 p.
15. McDonald JB (1984) Some Generalized Functions for the Size Distribution of Income. Econometrica 52: 647–664.
16. Beltrán del Rio M, Cocho G, Naumis GG (2008) Universality in the tail of musical note rank distribution. Physica A 387: 5552–5560.
17. Bazzaz FA (1975) Plant Species Diversity in Old-Field Successional Ecosystems in Southern Illinois. Ecology 56: 485–488.
18. Destexhe A, Contreras D, Steriade M (1999) Spatiotemporal Analysis of Local Field Potentials and Unit Discharges in Cat Cerebral Cortex during Natural Wake and Sleep State. Journal of Neuroscience 19: 4595–4608.
19. Academic Ranking of World Universities 2007 http://ed.sjtu.edu.cn/rank/2007/ranking2007.htm.
20. Mansilla R, Köppen E, Cocho G, Miramontes P (2007) On the behavior of journal impact factor rank-order distribution. Journal of Informetrics 1: 155–160.
21. Spanish National Statistics Institute (2003).
22. Salgado H, et al. (2006) RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. Nucleic Acds Res 34: (Database issue) D394–397.
23. Albert R, Barabasi A-L (2000) Topology of Evolving Networks: Local Events and Universality. Phys Rev Lett 85: 5234–5237.
24. Naumis GG, Cocho G (2008) Tail universalities in rank distributions as an algebraic problem: The beta-like function. Physica A 387: 84–96.
25. Li W (1991) Expansion-modification Systems: A model for Spatial 1/f Spectra. Phys Rev A 43: 5240–5260.
26. Czirok A, Mantegna RN, Havlin S, Stanley HE (1995) Correlations in Binary sequences and a Generalized Zipf's analysis.
27. Frisch U (1995) Turbulence, The Legacy of A. N. Kolmogorov. Cambridge, UK: Cambridge University Press. 296 p.
28. Kadanoff LP (2000) Statistical Physics: Statics, Dynamics and Renormalization. Singapore: World Scientific Publishing Co. 483 p.
29. Concise Oxford Dictionary 9th Edition. Oxford, UK: Oxford University Press.
30. Popescu I (2003) On a Zipf's law extension to impact factors. Glottometrics 6: 83–89.
31. Mexican National Institute of Statistics, Geography and Informatics (2003).
32. Major Figures on 2000 Population Census of China (2001) China: China Statistical Press 2001.

Apéndice C

MÉTODOS DE AJUSTE

Para encontrar los ajustes de DBDP se utilizó código escrito en lenguaje interpretado de MatLab. La entrada del programa es un vector unidimensional de cualquier longitud, el programa lo interpreta como una colección de N datos igualmente espaciados entre 0 y 1 y ajusta una DBDP, utilizando el área del vector para asignar a la DBDP una normalización, sin tratar a ésta como un tercer parámetro libre. A continuación reproducimos el código de dicho programa:

```
function [Resul, Bond]=betabeta(A)

B=(1/length(A):1/length(A):1);
A=A';
size(A);
size(B);

area=(  sum(A(2:length(A)-1)) +2*A(1)  +A(length(A))/2
        -A(2)/2 )/length(A);


Tam=1;
ord=B;

Fub=fittype('(Area/beta(-a+1,b+1))*((x)^(-a))*((T-x)^b)',
'problem',{'Area' 'T'});options=fitoptions('Exclude',
excludedata(B,A,'box',[0.000001 .999999  0 Inf]),'Lower'
,[ -Inf -0.99999 ],'Upper',[ Inf   Inf],'method',
'NonlinearLeastSquares','Robust','off','StartPoint',[.5,2],
'MaxFunEvals',100000,'MaxIter',100000,'Algorithm',
'Gauss-Newton','DiffMaxChange',.01, 'DiffMinChange',.0000000001);

[Resul,Bond] =fit(ord','A',Fub,options,'problem',{area Tam});
```

Si la cantidad de datos a a justar no es lo suficientemente grande entonces el área calculada computacionalmente como la suma de los rectángulos subyacentes puede no ser precisa (el error en área es $\Delta A \propto \frac{1}{2^N}$, con N el número de intervalos.) y entonces consideramos la normalización de la DBDP como un tercer parámetro libre. A continuación, el código del programa que utilizamos en dichos casos:

```
function [Resul, Bond]=betavect(A)
B=(1:length(A));
A=A';

size(A);
size(B);


 Tam=max(B);
ord=B;
Fub=fittype('N*((x)^(-alfa))*((T-x)^beta)','problem','T');
%Fub=fittype('()((x)^(-alfa))*((T-x)^beta)','problem','T');

options=fitoptions('Exclude',excludedata(B,A,'box',
[0 length(A)-1 0 Inf]),'Lower',[0 -Inf -Inf ],
'Upper',[Inf Inf Inf],'method','NonlinearLeastSquares',
'Robust','off','StartPoint',[1,.5,2],'MaxFunEvals',100000,
'MaxIter',100000,'Algorithm','Gauss-Newton',
'DiffMaxChange',.01,'DiffMinChange',.0000000001);

[Resul,Bond]=fit(ord','A',Fub,options,'problem',{Tam});
```

Bibliografía

[1] D. Sornette. "Critical Phenomena in Natural Sciences" Ed. Springer. 2nd Ed. pp. 163-196. (2006). ISBN 978-3540308829.

[2] J.B. McDonald. "Some generalized functions for the size distribution of income" *Econometrica*. 52-3. (1984).

[3] B. Russell. "History of Western Philosophy". Ed. Allen & Unwin. pp. 462-463. (2000). ISBN 0-415-22854-9.

[4] M.E.J. Newman "Power Laws, Pareto distribution and Zipf's law". *Contemporary Physics*, Vol.46, No5, (Septiembre-Octubre 2005) 323-351

[5] D. Sornette "Multiplicative processes and power laws". *Physical Review E.* 57:44, 4811-4813, American Physical Society, 4/1998.

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

[7] Le Quan Ha, E. I. Sicilia-Garcia, Ji Ming, F. J. Smith. "Extension of Zipf's law to words and phrases" *Proceedings of the 19th international conference on computational linguistics*. Vol. 1

[8] M. Mitzenmacher "A brief history of generative models for power laws and log-normal distributions". *Internet Mathematics*. Vol. 1, No. 2: 226-251.

[9] R. Durret, J. Schweinsberg "Power laws for family sizes in duplication model" `arXiv:Mat.PR`/0406216

[10] A. Chakraborti, M. Patriarca "A variational principle for Pareto's law" `arXiv:cond-Mat`/0605325

[11] G. Frenkel, E. Katzav, M Schwartz, N. Sochen. "Distribution of Anomalous Exponents of Natural Images" *Phys. Rev. Let.* .97, 103902 (2006)

[12] A. Fujihara, A. Othsuki, H. Yamamoto."Power-law tails in nonstationary stochastic processes with asymmetrically multiplicative interactions" *Phys. Rev. E* .70, 031106 (2004)

[13] L. A. N. Amaral, A. Scala, M. Barthélémy "Classes of small-world networks" *PNAS*10.1073/pnas.200327197

[14] M. Schroeder "Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise" Ed. W. H. Freeman (1992)

[15] V. F. Pisarenko, D. Sornette. "New statistic for financial return distributions: power-law or exponential?" *physics/0403075* (Marzo 2004)

[16] T. Bochud, D. Challet "Optimal approximations of power-laws with exponentials" `arXiv:physics`/0605149

[17] "A.L.I.C.E. Artificial Intelligence Foundation" `http://alicebot.org/superbot.html`

[18] M. Goldstein, S. Morris, G. Yen "Problems with Fitting to the Power-Law Distribution" `arXiv:cond-Mat`/0402322

[19] B. Manaris, P. Machado, C. McCauley, J. Romero, and D.Krehbiel, *Lecture Notes in Computer Science, Applications of Evolutionary Computing*, LNCS 3449, (Springer-Verlag, Berlin, 2005) pp. 498-507.

[20] B. Manaris, J. Romero, P. Machado, D. Krehbiel, T. Hirzel, W. Pharr, and R.B. Davis, *Computer Music Journal* 29 (2005) 55.

[21] M. Abramowitz, I. Stegun "Handbook of mathematical functions: with formulas, graphs, and mathematical tables" Volúmen 55 de "Applied mathematics series", Ed. Dover, 1976

[22] W. Li , *Phys. Rev. E* 43, 5240 (1991), http://www.nslij-genetics.org/wli/zipf/ (2003).

[23] R. Álvarez-Martínez, G. Martínez-Mekler,G. Cocho "Order-Disorder Transition in Conflicting Dynamics Leading to Rank-Frecuency Generalized Beta Distributions." Mandado a *Physica A*.

[24] M. Beltrán del Río, et al., "Universality in the tail of musical note rank distribution", *Physica A* (2008),doi:10.1016/j.physa.2008.05.031

[25] M. Beltrán del Río, G. Cocho Rank-Size Distribution of Notes in Harmonic Music: Hierarchic Shufflling of Distributions",*"Lecture Notes of the Institute for Computer Sciences, Social Infor- matics and Telecommunications Engineering"*, Vol.5,2222-2228,Springer Berlin Heidelberg

[26] F.D.M. Haldane ""Fractional statistics" in arbitrary dimentions: A generalized Pauli principle" *Physical Review Letters* Vol67-Num8 (1991).

[27] M. Bretz, R. Zaretzki, S.B. Field, N. Mitarai and F. Nori,*Europhysics Lett.* 74 (2006) 1116 .

[28] G. Audi, O. Bersillon, J. Blachot and A. H. Wapstra,*Nuclear Physics* A624 (1997) 124.

[29] S. Fortunato, A. Flammini, and F. Menczer, *Phys. Rev. Lett.* 96 (2006) 218701.

[30] A.C.C. Yang, S.S. Hseu, H.W Yien, A.L. Goldberger, and C.K.Peng, *Phys. Rev. Lett.* 90 (2003) 108103.

[31] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, A.L. Barabasi,*Nature* 407 (2000) 651.

[32] G.G. Naumis, G. Cocho, *Physica A* 387 (2008) 84.

[33] E.W. Montroll and M.F. Shlesinger *J. of Statistical Physics* 32 (1983) 209.

[34] R. Mansilla,G. Cocho Complex Systems. 12 (2000) 207.

[35] J. Laherrere and D. Sornette, *Eur. Phys. J.B.* 2 (1998) 525 .

[36] M.E.J. Newman, *Contemporary Physics* 46, (2005), 323-351

[37] G.G. Naumis, G. Cocho, *New J. Phys.* 9 (2007) 286.

[38] F. A. Bazzaz "Plant Species Diversity in Old-Field Successional Ecosystems in Southern Illinois" *Ecology*, Vol. 56, No. 2 (Early Spring, 1975), pp. 485-488

[39] M. Aldana, G. Cocho, H. Larralde, G. Martinez-Mekler "Translocation Properties of Primitive Molecular Machines and Their Relevance to the Structure of the Genetic Code". *Journal of Theoretical Biology* 220-1 27-45 2003

[40] G. Tarjus, et al. "Random sequential adsorption of polydisperse mixtures: asymptotic kinetics and structure" 1991 *J. Phys. A: Math. Gen.* 24 L913-L917

[41] N. Brilliantov, V. Andrienko, Yu. A. Krapivsky "Polydisperse adsorption: Pattern formation kinetics, fractal properties, and transition to order" *J. Phys. Rev.* E 58, 3530-3536

[42] R. Mansilla, E. Köppen, G. Cocho, P. Miramontes "On the Behavior of Journal Impact Factor Rank-Order Distribution" `arXiv:cs`/0610091v4 [cs.IR]

[43] L.G. Moyano, C. Tsallis, M. Gell-Mann, *Europhys. Lett.* 72, 355 (2006)

[44] B. Lewin "Genes VIII" Ed. Pearson Prentice Hall, 2004.

[45] D. Sornette "Critical Phenomena in Natural Sciences" Capítulo 3. Ed. Springer-Verlag (Springer series in Synegetics, ISSN 0172-7389) (2000)

[46] S.P. Hubbel. "A unified theory of biogeography and relative species abundance and its application to tropical rain forests and coral reefs" *Coral Reefs* 16, Suppl.:S9-S21 (1997)

[47] S.N. Majumdar "Brownian functionals in Physics and Computer Science" *Current Science*, Vol. 89, No. 12, 25 December 2005

[48] A.J. McKane, D. Alonso, R.V. Solé. "Analytic solution of Hubbell's Model of Local Community Dynamics" `arXiv:cs`/0305022 v1 (2003)

[49] G. Martínez-Mekler ,R.A. Martínez ,M. Beltrán del Río ,R. Mansilla,P. Miramontes, et al. "Universality of Rank-Ordering Distributions in the Arts and Sciences." *PLoS ONE* 4(3) (2009): e4791. doi:10.1371/journal.pone.0004791