MAXENT
Inferencia Bayesiana con Máximo de Entropía

Jorge Hermosillo

September 2023

1 Inferencia Bayesiana

La *Inferencia Bayesiana* tiene una una larga historia. El nombre proviene de Thomas Bayes (1701 - 1761), quien resolvió el problema inverso del cáclulo de Bernoulli. Pero fue en realidad Pierre-Simon Laplace (1749-1827) quien formalizó las bases de lo que hoy conocemos como el Teorema de Bayes.

1.1 Enfoque frecuentista de la probabilidad

En principio, conocer la probabilidad de un evento A requiere saber el número de ocurrencias M de A y el número total de resultados posibles donde A puede, o no, ocurrir N. $p(A) = \frac{M}{N}$. Sin embargo, en muchos casos esto es imposible; por ejemplo, ¿cómo pretender calcular la probabilidad de muerte por alguna enfermedad? En la realidad, sólo tenemos acceso a un cierto número de m observaciones del evento A como parte de un número de resultados n. La cantidad que obtenemos:

$$f(A) = \frac{m}{n} \tag{1}$$

es en realidad la frecuencia de ocurrencia de A. Ante este problema, Jacob Bernoulli (1654 - 1705) propuso una forma de relacionar la probabilidad p con la frecuencia f de A:

$$P(m|p,n) = \binom{n}{m} p^m (1-p)^{n-m} \tag{2}$$

Bayes, resolvió el problema inverso:

$$P(dp|m,n) = \frac{(n+1)!}{m!(n-m)!} p^m (1-p)^{n-m} dp \tag{3}$$

donde dp define una ventana de probabilidades de A: $p < p(A) < p+dp$ (Giffin).

1.2 Enfoque subjetivo de la probabilidad

Laplace, quien redescubrió la expresión de Bayes (3), propuso un cambio en el paradigma: las probabilidades no son cosas físicas, intrínsecas al fenómeno, sino estados de conocimiento sobre el fenómeno; las probabilidades son subjetivas. En este sentido, si dos personas tienen el mismo concomiento acerca del fenómeno, deberían asignarle el mismo valor de probabilidad.

El aspecto subjetivo del enfoque bayesiano es un cambio de paradigma en relación con la estadística frecuentista. A Laplace debemos la frase de que "*la probabilidad no es mas que sentido común reducido a cálculo*".

Laplace formalizó estas ideas de manera elegante en el teorema de Bayes. Su razonamiento iba en este sentido. Supongamos un conjunto de eventos $X = \{x_1, x_2, ..., x_n\}$. Las posibles causas, o hipótesis, que dieron origen a estos eventos pueden estar representadas por $H = \{h_1, h_2, ..., h_N\}$. Laplace propuso que la probabilidad de una hipótesis h_i dado X era la razón entre la probabilidad condicional $P(X|h_i)$ y la probabilidad de X. El término $P(X|h_i)$ es la plausibilidad (o verosimilitud - *likelihood*) de que la hipótesis h_i explique los datos X. Sin embargo, ¿que ocurre si las hipótesis no son todas equiprobables? Hay que pesar cada término $P(X|h_i)$ por el grado de verosimilitud de la hipótesis, en función de nuestro conocimiento *a priori* (I - de información). Por lo tanto, si tomamos en cuenta toda información I que disponemos a priori para evaluar la plausibilidad de h_i, podemos ponderar el numerador por un término $P(h_i|I)$, dando como resultado:

$$P(h_i|XI) = \frac{P(h_i|I)P(X|Ih_i)}{\Sigma_{j=1}^{N} P(h_j|I)P(X|Ih_j)} = \frac{P(h_i|I)P(X|Ih_i)}{P(X|I)} \tag{4}$$

1.3 Predicción vs. inferencia

¿Qué método es el correcto? ¿Debemos apegarnos a los cánones de la estadística frecuentista si queremos hacer predicciones sobre aspectos "reales" del mundo? O ¿deberíamos tomar en cuenta conocimiento a priori en nuestros modelos, para saber si los datos aportan nueva información, y eventualmente, nos permiten cambiar nuestros puntos de vista?

En la Inglaterra del siglo XIX, dos posturas filosóficas se debatían el enfoque correcto acerca de la función del método científico. El empirismo, que era dominante, afirmaba que la experiencia debía ser la única guía. El problema es que esta era subjetiva, una creación de la mente del observador (postura idealista). En un planteamiento similar, que hoy podríamos calificar de "cimentado", el fenomenalismo desarrollado por Karl Pearson (1892) y Ernst Mach (1883) afirmaba que nada puede suponerse existente que no pueda reducirse a una descripción de sensaciones. Por otra parte, el realismo era una teoría del conocimiento que sostenía que el mundo externo existe independiente del observador y que la función del método científico es descubrir sus propiedades (Geisser, 1978).

Hoy en día, este debate se manifiesta en forma más sutil entre la estadística "ortodoxa" y la bayesiana. La primera concentra la atención al 100% en la predicción física, mientras que el enfoque bayesiano se centra en la inferencia. Para apreciar la distinción entre predicción física e inferencia es esencial reconocer que están implicadas proposiciones a dos niveles diferentes. En la predicción física intentamos describir el mundo real; en la inferencia sólo describimos nuestro estado de conocimiento sobre el mundo. Un filósofo diría que la predicción física opera a nivel ontológico y la inferencia a nivel epistemológico. En el nivel epistemológico, un cálculo bayesiano nos da sólo las mejores predicciones que se pueden hacer a partir de la información que se ha utilizado en el cálculo. Pero siempre es posible que en el mundo real existan factores de control adicionales de los que no éramos conscientes (Skilling).

En otras palabras, la estadística bayesiana asume que los modelos que usamos para hacer predicciones pueden, y suelen, ser incompletos, en el sentido de que en el mundo real pueden haber variables ocultas de las que nuestro modelo no da cuenta. En virtud de ello, nuestras predicciones pueden ser erróneas. ¿Debemos rechazar el modelo por ello? Pero, ¿de qué otra forma habríamos podido conocer esos factores desconocidos? Sólo cuando fallan nuestras predicciones epistemológicas aprendemos cosas nuevas sobre el mundo real; ésos son precisamente los casos en los que la teoría de la probabilidad desempeña su función más valiosa como lo han demostrado innumerables ejemplos en Física, particularmente en el campo de la Mecánica Cuántica (Skilling, Jaynes).

Así pues, para el enfoque bayesiano, la teoría de la probabilidad no es un oráculo que nos dice cómo debe ser el mundo; es una herramienta para aprender. En este sentido, replantea las preguntas: ¿Es nuestro estado de conocimiento adecuado para describir el mundo? y ¿Para qué aspectos del mundo es nuestra información adecuada para hacer predicciones? (Skilling).

1.4 La probabilidad como lógica

En la visión subjetiva de la probabilidad, las distribuciones que utilizamos para la inferencia no describen ninguna propiedad del mundo, sino sólo un cierto estado de información sobre el mundo. No se trata sólo de una posición filosófica, sino que nos proporciona importantes ventajas técnicas debido a la mayor flexibilidad con la que podemos utilizar la teoría de la probabilidad.

En 1988, Edwin Jaynes (1922 - 1998) llevó este paradigma más allá del campo estadístico, proponiendo la probabilidad bayesiana como una extensión de la lógica, y como el mecanismo formal idóneo de la inferencia científica (Jaynes).

En este marco conceptual, las variables de un modelo bayesiano son proposiciones que pueden tener un valor lógico entre 0 y 1. Esta concepción de la inferencia bayesiana, fue motivo de inspiración para algunos investigadores en Inteligencia Artificial que propusieron este marco epistemológico y formal para diseñar agentes inteligentes (Bessiere). En este marco teórico, es posible diseñar agentes de muy diversa índole y complejidad, con la ventaja de poder explicar

los comportamientos del agente y vencer algunas de las críticas epistemológicas contra el razonamiento simbólico (Harnad), sentando las bases para el desarrollo de modelos computacionales del comportamiento animado (Bessiere).

Sin embargo, existe un escollo de carácter pragmático. Observando el Teorema de Bayes (4) uno de los principales problemas es ¿cómo determinar la distribución a priori de nuestras hipótesis? ¿Cómo tomar en cuenta información a priori sin comprometer las predicciones del modelo con datos inexistentes? Más aún, bajo el enfoque subjetivo, una probabilidad es una construcción teórica, a nivel epistemológico, que asignamos para representar un estado de conocimiento, o que calculamos a partir de otras probabilidades según las reglas de la teoría de la probabilidad. Una frecuencia es una propiedad del mundo real, en el plano ontológico, que medimos o estimamos. ¿Cómo conciliar estos aspectos aparentemente antagónicos y problemáticos desde un punto de vista formal? En 1988, Jaynes propuso el principio del máximo de entropía (MAXENT) como un mecanismo formal, teóricamente sólido y epistemológicamente válido, en el sentido de que arroja datos consistentes tanto en la Mecánica Estadística, como en la Teoría de la Información, en la Biología y en la Astrofísica entre otros campos del conocimiento.

2 MAXENT

2.1 Entropía

En 1948, Claude Shannon (1916 - 2001) propuso la entropía como una medida de la cantidad de información promedio que un determinado canal puede transmitir de manera óptima en presencia de ruido. El supuesto es que la salida del canal depende probabilísticamente de la entrada: $p(Y|X)$, donde X es el mensaje codificado transmitido y Y el mensaje codificado recibido.

Desde un punto de vista estadístico, un mensaje puede verse como una variable aleatoria (VA) X, un conjunto de símbolos x formando una cadena a transmitir, cada símbolo pudiendo ocurrir con cierta probabilidad $p(x)$. Como en general existe una dualidad entre la compresión de un mensaje, que se logra eliminando toda redundancia, y la precisión de la transmisión, que se logra añadiendo redundancia controlada para que la entrada pueda recuperarse incluso en presencia de ruido, el objetivo es codificar el mensaje de tal manera que ocupe el mínimo espacio y, al mismo tiempo, contenga suficiente redundancia para poder detectar y corregir errores.

De esta forma, la entropía puede considerarse como la longitud media del mensaje necesario para transmitir un resultado de esa variable. En este mismo sentido, pero de manera general, la entropía $H(p) = H(X)$ es la cantidad de información de una variable aleatoria, y se define matemáticamente como (Manning):

$$H(p) = H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{5}$$

2.2 Principio de máximo de entropía

Alternativamente, la noción de entropía puede verse como una *medida de cuánto se desconoce acerca de un sistema*. Esta noción de *conocimiento* sobre un sistema, puede traducirse en la pregunta: ¿qué distribución de probabilidad gobierna una VA X? Pongamos un ejemplo clásico: nos presentan un dado cualquiera y nos dicen que se lanzó $N = 1000$ veces. ¿Podemos estimar la frecuencia con que cada cara cayó en esos 1000 lanzamientos (Ecuación 1)? Instintivamente, o por algún tipo de sentido común quizá, apostaríamos por una distribución uniforme, prediciendo $\frac{1}{6}$ para cada cara. Si nos piden justificar la respuesta, no nos quedaría más remedio que argumentar el "Principio de Indiferencia" (*Principle of Insufficient Reason*) de Bernoulli, que consiste en decir que, ante la ausencia de cualquier otro elemento de información, no hay razones para suponer que alguna cara deba ser más favorecida que otra. ¿Cómo podemos formalizar este principio de indiferencia?

Se ha dicho, desde hace mucho tiempo, que la expresión (5) define una especie de medida en el espacio de distribuciones de probabilidad, de tal forma que aquellas con alta entropía se prefieren sobre las otras. Esto se ha traducido en declaraciones como: las distribuciones de más alta entropía representan mayor "desorden", son "más suaves", "más probables", "menos predecibles", que "presuponen menos" según la interpretación de Shannon de la entropía como una medida de información. Jaynes propuso una forma muy elegante y convincente de demostrar estos argumentos, utilizando el teorema de concentración de entropía (Jaynes, 1982).

Supongamos que un experimento aleatorio puede producir n valores posibles en cada ensayo; por lo que en N ensayos, tenemos n^N resultados posibles. Cada resultado produce un conjunto de muestras $\{N_i\}$ y frecuencias $\{f_i = N_i/N, 1 \leq i \leq n\}$, con entropía

$$H(f_1, f_2, \ldots, f_n) = -\sum_{i=1}^{n} f_i \ln(f_i) \tag{6}$$

Por ejemplo:

1. *Dado cargado en general*: Un dado con n caras se lanza N veces, cada cara saliendo N_i veces.
2. *Mecánica estadística*: Un sistema contiene N moléculas, de las cuáles N_i se encuentran en el estado cuántico i.
3. *Comunicación*: Recibimos un mensaje de N símbolos, elegidos de un alfabeto de n letras, la i-ésima letra ocurriendo N_i veces.

4. *Serie de tiempo*: La naturaleza genera N instancias de la serie de tiempo $Y \equiv \{y_0, y_1, \cdots, y_T\}$, de las cuáles n secuencias diferentes son posibles $\{Y^{(1)}, \cdots, Y^{(n)}\}$. La secuencia $Y^{(i)} = \{y_0^{(i)}, \cdots, y_T^{(i)}\}$ se produce N_i veces.

Consideremos una subclase C de todos los posibles resultados que se pueden observar en N ensayos, compatibles con m restricciones ($m < n$) linealmente independientes de la forma

$$\sum_{i=1}^{n} A_{ij} f_i = d_j \qquad\qquad 1 \leq j \leq m. \tag{7}$$

Conceptualmente, esto quiere decir que se han medido m cantidades físicas, cuya "naturaleza" está codificada en A_{ij} y $D = \{d_1, \cdots, d_m\}$ es nuestro conjunto de datos. Por ejemplo, una restricción para el lanzamiento de un dado no cargado es que la suma de sus frecuencias sea 1. El problema es estimar los valores de frecuencia f_i a partir de los datos observados d_j, es decir, a partir de la información que nos dan las restricciones. Jaynes analizó la base combinatoria de usar (6) para estimar las f_i.

Una fracción F de los resultados en la clase C producirá una entropía en el rango

$$H_{max} - \Delta H \leq H(f_1, \ldots, f_n) \leq H_{max} \tag{8}$$

donde H_{max} se determina por un algoritmo de optimización (Jaynes, 1982). La relación entre F y ΔH está dada por el teorema de concentración de entropía. Este teorema afirma que, asintóticamente, $2N\Delta H$ sigue una distribución χ^2_k con $k = n - m - 1$ grados de libertad sobre la clase C, independientemente de la naturaleza de las restricciones

$$2N\Delta H = \chi^2_k(1 - F) \tag{9}$$

donde $1 - F$ es el nivel de significancia (el área de la cola superior de la curva).

El teorema tiene una base combinatoria sobre el número de resultados posibles que puede generar un conjunto de frecuencias observadas (la fracción F en la clase C). En el caso del dado, por ejemplo, el número total de resultados posibles en N ensayos es 6^N. Dado un conjunto particular de frecuencias $\{f_1, \cdots, f_n\}$, el número de resultados que arrojaría este mismo conjunto se calcula

$$W(f_1, \cdots, f_n) = \frac{N!}{(Nf_1)! \cdots (Nf_n)!} \tag{10}$$

y conforme $N \to \infty$ se tiene la aproximación

$$N^{-1} \log W \to H(f_1, \cdots, f_n)$$

Dadas dos conjuntos de frecuencias $\{f_i\}$ y $\{f'_i\}$ la razón (número de maneras en que se puede obtener f_i)/(número de maneras en que se puede obtener f'_i) es asintóticamente

$$\frac{W}{W'} \sim A \exp(N[H - H']) \tag{11}$$

Sigamos con el ejemplo de un dado. Suponiendo un dado no cargado, no tenemos restricción alguna; tan solo sabemos que la suma de las frecuencias debe ser 1 ($\sum f_i = 1$). En este caso, el máximo de entropía $H_{max} = \ln 6 = 1.792$, que corresponde a la distribución uniforme. Aplicando el teorema de concentración de entropía, tenemos $6 - 1 = 5$ grados de libertad. Con un nivel de significancia del 5%, tenemos $\chi^2_5(0.05) = 11.07$. Así, el 95% de todos los resultados posibles tienen una entropía en el rango $2N\Delta H = 11.07$, o

$$1.786 \leq H \leq 1.792$$

Más aún, considerando el 99.5% de todos los resultados posibles ($\chi^2_5(0.05) = 16.75$), estos tendrían una entropía en el rango

$$1.783 \leq H \leq 1.792 \tag{12}$$

Supongamos ahora que tenemos como nueva información que después de 1000 lanzamientos, el valor esperado del dado es

$$\sum_{i=1}^{6} i f_i = 4.5 \tag{13}$$

que es un caso particular de (7) . Aplicando el método MAXENT[1] de (Jaynes, 1982) se obtiene:

$$\{f_1, \cdots, f_6\} = \{0.0543, 0.0788, 0.1142, 0.1654, 0.2398, 0.3475\} \tag{14}$$

con

$$H_{max} = 1.61358,$$

muy por debajo del rango (12).

Nuevamente, aplicando el teorema de concentración de entropía, tenemos ahora $6 - 1 - 1 = 4$ grados de libertad. Buscando en las tablas $\chi^2_4(0.05) = 9.488$, tenemos que para el 95% de los resultados posibles $\Delta H = 9.488/2000 = 0.00474$ o bien

$$1.609 \leq H \leq 1.614$$

Considerando el 99.99% de los resultados posibles, este rango se amplía

$$1.602 \leq H \leq 1.614$$

En un caso extremo, sólo uno en 10^8 resultados posibles tendría una entropía fuera de

$$1.592 \leq H \leq 1.614 \tag{15}$$

Lo que todo esto nos dice, es que dada *información incompleta (o parcial)*, la distribución que maximiza la entropía no sólo es la que se puede realizar en la

[1] Se demuestra el método en el Apéndice A.

mayor cantidad de formas posibles, sino que para N grande, la gran mayoría de todas las distribuciones *posibles* compatibles con nuestra información tienen una entropía muy cercana a la máxima.

Jaynes nos da un ejemplo de esta afirmación, comparando el número de resultados posibles de una distribución que rechaza el máximo de entropía, pero que cumple la restricción (13)

$$f_i' = \binom{5}{i-1} p^{i-1}(1-p)^{6-i}, \qquad 1 \leq i \leq 6 \tag{16}$$

Dicha distribución tiene una entropía $H' = 1.4136 = H_{max} - 0.200$, muy lejos del límite (15). Ahora tenemos $2N\Delta H = 400 = \chi^2_4(1-F)$, donde $1-F \approx 2.94 \times 10^{-84}$ (Jaynes, 1982). Esto quiere decir que, en 1000 lanzamientos, menos de uno en 10^{83} de los resultados posibles es compatible con la restricción (13). Desde otro punto de vista, de acuerdo con (11) por cada forma en que (16) se puede obtener, existen alrededor de $\exp(N\Delta H) = \exp(200)$, o más de 10^{86} maneras, en que la distribución del máximo de entropía (14) se puede realizar (cerca de 10^{62} maneras por cada microsegundo de la edad del universo). Incluso para N pequeño, las fórmulas son exactas (no asintóticas). Para $N = \sum N_k = 50$, por ejemplo, considerando $N_k = \{3, 4, 6, 8, 12, 17\}$ y $N_k' = \{0, 1, 7, 16, 18, 8\}$, $W/W' = (7!\,16!\,18!)/(3!\,4!\,6!\,12!\,17!) = 38\,220$ maneras más en que MAXENT $\{N_k\}$ se puede realizar.

En conclusión, MAXENT se basa en la idea de que *cuando hacemos inferencias sobre la base de información incompleta, deberíamos extraerlas de aquella distribución de probabilidad que tenga la máxima entropía permitida por la información de la que disponemos* (Jaynes 1982, p.940). En otras palabras, la distribución que maximiza la entropía de la información disponible es la estadísticamente más probable. Cuando se ha tenido en cuenta toda la información conocida, un sistema con la máxima entropía informativa es el estado más probable porque es el sistema en el que se ha definido la menor cantidad de información.

En este sentido, MAXENT proporciona una forma de asignar probabilidades a un fenómeno de tal manera que representan mejor nuestro estado de conocimiento, sin comprometerse con información no disponible. En otras palabras, la entropía traduce la información en asignación de probabilidades, y la elección del modelo con mayor entropía evita la introducción arbitraria de información desconocida.

2.3 MAXENT, Tranformada de Fourier y Autoregresión

Las consideraciones anteriores aplican por igual a series de tiempo con valores reales. Note que en el caso de los N lanzamientos de un dado, no existen correlaciones entre los resultados de cada ensayo. MAXENT no asume correlación alguna en ese caso, es decir, no introduce información inexistente. En el caso de las series de tiempo, sin embargo, Jaynes considera que además de contar con restricciones del tipo (7) la serie de tiempo puede proporcionar información sobre

propiedades mutuas entre los datos en distintos instantes de tiempo, induciendo correlaciones de las que MAXENT puede dar cuenta esta vez.

Para el presente análisis, usaremos la siguiente convención de escritura:

$A \equiv$ valor real, conocido o desconocido.

$A' \equiv$ valor numérico proporcionado como planteamiento del problema: los "datos".

$\widehat{A} \equiv$ valor estimado que calculamos, usualmente el valor promedio sobre una distribución MAXENT.

En el caso de una serie de tiempo, las restricciones a las que está sometida la entropía representan toda la información sobre la serie de tiempo. Si estas restricciones toman la forma de valores promedio de m cantidades A_k, el conjunto de datos $D = \{A'_1, \cdots, A'_m\}$ impone restricciones

$$A'_k = \int dy_0 \cdots \int dy_T p(p_0 \cdots y_T) A_k(y_0 \cdots y_T) \qquad 1 \leq k \leq m \tag{17}$$

Siguiendo el algoritmo MAXENT, se llega a la expresión del máximo de entropía

$$H_{max} = S(A'_1 \cdots A'_m) = \ln Z + \lambda A'$$

Conociendo las funciones Z y A' se puede calcular $\lambda_k = \partial S / \partial S'_k$.

Jaynes, demostró que MAXENT se puede aplicar al análisis de una serie de tiempo para la que se miden valores R'_k de la autocovarianza[2]:

$$R_k(y_0 \cdots y_T) = \frac{1}{T+1} \sum_{j=0}^{T-k} y^\star y_{j+k} \qquad\qquad 0 \leq k \leq m \tag{18}$$

para $m+1$ retrasos, donde $m < T$, y donde se plantea $R_{-k} = R^\star_k$.

Sin ninguna otra información, la densidad de probabilidad con máxima entropía que genera la autocovarianza correcta contendrá los multiplicadores de Langrange $\{\lambda_{-m} \cdots \lambda_0 \cdots \lambda_m\}$. En virtud de que los coeficientes A_k se pueden definir de manera arbitraria, Jaynes propone hacerlos independientes de T

$$A_k = \frac{T+1}{2} R_k \qquad\qquad -m \leq k \leq m$$

Siguiendo el método de Jaynes, se determinan los λ_k a partir de $A'_k = -\partial \ln Z / \partial \lambda_k$ o bien

$$R''_k = \frac{1}{2\pi} \int_0^{2\pi} \frac{e^{ik\theta} d\theta}{g(e^{i\theta})} \qquad\qquad -m \leq k \leq m \tag{19}$$

Cuando k está en la región de información proporcionada ($-m \leq k \leq m$), (19) representa las restricciones $R''_k = R'_k$ que determinan los multiplicadores

[2] Ver Apéndice B

de lagrange λ_k. Cuando $|k| > m$, (19) representa la extrapolación predicha por MAXENT de la función de covarianza: $R''_k = \widehat{R}_k$.

Suponiendo que la información proporcionada es un intervalo I; es decir, nos dan R'_k para $k \in I$, entonces, si definimos

$$g(z) \equiv \sum_{k \in I} \lambda_k z^k$$

La interpretación de (19) es la siguiente:

$$R''_k = \left\{ \begin{array}{ll} \text{datos } R'_k, & k \in I \\ \text{predicción } \widehat{R}_k, & \text{de lo contrario} \end{array} \right\} \tag{20}$$

Una vez que se han determinado los λ_k, la predicción MAXENT de cualquier propiedad de la serie de tiempo se obtiene de su distribución. En este caso, la función de partición es

$$\ln Z = -\frac{1}{2} \sum_{j=0}^{T} \ln g_j + (const)$$

donde g_j son los eigenvalores de la matriz Λ, en su forma Toeplitz, de $T+1$ filas y columnas

$$\Lambda = \left\{ \begin{array}{ll} \lambda_{j-i} & |j-i| \leq m \\ 0 & |j-i| > m \end{array} \right\}$$

La distribución de MAXENT es

$$p(y_0 \cdots y_T) \propto \exp\left[-\frac{1}{2} (y^\dagger \Lambda y) \right]$$

donde y es el vector columna $(y_0 \cdots y_T)$ y $y^\dagger$ es su vector fila Hermitiano (complejo conjugado transpuesto) $(y_0^\star \cdots y_T^\star)$.

Un resultado importante que se deriva del proceso mismo de MAXENT (sin haber supuesto ningún otro tipo de información) es que esta distribución es una Gaussiana multivariada.

A partir de estos resultados, es posible calcular el espectro de potencias de la serie de tiempo, usando la autocorrelación:

$$P(f) = \sum_{k=-\infty}^{k=\infty} R_k \exp(+i2\pi f k) \qquad |f| \leq \frac{1}{2}$$

pero (19) es la inversión de esta serie de Fourier, y por inspección, se tiene $P(f) = 1/g$; por lo que el espectro predicho es

$$\widehat{P}(f) = \frac{1}{\sum_{k \in I} \lambda_k e^{-i2\pi f k}} \qquad |f| \leq \frac{1}{2} \tag{21}$$

A Demostración de MAXENT

Para demostrar MAXENT recurrimos a (Jaynes, 1978). La demostración es una aplicación de métodos variacionales utilizando multiplicadores de lagrange.

El método de Lagrange de optimización de una función $f(x)$ sujeta a una restricción $g(x) = c$ parte del principio de que el máximo de f se encuentra calculando el gradiente de f, ∇f. Este gradiente nos indica los máximos de f cuando es $\nabla f = 0$. En dimensiones $dim(f) > 2$ el gradiente de f define contornos de nivel, donde el valor de la derivada es el mismo. Dado que la restricción $g(x) = c$ debe cumplirse, esto significa que g debe ser tangente a f en algún punto. En ese punto los gradientes de f y g son paralelos. Así, el valor máximo de $f(x)$ que cumple con la restricción $g(x) = c$ es este punto de contacto. La condición numérica que debe verificarse, es $\nabla f = \lambda \nabla g$. En este caso, el langrangiano es $\mathcal{L}(x, \lambda) = f(x) - \lambda g(x) + c$.

El problema es estimar el valor λ calculando:

$$\nabla \mathcal{L} = 0.$$

Consideremos una variable aleatoria X con n resultados posibles $D = \{f(x_1), \cdots f(x_n)\}$. Vamos a considerar que podemos calcular el valor esperado

$$d = \sum_{i=1}^{n} p_i f(x_i)$$

para la cuál, las probabilidades p_i se desconocen; tan sólo podemos plantear que $\sum_i p_i = 1$. La entropía de este sistema es

$$H(p_1 \cdots p_n) = -\sum_{i=1}^{n} p_i \ln p_i$$

El langrangiano que nos interesa resolver es

$$\mathcal{L}(p_i, \mu, \lambda) = -\sum_{i=1}^{n} p_i \ln p_i - \sum_{i=1}^{n} \mu p_i f(x_i) - \sum_{i=1}^{n} \lambda p_i + d \cdot \mu + \lambda \tag{22}$$

Obteniendo el gradiente e igualando a 0

$$\frac{\partial \mathcal{L}}{\partial p_i} = -(\ln p_i + 1) - \mu f(x_i) - \lambda = 0 \tag{23}$$

$$\frac{\partial \mathcal{L}}{\partial \mu} = -\sum_{i=1}^{n} p_i f(x_i) + d = 0 \tag{24}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = -\sum_{i=1}^{n} p_i + 1 = 0 \tag{25}$$

vemos que (24) corresponde a d, el valor esperado que se da como dato del problema, (25) corresponde a la restricción sobre la suma de las probabilidades, y por último, de (23) podemos escribir

$$p_i = e^{-1-\lambda-\mu f(x_i)}$$

Sustituyendo en las otras ecuaciones, tenemos

$$\sum_{i=1}^{n} e^{-1-\lambda-\mu f(x_i)} = 1 \tag{26}$$

$$\sum_{i=1}^{n} f(x_i) e^{-1-\lambda-\mu f(x_i)} = d \tag{27}$$

Pongamos

$$Z(\mu) = \sum_{i=1}^{n} e^{-\mu f(x_i)}$$

llamada función de partición, en alusión al concepto utilizado en Mecánica Estadística; Z es una constante de normalización. De (26) tenemos

$$\lambda = \ln \sum_{i=1}^{n} e^{-\mu f(x_i)} - 1 = \ln Z(\mu) - 1$$

es decir,

$$Z = e^{1+\lambda} \Rightarrow \frac{1}{Z} = e^{-1-\lambda}$$

Pero,

$$\frac{\partial Z}{\partial \mu} = -\sum_{i=1}^{n} f(x_i) e^{-\mu f(x_i)}$$

Por lo que (27) se escribe

$$d = -\frac{1}{Z}\frac{\partial Z}{\partial \mu} = \frac{1}{Z}\sum_{i=1}^{n} f(x_i) e^{-\mu f(x_i)}$$

y podemos reescribir

$$p_i = \frac{1}{Z}\exp(-\mu f(x_i)) \tag{28}$$

Sustituyendo estos valores en la expresión de la entropía, tenemos

$$H_{max} = -\sum_{i=1}^{n}\left[\frac{1}{Z}e^{-\mu f(x_i)} \ln(\frac{1}{Z}e^{-\mu f(x_i)})\right] = \underbrace{\sum_{i=1}^{n}\frac{1}{Z}e^{-\mu f(x_i)}}_{=1}[\ln Z + \mu f(x_i)]$$

luego

$$H_{max} = \ln Z + \mu \underbrace{\frac{1}{Z}\sum_{i=1}^{n} f(x_i)e^{-\mu f(x_i)}}_{=d}$$

por lo que finalmente

$$H_{max} = \ln Z + \mu \cdot d \tag{29}$$

De la ecuación (29) calculamos μ de

$$\frac{\partial H_{max}}{\partial \mu} = \frac{\partial}{\partial \mu} \ln Z + d = 0$$

o bien

$$d = -\frac{\partial}{\partial \mu} \ln Z \tag{30}$$

Estos cálculos se pueden extrapolar a un mayor número m de restricciones, suponiendo de manera general que éstas se pueden definir como (7).

En este caso, tendremos $\{\lambda_1 \cdots \lambda_m\}$ multiplicadores de langrange, y la función de partición es

$$Z(\lambda_1 \cdots \lambda_m) = \sum_{i=1}^{n} \exp\left(-\sum_{k=1}^{m} \lambda_k A_{ki}\right). \tag{31}$$

Entonces

$$H_{max} = S(f_1 \cdots f_n) = \ln Z + \sum_{k=1}^{m} \lambda_k d_k, \tag{32}$$

y los parámetros λ_k se obtienen de

$$d_k = \frac{\partial}{\partial \lambda_k} \ln Z, \tag{33}$$

Por lo que ahora, la distribución de las frecuencias $f(x_i)$ es

$$f_i = \frac{1}{Z} \exp\left(-\sum_{k=1}^{m} \lambda_k A_{ki}\right) \tag{34}$$

Otras distribuciones $\{f_i'\}$ permitidas por las restricciones (7) tendrán varias entropías menores a H_{max}.

B Covarianza, correlación y complejos conjugados

En términos generales, la autocovarianza se calcula como

$$Cov(r_t, r_{t-s}) = \frac{1}{N-1}\sum_{t-1}^{T}(r_t - \bar{r})(r_{t-s} - \bar{r}),$$

que se relaciona con la autocorrelación de la siguiente forma:

$$\rho_s = \frac{Cov(r_t, r_{t-s})}{Var(r_t)}$$

Complejos conjugados.

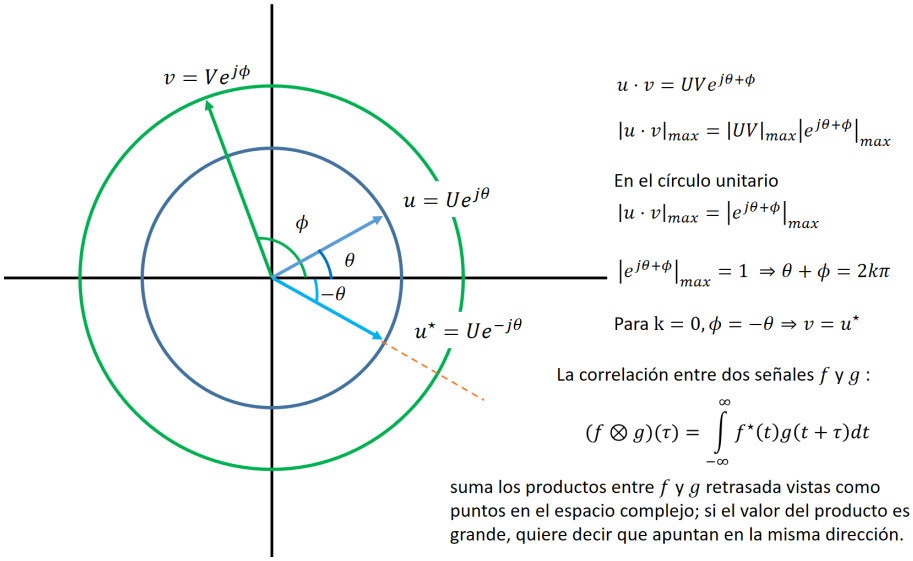

Figure 1: Interpretación gráfica de la correlación.