Memorias de la XVII Reunión de la Asoc. Lat. Prod. Animal, La Habana, Cuba (643) pp 1936-1939, 2001.

20-23 Nov., 2001

G 43. UNA APROXIMACIÓN BAYESIANA PARA LA ESTIMACIÓN DE COMPONENTES DE VARIANZA EN UN MODELO ADITIVO HETEROCEDÁSTICO. II. UN ESTUDIO DE SIMULACIÓN.

A BAYESIAN APPROACH TO ESTIMATE VARIANCE COMPONENTS IN A HETEROSKEDASTIC ADDITIVE GENETIC MODEL. II. A SIMULATION STUDY.

Alejandro Jara, Mauricio Elzo[1] y Nelson Barría

Departamento Fomento de la Producción Animal, Universidad de Chile, Santa Rosa 11,735, La Granja, Santiago, Chile. Tél: (56-2) 678-5572, Fax: (56-2) 678-5611, e-mail: ajara@uchile.cl

[1] Department of Animal Sciences, University of Florida, FL, USA.

RESUMEN

En este trabajo, se evalúan, a través de simulación estocástica, algunas propiedades frecuentistas de una aproximación Bayesiana para la estimación de componentes de varianza en modelos animales univariados, permitiendo diferencias en los promedios genético-aditivos y heterogeneidad de varianzas aditivas y ambientales a través de grupos raciales bajo un modelo aditivo de herencia. El número de iteraciones (20 a 109) y el tiempo de CPU (28 a 49 min) requerido para alcanzar la convergencia en la estimación de los componentes de varianza, en un grupo de datos de pequeño tamaño (22.978 animales y 22.178 registros fenotípicos) de una población de razas parentales y varianza ambiental homogénea, sugieren que los procedimientos son factibles desde un punto de vista computacional. La convergencia y el error cuadrático medio de las estimaciones mejoraron con la incorporación de información a priori, sugiriendo que el uso de una distribución a priori no conjugada para las varianzas aditivas mejora la exactitud de las estimaciones.

Palabras Clave: varianzas aditivas heterogéneas / varianza de segregación / estimación de Bayes / máximo a posteriori

ABSTRACT

In this work some frequentist properties of a Bayesian approach to estimate variance components in univariate animal models allowing for different additive genetic means and heteroskedasticity of additive and error variances across breed groups under an additive model of inheritance were examined via simulation. Number of iterations (20 to 109) and computing times (28 to 499 min) to achieve convergence when estimating the variance components in a small data set (22,978 animals and 22,178 phenotypic records) of a two strain population and homogeneous error variance suggest that these procedures are computationally feasible. Convergence and mean squared error of the estimates improved by the incorporation of prior information suggesting that the use of a nonconjugate informative prior distribution for the additive genetic variances improve the accuracy of the estimates.

Keywords: heterogeneous additive genetic variance / segregation variance / Bayes estimation / maximum a posteriori

INTRODUCCION

Un método comúnmente utilizado para mejorar poblaciones ganaderas es la migración de genes desde poblaciones genéticamente superiores. En este tipo de situaciones, los procedimientos empleados para la evaluación genética, deben considerar las diferencias genéticas entre las poblaciones involucradas. Los métodos desarrollados con este fin, se basan en el mejor predictor lineal insesgado (BLUP), el que requiere del conocimiento de los parámetros de dispersión asociados con estos modelos (Elzo y Bradford, 1985; Elzo y Famula, 1985; Cantet y Fernando, 1995). Una posibilidad, es estimar los parámetros de dispersión a través de métodos máximo-verosímiles (ML, REML) y utilizar estas estimaciones en lugar de los valores verdaderos para la evaluación genética. Elzo (1994) y Cantet (1999, comunicación personal), desarrollaron procedimientos REML para la estimación de componentes de varianza en presencia de heterogeneidad de varianzas aditivas. Sin embargo, la estimación y predicción en dos etapas solo puede ser justificada cuando la función de verosimilitud es simétrica y curva. En otro caso, las predicciones obtenidas a través de REML pueden ser pobres. Debido a que las propiedades de la metodología REML se sustentan en justificaciones asintóticas, tales procedimientos pueden producir estimaciones poco confiables de los componentes de varianza dependiendo del tamaño muestral, y el tipo y distribución de las cruzas involucradas. Jara

et al., (2001, en este volumen), desarrollan una aproximación Bayesiana para la estimación de componentes de varianza, considerando diferencias en los promedios genéticos y heterogeneidad de varianzas a través de grupos raciales. Los objetivos de este trabajo fueron examinar, a través de simulación estocástica, algunas propiedades frecuentistas y algunas propiedades computacionales de los estimadores Bayesianos propuestos.

MATERIALES Y METODOS

Se consideró dos poblaciones parentales, H y O. Se simularon 22.178 registros de producción total por lactancia en un proceso de absorción de la raza O a la raza H, durante 5 generaciones. El proceso se inició con 150 toros H no relacionados, cruzados con una población de 3.500 vacas O no emparentadas. El número de hijas por toro, *ndau*, se simuló desde una distribución de Poisson(25). Este número se alcanzó, a través del cruzamiento aleatorio de cada toro con *ndau* vacas provenientes desde la población original de vacas O. Las siguientes generaciones se generaron de una forma similar, utilizando 150 toros H no relacionados por generación cruzados con la generación anterior de vacas híbridas. Se generaron 10 réplicas de los datos. Los valores aditivos de las vacas (O) y toros (H) fueron simulados desde una distribución N(99, 264.258) y N(385, 384.327), respectivamente. Los registros fenotípicos se generaron adicionando, al valor aditivo de cada vaca, el efecto de uno de 1.200 niveles del factor Predio-Año-Estación, muestreado independientemente desde una distribución N(4.968, 749.654), y un efecto ambiental muestreado en forma independiente desde una distribución N(0, 792.774). En las siguientes generaciones, los registros se simularon en forma similar, con la excepción de que los valores aditivos de cada animal se generaron a través del promedio de los valores aditivos de los padres, al que se le adicionó el efecto de muestreo mendeliano muestreado desde una distribución N(0, $\mathbf{G}_{eii}$). La varianza de segregación, se asumió igual a 120.069 kg^2.

Los componentes de varianza fueron estimados utilizando el algoritmo descrito por Jara et al. (2001, en este volumen). Estimaciones REML homogéneas de los componentes de varianza fueron utilizadas como valores iniciales. Se consideró 5 tipos de distribuciones a priori, las que tuvieron valores diferentes para los parámetros que describen la incertidumbre sobre el conocimiento previo sobre los parámetros de dispersión. El primer tipo, (PI), fue $v_i = -2$ y $S^2_i = 0$ ($i = a_1, a_2, a_3, e$), correspondiendo a distribuciones a priori planas. Los tipos 2 y 3 (PII and PIII) se especificaron tal que $100\ [v_i / (v_i + db_i)] = 1$ y 10, respectivamente, donde, db_i es el "grado de credibilidad" contribuida por los datos sobre la varianza respectiva. En nuestro caso, para las varianzas aditivas, db_i se definió como la suma de los ponderadores que determinan la contribución de cada varianza a la varianza de los efectos de muestro mendeliano, debido a que no todos los animales contribuyen a la estimación de todos los parámetros genéticos de dispersión. Para estas tres a priori, los componentes de varianza utilizados en la simulación de los datos se utilizaron como valores de los hiperparámetros de las distribuciones a priori. Las últimas dos a priori se especificaron con el objeto de lograr distribuciones a priori propias pero no informativas, utilizando valores bajos y altos para las varianzas aditivas a priori. Las a prioris vagas (PIV y PV) se obtuvieron utilizando v_i=4,000001, de tal forma que las distribuciones fuesen tan planas como fuese posible pero con varianza finita. La especificación de los valores a de los hiperparámetros se resume en la Tabla I.

Tabla I. Distribuciones a Priori para la varianza aditiva de la raza O (σ^2_{a1}), de la raza H (σ^2_{a2}), de la varianza de segregación (σ^2_{a12}), y de la varianza del error (σ^2_e), utilizada en la estimación de los parámetros de dispersión (S^2_i, varianza a prior, v_i, grado de creencia a priori).

Varianza	σ^2_{a1}		σ^2_{a12}		σ^2_{a2}		σ^2_e	
A Priori	S^2_{a1}	v_{a1}	S^2_{a12}	v_{a12}	S^2_{a2}	v_{a2}	S^2_e	v_e
PI	0	-2	0	-2	0	-2	0	-2
PII	264.258	53	120.069	46	384.327	84	792.774	232
PIII	264.258	588	120.069	503	384.327	922	792.774	2.548
PIV	211.406	4,000001	60.388	4,000001	317.817	4,000001	852.454	4,000001
PV	317.110	4,000001	177.980	4,000001	447.298	4,000001	734.862	4,000001

Con el objeto de acelerar la convergencia, el componente REML del algoritmo se calculó utilizando las siguientes expresiones:

$$\tilde{\sigma}^{2[t]}_{ai} = \left[\frac{\hat{u}' A_i^{*} \hat{u}}{tr\left[G^{*}_{ei}\right] - tr\left(A_i^{*} C^{uu}\right)_{\alpha_i^{[t-1]}}} \right]^{[t]} \qquad \tilde{\sigma}^{2[t]}_{ei} = \left[\frac{\hat{e}'\left(y - X\hat{b} - ZQ\hat{g} - Z\hat{u}\right)}{n - rank(X)} \right]^{[t]}$$

donde, el escalar $\lambda_i^{[t-1]}$, es la razón entre la varianza del error y la respectiva varianza genética en la iteración [t-1]. Los cálculos se llevaron a cabo, utilizando un programa Fortran, compilado sin ninguna optimización con el programa Microsoft Power Station, y un PC Compaq Presario modelo 1.277. Este programa utilizó las subrutinas FSPAK (Pérez-Encizo et al., 1994) para obtener la inversa de la matriz de coeficientes de las ecuaciones de los modelos mixtos. Se asumió que la convergencia se alcanzó cuando la razón de la diferencia entre las sumas de cuadrados entre dos iteraciones sucesivas relativa a la suma de cuadrados de las varianzas estimadas en la iteración previa fue menor que 10^{-7}.

RESULTADOS Y DISCUSION

El promedio y el rango de los estimadores MAP de la varianza genético aditiva de la raza O (σ^2_{a1}), raza H (σ^2_{a2}), la varianza de segregación (σ^2_{a12}), y la varianza del error (σ^2_e), bajo los 5 tipos de distribuciones a priori para los componentes de varianza, se muestran en la Tabla II. El mínimo y el máximo de las estimaciones, expresado como porcentaje del valor parametral verdadero, fueron 28,46 y 151,06%, 66,14 y 116,57%, 91,85 y 110,92%, 30,27 y 147,34%, y 44,40 y 149,62% para las a priori PI, PII, PIII, PIV y PV, respectivamente.

Tabla II. Promedio (rango) de las estimaciones MAP de la varianza aditiva de la raza O (σ^2_{a1}), la varianza de segregación (σ^2_{a12}), la varianza aditiva de la raza H (σ^2_{a2}), y la varianza del error (σ^2_e), utilizando diferentes distribuciones a priori (P) para los componentes de varianza.

P	σ^2_{a1}	σ^2_{a12}	σ^2_{a2}	σ^2_e
PI	274.493 (246.135 ; 311.410)	109.035 (34.175 ; 181.377)	387.677 (346.509 ; 439.399)	792.334 (760.913 ; 812.716)
PII	273.796 (243.741 ; 308.034)	115,760 (79.410 ; 162.898)	385.329 (350.737 ; 426.516)	791.971 (764.842 ; 812.174)
PIII	270.973 (246.552 ; 293.108)	118,206 (110.278 ; 126.824)	384.257 (364.127 ; 411.440)	792.678 (772.751 ; 806.127)
PIV	273.220 (244.741 ; 309.852)	104,043 (36.342 ; 176.910)	388.353 (347.298 ; 438.314)	793.424 (762.108 ; 814.021)
PV	274.629 (245.757 ; 311.012)	113.358 (53.3314 ; 179.653)	386.452 (346.850 ; 434.321)	791.783 (761.207 ; 812.636)

La diferencia entre los valores estimados y los reales de los componentes de varianza, evaluada a través de la prueba de T, indicó que todos los estimadores son insesgados. Además, el comportamiento del estimador sobre muestreos repetidos, indicó que la incorporación de información a priori sobre los componentes de varianza produce una mejora substancial de las estimaciones, lo que concuerda con los resultados reportados en la literatura por Hoeschelle et al. (1987), en modelos umbrales homocedásticos. La Tabla III, muestra que el error cuadrático medio (MSE) de los componentes de varianza. Bajo PII, PIII, PIV y PV, el error cuadrático medio, expresado como porcentaje del MSE de las estimaciones obtenidas utilizando PI, fueron 93, 45, 94 y 98%, 29, 1, 98 y 71%, 65, 27, 98 y 88%, y 77, 37, 99 y 95 % para la varianza aditiva de la raza O, para la varianza de segregación, para la varianza aditiva de la raza H y para la varianza del error, respectivamente.

Tabla III. Error cuadrático medio de las estimaciones Máximos a Posteriori de la varianza aditiva de la raza O (σ^2_{a1}), varianza de segregación (σ^2_{a12}), varianza aditiva de la raza H (σ^2_{a2}), y varianza del error (σ^2_e), utilizando diferentes distribuciones a priori para los parámetros de dispersión.

A Priori	σ^2_{a1}	σ^2_{a12}	σ^2_{a2}	σ^2_e
PI	407.344.785	2.389.636.951	722.929.614	243.857.301
PII	379.380.542	691.589.424	472.535.783	188.414.321
PIII	184.900.584	31.577.477	191.886.269	89.562.829
PIV	381.638.304	2.342.622.377	705.007.762	240.629.419
PV	400.345.018	1.707.273.406	636.183.265	232.499.363

Por otra parte, como se muestra en la Tabla V, la aproximación a la convergencia tiende a acelerarse en la medida en que la información a priori sobre los componentes de varianza aumenta. El promedio y el rango de las iteraciones requeridas para alcanzar la convergencia disminuyó, reduciendo el tiempo promedio computacional de 322 minutos bajo la a priori PI a 225, 93, 293 y 284 minutos bajo las a priori II, III, IV y V, respectivamente. El tiempo

computacional requerido para los análisis de los registros simulados fue probablemente mayor al necesario debido a que el programa computacional utilizado no fue optimizado por velocidad, y presenta verificaciones en varios puntos del procedimiento computacional los que adicionan tiempo en cada iteración. La principal dificultad del algoritmo propuesto fue, en el paso de E, requiere del cálculo de la esperanza de formas cuadráticas y, de esta forma, la inversa de la matriz de coeficientes de las ecuaciones de los modelos mixtos. Una alternativa es aproximar estas esperanzas a través de la implementación de un algoritmo Monte Carlo EM (MCEM) (Guo y Thompson, 1992). Varias estrategias se han sugerido para reducir la varianza de muestreo en el algoritmo MCEM bajo modelos lineales homocedásticos. Sin embargo, se requiere de trabajo adicional para comparar el comportamiento de estos algoritmos bajo el modelo considerado en este trabajo.

Tabla V. Promedio (rango) del número de iteraciones y del tiempo computacional necesario para alcanzar la convergencia de las estimaciones, utilizando diferentes distribuciones a priori para los componentes de varianza.

A Priori	Iteraciones	Tiempo Computacional (segundos)
PI	70,6 (26 ; 109)	19.327,16 (6.851,04 ; 29.909,08)
PII	50,2 (38 ; 56)	13.504,25 (10.078,23 ; 15.212,00)
PIII	21,1 (20 ; 23)	5.600,15 (5.259,52 ; 6.054,74)
PIV	68,0 (29 ; 99)	17.585,79 (7.170,39 ; 25.105,52)
PV	62,3 (27 ; 82)	17.037,31 (7.115,55 ; 24.599,89)

CONCLUSIONES

Si existe información a priori pertinente sobre las varianzas genético-aditivas, ésta puede ser incorporada a través de la distribución gamma invertida, una a priori no conjugada para este modelo, debido a que esta produce estimaciones insesgadas y con menor error cuadrático medio.

AGRADECIMIENTOS

Este trabajo fue financiado por los proyectos N° 1000794 y N° 7000794 del Fondo Nacional de Investigación Científica y Tecnológica de Chile, FONDECYT.

REFERENCIAS BIBLIOGRAFICAS

1. Cantet, R.J.C., and R.L. Fernando R.L. 1995. Prediction of breeding values with additive animal models for crosses from two populations. Genet. Sel. Evol. 27: 323-334.
2. Elzo, M.A. 1994. Restricted maximum likelihood procedures for the estimation of additive and nonadditive genetic variances and covariances in multibreed populations. J. Anim. Sci. 72: 3055-3065.
3. Elzo, M.A., and G.E. Bradford G.E. 1985. Multibreed sire evaluation procedures across countries. J. Anim. Sci. 60: 953-963.
4. Elzo, M.A., and T.R. Famula. 1985. Multibreed sire evaluation procedures within a country, J. Anim. Sci. 60: 942-952.
5. Guo, S.W., and E.A. Thompson. 1992. Monte Carlo estimation of variance components models for large complex pedigrees, IMA J. Math. Appl. Med. Biol. 8: 171-189.
6. Hoeschele, I., D. Gianola, and J.L. Foulley. 1987. Estimation of variance components with quasi-continuous data using Bayesian methods, J. Anim. Breed. Genet. 104: 334-349.
7. Jara, A., M.A. Elzo y N. Barría. 2001. Una aproximación Bayesiana para la estimación de componentes de varianza en un modelo aditivo feterocedástico. I. Aspectos Teóricos. ALPA, La Habana, Cuba.
8. Perez-Enciso, M., I. Misztal, and M.A. Elzo. 1994. FSPAK: An interface for public domain sparse matrix subroutines, in: Proceedings of the 5th World Congress on Genetics Applied to Livestock Production, 7-12 August 1994, University of Guelph, Guelph, vol. 22, pp. 87-88.