Cálculo de la inversa de la matriz de coeficientes de las ecuaciones de modelo mixto utilizando muestreo de Gibbs y aplicaciones

L.A. García-Cortés y D. Sorensen. *National Institute of Animal Science. Dept. for Research in Pigs and Horses. P.O.Box 39, DK-8830 Tjele, Denmark*

C. Moreno, L. Varona y J. Altarriba. *Unidad de Genética Cuantitativa y Mejora. Facultad de Veterinaria. Universidad de Zaragoza*

Resumen

En este trabajo se presenta un procedimiento para obtener la inversa de una matriz utilizando el muestreo de Gibbs y se discute su utilidad para reducir el ruido de Monte Carlo asociado a la implementación práctica del método de Monte Carlo EM introducido en 1992 por Guo y Thompson.

Introduccion

La varianza de la distribución marginal de un efecto fijo o aleatorio, si se utiliza un muestreo de Gibbs que asume los componentes de varianza conocidos, coincide con la varianza de error de estimación o predicción de dichos efectos (Wang et al, 1993). Esto significa que, en ciertas circunstancias, el muestreo de Gibbs puede proporcionar elementos de la inversa de la matriz de coeficientes de las ecuaciones de modelo mixto, en este caso los elementos de la diagonal. En este trabajo se propone un método general para invertir una matriz usando el muestreo de Gibbs.

El algoritmo para invertir una matriz.

Supongamos que estamos interesados en la inversa de una matriz simétrica definida positiva, llamada C. Si consideramos cualquier vector $\mathbf{x} \sim \xi(\mathbf{0}, \mathbf{C}^{-1})$, es decir, cualquier distribución con media $\mathbf{0}$ y varianza , entonces $\mathrm{E}(\mathbf{xx}') = \mathrm{Var}(\mathbf{x}) + \mathrm{E}(\mathbf{x})\mathrm{E}(\mathbf{x}') = \mathbf{C}^{-1}$

La única condición para que exista **x** es que $\mathbf{C}^{-1}$ sea simétrica y definida positiva, sólo así puede interpretarse como una matriz de varianzas y covarianzas. En general, $\mathrm{E}(x_i x_j) = \mathrm{Cov}(x_i, x_j) + \mathrm{E}(x_i)\mathrm{E}(x_j) = \mathbf{C}_{ij}^{-1}$, donde $\mathbf{C}_{ij}^{-1}$ es un elemento de $\mathbf{C}^{-1}$. Si realizamos varias extracciones de la variable aleatoria **x** utilizando cualquier método de Monte Carlo, obtenemos así la inversa de C:

$$\mathbf{C}^{-1} \cong \sum_{k=1,n} \mathbf{x}_k \mathbf{x}_k' / n \qquad [1]$$

donde $\mathbf{x}_k$ es la k-ésima extracción de **x**. A partir de [1] no es difícil desarrollar fórmulas sencillas para contener las trazas necesarias que aparecen en las expresiones habituales. Para obtener la inversa, necesitamos un método de Monte Carlo para extraer observaciones de **x**. Aquí proponemos utilizar un muestreo de Gibbs.

El muestreo de Gibbs en la extracción de $\mathbf{x} \sim \xi(\mathbf{0}, \mathbf{C}^{-1})$.

Utilizamos una normal multivariante debido a que **x** representa una distribución de cualquier forma. En el caso de la distribución normal, las distribuciones condicionales son muy sencillas de desarrollar y de calcular. Así, la distribución conjunta de **x** es:

$$f(\mathbf{x}) = cte * exp\left(-0.5\ \mathbf{x}'\mathbf{C}\mathbf{x}\right)$$

Nótese que lo importante de esta expresión es que sólo aparece C y no C^{-1}. A partir de aquí, las condicionales nunca incluirán elementos de la inversa. En la distribución condicional de x_i respecto del resto de variables, sólo los términos que incluyen dicha variable han de ser tenidos en cuenta:

$$f(x_i|x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n) \propto exp\left[-0.5\left(x_i^2\, c_{ii} + 2\, x_i \sum_{j\neq i} x_j c_{ij}\right)\right]$$

Que puede demostrarse que corresponde a la distribución normal con media $-\sum_{j\neq i} x_j c_{ij} / c_{ii}$ y varianza $1 / c_{ii}$. Es decir, $x_i|x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n,\mathbf{y} \sim N(-\sum_{j\neq i} x_j c_{ij} / c_{ii}, \sigma_e^2 / c_{ii})$ [2]

Nótese que en este caso la media de la distribución marginal es conocida (cero), así que si comenzamos la cadena de Markov en el punto $\mathbf{x} = \mathbf{0}$, partimos de la moda de la distribución conjunta, ventaja de la que normalmente carecen otros algoritmos basados en el muestreo de Gibbs, y que obliga en ocasiones a desechar gran cantidad de extracciones al comienzo. En este caso, aunque no esta claro si es necesario desechar alguna extracción, nunca será un número tan grande como si desconociesemos las medias marginales de las distribuciones. Como en el caso del metodo de remuestreo presentado en García Cortés et al (1992), pueden utilizarse otro tipo de distribuciones condicionales, por ejemplo Bernouilli: una distribución de Berouilli con media m y varianza s, presenta un 50% de la probabilidad en el punto $x_i=m-s^{1/2}$ y un 50% en el punto $x_i=m+s^{1/2}$. Otros valores de x_i no estan definidos.

Estudio del ruido de Monte Carlo asociado a la estimacion REML de componentes de varianza.

Asumimos el modelo $\mathbf{y} = \mathbf{Xb} + \mathbf{a} + \mathbf{e}$, respondiendo a la notación habitual de un modelo animal. Las variables aleatorias son $\mathbf{y} \sim N(\mathbf{Xb}, \mathbf{A}\sigma_a^2 + \mathbf{I}\sigma_e^2)$, $\mathbf{a} \sim N(\mathbf{0}, \mathbf{A}\sigma_a^2)$ y $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$.

En una iteración de EM, se calcula el componente de varianza aditivo con la siguiente expresión:

$$q\hat{\sigma}_a^2 = \hat{\mathbf{a}}'\mathbf{A}^{-1}\hat{\mathbf{a}} + tr(\mathbf{A}^{-1}\mathbf{C}^{aa}) \qquad [3]$$

donde q es el número de animales, $\hat{\mathbf{a}}$ es el predictor BLUP de $\mathbf{a}$, y $\mathbf{C}^{aa}$ es el bloque aleatorio de la inversa de la matriz de coeficientes. Estimaremos $q\hat{\sigma}_a^2$ en [3] mediante 4 métodos de Monte Carlo

Metodo 1. El método Monte Carlo EM descrito por Guo y Thompson. Utilizando un algoritmo de muestreo de Gibbs como el descrito en Wang et al. (1993), se utiliza la expresión $\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}$ en cada paso de la cadena. Nótese que nuevos valores $\mathbf{a}$ son simulados en cada paso de la cadena. Así:

$$q\hat{\sigma}_a^2 = E(\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}) = \sum_{k=1,B} \mathbf{a}'_k\, \mathbf{A}^{-1}\mathbf{a}_k / B \qquad [4]$$

donde B es el número de extracciones de $\mathbf{a}$, es decir, la longitud de la cadena. En este caso las condicionales son: $x_i|x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n,\mathbf{y} \sim N(RHS_i - \sum_{j\neq i} x_j c_{ij} / c_{ii}, \sigma_e^2 / c_{ii})$ [5]

Método 2.- La propuesta de Thompson (1994) en el 5WCGALP. La primera idea para reducir el ruido de Monte Carlo presente en [4] fue utilizar la esperanza de las distribuciones condicinales en lugar de los valores de $\mathbf{a}$. El muestreo de Gibbs se realiza de la misma forma que en el método anterior (ecuación 5), pero [4] se calcula de otro modo más eficaz.

$$E(x_i) = RHS - \sum_{j\neq i} x_j c_{ij} / c_{ii}$$

Entonces x_{ik} puede expresarse como: $x_{ik} = E(x_{ik}) + e_{ik}$,con $Var(e_{ik}) = \sigma_e^2 / c_{ii}$, y

$$q\hat{\sigma}_a^2 = \sum_{k=1,q} \mathbf{a}'^*_k\, \mathbf{D}^{-1}\mathbf{a}^*_k / q + \hat{\sigma}_e^2 tr(\mathbf{D}^{-1}(diag C_{aa})^{-1})$$

donde $\text{diag}\mathbf{C}_{aa}$ es la diagonal del bloque aleatorio de $\mathbf{C}$, $\mathbf{A}^{-1} = \mathbf{L}^{-1}\mathbf{D}^{-1}\mathbf{L}^{-t}$, y $\mathbf{a}^*$ es un vector que contiene los elementos $a_i^* = E(a_i) - 0.5*(a_{p(i)} + a_{m(i)})$. Más detalles en Thonpson (1994)

Método 3.- Usando el muestreo de Gibbs para calcular la inversa en $\mathbf{C}^{aa}$ en [3]. Usando el algoritmo propuesto aquí en la fórmula habitual para calcular $q\hat{\sigma}_a^2$ en [3], el muestreo de Gibbs es ligeramente diferente, usando [2] en lugar de [5].

$$\mathbf{C}^{aa} = \sum_{k=1,B} \mathbf{a}^+\mathbf{a}'^+ / B \qquad [6]$$

donde $\mathbf{x}' = \begin{bmatrix}\mathbf{b}'^+ & \mathbf{a}'^+\end{bmatrix}$. Entonces, sustituyendo [6] en [3]:

$$q\hat{\sigma}_a^2 = \hat{\mathbf{a}}'\mathbf{A}^{-1}\hat{\mathbf{a}} + \hat{\sigma}_e^2 \sum_{j=1,n} \mathbf{a}'^+ \mathbf{A}^{-1}\mathbf{a}^+ / n \qquad [6]$$

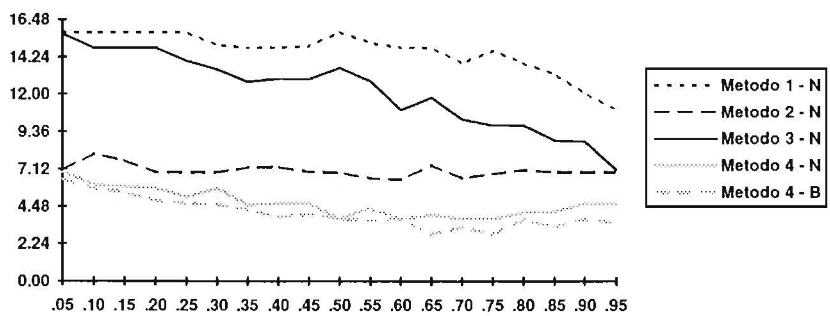
Método 4. Combinando las ideas de los métodos 2 y 3. Es decir, usando el muestreo de Gibbs como en el método 2 pero calculando el componente basado en el esperanza de las condicionales y no en los valores extraídos. Aqui utilizamos la variable $\mathbf{a}^{*+}$, que representa lo mismo que $\mathbf{a}^*$ salvo que el muestreo ha sido realizado de forma diferente.

$$q\hat{\sigma}_a^2 = \hat{\mathbf{a}}'\mathbf{A}^{-1}\hat{\mathbf{a}} + \hat{\sigma}_e^2 \sum_{j=1,n} \mathbf{a}'^{*+}_j \mathbf{D}^{-1}\mathbf{a}^{*+}_j / n + \hat{\sigma}_e^2 tr(\mathbf{D}^{-1} diag(\mathbf{C})^{-1})$$

Notese que en este caso el ruido de Monte Carlo sólo esta presente en la segunda parte de la fórmula. Este cuarto método ha de tener el menor ruido de Monte Carlo.

RESULTADOS

Asumiendo el modelo animal descrito antes, simulamos una población de 50 animales (10 generaciones de 5 animales cada una) y 5 efectos fijos asignados al azar. Entonces, calculamos el valor exacto de $\hat{\mathbf{a}}'\mathbf{A}^{-1}\hat{\mathbf{a}} + tr(\mathbf{A}^{-1}\mathbf{C}^{aa})$ y las aproximaciones utilizando los 4 métodos propuestos. La longitud de la cadena fue de 500.000 y consideramos sólo un punto de cada 50. Así, disponemos de 10.000 valores casi incorrelados que pueden ser comparados frente al valor exacto. La media de estos 10.000 valores fue igual al valor exacto, pero calculamos el ruido de Monte Carlo en los 4 métodos calculando la variable "error relativo en porcentaje". Es decir, el valor exacto menos el obtenido en cada punto y dividido por el valor exacto. De esos 10.000 errores relativos, se presentan en la gráfica las medias de los valores absolutos cuando la heredabilidad simulada se cambia desde 0.05 hasta 0.95. Se presenta el cuarto método utilizando la distribución de Bernouilli y los 4 métodos usando una distribución normal.



Bibliografía citada

García-Cortés, L.A., Moreno, C., Varona, L. and Altarriba, J., 1992. J.Anim. Breed. Gen., 109:358
Guo, S.W. and Thompson, E.A., 1992. IMA J. of Math. Appl. in Med. and Biol., 8:171-189
Thompson, R., 1994: Proc. 5th World Cong. on Genet. Appl. to Livest. Prod., 18:337
Wang, C.S., Rutlegde, J.J. and Gianola, D., 1993. Gen. Sel. Evol. 25:41-62