Técnicas de muestreo paralelo para distribuciones finales de datos circulares

Edoardo Isaías Sánchez Ibáñez (edoardosanchez16@gmail.com)

Universidad Autónoma Metropolitana, Unidad Iztapalapa; Maestría en Ciencias, Matemáticas Aplicadas e Industriales, (**MCMAI**).

Asesor: Dr. Gabriel Núñez Antonio (gab.nuneza@gmail.com)

Objetivos

La modelación de problemas reales muchas veces presentan un alto costo computacional para el estudio de los parámetros de interes, por lo que una forma de subsanar eso es mediante la inferencia vía un método en paralelo. En este trabajo se considerá el análisis de distribuciones finales vía métodos de muestreo en paralelo. Particularmente bajo este enfoque se muestra la manera de realizar inferencia para *datos circulares* bajo un modelo **Normal Proyectado**.

Estadística bayesiana

La estadística bayesiana es una metodología que emplea la teoría de la información y teoría de la decisión. Este enfoque nos permite:

- Incorporar información del investigador al análisis.
- La distribucón inicial se actualiza bajo la luz de los datos y da origen a la distribucón final.
- Se puede hablar de regiones de probabilidad para la inferencia realizada sobre los parámetros.

Distribución Final

El estudió de la distribución final vía el teorema de *Bayes* se obtiene la distribución final:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_{\Omega} p(y|\theta)p(\theta)\mathrm{d}\theta}.$$

Hay que notar que la integral $\int_{\Omega} p(y|\theta)p(\theta)\mathrm{d}\theta$ en muchas circunstancias puede ser difícil de trabajar, analítica o numéricamente.

Datos circulares

Los datos que presentan dirección o ángulo respecto a un sistema orientado se denominan *datos direccionales*.
Se pueden representar como puntos en la superficie de un *hiperespacio,* tal que $\mathbb{S}^{r-1} = \left\{\boldsymbol{x} \in \mathbb{R}^r : \boldsymbol{x}^T\boldsymbol{x} = 1\right\}$.
Cuando $r = 2$, es decir, $\mathbb{S}^1 = \left\{\boldsymbol{x} \in \mathbb{R}^2 : \boldsymbol{x}^T\boldsymbol{x} = 1\right\}$, los datos reciben el nombre de **datos circulares**. Existen otras formas de representarlos:

$$\boldsymbol{x} = (\cos\alpha, \operatorname{sen}\alpha)^T \quad \text{ó} \quad z = e^{i\alpha} = \cos\alpha + i\operatorname{sen}\alpha.$$

Su representación cambia

Se requieren métodos adecuados para representarlos, a continuación mostramos uno de ellos:

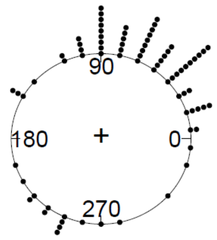

Figura: Representación como puntos en el círculo para datos de 76 tortugas

No los podemos tratar de la misma manera

Como se imaginaran, no podemos tratarlos de la misma manera que a los datos lineales pues nos puede llevar a ciertas paradojas

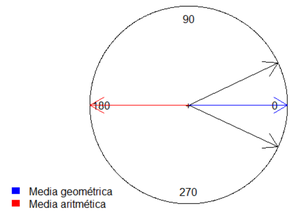

Figura: Diferencia entre la media geométrica y aritmética

El modelo *Normal Proyectado*

Definición

La función de densidad de probabilidad de una distribución Normal proyectada, para un ángulo aleatorio θ, está dada por la siguiente expresión

$$NP(\theta|\boldsymbol{\mu}, \Lambda) = \frac{\varphi(\theta|\boldsymbol{\mu}, \Lambda) + |\Lambda|^{-1/2} D(\theta)\Phi(D(\theta))\phi\left(|\Lambda|^{-1/2}(\boldsymbol{u}'\Lambda\boldsymbol{u})^{-1/2}\boldsymbol{\mu}'\boldsymbol{u}\right)}{\boldsymbol{u}'\Lambda\boldsymbol{u}}\mathbb{I}_{(0,2\pi)}$$

donde, $\varphi(\cdot|\mathbf{0}, \Lambda)$, denota la función de densidad de una Normal bivariada, $\Phi(\cdot)$ y $\phi(\cdot)$ denotan las funciones de distribución y de densidad de una Normal estándar, respectivamente y $\boldsymbol{u} = (\cos\theta, \operatorname{sen}\theta)'$.

Muestreo en paralelo

En el siglo V a.C, el gran estratega militar *Sun Tzu* escribio un tratado "*El arte de la guerra*".

"*Divide y vencerás*"

Esquemas del muestreo en paralelo

Vamos a explorar métodos de muestreo como el denominado *Consensus Monte Carlo (CMC)* y el método *LISA* (*Likelihood Inflating Sampling Algorithm*). Estos métodos consisten en lo siguiente:
Considerar el conjunto $\boldsymbol{X}$ el cual representa todos los n datos asociados al modelo $f(\boldsymbol{X}, \theta)$. Construir una partición con S subconjuntos $\boldsymbol{X}^{(j)}$, tales que,
$\bigcup_{j=1}^{S} \boldsymbol{X}^{(j)} = \boldsymbol{X}$, además $|\boldsymbol{X}^{(j)}| = n_j$, de forma tal que
$\sum_{j=1}^{S} n_j = n$.

Consensus Monte Carlo (CMC)

- CMC:
  De acuerdo a Scott, *et al.*, (2016)la distribución final asociada a cada subconjunto $\boldsymbol{X}^{(j)}$, se puede escribir como:

$$p_j\left(\theta|\boldsymbol{X}^{(j)}\right) \propto f\left(\boldsymbol{X}^{(j)}|\theta\right) p(\theta)^{1/S}. \qquad (1)$$

  De esta manera la *priori* para θ resulta ser:
  $p(\theta) = \prod_{j=1}^{S} p(\theta)^{1/S}$.
  De la ecuación (1) se tiene que la *distribución final completa* resulta ser:

$$P_{full}(\theta|\boldsymbol{X}) \propto \prod_{j=1}^{S} p_j\left(\theta|\boldsymbol{X}^{(j)}\right). \qquad (2)$$

LISA (*Likelihood Inflating Sampling Algorithm*)

- LISA:
  Ahora bien, para el método de LISA de acuerdo a Entezari, *et al.*, (2018), la distribución final por lotes está dada por:

$$p_j\left(\theta|\boldsymbol{X}^{(j)}\right) \propto f\left(\boldsymbol{X}^{(j)}|\theta\right)^S p(\theta). \qquad (3)$$

Ejemplo: Inferencia para μ_1 y μ_2 de un modelo $NP(\theta|\boldsymbol{\mu}, \boldsymbol{I})$

- Se simulo una muestra de tamaño $n = 1000$, $\theta \sim NP(\boldsymbol{\mu}, \boldsymbol{I})$ con $\boldsymbol{\mu} = (2, 1)$.
- Método por Nuñez Antonio y Gutierrez Peña (2005) se introducen n variables latentes r_i, $i = 1 \ldots n$.
- Vía el algoritmo *Gibbs - Sampler*, dadas las condicionales completas $f(\mu|r, \cdot)$ y $f(r_i|\mu, \cdot)$.
- CMC: 5 lotes de 200 datos c/u.

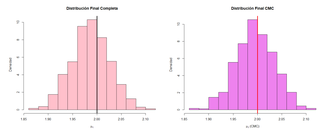

Figura: Inferencia sobre μ_1, vía Gibbs-sampler con y sin partición

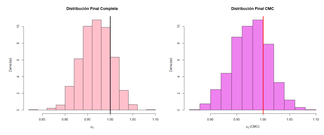

Figura: Inferencia sobre μ_2, vía Gibbs-sampler con y sin partición

Tablas comparativas

	μ_1	μ_1 CMC
Valor real	2.00	2.00
Mediana	1.987	1.995
Intervalo de probabilidad	(1.910, 2.062)	(1.918, 2.074)
Tiempo de cómputo	3.88	0.79

Cuadro: Inferencia para μ_1

	μ_2	μ_2 CMC
Valor real	1.00	1.00
Mediana	0.969	0.977
Intervalo de probabilidad	(0.901, 1.027)	(0.906, 1.043)
Tiempo de cómputo	3.88	0.79

Cuadro: Inferencia para μ_2

Conclusión

Al parecer métodos como el **CMC** pueden ser apropiados desde el punto de vista computacional para realizar inferencia en modelos complejos como aquellos que incorporan variables latentes.

Bibliografía

- Nuñez - Antonio Gabriel, and Gutiérrez - Peña Eduardo. *A Bayesian Analysis of Directional Data using the Projected Normal Distribution*.
- Reihaneh Entezari, Radu V Craiu, and Jeffrey S Rosenthal. *Likelihood inflating sampling algorithm*.
- Steven L Scott, Alexander W Blocker, Fernando V Bonassi, High A Chipman, Edward I George, and Robert E McCulloch. *Bayes and big data: The Consensus Monte Carlo Algorithm*.