Explaining the Gibb Sampler

George Casella, Edward I. George

Entre los métodos computacionales que han ayudado al desarrollo de la estadística tenemos al Muestreador de Gibb, que es una técnica que genera variables aleatorias indirectamente de distribuciones marginales sin tener que calcular la densidad. Este algoritmo se basa en las propiedades principales de las Cadenas de Markov. Aunque normalmente realcionado con la estadística Bayesiana, el Muestreador de Gibb también es útil en la visión clásica de la estadística.

Supongamos que tenemos una distribución conjunta $f(x, y_1, y_2, ..., y_p)$

$$f(x) = \int \cdots \int f(x, y_1, y_2, ..., y_p) dy_1, dy_2, ..., dy_p$$

Y nos interesan las características de la densidad marginal como la media o la varianza de x, con el Muestreador de Gibb podemos generar una muestra $X_1, ..., X_m \sim f(x)$ sin requerir calcular $f(x)$ directamente y obteniendo la media o la varianza con suficiente precisión.

Para explorar con detalle como funciona el Muestreador de Gibb, se toman dos variables aleatorias (X, Y) y el Muestreador de Gibb genera una muestra de $f(x)$ muestreando las distribuciones condicionales $f(x|y)$ y $f(y|x)$ que normalmente son conocidas en los modelos estadísticos. Esto se logra generando una "secuencia de Gibb"de variables aleatorias donde los valores iniciales son especificados y el resto se obtiene de manera iterativa generando así valores para

$$X'_j \sim f(x|Y'_j = y'_j)$$
$$Y'_{j+1} \sim f(y|X'_j = x'_j)$$

Esto es lo que se llama muestreo de Gibb, si $k \to \infty$ la distribución de X'_k convergerá con la verdadera distribución marginal de X ($f(x)$).

El Muestreador de Gibb se puede pensar como una implementación práctica del conocimiento de que el conocimiento de las distribuciones marginales es suficiente para conocer la distribición conjunta y aunque esto parezca claro para casos bivariados no es tan directo para los casos multivariados.

Suponemos dos variables aleatorias X, Y, de las cuales sabemos sus distribuciones condicionales $f_{X|Y}(x|y)$ y $f_{Y|X}(y|x)$. A partir de estas podríamos calcular la función marginal de X y la distribución conjunta de ambas variables, mediante el siguiente argumento:

$$f_X(x) = \int f_{XY}(x, y) \quad dy$$

donde $f_{XY}(x, y)$ aún es desconocida. Si usamos el hecho que $f_{XY}(x, y) = f_{X|Y}(x|y)f_Y(y)$ tendríamos que,

$$f_X(x) = \int f_{X|Y}(x|y)f_Y(y) \quad dy$$

asimismo, si sustituimos $f_Y(y)$,

$$\begin{aligned} f_X(x) &= \int f_{X|Y}(x|y)f_{Y|X}(y|t)f_X(t) \quad dtdy \\ &= \int [\int f_{X|Y}(x|y)f_{Y|X}(y|t)dy]f_X(t)dt \\ &= \int h(x, t)f_X(t)dt \end{aligned}$$

Esto se llama una ecuación integral con un punto fijo que tiene como solución $f_X(x)$. Esta ecuación es una forma limitada de la iteración de Gibbs, ilustrando como las distribuciones condicionales producen una distribución marginal. Aunque la distribución conjunta de X, Y determinan las condicionales y las marginales, no siempre las condicionales determinen de manera tan directa la distribución marginal.

En cuantas más variables existan, el problema se vuelve más complejo pues la relación entre las condicionales, marginales y conjuntas se vuelve más intrincada. Por ejemplo, la relación $condicional \times marginal = conjunta$ no

se sontiene para todas las condicionales y marginales. Pero se pueden hacer varios conjuntos de variables para construir las ecuaciones integrales con un punto fijo para calcular la distribución marginal de interés.

Supongamos que tenemos las variables aleatorias X, Y, Z y queremos la distribución $f_X(x)$, la ecuación integral de punto fijo si tomamos (Y, Z) como una sola variable, lo que resultaría en,

$$f_X(x) = \int [\int \int f_{X|YX}(x|y,z) f_{YZ|X}(y,z|t) dydz] f_X(t) dt$$

De esta manera, muestreando iterativamente de $f_{X|YZ}$ y $f_{YZ|X}$ resultarían en una serie de variables aleatorias que convergen en $f_X(x)$. Por otro lado, el Muestreador de Gibb muestrearía iterativamente las distribuciones $f_{X|YZ}, f_{Y|XZ}, f_{Z|X}$ y en la j-ésima iteración tendríamos que,

$$\begin{aligned} X'_j &\sim f(x|Y'_j = y'_j, Z'_j = z'_j) \\ Y'_{j+1} &\sim f(y|X'_j = x'_j, Z'_j = z'_j) \\ Z'_{j+1} &\sim f(z|X'_j = x'_j, Y'_{j+1} = y'_{j+1}) \end{aligned}$$

Este esquema de iteraciones nos produce una secuencia de Gibbs,

$$Y'_0, Z'_0, X'_0, Y'_1, Z'_1, X'_1, ...$$

con la propiedad de que ente más grande es la k, $X'_k = x'_k$ es un punto de la distribución marginal $f(x)$ y resolverá la ecuación integral con punto fijo.

En la estadística bayesiana, el Muestreador de Gibbs se utiliza para calcular la distribución posterior mientras que en la estadística clásica se utiliza para calcular la función de verosimilitud. Es importante mencionar que tanto el Muestreador de Gibbsy el algoritmo EM tienen en común el uso de una estructura subyacente, o variables no observables.

La utilidad del Muestreador de Gibbs es más evidente con problemas de mayor complejidad pues ahorra muchos cálculos engorrosos de una manera más elegante y con igual de precisión; además de su potencial práctico.

Full backward non-homogeneous semi-Markov processes for disability insurance models: A Catalunya real data application

Guglielmo D'Amicoa, Montserrat Guillen, Raimondo Manca

Los procesos semi-Markov han sido utilizados en contextos financieros, actuariales y de demografía. Éstos procesos se refieren a aquellos procesos aleatorios que evolucionan con el tiempo y cuyas realizaciones en cualquier momento dado del tiempo tiene un estado definido. Por lo que, la generalización de las probabilidades de trancisión de los procesos semi-markovianos no homogéneos se obtiene introduciendo la reversibilidad, pues en este caso las probabilidades de trancisión dependen del tiempo en el que el proceso entró en un cierto estado, no como en un proceso semi-markoviano homogéneo donde se entra al sistema en el estado inicial al tiempo inicial. La recurrencia en el tiempo en los procesos reversibles se pueden considerar al inicio o al final del horizonte de tiempo considerado.

Suponemos las siguientes variables aleatorias (J_n, T_n) como un proceso de renovación de Markov no homogéneo, J_n representa el estado a la n-ésima trancisión y T_n el tiempo a la n-ésima trancisión, de este modo se define $X_n = T_{n+1} - T_n$ como el proceso de tiempo de llegada. Con esta información, se puede definir lo siguiente:

- $Q_{ij}(s,t) = P(J_{n+1} = j, T_{n+1} \leq t | J_n = i, T_n = s)$.
  La probabilidad de que en la $n+1$-ésima realización el proceso esté en el j-ésimo estado en un tiempo menor o igual a t si en la realización anterior estaba en el i-ésimo estado en el tiempo s. ij son los estados en los que está el proceso y (s,t) los tiempos del mismo.
- $H_i(s,t) = P(T_{n+1} \leq t | J_n = i, T_n = s) \Rightarrow H_i(s,t) = \sum_{j=1}^{m} Q_{ij}(s,t)$.

Probabilidad que el proceso salga del estado i en el intervalo del tiempo s a t en una sola realización.

- $F_{ij}(s,t) = P(T_{n+1} \leq t | J_n = i, J_{n+1} = j, T_n = s)$.
  La función de distribución del tiempo de espera en cada estado i dado que el estado en la realización siguiente es conocido.

La mayor diferencia entre los procesos de Markov no-homogéneos discretos y los procesos semi-Markov reside en las funciones $F_{ij}(s,t)$. En los primeros ésta tendría que comportarse como una función de distribución geométrica, mientras que con el proceso semi-Markov no-homogéneos discretos ésta puede ser de cualquier tipo.

Ahora definimos el proceso de conteo de realizaciones como,

$$N(t) = sup\{n \in \mathbb{N} : T_n \leq t\}$$

Ya que tenemos establecido el concepto de proceso de conteo se puede definir el proceso semi-Markov no-homogéneo discreto, $Z(t) = J_{N(t)}$ como el estado ocupado por el proceso a cada momento del mismo. Por lo que las probabilidades de trancisión serían,

$$\phi_{ij}(s,t) = d_{ij}(s,t) + \sum_{\beta=1}^{m} \sum_{\vartheta=s+1}^{t} b_{i\beta}(s,\vartheta)\varphi_{\beta,j}(\vartheta,t)$$

La primera parte de la fórmula $d_{ij}(s,t)$ se refiere a la probabilidad de que el proceso no transicione al tiempo t dado que entró en el estado i al tiempo s, esto ocurre solo cuando $i = j$, es decir, es referente al tiempo de transición no al estado. La segunda parte $(b_{i\beta}(s,\vartheta))$, representa la probabilidad de que el sistema entre al estado β justo en el tiempo ϑ dado que entró al estado i en el tiempo s. Después de la trancisión, el sistema llegará al estado j en el tiempo t siguiendo cualquiera de las posibles trayectorias que van del estado β al tiempo ϑ.

Â¿Se podría interpretar las variables β, ϑ como variables latentes que nos podrían dar información sobre las variables que sí observamos? Eso podría dar a entender que en la segunda sumatoria donde la variable ϑ empieza el conteo al tiempo $s+1$ y termina al tiempo t, es decir recorre la trayectoria entre los tiempos que conocemos.

Ahora bien, definimos $B(t) = t - T_{N(t)}$ como el proceso reversible que se denota como,

$$^{b}\phi_{ij}(l, s; t) = P(Z(t) = j | Z(s) = i, B(s) = s - l)$$

$$\phi^{b}_{ij}(s; l', t) = P(Z(t) = j, B(t) = t - l' | Z(s) = i)$$

Estos son las probabilidades de transición del proceso semi-Markoviano con el tiempo de recurrencia reversible al inicio y al final, respectivamente.

En la primera ecuación sabemos que el sistema está en el estado i al tiempo s, sabemos también que entró a ese estado en el tiempo l por lo que $s - l$ representa el tiempo reversible inicial; así que lo que se busca la probabilidad de estar en el estado j al tiempo t. En la segunda ecuación sabemos que el sistema entró al estado i al tiempo s, en este caso el objeto de interés es saber la probabilidad de estar en el estado j al tiempo t entrando a este estado en el tiempo l'; el tiempo reversible final es $t - l'$.

Si definimos un proceso reversible en el tiempo inicial y final se tiene que,

$$^{b}\phi^{b}_{ij}(l, s; l', t) = P(Z(t) = j, B(t) = t - l' | Z(s) = i, B(s) = s - l)$$

De igual modo que con el proceso sin la reversibilidad, se definen las siguientes probabilidades de transición:

$$^{b}\phi_{ij}(l, s; t) = d_{ij}(l, s; t) + \sum_{\beta=1}^{m} \sum_{\vartheta=s+1}^{t} b_{i\beta}(l, s; \vartheta) \varphi_{\beta j}(\vartheta, t)$$

$$\phi^{b}_{ij}(s; l', t) = d_{ij}(s, t)\mathbf{1}_{\{l'=s\}} + \sum_{\beta=1}^{m} \sum_{\vartheta=s+1}^{l'} b_{i\beta}(s, \vartheta) \varphi^{b}_{\beta j}(\vartheta; l', t)$$

$$^{b}\phi^{b}_{ij}(l, s; l', t) = d_{ij}(l, s; t)\mathbf{1}_{\{l'=s\}} + \sum_{\beta=1}^{m} \sum_{\vartheta=s+1}^{l'} b_{i\beta}(l, s; \vartheta) \varphi^{b}_{\beta j}(\vartheta; l', t)$$

En las últimas dos ecuaciones donde se tiene el término $\mathbf{1}_{\{l'=s\}}$, se refiere a que la expresión es igual a 1 si y solo si $\{l' = s\}$, sino es igual a 0.

La primera ecuación se refiere a la probabilidad de que el sistema esté en

el estado j en el tiempo t dado que estaba en el estado i en el tiempo s entrando a ese estado al tiempo l, si $l = s$ entonces tenemos el proceso sin reversibilidad.
La segunda ecuación da como resultado la probabilidad de que el sistema llegue al estado j al tiempo l' y permanecerá allí hasta el tiempo t, dado que entró al estado i al tiempo s . La primera parte $d_{ij}(s,t)\mathbf{1}_{\{l'=s\}}$ significa la probabilidad de que no exista trancisión de estados entre el tiempo s, t por lo que el tiempo reversible final $t - l'$ debe ser exactamente igual que $t - s$ y solo tiene sentido cuando $i = j$. La segunda parte se refiere significa que el sistema no se mueve del tiempo s al tiempo ϑ y que, justo en este tiempo, salta al estado β; después siguiendo cualquiera de las trayectorias posibles, el sistema llega al estado j en el tiempo l' y se queda allí hasta el tiempo t.

Es importante mencionar que teniendo todos los valores del proceso reversible final tenemos las posibles probabilidades de transición del proceso sin reversibilidad, es decir,

$$\phi_{ij}(s,t) = \sum_{l'=s}^{t} \phi_{ij}^{b}(s; l', t)$$

Por último, la tercera ecuación expresa la probabilidad de que el sistema entre al estado j al tiempo al tiempo l' y se queda sin transicionar hasta el tiempo t, dado que entró al estado i al tiempo l y se quedó allí hasta el tiempo s. El primer término $d_{ij}(l, s; t)\mathbf{1}_{\{l'=s\}}$ es, de manera análoga, la probabilidad de no tener trancisiones de estados entre los tiempos l a t, es decir, quedarse en el estado i dado que no ocurrieron transiciones entre los tiempos l a s. Esta probabilidad es distinta a 0 si y solo si $i = j$ y $l' = s$. La segunda parte de la ecuación representa la probabilidad de hacer la siguiente transición al tiempo l del estado i a cualquier estado β al cualquier tiempo ϑ, para después seguir cualquier trayectoria para llegar al estado j al tiempo l' sin volver a moverse hasta el tiempo t.

Este modelo es el que se utiliza en algunos cálculos referentes a modelos de incapacidad que consideran tiempos reversibles iniciales y finales. Estos modelos tienen los siguientes estados, con sus respectivas transiciones:

- Activo
    - Activo

        - Pensionado
        - Incapacitado
        - Muerte
    - Pensionado:
        - Pensionado
        - Incapacitado
        - Muerte
    - Incapacitado:
        - Incapacitado
        - Muerte
    - Muerte

Es decir, que la muerte es un estado absorbente.

El experimento a desarrollar en el artículo es sobre una población de 150,000 asegurados de la cobertura de invalidez en la región de Cataluña,España durante 30 años. La condición de invalidez es verificada por un perito médico y corresponde a los padecimientos establecidos en la póliza.

Se aplica el modelo de procesos semi-Markov discretos no-homogéneos con reversibilidad incial y final a la muestra. Dado que el número de transiciones no era suficiente para intervalos de un año de edad, se construyeron grupos de cinco años de edad con los cuatro estados descritos. Los resultados mostraron diferencias según el tiempo de reversibilidad aplicado, inicial o final. Este modelo sirve exclusivamente para ver las transiciones entre estados, por lo que el siguiente paso sería incluir la modelación de costos asociados a cada estado.

Parametric Modelling of cost data, some simulation evidence

Briggs, A. and Nixon, R. and Dixon, S. and Thompson, S.

Los estudios concernientes al análisis de costos medios de algún padecimiento se complican cuando se llevan a cabo con datos observados pues éstos se pueden ver muy sesgados con datos de unos pocos pacientes con costos muy altos.Para lidiar con estas complicaciones se han propuestos varias soluciones como usar métodos no paramétricos, transformar los datos, tomar la media muestral, etc. Cada una de estas soluciones conlleva sus respectivas críticas, por lo que no hay consenso sobre el mejor camino a seguir.

Es por esto que el propósito de este artículo es explorar dos particulares opciones para calcular estimadores de la población de la media de los costos de tratamientos hospitalarios. La primera opción consiste en observar el comportamiento de los datos cuando se someten a restricciones de parámetros supuestos. Para la segunda opción, se repite la comparación utilizando tres muestras con datos de costos hospitalarios y sacando los estimadores empíricos. El objetivo es evidenciar el beneficio en eficiencia que se obtiene utilizando los estimadores adecuados y también lo costoso que sería lo contrario.

Para el primer acercamiento al problema, usaremos dos distribuciones: Gamma y log-normal, pues ambas se utilizan para modelar datos sesgados positivamente. Al utilizar estimadores de máxima verosimilitud (EMV) tenemos, para la distribución Gamma que su EMV es la media muestral y para la distribución log-normal es $exp(lm+lv/2)$ donde lm y lv son la media y varianza en escala logarítmica, respectivamente.

En el experimento, para ambas distribuciones, la media de la población fue designada de 1000 con cinco opciones de coeficientes de variación (CoV=

0.20,0.50,1.0,1.5,2.0) para definir los parámetros de la distribución. A su vez, se hicieron experimentos con cinco distintos tamañ̃os de muestra (n= 20,50,200,500,2000) para cada CoV, lo cual resulta en 50 experimentos y para cada uno de ellos se realizaron 10,000 realizaciones.

Para observar el sesgo y la precisión de los estimadores se calcula la Raíz del Error Mínimo Cuadrado (REMC), como esperado este coeficiente decrece entre es más bajo es el coeficiente de variación y mayor es la muestra sin importar el EMV que se utilice.

Cuando los datos son log-normales el mejor estimador es su propio EMV y exhiben menor REMC que con el estimador de la media muestral; en cambio, cuando el estimador log-normal es aplicado a datos que se distribuyen Gamma los resultados son terribles, sobre todo cuando el CoV es más grande. Esto se debe a que el estimador log-normal está mucho más sesgado a cambios en el coeficiente de variación, mientras que la media muestral no. Estos resultados se repiten cuando medimos el intervalo de confianza, entre menos coeficiente de variación y mayor número de muestra, mayor es el intervalo de confianza.

El segundo enfoque es usar tres bases de datos observados de costos hospitalarios y sacar estimadores empíricos, pues dificilmente se obtienen datos que se comporten como una distribución. Estas bases de datos a nivel costo-paciente son,

- Datos CPOU: Datos obtenidos de la Unidad de Observación para Dolores de Pecho en un hospital escuela. Se reclutaron 972 pacientes con costos a precios de 2001-2002 de las primeras 6 horas de hospitalización, duración de la misma, medicinas, estudios, procedimientos, etc.

- Datos de Fluidos IV: Estos datos fueron obtenidos mediante dos protocolos de atención con fluidos intravenosos aplicado por paramédicos en pacientes con traumas severos antes de llegar al hospital. Se obtuvieron datos de costos hasta 6 meses después del trauma, costos de ambulancia, de fluidos, cuidado ambulatorio para 1191 pacientes a costos de 1997-1998.

- Datos de Paramédicos: Estos datos se consiguieron a traveés de un estudio controlado de paramédicos y técnicos de ambulancia para pacientes con traumas. La muestra de pacientes es de 1852 con datos de 1996-1997 con costos hasta de 6 meses después del trauma inicial incluyendo costos de ambulancia y tratamientos, hospitalización y cuidados ambulatorios.

En un primer análisis exploratorio de los datos vemos que la curtosis y el sesgo de los datos es muy grande comparado con los valores de la normal, es también interesante que la desviación estándar es lo doble a la media para las tres bases de datos. Dado estos resultados se transforman los datos mediante el logaritmo natural, lo que hace que los resultados de media, varianza, sesgo y curtosis se vean más normales.

Tomando otra vez el experimento de simulación, de nuevo con 10,000 realizaciones, que consiste en la extracción de datos de manera aleatoria sin reemplazo de las bases de datos, modificando el tamaño de la muestra, es decir, (n=20,50,300,500) se vuelven a usar los estimadores anteriores: media muestral y $exp(lm + lv/2)$ con sus recpectivos intervalos de confianza. Una vez realizado este experimento se analizan los resultados con muchas similitudes a los resultados anteriores.

La REMC, como esperado, disminuye entre mayor en el tamaño de la muestra. Es importante mencionar que cuando el tamaño de la muestra es más chico para los dos estimadores en cada una de las bases las REMC's son bastante cercanas; sin embargo, cuando el tamaño aumenta el Teorema del Límite Central empieza a ser relevante y la media muestral se vuelve más certera. Sin embargo, lo contrario le ocurre al estimador log-normal, pues sus intervalos de confianza se deterioran rápidamente mientras más grande es la muestra. Esto es porque el Teorema del Límite Central es crucial en la validación de estimadores como la media muestra, el incremento en el tamañono es garantía para otros estimadores paramétricos cuyos supuestos no aguanten este incremento.

Normalmente se diría que el Teorema del Límite Central aplica para cualquier muestra mayor a 30 observaciones, independientemente de su distribución original; sin embargo, esta regla de la práctica no aplica con distribuciones asimétricas como podría ser este el caso.

Como se puede ver cuando se tienen datos de costos hospitalarios es poco probable que estos se comporten como una distribución paramétrica. La simple construcción de estos datos nos da la pista, pues la mayoría de ellos son la suma de costos más pequeños (costos de ambulancia, tratamientos, medicinas, etc.) lo cual significa que es una suma de distintas distribuciones. Es por esto que no podemos confiar que los datos solos nos den mucha información sobre la forma de la distribución.

Con los experimentos realizados a partir de los datos observados confirman que cuando se conoce la forma de la distribución de los datos de costos, el uso del estimador correcto es una gran ganancia en eficiencia; al igual que el uso del estimador incorrecto supone resultados totalmente engañoso. La literatura sobre riesgos recomiendan una cuidadosa modelación paramétrica de los datos para escoger el mejor estimador, aunque se recomienda un número mayor de muestra (entre 10,000-50,000 observaciones), este número es difćil de obtener en cualquier protocolo de estudios.

Por la dificultad de conseguir una muestra de datos de costos hospitalarios suficientemente grande para una modelación paramétrica adecuada, la manera de una correcta estimación con datos basados en la experiencia siguie siendo un reto.