Memorias de la XVII Reunión de la Asociación Latinoamericana de Producción Animal, Ciudad de la Havana, Cuba (G36) pp 1964-1967, 2001.
20-23 Nov., 2001

G 36. UNA APROXIMACIÓN BAYESIANA PARA LA ESTIMACIÓN DE COMPONENTES DE VARIANZA EN UN MODELO ADITIVO HETEROCEDÁSTICO. I. ASPECTOS TEÓRICOS.

A BAYESIAN APPROACH TO ESTIMATE VARIANCE COMPONENTS IN A HETEROSKEDASTIC ADDITIVE GENETIC MODEL. I. THEORETICAL ASPECTS.

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

Este trabajo presenta 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. Los estimadores fueron derivados utilizando la teoría del algoritmo EM generalizado y se basaron en la moda de la distribución posterior conjunta de los componentes de varianza utilizando la distribución gamma invertida, una a priori no conjugada para las varianzas aditivas en este modelo, para describir la incertidumbre a priori sobre los componentes de varianza. Los estimadores combinan información de la distribución a priori y de un estadístico basado en los datos, el estimador REML.

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

ABSTRACT

This work presents 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. The estimators were derived using the theory of generalized Expectation-Maximization algorithm and based on the mode of the joint posterior distribution of the variance components using the inverse gamma distribution, a nonconjugate prior for the heterogeneous additive genetic variances under this model, to describe the prior uncertainty about the variance components. The estimators combines in an optimal way information from the prior distribution and from a data based statistic, the REML estimator.

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

INTRODUCCION

En las últimas décadas, ha aumentado el interés por evaluar desde un punto de vista genético poblaciones de animales puros e híbridos simultáneamente. Esto ha llevado al desarrollo de aspectos teóricos relativos a la semejanza entre parientes y a la modelación y predicción de valores genéticos en este tipo de poblaciones (Elzo, 1990a; Lo et al., 1993). 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. 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. De esta forma, el objetivo del este trabajo fue desarrollar estimadores Bayesianos para los componentes de varianza considerando diferencias en los promedios genéticos y heterogeneidad de varianzas a través de grupos raciales.

MATERIALES Y METODOS

Se consideró el siguiente modelo lineal mixto: $\mathbf{y} = \mathbf{Xb} + \mathbf{ZQg} + \mathbf{Zu} + \mathbf{e}$, donde, $\mathbf{y}$ es el vector de las observaciones ($\mathbf{y} \mid \mathbf{b}, \mathbf{g}, \mathbf{u}, \phi_e \sim N(\mathbf{Xb} + \mathbf{ZQg} + \mathbf{Zu}, \mathbf{R})$, con $\mathbf{R}$ diagonal), $\mathbf{b}$ es el vector de "efectos fijos" ($f(\mathbf{b}) \propto$ cte.), $\mathbf{g}$ es el vector de promedios genéticos de las poblaciones parentales ($f(\mathbf{g}) \propto$ cte.), $\mathbf{u}$ es el vector de los valores aditivos ($\mathbf{u} \mid \phi_a \sim N(\mathbf{0}, \mathbf{G})$), expresados como desvío de su promedio genético de grupo, $\mathbf{e}$ es el vector de errores, $\mathbf{X}$, $\mathbf{Q}$ y $\mathbf{Z}$ son las respectivas matrices de diseño y, ϕ_e y ϕ_a es el vector de nrv varianzas ambientales y ngv varianzas aditivas, respectivamente. La matriz de covarianzas entre los efectos genético-aditivos, puede escribirse como:

$$\mathbf{G} = (\mathbf{I} - \mathbf{P}')^{-1}\mathbf{G_e}(\mathbf{I} - \mathbf{P})^{-1} = (\mathbf{I} - \mathbf{P}')^{-1}\left(\sum_{i=1}^{ngv}\mathbf{D}_i\phi_{ai}\right)(\mathbf{I} - \mathbf{P})^{-1}$$

donde, $\mathbf{P}$ es una matriz que relaciona la progenie y los padres, $\mathbf{G_e}$ es la matriz de covarianzas entre los efectos de muestreo mendeliano y $\mathbf{D_i}$ es una matriz diagonal, cuyos elementos son los ponderadores que determinan la contribución de cada varianza a la varianza de muestre mendeliano. De esta forma, los elementos de la matriz diagonal $\mathbf{G_e}$ son funciones lineales de las varianzas aditivas de las poblaciones parentales (Elzo, 1990; Lo et al., 1993) y las varianzas que surgen de la segregación de alelos entre poblaciones con diferentes frecuencias génicas (Lo et al., 1993). Para simplificar la notación, denotaremos a ϕ_a como el vector columna de ngv varianzas aditivas, donde ngv es igual a m(m+1)/2, y m es el número de poblaciones parentales.

En la asignación de distribuciones a priori informativas para componentes de varianza, la forma más utilizada es la distribución gamma invertida, IG(□,□), dentro de la cual la distribución de chi-cuadrado invertida es una caso especial. Esto último se justifica por el hecho que la distribución posterior de un componente de varianza en un modelo lineal es, generalmente, proporcional a esta distribución. Sin embargo, para el modelo considerado en este trabajo, las densidades posteriores de las varianzas aditivas no son proporcionales a ninguna distribución conocida. De esta forma, para los parámetros de dispersión genéticos, la distribución IG representa una a priori no conjugada, debido a que las distribuciones a priori y a posteriori no tienen la misma forma funcional, mientras que para los parámetros de dispersión ambientales ésta representa una a priori conjugada natural. Las respectivas distribuciones a priori para las ngv varianzas genéticas y nrv varianzas ambientales fueron,

$$p(\phi_a \mid \mathbf{v_u}, \mathbf{s_u}) = \prod_{i=1}^{ngv}(\phi_{ai} \mid v_{ai}, S_{ai}^2) = \prod_{i=1}^{ngv}\left[\frac{1}{\Gamma\left(\frac{1}{2}v_{ai}\right)}\left(\frac{1}{2}v_{ai}S_{ai}^2\right)^{\frac{v_{ai}}{2}}(\phi_{ai})^{-\frac{1}{2}(v_{ai}+2)}\exp\left\{\frac{-\frac{1}{2}v_{ai}S_{ai}^2}{\phi_{ai}}\right\}\right]$$

$$p(\phi_e \mid \mathbf{v_e}, \mathbf{s_e}) = \prod_{i=1}^{nrv}(\phi_{ei} \mid v_{ei}, S_{ei}^2) = \prod_{i=1}^{nrv}\left[\frac{1}{\Gamma\left(\frac{1}{2}v_{ei}\right)}\left(\frac{1}{2}v_{ei}S_{ei}^2\right)^{\frac{v_{ei}}{2}}(\phi_{ei})^{-\frac{1}{2}(v_{ei}+2)}\exp\left\{\frac{-\frac{1}{2}v_{ei}S_{ei}^2}{\phi_{ei}}\right\}\right]$$

En estas expresiones, $\mathbf{s_u}$ ($\mathbf{s_e}$) es un vector cuyos elementos, S_{ai}^2 (S_{ei}^2), pueden ser interpretados como los valores a priori de ϕ_{ai} (ϕ_{ei}), $\mathbf{v_u}$ ($\mathbf{v_e}$) es un vector cuyos elementos, v_{ai} (v_{ei}), pueden interpretarse como el grado de credibilidad a priori sobre el valor a priori del correspondiente componente de varianza, y $\Gamma(.)$ es la función gamma.

En este trabajo, se presenta un algoritmo para obtener los elementos de la moda conjunta de los parámetros de dispersión. La derivación de este tipo de estimador Bayesiano, denominado como estimador máximo a posteriori (MAP), se basa en el hecho de que la moda de la distribución posterior conjunta de los componentes de varianza se puede obtener desde la distribución posterior conjunta de todos los parámetros desconocidos sin la utilización de integración directa de los parámetros de posición, utilizando la teoría del algoritmo EM generalizado (Dempster et al., 1977) para maximizar la distribución posterior más que la función de verosimilitud.

RESULTADOS Y DISCUSION

Los estimadores MAP de los componentes de varianza para el modelo aditivo heterocedástico, se pueden obtener a través del siguiente esquema iterativo:

$$\phi_{ai}^{[t+1]} = \left[\frac{\hat{u}' A_i^* \hat{u} + tr\left(A_i^* C^{uu}\right) + v_{ai} S_{ai}^2}{tr\left[G_{ei}^*\right] + v_{ai} + 2} \right]^{[t]} \qquad \phi_{ei}^{[t+1]} = \left[\frac{\hat{e}_i' \hat{e}_i + tr(W_i C W_i') + v_{ei} S_{ei}^2}{n_i + v_{ei} + 2} \right]^{[t]}$$

donde, $[\ .\ ]^{[t]}$ denota la correspondiente expresión, evaluada en $\phi_a = \phi_a^{[t]}$ y $\phi_e = \phi_e^{[t]}$, respectivamente, $A_i^* = (I - P')G_{ei}^* D_i^{-1} G_{ei}^* (I - P)$, D_i^{-1} es una inversa generalizada de D_i, $\hat{u} = E\left(u | y, \phi^{[t]}\right)$, C^{uu} es la submatriz de la inversa de las ecuaciones de los modelos mixtos de Henderson correspondiente a los valores aditivos, $\hat{e}_i = E\left(e_i | y, \phi^{[t]}\right)$, W_i es la matriz de incidencia correspondiente al estrato de datos definido, C es la inversa de las ecuaciones de los modelos mixtos de Henderson y $G_{eiii}^* = D_{iii}\phi_{ai} \Big/ \sum_{j=1}^{ngv} D_{iii}\phi_{aj}$.

Los estimadores propuestos pueden ser explicados como una combinación lineal de dos fuentes de información: un estadístico basado en los datos, el estimador REML ($\tilde{\sigma}_i^2$), y el parámetro de posición de la distribución a priori del componente de varianza respectivo, tal que:

$$\phi_{ai}^{[t+1]} = \left[\frac{tr\left[G_{ei}^*\right]\tilde{\sigma}_{ai}^2 + v_{ai} S_{ai}^2}{tr\left[G_{ei}^*\right] + v_{ai} + 2} \right]^{[t]} \qquad \phi_{ei}^{[t+1]} = \left[\frac{n_i \tilde{\sigma}_{ei}^2 + v_{ei} S_{ei}^2}{n_i + v_{ei} + 2} \right]^{[t]}$$

donde,

$$\tilde{\sigma}_{ai}^{2[t]} = \left[\frac{\hat{u}' A_i^* \hat{u} + tr\left(A_i^* C^{uu}\right)}{tr\left[G_{ei}^*\right]} \right]^{[t]} \qquad \tilde{\sigma}_{ei}^{2[t]} = \left[\frac{\hat{e}_i' \hat{e}_i + tr(W_i C W_i')}{n_i} \right]^{[t]}$$

De esta forma, cuando la información sobre el parámetro de dispersión contenida en los datos es grande, el componente REML del estimador domina. En otro caso, el parámetro de posición de la distribución a priori es ponderado en mayor forma. De esta forma, el método propuesto debe diferir considerablemente de las estimaciones REML cuando existe poca información en los datos sobre el parámetro de dispersión.

Por simplicidad, efectos genéticos no-aditivos no fueron considerados en este trabajo. Sin embrago, si este tipo de variación genética no puede ser completamente ignorada, el modelo debe ser modificado para tomar en cuenta estos efectos. Lo et al. (1995), desarrollaron un modelo bajo dominancia para dos razas y cualquier cruza involucrada. Este modelo requiere de 25 componentes de varianza, los que se reducen a 12 en ausencia de consanguinidad. El número de varianzas aumenta cuando se consideran más de dos razas. La necesidad para estimar estos componentes de variación representa una desventaja para el uso de este modelo para situaciones generales. Por otra parte, no existe aun, ningún algoritmo eficiente para invertir la matriz de covarianzas de los efectos genotípicos. Sin embargo, la utilización de supuestos apropiados podría permitir la incorporación de estos efectos. Elzo (1990b), modela los efectos genéticos no-aditivos, en términos de interacciones entre los alelos de padres y madres de diferentes grupos raciales en uno o dos loci y desarrolló procedimientos recursivos para calcular la inversa de la matriz de covarianzas, en modelos de subclases y de regresión. Bajo este tipo de supuestos, las covarianzas entre interacciones se pueden explicar por combinaciones lineales de las varianzas de interacciones alélicas intra- y entre-locus, intra- y entre-razas. Entonces, la inclusión de estos efectos puede ser fácilmente manejada por el procedimiento descrito aquí.

Los estimadores de las varianzas del error heterogéneas en este trabajo fueron derivados en una forma general, sin ninguna especificación sobre la fuente de heterogeneidad ambiental. Sin embargo, con relación a este punto, la discusión se centrará sobre la estimación de varianzas ambientales a través de grupos raciales. En este caso, podría ser necesario estimar un número muy grande de varianzas ambientales, y cada parámetro individual podría no ser bien estimado en la medida que exista poca información por subpoblación. Posibles alternativas podrían ser, estratificar los grupos raciales en grupos de mayor tamaño, o, como propone Elzo (1994), asumir que las varianzas ambientales dependen solo de la composición racial del individuo y sus padres. En este último caso, las varianzas

ambientales pueden ser calculadas como combinaciones lineales de las varianzas intra- y entre- razas, reduciendo el número de parámetros de dispersión al mismo número de varianzas genético-aditivas (nrv=ngv). Otra posibilidad, es que las estimaciones sean regresadas sobre una varianza ambiental común. Gianola et al. (1992), derivan métodos Bayesianos para la estimación de varianzas subpoblacionales que combinan, en una forma óptima, la información dentro de cada subpoblación con estimaciones obtenidas a través de las subpoblaciones, y sugieren la estimación del hiperparámetro de la varianza del error, S^2_e, utilizando estimaciones REML como si las varianzas fuesen homogéneas. Weigel y Gianola (1992), muestran, a través de un estudio de simulación, que esta estrategia mejora la convergencia y la exactitud de los componentes de varianza estimados.

CONCLUSIONES

La aproximación Bayesiana desarrollada en el presente trabajo, permite la estimación de componentes de varianza en características gobernadas por efectos aditivos en sistemas abiertos de cruzamiento. Los estimadores permiten una asignación diferencial de grados de creencia para describirla incertidumbre a priori sobre los parámetros de dispersión. Aunque efectos genéticos no-aditivos no fueron considerados en los procedimientos descritos en este trabajo, su inclusión pude ser fácilmente manejada a través de esta aproximación a través de la utilización de supuestos sobre el número de loci involucrados.

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. Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Satist. Soc. B. 39: 1-38.
2. Elzo, M.A. 1990a. Recursive procedures to compute the inverse of the multiple trait additive genetic covariance matrix in inbred and noninbred multibreed populations. J. Anim. Sci. 68: 1215-1228.
3. Elzo, M.A. 1990b. Covariances among sire by breed group of dam interaction effect in multibreed sire evaluation procedures. J. Anim. Sci. 68: 4079-4099.
4. 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.
5. Gianola, D., J.L. Foulley, R.L. Fernando, C.R. Henderson, and K.A. Weigel. 1992. Estimation of heterogeneous variances using empirical Bayes methods: theoretical considerations. J. Dairy Sci. 75: 2805-2823.
6. Lo, L.L., R.L. Fernando, and M. Grossman M. 1993. Covariance between relatives in multibreed populations: additive model. Theor. Appl. Genet. 87: 423-430.
7. Lo, L.L., R.L. Fernando, R.J.C. Cantet, and M. Grossman. 1995. Theory for modelling means and covariances in a two-breed population with dominance inheritance,.Theor. Appl. Genet. 90: 49-62.
8. Weigel, K.A., and D. Gianola. 1992. Estimation of heterogeneous within-herd variance components using Bayes methods: a simulation study. J. Dairy Sci. 75: 2824-2833.