Inferencia bayesiana sobre los parámetros de dispersión genéticos y ambientales en modelos animales con efectos maternos

Tesis presentada para optar al título de Doctor de la Universidad de Buenos Aires, Área Ciencias Agropecuarias

Sebastián Munilla Leguizamón
Ing. Agr. - FAUBA - 2004

Lugar de trabajo: Facultad de Agronomía, Universidad de Buenos Aires

Escuela para Graduados Ing. Agr. Alberto Soriano
Facultad de Agronomía – Universidad de Buenos Aires

COMITÉ CONSEJERO

Director de tesis
Rodolfo Juan Carlos Cantet
Ing. Agr. (Universidad de Buenos Aires, Argentina)
MSc. (Montana State University, Estados Unidos de América)
MSc. (University of Illinois, Estados Unidos de América)
Ph.D (University of Illinois, Estados Unidos de América)

Consejera de Estudios
Zulma Gladis Vitezica
Ing. Agr. (Universidad de Buenos Aires, Argentina)
MSc. (Universidad de Buenos Aires, Argentina)
Docteur (Institut National Agronomique Paris-Grignon, Francia)

JURADO DE TESIS

Director de tesis
Rodolfo Juan Carlos Cantet
Ing. Agr. (Universidad de Buenos Aires, Argentina)
MSc. (Montana State University, Estados Unidos de América)
MSc. (University of Illinois, Estados Unidos de América)
Ph.D (University of Illinois, Estados Unidos de América)

JURADO
Daniel Omar Maizon
Méd. Vet. (Universidad de Buenos Aires)
MSc. (Universidad de Buenos Aires, Argentina)
Ph.D (Cornell University, Estados Unidos de América)

JURADO
Ignacio Aguilar García
Ing. Agr. (Universidad de la República, Uruguay)
MSc. (University of Georgia, Estados Unidos de América)
Ph.D (University of Georgia, Estados Unidos de América)

JURADO
Fernando Flores Cardoso
Méd. Vet. (Universidade Federal de Pelotas, Brasil)
MSc. (Universidade Federal de Pelotas, Brasil)
MSc. (Michigan State University, Estados Unidos de América)
Ph.D (Michigan State University, Estados Unidos de América)

Fecha de defensa de la tesis: 31 de octubre de 2011

Los hombres nobles eluden un esfuerzo realizando otro mucho mayor. Por no arrancar una rosa, construyen un palacio. Por no escuchar un reproche, ejercen la rectitud toda la vida. Por no bajarse del caballo, conquistan el Asia (Alejandro Dolina, "Crónicas del ángel gris")

A mi dulce señora Sofía de La Marque, Princesa de la Luna.

A la memoria de mis abuelos.

AGRADECIMIENTOS

Quiero reconocer a todas aquellas personas e instituciones que de mil maneras diferentes han contribuido a que esta tesis haya visto la luz. Si algún mérito tiene este trabajo, es gracias a todos ellos.

- A los miembros del jurado, Daniel Maizon (Argentina), Ignacio Aguilar (Uruguay) y Fernando Cardoso (Brasil), por tomarse el trabajo desinteresado de leer el manuscrito y venir a Buenos Aires a escuchar y discutir sobre los resultados de esta investigación.

- A la sociedad Argentina en su conjunto, por haberme permitido desarrollar mis estudios de posgrado, financiando mi manutención mediante dos becas de estudio. Mi objetivo profesional no es otro que el de retribuir semejante generosidad.

- A las instituciones que me otorgaron dichas becas, la Universidad de Buenos Aires y el Consejo Nacional de Investigaciones Científicas y Técnicas.

- Al Dr. Claudio Fioretti, de la empresa Estancias y Cabañas Las Lilas S.A., Argentina, y al Dr. Chris Morris, del AgResearch Crown Research Institute, Nueva Zelanda, por facilitarme los datos de campo y experimentales utilizados en esta tesis.

- A mi hogar académico, el Departamento de Producción Animal de la Facultad de Agronomía de la Universidad de Buenos Aires, por la disponibilidad de sus instalaciones y recursos.

- A todo el personal de la Escuela para Graduados "Alberto Soriano".

- A los colegas y amigos de las cátedras de Anatomía y Fisiología Animal, Bovinos de Carne y Nutrición y Alimentación Animal, por tantos momentos vividos durante este tiempo.

- A mi amigo personal, Mario Javier Cosentino, y sus infaltables mates de las cinco.

- A mi familia académica, el grupo de Mejoramiento Genético Animal de la FAUBA. Todos y cada uno de ellos ha constituido un pilar fundamental en todos estos años.

    - A Ana Birchmeier, quien me enseñó mucho de lo que aprendí.

    - A Valeria Schindler, por sus incansables gestiones para incorporarme formalmente a la Cátedra.

    - A mis hermanos académicos, Eduardo Cappa, Lenin Ron, Oscar Rhodas, María José Suárez, Yeni Bernal, José Luis Gualdrón, Natalia Forneris y Juan David Corrales, por su amistad.

- A mi consejera de estudios, Zulma Vitezica, por su tiempo y disposición.

- Y, por sobre todo, a mi padre académico, Rodolfo J. C. "Fito" Cantet. El uso de la palabra "padre" no es arbitrario. Fito fue muchísimo más que un Director de tesis, pues no sólo me orientó, acompañó y auxilió durante cada una de las etapas de este trabajo, sino que a él le debo mi formación integral como investigador científico y, por decantación, como persona. Siempre admiré su compromiso con el trabajo, con la ciencia y con el país. Su ejemplo me inspira, y espero ser un fiel depositario de los conocimientos que me transmitió. Por todo esto, le estaré eternamente agradecido. Es una deuda que jamás podré saldar.

- Por último, quiero agradecer a esa enorme red de contención que constituyen mi familia, mis amigos, mis padres y, muy especialmente, mi mujer, Sofi, que siempre caminó a mi lado en este sendero.

Declaro que el material incluido en esta tesis es, a mi mejor saber y entender, original producto de mi propio trabajo (salvo en la medida en que se identifique explícitamente las contribuciones de otros), y que este material no lo he presentado, en forma parcial o total, como una tesis en ésta u otra institución.

Sebastián Munilla Leguizamón

PUBLICACIONES DERIVADAS

1. Munilla Leguizamón, S. y R. J. C. Cantet. 2010. Equivalence of multibreed animal models and hierarchical Bayes analysis for maternally influenced traits. *Genet. Sel. Evol.*, 42(20): 1–12.

2. Munilla Leguizamón, S. y R. J. C. Cantet. 2010. Estimation of residual dam-offspring correlation for a maternal animal model through a Griddy Gibbs Sampler. En *9th World Congress on Genetics Applied to Livestock Production*, Leipzig, Alemania.

3. Munilla, S. y R. J. C. Cantet. 2011. Bayesian conjugate analysis using a generalized inverted Wishart distribution accounts for differential uncertainty among the genetic parameters – an application to the maternal animal model. *J. Anim. Breed. Genet*. (En prensa).

ÍNDICE GENERAL

página

ÍNDICE DE TABLAS

TABLA

TABLA

página

ÍNDICE DE FIGURAS

FIGURA página

ABREVIATURAS

AM Modelo animal.

BLUP Predictor lineal insesgado de mínima varianza.

cap. Abreviatura de 'capítulo'.

cdf Función de densidad acumulada (en inglés: *cumulative density function*).

cf. Abreviatura del latín *confer* ('compare' o 'consulte'). Se utiliza para indicar 'para más detalle, consúltese…'.

CVC Componentes de (co)varianza.

EE Error estándar.

e.g. Acrónimo del latín *exempli gratia* ('dado como ejemplo'). Se utiliza para indicar 'véase, por ejemplo…'.

ESS Tamaño efectivo de muestra.

GGS Algoritmo de estimación conocido en la literatura como '*Griddy Gibbs sampler*' (*cf.* Ritter y Tanner, 1992).

GS Muestreo de Gibbs. Refiere al algoritmo de estimación en el contexto de un análisis bayesiano.

h^2 Heredabilidad en sentido estricto.

h^2_T Heredabilidad total. En el contexto de un carácter bajo la influencia de efectos maternos, se utiliza para definir la relación entre la respuesta a la selección esperada y el diferencial de selección realizado.

i.e. Acrónimo del latín *id est* ('esto es'). Se utiliza para indicar 'es decir, …'.

IW Wishart invertida.

MAM Modelo animal con efectos maternos.

MBAM Modelo animal multirracial.

MCMC Cadenas de Markov y simulación de Monte Carlo. Refiere a un conjunto de métodos de simulación por Monte Carlo basados en la teoría de las cadenas de Markov.

ML Máxima verosimilitud.

MME Ecuaciones del modelo mixto.

NMV Distribución normal multivariada.

pág. Abreviatura de ‘página’.

PSRF Factor potencial de reducción de escala. Estadístico que se utiliza para monitorear convergencia a partir de múltiples cadenas MCMC (*cf*. Gelman y Rubin, 1992)

REML Máxima verosimilitud restringida.

$S\chi_{\upsilon}^{-2}$ Distribución Chi-cuadrada escalada invertida con parámetros (υ, S).

Título: Inferencia bayesiana sobre los parámetros de dispersión genéticos y ambientales en modelos animales con efectos maternos

RESUMEN

Los modelos ‘modelos animales con efectos maternos’ (MAM) son modelos lineales mixtos que se utilizan para ajustar registros de caracteres bajo la influencia de efectos maternos. Uno de los desafíos más importantes en el marco de los MAM es la estimación de los parámetros de dispersión o ‘componentes de (co)varianza’ (CVC). En esta tesis se introducen desde una perspectiva bayesiana contribuciones teóricas y metodológicas con relación a la estimación de CVC para MAM sujetos a estructuras de covarianza novedosas. En primer lugar, se describe una implementación del análisis bayesiano jerárquico vía el algoritmo del muestreo de Gibbs. Luego, se considera una especificación conjugada diferente para la distribución a priori de la matriz de covarianza genética, basada en la distribución Wishart invertida generalizada, y se presenta una estrategia para determinar los correspondientes hiperparámetros. Esta estrategia fue comparada contra otras especificaciones a priori mediante un estudio de simulación estocástica, y produjo estimaciones precisas de los parámetros genéticos, con menores errores estándares y mejor tasa de convergencia. En segundo lugar, se presenta una formulación alternativa del MAM que incluye un parámetro de correlación ambiental entre pares de observaciones madre–progenie, y se desarrolla un procedimiento de estimación basado en un algoritmo de muestreo por grilla. El procedimiento fue programado y ejecutado exitosamente, y se obtuvo la primera estimación del parámetro de correlación con datos de campo para peso al destete en bovinos de carne. Por último, se considera el problema de la estimación de CVC en una población multirracial, donde en general es necesario especificar una estructura de covarianza heterogénea para los valores de cría. En particular, se demuestra que el modelo basado en la descomposición de la matriz de covarianza genética es equivalente al que deriva de la teoría genética cuantitativa. Además, se extiende el modelo para incluir efectos maternos y se describe la implementación de un análisis bayesiano jerárquico con el objetivo de estimar los CVC. El procedimiento fue implementado con éxito en datos experimentales de peso al destete y se obtuvieron por primera vez estimaciones para el conjunto completo de CVC.

Palabras claves: Efectos maternos, componentes de (co)varianza, estimación de parámetros, análisis bayesiano jerárquico, muestreo de Gibbs, distribución Wishart invertida generalizada, correlación madre–progenie, poblaciones multirraciales.

Title: Bayesian inference about genetic and environmental dispersion parameters in maternal animal models

ABSTRACT

'Maternal animal models' (MAM) are mixed linear models used to fit records on maternally influenced traits. The estimation of the dispersion parameters or '(co)variance components' (CVC) under a MAM is a challenging task. In this thesis, theoretical and methodological developments in connection with CVC estimation for MAM subject to novel covariance structures are introduced under a Bayesian perspective. First, a standard hierarchical Bayes analysis via the Gibbs sampler is described. Next, a different conjugate specification for the prior distribution of the genetic covariance matrix through the generalized inverted Wishart distribution is considered, and a strategy to elicit the corresponding prior parameters is further developed. Using simulated data, the strategy returned accurate estimates and reduced standard errors when compared with other prior specifications, while improving the convergence rates. Second, an alternative formulation for the MAM including a residual dam–offspring correlation parameter is introduced, and an estimation procedure based on a Griddy Gibbs sampler algorithm is further developed. The procedure was successfully executed, and the correlation parameter was estimated for the first time using weaning weight records of beef calves. Finally, the problem of CVC estimation in a multibreed population is considered, where specifying a heterogeneous genetic covariance structure is usually mandatory. In this regard, it is shown that a model based on the decomposition of the genetic covariance matrix is equivalent to the one that arises using quantitative genetics arguments. Furthermore, the model is extended to include maternal effects and a hierarchical Bayes analysis implementation is described. The implementation using experimental weaning weight records was accomplished successfully and, for the first time, estimates of the full set of CVC were obtained.

Key words: Maternal effects, (co)variance components, parameter estimation, hierarchical Bayes analysis, Gibbs sampling, generalized inverted Wishart distribution, dam–offspring correlation, multibreed populations.

1
Introducción general

1.1. INTRODUCCIÓN

Uno de los objetivos más importantes del mejoramiento genético animal es la predicción del mérito genético de los individuos candidatos a la selección para determinados caracteres de importancia económica. Actualmente, mediciones de estos caracteres tomadas sobre los individuos se ajustan empleando una serie de modelos lineales mixtos particulares a la disciplina, los 'modelos animales' (AM). Bajo un AM, el mérito genético o, más estrictamente, el 'valor de cría' de un individuo es tratado como un efecto aleatorio sobre el fenotipo, y su valor realizado se estima (se 'predice', en la concepción original de Henderson *et al*., 1959) a través de su 'predictor lineal insesgado de mínima varianza' (BLUP, según sus siglas en inglés) (*cf*. Henderson, 1984, cap. 5). Las predicciones BLUP, por su parte, se obtienen resolviendo un sistema de ecuaciones conocido como las 'ecuaciones del modelo mixto' (MME) (*cf*. Henderson, 1984, cap. 3).

Para algunos caracteres de interés, el valor medido sobre el individuo, o 'valor fenotípico', puede atribuirse a dos o más caracteres componentes. En aquellas especies en las que miembros de una familia dependen o están próximos los unos a los otros, individuos emparentados pueden contribuir con alguno de estos caracteres componentes (Willham, 1963) y ejercer, en consecuencia, un efecto sobre el valor fenotípico. En particular, cuando el efecto es contribuido por la madre del individuo, se lo denomina 'efecto materno'. Considérese, por ejemplo, el carácter peso al destete en mamíferos. En este caso, el ambiente que provee una madre a su cría durante todo su desarrollo predestete incide notoriamente sobre la performance de esta última (*e.g*. Koch, 1972). Los atributos de una madre que podrían afectar el peso al destete de su cría incluyen, entre otros, el desarrollo durante la gestación, la inmunidad transmitida, la producción de leche y las relaciones de comportamiento madre-hijo. En la Figura 1.1 se esquematiza la contribución genética de los progenitores a su cría. Nótese que desde el punto de vista del individuo, el efecto materno es de naturaleza ambiental. Sin embargo, desde el punto de vista de la madre, puede atribuirse tanto a causas ambientales como genéticas.

Los efectos maternos son de gran importancia para el mejoramiento genético animal fundamentalmente por dos razones. En primer lugar, porque la presencia de efectos maternos puede alterar los resultados esperados en un programa de selección artificial, dado que la tasa y dirección de la respuesta a la selección dependerán de la herencia de caracteres componentes que no están directamente bajo presión de selección (Kirkpatrick y Lande, 1989). En segundo lugar, porque los efectos maternos constituyen muchas veces un objetivo de selección en sí mismos, particularmente en aquellos casos en los que la aptitud materna no puede medirse en forma directa.

En términos del AM, los efectos maternos se ajustan incluyendo un nuevo factor aleatorio en la ecuación del modelo, que recibe entonces el nombre de 'modelo animal con efectos maternos' (MAM). Formalmente, el MAM se basa en la extensión de la teoría de covarianza entre parientes (Cockerham, 1954; Kempthorne, 1954) desarrollada por Willham (1963). La formulación en términos de un modelo lineal mixto, por su parte, se debe a Quaas y Pollak (1980). Bajo el MAM, las predicciones BLUP de los valores de cría de los caracteres componentes, de aquí en más 'directo' y 'materno', se obtienen resolviendo una versión modificada de las MME.

Estas MME dependen de un conjunto de parámetros asociados a las distribuciones de probabilidad de las variables aleatorias definidas bajo el modelo de análisis. Estos parámetros reciben el nombre genérico de 'componentes de (co)varianza' (CVC) (*cf*. Searle *et al*., 1992). Estrictamente, las propiedades estadísticas del predictor BLUP se

basan en asumir que los CVC son conocidos. En la práctica, sin embargo, los CVC deben estimarse previamente a partir de los mismos datos.

En general, la estimabilidad de los CVC dependerá de la estructura de dispersión o, más formalmente, de la 'estructura de covarianza' que el investigador asigne a las variables aleatorias del modelo de análisis. En muchos casos, el investigador debe hallar la estructura de covarianza que mejor describa sus observaciones. En el contexto de los AM, sin embargo, la definición de la estructura de covarianza de los valores de cría se basa en la teoría genética cuantitativa fundada por Fisher (1918). Esta teoría constituye, de hecho, el marco teórico del presente trabajo.

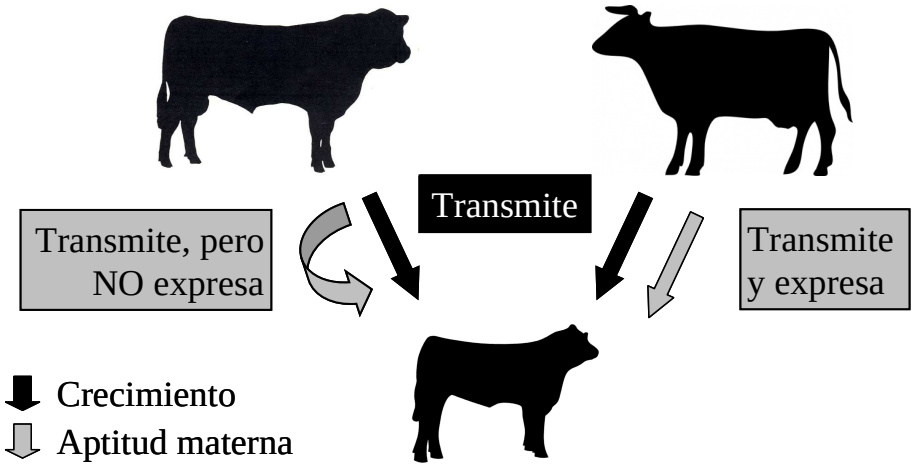


Figura 1.1. Contribución genética de los progenitores a su cría para el carácter peso al destete. Ambos padres transmiten a su progenie genes que afectarán su crecimiento (flechas negras). Además, transmitirán genes asociados con la aptitud materna (flechas grises); naturalmente, estos últimos sólo se expresarán si la cría es una hembra y, a futuro, una madre. La madre, por su parte, no sólo transmitirá genes para la aptitud materna, sino que expresará el carácter y ejercerá así una influencia sobre el crecimiento de su cría (efecto materno).

1.2. MARCO TEÓRICO: LA TEORÍA GENÉTICA CUANTITATIVA

La gran mayoría de los caracteres de importancia económica para la producción animal presentan una distribución continua de valores fenotípicos. Si bien la segregación de los genes que afectan los caracteres de esta naturaleza se rige por las leyes de transmisión mendelianas, tal distribución se explica básicamente por dos factores que interactúan entre sí en la expresión del fenotipo: la segregación de un número grande de genes y la influencia que el ambiente ejerce sobre la expresión de los mismos. Fisher (1918) y Wright (1921a) formalizaron matemáticamente esta idea, y sentaron así las bases de la teoría genética cuantitativa.

El núcleo de la teoría radica en la partición de la variabilidad total observada para un carácter en una determinada población, o 'varianza fenotípica', en sus componentes genética y ambiental. La variabilidad genética, luego, puede atribuirse a diferentes componentes causales, los denominados 'efectos génicos'. Estos son: 1. la 'aditividad' (el efecto independiente de un alelo sobre el carácter); 2. la 'dominancia' (el efecto de la interacción entre alelos dentro de un mismo locus); y 3. la 'epistasis' (el efecto de las

interacciones entre alelos en diferentes loci). Es importante destacar que esta definición de los efectos génicos es una abstracción estadística, y no necesariamente refiere a los mecanismos de acción génica. De hecho, la existencia de epistasis y dominancia a nivel de la acción génica es compatible con altos niveles de variabilidad genética aditiva (Hill *et al*., 2008).

En el marco de la teoría genética cuantitativa, los efectos génicos aditivos sumados sobre todos los loci que gobiernan el carácter constituyen el valor de cría (Falconer y Mackay, 1996, cap. 7). En la práctica, los programas de mejoramiento genético animal se basan en el principio de que el valor fenotípico de un individuo provee información respecto al valor de cría subyacente. El punto crucial para predecir el valor de cría a partir de la información fenotípica radica en la conexión existente entre la similitud de individuos emparentados y las diferentes fuentes de variabilidad asociadas a los efectos génicos. Una vez más, Fisher (1918) fue pionero en desarrollar esta idea. Más tarde, Cockerham (1954) y Kempthorne (1954), en forma independiente, generalizaron la partición de la variabilidad genética en sus múltiples componentes, y establecieron así la teoría de covarianza entre parientes.

En particular, la componente aditiva de la variabilidad genética, o 'varianza aditiva', es el principal determinante del grado en el que los individuos emparentados se asemejan los unos a los otros (Lynch y Walsh, 1998). En este contexto, un concepto de trascendental importancia es el de la 'heredabilidad en sentido estricto' (de aquí en más, 'heredabilidad' o h^2), que resulta del cociente entre la varianza aditiva y la varianza fenotípica. En un programa de selección artificial, la heredabilidad determina la precisión con la cual el valor de cría puede predecirse a partir de la información fenotípica y, en consecuencia, gobierna la tasa de respuesta a la selección del carácter (*cf*. Visscher *et al*., 2008).

En el caso de un carácter bajo la influencia de efectos maternos, se manifiestan nuevas fuentes causales de variabilidad fenotípica asociadas a los efectos génicos y ambientales de la componente materna del carácter (Figura 1.2). Willham (1963) desarrolló una expresión para la covarianza entre parientes en este contexto, y extendió así la teoría existente. La respuesta a la selección será función ahora de la contribución de los efectos génicos aditivos de las componentes directa y materna, y de la correlación genética que exista entre ambas componentes. Esta relación funcional suele sintetizarse bajo el concepto de 'heredabilidad total' (h^2_T) en la literatura de efectos maternos (*e.g*. Meyer, 1992), aunque estrictamente h^2_T define el cociente entre la respuesta a la selección esperada y el diferencial de selección realizado en un contexto de selección fenotípica (Eaglen y Bijma, 2009). En todo caso, la expresión de Willham (1963) permite obtener estimaciones de las diferentes fuentes de variabilidad genética ajustando las correlaciones observadas para diferentes relaciones de parentesco contra las esperadas de acuerdo a la teoría genética cuantitativa. En la sección siguiente se revisará la literatura en este sentido.

1.3. ESTIMACIÓN DE CVC BAJO EL MAM

La estimación de CVC en modelos con efectos maternos es, en esencia, problemática (Meyer, 1992). Conceptualmente, la posibilidad de identificar y cuantificar la magnitud de las diferentes fuentes de variabilidad fenotípica depende del contraste entre las correlaciones teóricas y observadas para individuos emparentados. En caracteres bajo la influencia de efectos maternos, sin embargo, se presentan los siguientes problemas (Will-

ham, 1980): 1. los efectos maternos están confundidos con la contribución directa de la madre a su progenie; 2. puede existir una correlación genética entre los efectos directos y maternos; 3. los efectos maternos están una generación retrasados respecto a los efectos directos en su expresión; 4. la expresión de los efectos maternos está limitada a un sexo; y 5. los efectos maternos ocurren generalmente tarde en la vida de las hembras. Estas consideraciones implican que una correcta estimación de CVC bajo un modelo con efectos maternos requiere una estructura de información de parentescos que involucre, como mínimo, datos de performance de individuos emparentados por vía materna en generaciones sucesivas, por un lado, y datos de performance de individuos emparentados exclusivamente por vía paterna, por otro. En esta sección repasaremos, en orden histórico, los métodos que se sucedieron en la estimación de los CVC bajo el MAM.

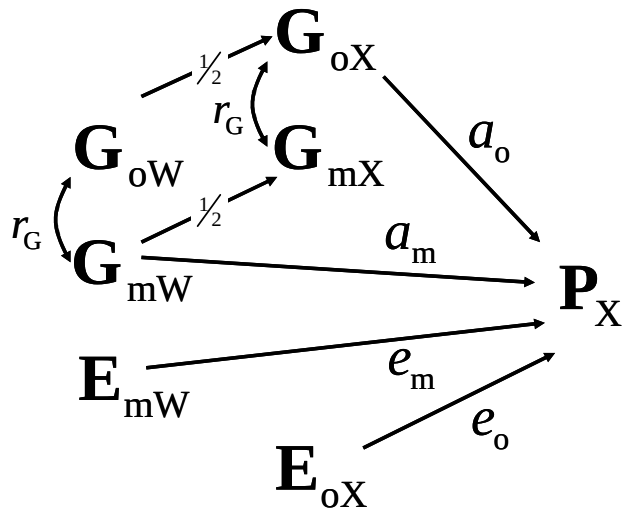


Figura 1.2. Diagrama de coeficientes de paso para un carácter bajo la influencia de efectos maternos. El diagrama de coeficientes de paso (Wright, 1921b) representa la contribución de los efectos génicos (G) y ambientales (E) al valor fenotípico (P) del individuo *X*, hijo de la madre *W*. Los subíndices "o" y "m" representan las vías directas y maternas de determinación, respectivamente. En general, las flechas en un sentido indican la influencia directa de una variable sobre otra, mientras que las flechas en ambos sentidos indican correlación entre las variables. Los símbolos a_o, a_m, e_o y e_m, por último, representan los coeficientes de paso. (Adaptado de Willham, 1963).

El primer método propuesto para estimar CVC bajo el modelo con efectos maternos de Willham fue el de mínimos cuadrados (*cf*. Graybill, 1961). Eisen (1967) presentó tres diseños de apareamientos experimentales concebidos específicamente para proveer un número suficiente de relaciones de parentesco como para obtener estimaciones de todos los parámetros. En este contexto, las estimaciones de los CVC se obtienen resolviendo el sistema de ecuaciones lineales que surge de regresar las covarianzas entre parientes observadas para un carácter de interés en los coeficientes apropiados de acuerdo al modelo de Willham. Una descripción detallada del método puede consultarse en el libro de Lynch y Walsh (1998, cap. 23). Ahora bien, la complejidad de los diseños de Eisen (1967) los hace más apropiados para animales de laboratorio que para especies domésticas (Thompson, 1976). En datos de campo, la información suele estar muy desbalanceada, lo cual limita la factibilidad del método de mínimos cuadrados para la estimación de CVC.

En estos casos, los métodos de estimación por máxima verosimilitud (ML), y en particular el método de máxima verosimilitud restringida (REML) (*cf*. Searle *et al*., 1992, cap. 6), han sido adoptados masivamente. Thompson (1976) introdujo estos métodos en la literatura de efectos maternos, y describió cómo aplicarlos en el contexto de los diferentes diseños de apareamientos propuestos, como los descriptos por Eisen (1967) y Bondari *et al*. (1978). Más tarde, la formulación del MAM en términos de un modelo lineal mixto (Quaas y Pollak, 1980) simplificó notablemente la implementación de los métodos de máxima verosimilitud, al menos desde un punto de vista conceptual (Meyer, 1989). Bajo un modelo lineal mixto, los CVC aparecen asociados a las estructuras de covarianza de los efectos aleatorios del modelo. Actualmente, existen múltiples paquetes estadísticos de uso general que permiten obtener estimaciones de CVC bajo el MAM mediante algoritmos REML, como, por ejemplo, el ASReml (Gilmour *et al*., 2006) o el WOMBAT (Meyer, 2007).

Como alternativa, se han propuesto y difundido también métodos de estimación de CVC basados en el enfoque bayesiano. Gianola y Foulley (1982) introdujeron los métodos bayesianos en el mejoramiento genético animal en el contexto del análisis de caracteres de umbral. Más tarde, Gianola y Fernando (1986) propusieron el enfoque como una estrategia conceptual general para abordar problemas en la disciplina. En particular, Cantet *et al*. (1992a) presentaron un análisis bayesiano de inferencia de CVC bajo el MAM, y revisaron algunos métodos de integración numérica para obtener las distribuciones marginales de los parámetros, un punto crucial de la metodología. En aquel entonces aún no se había difundido la aplicación del algoritmo del 'muestro de Gibbs' (GS) (*cf*. Casella y George, 1992), que simplifica enormemente la implementación del análisis bayesiano. Fueron Jensen *et al*. (1994) quienes finalmente describieron cómo estimar los CVC inherentes a un MAM mediante el GS en el marco de un análisis bayesiano jerárquico. Una extensa cobertura de los métodos bayesianos en el mejoramiento genético animal y, en particular, una descripción detallada del problema de la estimación de CVC bajo el MAM puede consultarse en el libro de Sorensen y Gianola (2002).

1.4. EL ENFOQUE BAYESIANO Y ALGUNOS DESAFÍOS EN EL CONTEXTO DE LA INFERENCIA DE CVC

Bajo el punto de vista bayesiano se asume que todas las incógnitas del modelo de análisis están sujetas a cierto grado de incertidumbre y, en consecuencia, se las trata como variables aleatorias (Sorensen y Gianola, 2002, cap. 5). De este modo, la incertidumbre se representa a través de una distribución de probabilidad. En particular, todo el conocimiento del que se dispone respecto a las incógnitas del modelo antes de la recopilación de las observaciones se representa mediante la distribución de probabilidad 'a priori'. La idea básica de los métodos de inferencia bayesianos, que se desprende directamente del teorema de Bayes, consiste en actualizar este 'conocimiento' a priori con la información que proveen las observaciones. Los resultados quedarán finalmente expresados mediante una distribución de probabilidad 'posterior' y, en consecuencia, se pueden interpretan en términos probabilísticos. Así, el enfoque bayesiano constituye una forma intuitiva de abordar un problema, donde las distribuciones de probabilidad a priori cifrarán grados de credibilidad respecto a ciertos valores que pueden tomar las incógnitas del modelo, y los datos luego permitirán actualizar dicha credibilidad, ya sea fortaleciéndola o debilitándola (Pearl, 2000).

En este trabajo se adherirá al enfoque bayesiano fundamentalmente por tres motivos. En primer lugar, porque el análisis bayesiano es más sencillo de implementar, es-

pecialmente en lo referente a la programación de algoritmos de estimación para modelos jerárquicos complejos (Misztal, 2008) como los que se discutirán aquí. En segundo lugar, porque en el contexto de la inferencia de parámetros de dispersión los resultados obtenidos son más informativos. Mientras los algoritmos REML devuelven estimaciones puntuales de los CVC, acompañadas en general de alguna medida de error basada en teoría asintótica, los algoritmos bayesianos proveen toda una distribución de probabilidad sobre la cual reportar los resultados; de hecho, en el caso trivial de asumir distribuciones a priori no informativas, el modo de la distribución posterior conjunta de los CVC se corresponderá con las estimaciones REML (Sorensen y Gianola, 2002). En tercer lugar, porque el enfoque bayesiano y, en particular, la "propiedad de 'memoria' del teorema de Bayes" (Gianola y Fernando, 1986) constituyen una forma intuitiva de abordar el problema de la estimación de CVC en el contexto de la disciplina, donde las bases de datos que se utilizan para realizar inferencias se actualizan constantemente.

El enfoque bayesiano no está exento de desafíos. La principal crítica que reciben los métodos bayesianos se refiere a la arbitrariedad con la que se asignan las distribuciones de probabilidad a priori en la mayoría de las implementaciones. Una excelente discusión al respecto, centrada en el problema de la inferencia en el mejoramiento genético animal, puede consultarse en la revisión de Blasco (2001). En este trabajo, en cambio, se abordará un desafío más puntual, con relación a la implementación del GS en el contexto de la estimación de CVC bajo el MAM: la restricción que impone el muestreo de los CVC genéticos a partir de la distribución multivariada Wishart invertida (IW).

En el marco de un análisis bayesiano jerárquico, la IW es la alternativa natural para modelar la distribución a priori de los CVC genéticos, parámetros asociados con la estructura de covarianza conjunta de los valores de cría directos y maternos. Sin embargo, presenta una limitante importante: mientras que contiene un conjunto completo de parámetros para modelar los posibles valores que el analista considera que pueden tomar los diferentes CVC genéticos a priori, la incertidumbre respecto a estos posibles valores está gobernada por un único parámetro escalar (Brown, 2002). Esta limitante genera básicamente dos problemas. En primer lugar, no permite modelar la incertidumbre diferencial que puede existir entre los parámetros de dispersión directos y maternos. En segundo lugar, al aplicar el GS se suelen observar altísimas correlaciones de muestreo entre los CVC genéticos, lo cual aumenta considerablemente la demanda en tiempo de cómputo del algoritmo para asegurar su convergencia. En uno de los capítulos de esta tesis se abordará este problema, y se considerará el uso alternativo de la distribución Wishart invertida generalizada (GIW).

1.5. ESTIMACIÓN DE COMPONENTES DE (CO)VARIANZA ASOCIADOS A NUEVAS FUENTES DE VARIABILIDAD FENOTÍPICA

Por otro lado, también se abordará en este trabajo el problema de la identificación y estimación de CVC asociados a fuentes de variabilidad fenotípica no contempladas en la formulación del MAM 'clásico' (*i.e.* el modelo basado en la formulación de Willham, 1963). En primer lugar, se tratará el problema de la posible existencia de una correlación de naturaleza ambiental entre los valores fenotípicos de una madre y su progenie. En segundo lugar, se discutirá la extensión de la formulación del MAM clásico para contemplar fuentes adicionales de variabilidad genética asociadas a la segregación de alelos con diferentes frecuencias génicas en poblaciones con individuos de diferente composición racial. Con la premisa de contextualizar estos problemas, y antes de formular explícitamente los objetivos generales de esta tesis, en esta sección se revisará la lite-

ratura pertinente. Motivará la discusión que sigue la referencia al carácter peso al destete en bovinos de carne. Es importante destacar, sin embargo, que la discusión y metodologías que se describirán en este trabajo aplican directamente a otras especies y a otros caracteres bajo la influencia de efectos maternos.

1.5.1. Correlación 'ambiental' madre–progenie

En bovinos de carne, estimaciones muy negativas de la correlación genética directa-materna para peso al destete son frecuentes en la literatura. Si bien cierta asociación genética adversa entre efectos directos y maternos es aceptada entre los investigadores, correlaciones tan negativas son tomadas con escepticismo (Meyer, 1997). En general, se acepta que las estimaciones están sesgadas por una asociación negativa de naturaleza ambiental entre efectos maternos en generaciones adyacentes (Baker, 1980). Esto implica que el ambiente materno provisto por una hembra impactaría sobre la futura aptitud materna de sus hijas, y que el MAM no contempla esta fuente de covariación de origen ambiental. Koch (1972) describió varios estudios experimentales aportando una fuerte evidencia sobre la existencia de este fenómeno. Aparentemente, este impacto estaría asociado al nivel nutricional que recibe una hembra durante etapas tempranas de su crecimiento, una condición que ha sido denominada 'síndrome de la ubre engrasada' ("*fatty udder syndrome*"), y que operaría incluso sobre animales criados bajo condiciones pastoriles (Koch, 1972). En el marco del MAM, varias formulaciones han sido propuestas para tener en cuenta este antagonismo de origen ambiental por vía ancestral materna.

En una primera aproximación al problema, Willham (1972) sugirió incluir en el modelo el efecto de abuela materna. Utilizando este modelo, Dodenhoff *et al*. (1998) obtuvieron estimaciones de la varianza de abuela materna para tres líneas de animales Hereford, y concluyeron que los efectos de abuela materna podrían ser importantes para el carácter peso al destete. Nótese que el modelo limita la relación recursiva entre madres a una generación. Quintanilla *et al*. (1999), por su parte, presentaron un modelo que considera la correlación entre efectos ambientales maternos permanentes entre madres en toda la línea de parentesco por vía materna, y obtuvieron estimaciones de este parámetro de correlación mediante métodos bayesianos. En ambos casos, los modelos alternativos redujeron la magnitud de la estimación de la correlación genética directa-materna respecto al MAM clásico, aunque Quintanilla *et al*. (1999) reportaron estimaciones sesgadas con datos simulados.

Una segunda línea de investigación, fundada sobre el trabajo de Falconer (1965), incorpora al modelo un término de regresión en el fenotipo de la madre para el mismo carácter (*e.g*. Robinson, 1996). Koerhuis y Thompson (1997), por ejemplo, ajustaron una serie de modelos de esta naturaleza a dos archivos de datos en pollos parrilleros. Meyer (1997), por su parte, hizo lo propio con datos experimentales y de campo de peso al destete en bovinos de carne de varias razas. Una vez más, estos modelos alternativos redujeron la magnitud de la estimación de la correlación genética directa-materna respecto al MAM clásico. Sin embargo, algunos autores han cuestionado que este enfoque altera la definición original de los efectos directos (*e.g*. Bijma, 2006).

En todo caso, si el valor fenotípico de la madre para un carácter ejerce una influencia en el valor fenotípico de su progenie para el mismo carácter, el efecto resulta en una correlación madre–progenie de naturaleza ambiental (Koerhuis y Thompson, 1997). En este sentido, Cantet (1990) sugirió introducir una covarianza entre los efectos ambientales maternos permanentes y el error del modelo. Este enfoque, sin embargo, genera una estructura de covarianza del error no lineal en los parámetros, lo cual dificulta la

estimación de los CVC mediante los métodos estadísticos utilizados comúnmente en el mejoramiento genético animal. Para evitar este problema, Bijma (2006) ajustó una serie de tiempo de medias móviles a esta estructura de la varianza del error, y obtuvo estimaciones insesgadas de la correlación genética directa-materna con datos simulados. Este procedimiento, sin embargo, sólo es válido cuando las madres con registro fenotípico tienen una única cría. En uno de los capítulos de este trabajo presentaremos un procedimiento de estimación aplicable en un contexto más amplio.

1.5.2. Estructura de covarianza genética en poblaciones multirraciales

Considérese, por otro lado, el problema de la estimación de CVC en una población animal formada originalmente a partir del apareamiento de individuos provenientes de poblaciones parentales con diferentes frecuencias génicas, tal como ocurre, por ejemplo, en razas bovinas de reciente formación. En poblaciones 'multirraciales' o 'compuestas' de esta naturaleza coexisten individuos pertenecientes o bien a alguna de las poblaciones parentales o bien a alguno de los diferentes grupos raciales formados por cruzamientos. En estos casos, al ajustar datos de performance es necesario extender la estructura de covarianza de los valores de cría para contemplar nuevas fuentes de variabilidad genética.

En este contexto y bajo el supuesto de herencia aditiva, Lo *et al*. (1993) derivaron la expresión de la varianza genotípica como una función lineal de las varianzas aditivas de cada población parental, por un lado, y una fuente adicional de variabilidad que surge de la diferencia en frecuencias alélicas entre las poblaciones parentales: la 'varianza de segregación' (*cf*. Wright, 1968; Lande, 1981). Cantet y Fernando (1995), por su parte, explicaron cómo utilizar esta expresión para predecir los valores de cría en una población compuesta de dos razas en el marco de un AM, y extendieron luego la formulación a caracteres correlacionados, en general, y al MAM, en particular. La estimación de los CVC genéticos bajo este último modelo es, sin embargo, complicada. Básicamente, el problema radica en que los CVC no pueden factorizarse de la matriz de covarianza genética. Para salvar este problema, García-Cortés y Toro (2006) sugirieron un modelo alternativo basado en la descomposición de la matriz de covarianza en sus componentes por origen racial. Estos autores ilustraron la equivalencia de su modelo mediante un pequeño ejemplo numérico, pero no presentaron la derivación formal. En el último capítulo de esta tesis abordaremos este problema, siempre en el contexto de la estimación de CVC bajo el MAM.

1.6. OBJETIVOS GENERALES Y NATURALEZA DE LA TESIS

En virtud de todo lo expuesto, los objetivos generales de la tesis son: 1. desarrollar métodos bayesianos de estimación de CVC en el contexto de modelos animales con efectos maternos sujetos a diferentes estructuras de covarianza; 2. Ajustar dichos modelos a datos de performance y obtener estimaciones de los CVC mediante los métodos propuestos.

El documento está organizado en seis capítulos, incluyendo este capítulo introductorio. En el Capítulo 2 se describe la estimación de CVC bajo el MAM clásico mediante el algoritmo GS. Este capítulo constituirá el marco de referencia para el resto del trabajo. En particular, las restricciones que impone el muestreo de los CVC genéticos a partir de la distribución IW serán expuestas y discutidas. Luego, en el Capítulo 3, se introduce formalmente la distribución Wishart invertida generalizada y se derivan resultados con

relación a su aplicación en el contexto de la estimación de CVC bajo el MAM. Como se verá oportunamente, una de las propiedades más interesantes de esta distribución es su flexibilidad para especificar las expectativas a priori del analista respecto a la incertidumbre en torno a los valores que pueden tomar los CVC genéticos. En el Capítulo 3 se presenta una estrategia basada en esta propiedad para determinar una especificación a priori 'experta' de la distribución de los CVC, y se evalúa luego mediante un estudio de simulación estocástica.

Por su parte, en los capítulos 4 y 5 se discuten en detalle modelos animales con efectos maternos que incluyen nuevos parámetros de dispersión respecto al MAM clásico, y se introducen y aplican métodos de inferencia bayesianos ideados para estimar los correspondientes CVC. En el Capítulo 4, en primer lugar, se presenta un MAM que incluye un parámetro de correlación en la estructura de covarianza del error entre pares de observaciones madre–progenie. Se introduce luego un procedimiento de estimación basado en un algoritmo conocido como 'Griddy Gibbs sampler' (*cf*. Ritter y Tanner, 1992) y, tras aplicarlo, se discute una estimación del parámetro obtenida a partir de un archivo de datos de peso al destete. En el Capítulo 5, por su parte, se formaliza la equivalencia entre diferentes modelos de análisis multirracial con estructuras de covarianza genética heterogénea, se extiende luego la formulación para incluir efectos maternos y se presentan, finalmente, estimaciones de todos los CVC involucrados, obtenidas a partir de un archivo de datos experimental de un cruzamiento Angus × Hereford.

El Capítulo 6, por último, contiene la discusión general y las conclusiones de la presente investigación. En general, en cada uno de los capítulos que integran este trabajo se describen en detalle los antecedentes, los objetivos específicos, los métodos empleados y los resultados obtenidos.

En este punto es necesario un comentario respecto a la naturaleza y el encuadre general de la presente investigación. Esta es una tesis metodológica. Los tres capítulos que presentan contribuciones originales, es decir, los capítulos 3–5, introducen métodos novedosos para abordar el problema de la estimación de CVC en modelos jerárquicos asociados al análisis de datos de performance para caracteres bajo efectos maternos. Es importante destacar que, en general, la formulación de estos modelos jerárquicos no es una contribución original de este trabajo. Todos ellos, en cambio, presentaban alguna limitación respecto a los métodos disponibles para abordar el problema de la inferencia paramétrica. En consecuencia, el énfasis de la tesis no está en comparar la superioridad relativa de estos modelos respecto al MAM clásico en términos de algún criterio estadístico. En última instancia, esto dependerá de que la estructura de un archivo de datos en particular provea suficiente información para estimar apropiadamente los nuevos parámetros.

Un último comentario. La programación de los algoritmos de inferencia llevó la mayor parte del tiempo y esfuerzo total dedicados a la ejecución del presente trabajo. Todos ellos fueron programados en el lenguaje Fortran 90. Dada su extensión, los códigos completos no serán incluidos en el documento. Sin embargo, pueden ser solicitados al autor. Por otro lado, en los capítulos correspondientes se describirá en detalle cómo adaptar los códigos a partir de la estructura general de un algoritmo de GS. Los códigos de algunos de estos algoritmos de GS están libremente disponibles, como, por ejemplo, los del programa MTGSAM (Van Tassell y Van Vleck, 1996) o aquellos de la colección de programas GIBBSF90 (Mistzal, 2002).

2
Inferencia bayesiana bajo el modelo animal con efectos maternos clásico

2.1. INTRODUCCIÓN

El modelo animal con efectos maternos que, en el contexto de este trabajo, se denominará 'clásico' (y se denotará 'MAM') corresponde al modelo lineal mixto que se utiliza con mayor frecuencia en el ámbito del mejoramiento genético animal para ajustar datos de caracteres bajo la influencia de efectos maternos. Bajo el MAM, la estructura de covarianza genética se basa en la extensión de la teoría de covarianza entre parientes (Cockerham, 1954; Kempthorne, 1954) desarrollada por Willham (1963), y por tal motivo suele citárselo en la literatura como el 'modelo de Willham' o más bien como el 'modelo de Willham reducido', dado que no incluye efectos de dominancia en su formulación. Estrictamente, Willham (1963) derivó la expresión de la varianza genotípica cuando un carácter puede interpretarse como la suma de dos o más caracteres componentes, contribuidos por individuos emparentados. La formulación del MAM en términos de un modelo lineal mixto, particularmente en términos de un modelo animal, se debe a Quaas y Pollak (1980).

En este capítulo se presenta el MAM 'clásico' y se describe luego en detalle la implementación de un análisis bayesiano jerárquico con el fin de estimar los parámetros de interés del modelo, en particular, los componentes de (co)varianza (CVC). El método se ilustra ajustando un archivo de datos de peso al destete en bovinos de carne. El objetivo de este capítulo es definir el marco de referencia sobre el que desarrollarán los próximos capítulos.

2.2. MÉTODOS

2.2.1. El MAM 'clásico'

La ecuación escalar básica del MAM para el dato del *i-ésimo* individuo, hijo de la *j-ésima* madre es la siguiente:

$$y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b} + a_{oi} + a_{mj} + e_{mj} + e_{oi}, \qquad [2.1]$$

donde

$\boldsymbol{x}_i^{\mathrm{T}}$: vector de incidencia de los efectos fijos. El superíndice 'T' indica 'traspuesta'.

$\boldsymbol{b}$: vector de efectos fijos.

a_{oi} : valor de cría directo del individuo *i*.

a_{mj} : valor de cría materno de *j*, la madre del individuo *i*.

e_{mj} : efecto ambiental materno permanente de *j*.

e_{oi} : error del modelo.

Bajo este modelo, los valores de cría directo y materno, el efecto ambiental materno permanente y el error son tratados como variables aleatorias no observadas que siguen una distribución normal. El modelo se completa entonces con las siguientes especificaciones:

$$E\begin{bmatrix} a_{oi} \\ a_{mj} \\ e_{mj} \\ e_{oi} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad Cov\begin{bmatrix} a_{oi} \\ a_{mj} \\ e_{mj} \\ e_{oi} \end{bmatrix} = \begin{bmatrix} \sigma^2_{a_o} & \frac{1}{2}\sigma_{a_o a_m} & 0 & 0 \\ \frac{1}{2}\sigma_{a_o a_m} & \sigma^2_{a_m} & 0 & 0 \\ 0 & 0 & \sigma^2_{e_m} & 0 \\ 0 & 0 & 0 & \sigma^2_{e_o} \end{bmatrix}, \qquad [2.2]$$

donde $E(\cdot)$ es el operador estadístico 'esperanza' y el símbolo $Cov(\cdot)$ representa una matriz de varianzas y covarianzas, o 'matriz de covarianza'.

Para *n* observaciones, la expresión matricial del MAM es (Quaas y Pollak, 1980):

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}_o\boldsymbol{a}_o + \boldsymbol{Z}_m\boldsymbol{a}_m + \boldsymbol{Z}_p\boldsymbol{e}_m + \boldsymbol{e}_o, \qquad [2.3]$$

donde

$\boldsymbol{y}$: vector ($n \times 1$) de registros fenotípicos.

$\boldsymbol{X}$: matriz ($n \times p$) de incidencia de los efectos fijos (que, sin perder generalidad, se asumirá es de rango completo).

$\boldsymbol{b}$: vector ($p \times 1$) de efectos fijos.

$\boldsymbol{Z}_o$, $\boldsymbol{Z}_m$ y $\boldsymbol{Z}_p$: matrices de incidencia de los efectos aleatorios, de órdenes ($n \times q$), ($n \times q$) y ($n \times d$), respectivamente. Aquí, q es el número de individuos en el pedigree, y d es el número de madres de individuos con registro fenotípico.

$\boldsymbol{a}_o$ y $\boldsymbol{a}_m$: vectores aleatorios de orden ($q \times 1$) de valores de cría directos y maternos, respectivamente.

$\boldsymbol{e}_m$: vector aleatorio ($d \times 1$) de efectos ambientales maternos permanentes.

$\boldsymbol{e}_o$: vector aleatorio ($n \times 1$) de errores.

En [2.3], todos los vectores aleatorios están definidos como desvíos de sus valores esperados y, en consecuencia, su esperanza es igual a cero. La estructura de covarianza, por su parte, es la siguiente:

$$Cov\begin{bmatrix} \boldsymbol{a}_o \\ \boldsymbol{a}_m \\ \boldsymbol{e}_m \\ \boldsymbol{e}_o \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}\sigma^2_{a_o} & \boldsymbol{A}\sigma_{a_o a_m} & 0 & 0 \\ \boldsymbol{A}\sigma_{a_o a_m} & \boldsymbol{A}\sigma^2_{a_m} & 0 & 0 \\ 0 & 0 & \boldsymbol{I}_d\sigma^2_{e_m} & 0 \\ 0 & 0 & 0 & \boldsymbol{I}_n\sigma^2_{e_o} \end{bmatrix}, \qquad [2.4]$$

donde $\boldsymbol{A}$ ($q \times q$) es la matriz de relaciones aditivas (Wright, 1922).

Nótese que si escribimos $\boldsymbol{a}^{\mathrm{T}} \equiv \left(\boldsymbol{a}_o^{\mathrm{T}}, \boldsymbol{a}_m^{\mathrm{T}}\right)$ y definimos luego

$$\boldsymbol{\Sigma} \equiv \begin{bmatrix} \sigma^2_{a_o} & \sigma_{a_o a_m} \\ \sigma_{a_o a_m} & \sigma^2_{a_m} \end{bmatrix}, \qquad [2.5]$$

entonces

$$Cov(\boldsymbol{a}) = \boldsymbol{\Sigma} \otimes \boldsymbol{A} \equiv \boldsymbol{G}. \qquad [2.6]$$

En [2.6], el símbolo $\otimes$ representa al operador matricial 'producto Kronecker'. Esta estructura Kronecker de la matriz de covarianza de los valores de cría (o 'matriz de covarianza genética', $\boldsymbol{G}$) ha sido clave en la amplia difusión del MAM entre los mejoradores animales, dado que facilita su inversión mediante las reglas de Henderson (1976).

Ahora ya estamos en condiciones de especificar la matriz de covarianza del vector de observaciones, $\boldsymbol{y}$:

$$\begin{aligned} Cov(\boldsymbol{y}) = \boldsymbol{Z}_o \boldsymbol{A} \boldsymbol{Z}_o^{\mathrm{T}} \sigma^2_{a_o} + \left(\boldsymbol{Z}_o \boldsymbol{A} \boldsymbol{Z}_m^{\mathrm{T}} + \boldsymbol{Z}_m \boldsymbol{A} \boldsymbol{Z}_o^{\mathrm{T}}\right) \sigma_{a_o a_m} + \\ + \boldsymbol{Z}_m \boldsymbol{A} \boldsymbol{Z}_m^{\mathrm{T}} \sigma^2_{a_m} + \boldsymbol{Z}_p \boldsymbol{Z}_p^{\mathrm{T}} \sigma^2_{e_m} + \boldsymbol{I} \sigma^2_{e_o}. \end{aligned} \qquad [2.7]$$

Definiendo, finalmente, $\boldsymbol{Z} \equiv \left(\boldsymbol{Z}_o, \boldsymbol{Z}_m\right)$ las correspondientes MME serán:

$$\begin{bmatrix} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z} & \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z}_p \\ \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z} + \boldsymbol{G}^{-1}\sigma^2_{e_o} & \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}_p \\ \boldsymbol{Z}_p^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{Z}_p^{\mathrm{T}}\boldsymbol{Z} & \boldsymbol{Z}_p^{\mathrm{T}}\boldsymbol{Z}_p + \boldsymbol{I}_d \sigma^2_{e_o} \sigma^{-2}_{e_m} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{a}} \\ \hat{\boldsymbol{e}}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \\ \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{y} \\ \boldsymbol{Z}_p^{\mathrm{T}}\boldsymbol{y} \end{bmatrix}. \qquad [2.8]$$

Resolviendo este sistema de ecuaciones se obtienen en simultáneo las estimaciones de los efectos fijos y las predicciones BLUP de los valores de cría directos y maternos, y de los efectos ambientales maternos permanentes. Nótese, sin embargo, que las MME [2.8] dependen de los parámetros de dispersión asociados a los efectos aleatorios del modelo, los CVC. En la próxima sección se describirá una metodología de estimación bayesiana de los CVC, basada en uno de los métodos de cadenas de Markov y simulación de Monte Carlo (MCMC), el muestreo de Gibbs (*cf*. Casella y George, 1992).

2.2.2. Estimación de CVC vía el algoritmo del muestreo de Gibbs

Considérese entonces la implementación de un análisis bayesiano jerárquico del modelo [2.3] con el objetivo de estimar los CVC (*e.g*. Sorensen y Gianola, 2002, cap. 13.3). En la primera etapa del análisis es necesario especificar la distribución condicional conjunta de las observaciones. Asúmase, en este caso, un proceso normal multivariado:

$$\boldsymbol{y} \mid \boldsymbol{b}, \boldsymbol{a}, \boldsymbol{e}_m, \sigma^2_{e_o} \sim NMV\left(\boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{a} + \boldsymbol{Z}_p \boldsymbol{e}_m, \boldsymbol{I}_n \sigma^2_{e_o}\right). \qquad [2.9]$$

En este contexto, los vectores $\boldsymbol{b}$, $\boldsymbol{a}$, y $\boldsymbol{e}_m$ constituyen los 'parámetros de posición' de la distribución condicional de las observaciones. Desde el punto de vista del análisis bayesiano, es necesario asignarles distribuciones a priori.

2.2.2.1. Distribuciones a priori

En primer lugar, se asumirá un proceso normal multivariado para el vector de efectos fijos. Como discuten Hobert y Casella (1996), la asignación una distribución normal con una gran varianza evita la ocurrencia de distribuciones posteriores impropias y, al mismo tiempo, permite reflejar incertidumbre respecto al verdadero valor de los parámetros. Siguiendo a Cantet *et al*. (2004), entonces,

$$\boldsymbol{b} \mid \boldsymbol{K} \sim NMV\left(\boldsymbol{0}, \boldsymbol{K}\right), \quad \boldsymbol{K} = diag\left\{k_i\right\}, \quad k_i \geq 1\times 10^7, \quad i = 1, \ldots, p. \qquad [2.10]$$

En segundo lugar, y de acuerdo a la teoría genética cuantitativa, se especificará una distribución normal multivariada para el vector de los valores de cría; *i.e.*,

$$\boldsymbol{a} \mid \boldsymbol{A}, \boldsymbol{\Sigma} \sim NMV\left(\boldsymbol{0}, \boldsymbol{\Sigma} \otimes \boldsymbol{A}\right). \qquad [2.11]$$

Finalmente, también se asumirá un proceso normal multivariado para el vector de los efectos ambientales maternos permanentes:

$$\boldsymbol{e}_m \mid \sigma^2_{e_m} \sim NMV\left(\boldsymbol{0}, \boldsymbol{I}_d \sigma^2_{e_m}\right). \qquad [2.12]$$

Luego, en el siguiente nivel de la jerarquía es necesario especificar distribuciones a priori para los CVC, es decir, para los escalares $\sigma^2_{e_o}$ y $\sigma^2_{e_p}$, y para la matriz $\boldsymbol{\Sigma}$. En este punto se asumirán distribuciones conjugadas Gamma invertidas: Chi-cuadradas para los escalares y Wishart invertida para la matriz $\boldsymbol{\Sigma}$. Así,

$$\begin{aligned} \sigma^2_{e_o} &\sim \upsilon_{e_o} S^2_{e_o} \chi^{-2}_{\upsilon_{e_o}}, \\ \sigma^2_{e_m} &\sim \upsilon_{e_m} S^2_{e_m} \chi^{-2}_{\upsilon_{e_m}}, \\ \Sigma &\sim IW\left(\upsilon, \boldsymbol{S}\right), \end{aligned} \qquad [2.13]$$

con $\boldsymbol{S} = \upsilon \boldsymbol{S}^*$.

Los parámetros de los que dependen las distribuciones a priori de los CVC se denominan ‘hiperparámetros’. Los hiperparámetros deben ser especificados por el analista y, en consecuencia, se utilizan para describir el conocimiento a priori o una opinión experta que éste tiene respecto a la distribución de los CVC. En el contexto del análisis bayesiano jerárquico aquí descripto, $S^2_{e_o}$ y $S^2_{e_m}$ representan valores ‘razonables’ para la varianza de los efectos ambientales maternos permanentes y para la varianza del error, respectivamente. Por su parte, $\boldsymbol{S}^*$ es una matriz (2 × 2) de valores ‘razonables’ para los CVC genéticos. El término ‘razonable’ aquí indica que estos hiperparámetros se interpretan como una sentencia sobre el valor esperado de los correspondientes CVC, al menos a priori. Por otro lado, los hiperparámetros υ, υ_{e_m} y υ_{e_o} se interpretan como los grados de credibilidad en dichos valores a priori y reflejan, en consecuencia, la incertidumbre que el analista asigna a los valores especificados. Nótese, en particular, que la incertidumbre respecto a los valores a priori de los tres parámetros de dispersión genéticos está modelada por un único parámetro escalar, υ.

2.2.2.2. Distribución condicional conjunta

Asúmase ahora que todas las incógnitas del modelo, *i.e.*, $\boldsymbol{b}$, $\boldsymbol{a}\,|\,\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}$, $\boldsymbol{e}_m\,|\,\sigma^2_{e_m}$, $\sigma^2_{e_m}$ y $\sigma^2_{e_o}$, son mutuamente independientes a priori. Entonces, y de acuerdo al teorema de Bayes, la distribución condicional conjunta será proporcional al producto de la verosimilitud y de las correspondientes distribuciones a priori (Box y Tiao, 1973). Es decir,

$$\begin{aligned} & p\left(\boldsymbol{b},\boldsymbol{a},\boldsymbol{\Sigma},\boldsymbol{e}_m,\sigma^2_{e_m},\sigma^2_{e_o}\,|\,\boldsymbol{y}\right)\propto \\ & \quad\propto p\left(\boldsymbol{y}\,|\,\boldsymbol{b},\boldsymbol{a},\boldsymbol{e}_m,\sigma^2_{e_o}\right)\times p\left(\boldsymbol{b}\,|\,\boldsymbol{K}\right)\times \\ & \quad\times p\left(\boldsymbol{a}\,|\,\boldsymbol{A},\boldsymbol{\Sigma}\right)\times p\left(\boldsymbol{\Sigma}\,|\,\upsilon,\boldsymbol{S}\right)\times \\ & \quad\times p\left(\boldsymbol{e}_m\,|\,\sigma^2_{e_m}\right)\times p\left(\sigma^2_{e_m}\,|\,\upsilon_{e_m},S^2_{e_m}\right)\times \\ & \quad\times p\left(\sigma^2_{e_o}\,|\,\upsilon_{e_o},S^2_{e_o}\right). \end{aligned} \qquad [2.14]$$

Explícitamente y tras agrupar factores afines (Sorensen y Gianola, 2002)

$$\begin{aligned} & p\left(\boldsymbol{b},\boldsymbol{a},\boldsymbol{\Sigma},\boldsymbol{e}_m,\sigma^2_{e_m},\sigma^2_{e_o}\,|\,\boldsymbol{y}\right)\propto \\ & \quad\propto exp\left\{-\left(\tfrac{1}{2}\right)\boldsymbol{b}^{\mathrm{T}}\boldsymbol{K}^{-1}\boldsymbol{b}\right\}\times \\ & \quad\times\left(\sigma^2_{e_o}\right)^{-\frac{1}{2}\left(\upsilon_{e_o}+n+2\right)} exp\left\{-\frac{\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}-\upsilon_{e_o}S^2_{e_o}}{2\sigma^2_{e_o}}\right\}\times \\ & \quad\times\left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}\left(q+\upsilon+3\right)} exp\left\{-\left(\tfrac{1}{2}\right)tr\left[\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{Q}+\boldsymbol{S}\right)\right]\right\}\times \\ & \quad\times\left(\sigma^2_{e_m}\right)^{-\frac{1}{2}\left(\upsilon_{e_m}+d+2\right)} exp\left\{-\frac{\boldsymbol{e}_m^{\mathrm{T}}\boldsymbol{e}_m-\upsilon_{e_m}S^2_{e_m}}{2\sigma^2_{e_m}}\right\}, \end{aligned} \qquad [2.15]$$

donde $\boldsymbol{e}=\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}-\boldsymbol{Z}\boldsymbol{a}-\boldsymbol{Z}_p\boldsymbol{e}_m$ y $\boldsymbol{Q}=\begin{bmatrix}\boldsymbol{a}_o^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{a}_o & \boldsymbol{a}_o^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{a}_m\\ \boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{a}_o & \boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{a}_m\end{bmatrix}$.

A partir de la expresión analítica [2.15] es posible identificar la distribución condicional posterior de cualquier parámetro de interés, manteniendo el resto de ellos constante. En el próximo apartado se derivarán las distribuciones condicionales de todas las incógnitas del MAM. Las derivaciones pueden seguirse con mayor grado de detalle en Sorensen y Gianola (2002, cap. 13.3) y Jensen *et al*. (1994).

2.2.2.3. Distribuciones condicionales posteriores

Defínase, en primer lugar, el vector de parámetros de posición según $\boldsymbol{\theta}^{\mathrm{T}}\equiv\left(\boldsymbol{b}^{\mathrm{T}},\boldsymbol{a}^{\mathrm{T}},\boldsymbol{e}_m^{\mathrm{T}}\right)$. Luego, la distribución condicional posterior de este vector es proporcional a

$$\begin{aligned} & p\left(\boldsymbol{\theta}\,|\,\boldsymbol{\Sigma},\sigma^2_{e_m},\sigma^2_{e_o},\boldsymbol{y}\right)\propto \\ & \quad\propto p\left(\boldsymbol{y}\,|\,\boldsymbol{b},\boldsymbol{a},\boldsymbol{e}_m,\sigma^2_{e_o}\right)\times p\left(\boldsymbol{b}\,|\,\boldsymbol{K}\right)\times \\ & \quad\times p\left(\boldsymbol{e}_m\,|\,\sigma^2_{e_m}\right)\times p\left(\boldsymbol{a}\,|\,\boldsymbol{A},\boldsymbol{\Sigma}\right). \end{aligned} \qquad [2.16]$$

Explícitamente,

$$\begin{aligned} & p\left(\boldsymbol{\theta} \mid \boldsymbol{\Sigma}, \sigma_{e_m}^2, \sigma_{e_o}^2, \boldsymbol{y}\right) \propto \\ & \propto exp\left\{-\frac{\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}}{2\sigma_{e_o}^2}\right\} \times exp\left\{-\left(\tfrac{1}{2}\right)\boldsymbol{b}^{\mathrm{T}}\boldsymbol{K}^{-1}\boldsymbol{b}\right\} \times \\ & \times exp\left\{-\frac{\boldsymbol{e}_m^{\mathrm{T}}\boldsymbol{e}_m}{2\sigma_{e_m}^2}\right\} \times exp\left\{-\frac{\boldsymbol{a}^{\mathrm{T}}\left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{A}^{-1}\right)\boldsymbol{a}}{2\sigma_{e_o}^2}\right\}. \end{aligned} \qquad [2.17]$$

Operando algebraicamente (*e.g*. Jensen *et al*., 1994), puede demostrarse que

$$\boldsymbol{\theta} \mid \boldsymbol{\Sigma}, \sigma_{e_m}^2, \sigma_{e_o}^2, \boldsymbol{y} \sim NMV\left(\hat{\boldsymbol{\theta}}, \boldsymbol{C}^{-1}\sigma_{e_o}^2\right). \qquad [2.18]$$

Aquí, $\hat{\boldsymbol{\theta}} = \boldsymbol{C}^{-1}\boldsymbol{r}$ es la solución a las MME en [2.8], con $\boldsymbol{C}^{-1}$ igual a la inversa de la correspondiente matriz de coeficientes, y $\boldsymbol{r}$ es el vector de 'términos a la derecha de las ecuaciones' (*right hand side*). Estrictamente, existe una pequeña diferencia con [2.8]: al computar la matriz de los coeficientes es necesario sumar k_i^{-1} al elemento diagonal correspondiente a cada efecto fijo, donde k_i es la cantidad a través de la cual se refleja gran incertidumbre a priori respecto al valor de los efectos fijos.

Por las propiedades de la distribución Normal multivariada se deduce que la distribución condicional posterior del *i*-ésimo elemento escalar del vector de parámetros de posición también seguirá una distribución Normal. Específicamente, defínase $\boldsymbol{\theta}_{-i}$ como el vector de parámetros de posición sin el *i*-ésimo elemento escalar. Entonces (*cf*. Sorensen y Gianola, Cap. 13.2.1),

$$\theta_i \mid \boldsymbol{\theta}_{-i}, \boldsymbol{\Sigma}, \sigma_{e_m}^2, \sigma_{e_o}^2, \boldsymbol{y} \sim N\left(\tilde{\theta}_i, c_{ii}^{-1}\sigma_{e_o}^2\right), \qquad [2.19]$$

tal que $\tilde{\theta}_i$ satisface

$$c_{ii}\tilde{\theta}_i = r_i - \boldsymbol{c}_{-i}^{\mathrm{T}}\boldsymbol{\theta}_{-i}, \qquad [2.20]$$

donde c_{ii} es el *i*-ésimo elemento diagonal de la matriz de coeficientes y $\boldsymbol{c}_{-i}$ corresponde a la *i*-ésima columna de la matriz sin dicho elemento.

Por su parte, la distribución condicional posterior de la varianza del error es proporcional a

$$\begin{aligned} & p\left(\sigma_{e_o}^2 \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}, \sigma_{e_m}^2, \boldsymbol{y}\right) \propto \\ & \propto p\left(\boldsymbol{y} \mid \boldsymbol{b}, \boldsymbol{a}, \boldsymbol{e}_m, \sigma_{e_o}^2\right) \times p\left(\sigma_{e_o}^2 \mid \upsilon_{e_o}, S_{e_o}^2\right). \end{aligned} \qquad [2.21]$$

Explícitamente,

$$\begin{aligned} & p\left(\sigma_{e_o}^2 \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}, \sigma_{e_m}^2, \boldsymbol{y}\right) \propto \\ & \propto \left(\sigma_{e_o}^2\right)^{-\frac{1}{2}\left(\upsilon_{e_o}+n+2\right)} exp\left\{-\frac{\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e} + \upsilon_{e_o}S_{e_o}^2}{2\sigma_{e_o}^2}\right\}. \end{aligned} \qquad [2.22]$$

Defínase luego

$$\tilde{S}^2_{e_o} \equiv \frac{\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e} + \upsilon_{e_o} S^2_{e_o}}{\tilde{\upsilon}_{e_o}}, \text{ con } \tilde{\upsilon}_{e_o} \equiv \upsilon_{e_o} + n. \qquad [2.23]$$

Entonces,

$$\begin{aligned} & p\left(\sigma^2_{e_o} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}, \sigma^2_{e_m}, \boldsymbol{y}\right) \propto \\ & \propto \left(\sigma^2_{e_o}\right)^{-\frac{1}{2}\left(\tilde{\upsilon}_{e_o}+2\right)} exp\left\{-\frac{\tilde{\upsilon}_{e_o}\tilde{S}^2_{e_o}}{2\sigma^2_{e_o}}\right\}. \end{aligned} \qquad [2.24]$$

Por inspección, la expresión [2.24] corresponde al 'núcleo' (*kernel*, en inglés) de una distribución Chi-cuadrada invertida con parámetros $\tilde{\upsilon}_{e_o}$ y $\tilde{\upsilon}_{e_o}\tilde{S}^2_{e_o}$. En consecuencia,

$$\sigma^2_{e_o} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}, \sigma^2_{e_m}, \boldsymbol{y} \sim \tilde{\upsilon}_{e_o}\tilde{S}^2_{e_o}\chi^{-2}_{\tilde{\upsilon}_{e_o}}. \qquad [2.25]$$

Argumentos similares permiten derivar la distribución condicional posterior de la varianza de los efectos ambientales maternos permanentes. Esta distribución será proporcional a:

$$\begin{aligned} & p\left(\sigma^2_{e_m} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}, \sigma^2_{e_o}, \boldsymbol{y}\right) \propto \\ & \propto p\left(\boldsymbol{e}_m \mid \sigma^2_{e_m}\right) \times p\left(\sigma^2_{e_m} \mid \upsilon_{e_m}, S^2_{e_m}\right). \end{aligned} \qquad [2.26]$$

Explícitamente,

$$\begin{aligned} & p\left(\sigma^2_{e_m} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}, \sigma^2_{e_o}, \boldsymbol{y}\right) \propto \\ & \propto \left(\sigma^2_{e_m}\right)^{-\frac{1}{2}\left(\upsilon_{e_p}+d+2\right)} exp\left\{-\frac{\boldsymbol{e}^{\mathrm{T}}_m\boldsymbol{e}_m + \upsilon_{e_m}S^2_{e_m}}{2\sigma^2_{e_m}}\right\}. \end{aligned} \qquad [2.27]$$

Definiendo

$$\tilde{S}^2_{e_m} \equiv \frac{\boldsymbol{e}^{\mathrm{T}}_m\boldsymbol{e}_m + \upsilon_{e_m}S^2_{e_m}}{\tilde{\upsilon}_{e_m}}, \text{ con } \tilde{\upsilon}_{e_m} \equiv \upsilon_{e_m} + d, \qquad [2.28]$$

entonces

$$\begin{aligned} & p\left(\sigma^2_{e_m} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}, \sigma^2_{e_o}, \boldsymbol{y}\right) \propto \\ & \propto \left(\sigma^2_{e_m}\right)^{-\frac{1}{2}\left(\tilde{\upsilon}_{e_m}+2\right)} exp\left\{-\frac{\tilde{\upsilon}_{e_m}\tilde{S}^2_{e_m}}{2\sigma^2_{e_m}}\right\}. \end{aligned} \qquad [2.29]$$

Por inspección, la expresión [2.29] corresponde al núcleo de una distribución Chi-cuadrada invertida con parámetros $\tilde{\upsilon}_{e_m}$ y $\tilde{\upsilon}_{e_m}\tilde{S}^2_{e_m}$. En consecuencia,

$$\sigma^2_{e_m} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}, \sigma^2_{e_o}, \boldsymbol{y} \sim \tilde{\upsilon}_{e_m}\tilde{S}^2_{e_m}\chi^{-2}_{\tilde{\upsilon}_{e_m}}. \qquad [2.30]$$

Resta, por último, obtener la distribución condicional posterior de la matriz de covarianza genética. Esta distribución será proporcional a:

$$\begin{aligned} p\left(\boldsymbol{\Sigma} \mid \boldsymbol{\theta}, \sigma^2_{e_m}, \sigma^2_{e_o}, \boldsymbol{y}\right) &\propto \\ &\propto p\left(\boldsymbol{a} \mid \boldsymbol{A}, \boldsymbol{\Sigma}\right) \times p\left(\boldsymbol{\Sigma} \mid \upsilon, \boldsymbol{S}\right). \end{aligned} \quad [2.31]$$

De acuerdo a Jensen *et al*. (1994), la expresión [2.31] puede escribirse explícitamente como

$$\begin{aligned} p\left(\boldsymbol{\Sigma} \mid \boldsymbol{\theta}, \sigma^2_{e_m}, \sigma^2_{e_o}, \boldsymbol{y}\right) &\propto \\ &\propto \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}(q+\upsilon+3)} exp\left\{-\left(\tfrac{1}{2}\right) tr\left[\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{Q}+\boldsymbol{S}\right)\right]\right\}. \end{aligned} \quad [2.32]$$

Esta expresión corresponde al núcleo de una distribución Wishart invertida; *i.e*.,

$$\boldsymbol{\Sigma} \mid \boldsymbol{\theta}, \sigma^2_{e_m}, \sigma^2_{e_o}, \boldsymbol{y} \sim IW\left(\upsilon + q, \boldsymbol{Q}+\boldsymbol{S}\right). \quad [2.33]$$

En resumen, nótese que bajo el análisis bayesiano jerárquico aquí descripto, las distribuciones condicionales posteriores de todas las incógnitas del MAM [2.3] pertenecen a familias de distribuciones de probabilidad conocidas. Específicamente, pertenecen exactamente a la misma familia que las correspondientes distribuciones a priori; es decir, son 'condicionalmente conjugadas' (*cf*. Daniels y Pourahmadi, 2002). Para poder realizar inferencias respecto a los CVC es necesario proceder al siguiente nivel de marginalización. Sin embargo, conseguirlo por medios analíticos es imposible y, en consecuencia, es necesario recurrir a métodos de aproximación o integración numérica. Una discusión de algunos de estos métodos en el contexto de la estimación de CVC bajo el MAM puede consultarse en Cantet *et al*. (1992a). En lo que resta de esta sección, en cambio, se presentará un método de integración numérica por simulación Monte Carlo que aplica cuando las distribuciones condicionales posteriores pertenecen a familias conocidas, el muestreo de Gibbs.

2.2.2.4. Muestreo de Gibbs

El GS es una técnica indirecta de muestreo de la distribución marginal de una variable aleatoria a partir del muestreo secuencial de una serie de distribuciones condicionales en dicha variable. A modo de breve explicación, se seguirá de cerca el trabajo de Casella y George (1992). Sea $f\left(x, y_1, \ldots, y_m\right)$ una función de densidad de probabilidad conjunta. Considérese luego la posibilidad de obtener alguna característica de la distribución marginal $f\left(x\right)$, tal como la esperanza o la varianza. El modo más directo de proceder sería calcular la distribución marginalizando la densidad conjunta, para obtener luego la característica deseada operando analíticamente. Muchas veces, sin embargo, las integrales a resolver son extremadamente difíciles de llevar a cabo. En esos casos, el GS permite generar una muestra de la distribución sin la necesidad de conocer su expresión analítica. Luego, con una muestra lo suficientemente grande puede estimarse cualquier característica de la distribución marginal con el nivel de precisión deseado. El único requisito es que sea posible muestrear de las distribuciones condicionales. El GS fue formalmente presentado por Geman y Geman (1984) en el contexto de procesamiento de imágenes y su difusión en la corriente principal de la estadística aplicada se debe principalmente al trabajo de Gelfand y Smith (1990).

En el marco del análisis bayesiano jerárquico para el MAM, la implementación del GS conlleva muestrear secuencialmente de las distribuciones [2.18], [2.25], [2.30] y [2.33]. Una vez que el algoritmo convergió, el muestreo secuencial de las distribuciones condicionales resultará en un muestreo de las distribuciones marginales posteriores de cada parámetro. Sea *m* el número de muestreos definido por el analista para garantizar la representatividad de la muestra de la distribución marginal de interés. Entonces, el algoritmo involucra los siguientes pasos:

1. Construir las MME [2.8], sumar k_i^{-1} al elemento diagonal de a cada efecto fijo y, finalmente, resolver el sistema de ecuaciones para inicializar $\hat{\boldsymbol{\theta}}$.
2. Muestrear $\boldsymbol{\theta}$ de $\boldsymbol{\theta} | \boldsymbol{\Sigma}, \sigma^2_{e_m}, \sigma^2_{e_o}, \boldsymbol{y} \sim NMV\left(\hat{\boldsymbol{\theta}}, \boldsymbol{C}^{-1}\sigma^2_{e_o}\right)$ o, alternativamente, muestrear en forma secuencial θ_i de $\theta_i | \boldsymbol{\theta}_{-i}, \boldsymbol{\Sigma}, \sigma^2_{e_m}, \sigma^2_{e_o}, \boldsymbol{y} \sim N\left(\tilde{\theta}_i, c_{ii}^{-1}\sigma^2_{e_o}\right)$ para todo *i*.
3. Calcular los residuales $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{Xb} - \boldsymbol{Za} - \boldsymbol{Z}_p\boldsymbol{e}_m$.
4. Computar $\tilde{S}^2_{e_o}$ y $\tilde{\upsilon}_{e_o}$ de acuerdo a [2.23].
5. Muestrear $\sigma^2_{e_o}$ de $\sigma^2_{e_o} | \boldsymbol{\theta}, \boldsymbol{\Sigma}, \sigma^2_{e_m}, \boldsymbol{y} \sim \tilde{\upsilon}_{e_o}\tilde{S}^2_{e_o}\chi^{-2}_{\tilde{\upsilon}_{e_o}}$.
6. Computar $\tilde{S}^2_{e_m}$ y $\tilde{\upsilon}_{e_m}$ de acuerdo a [2.28].
7. Muestrear $\sigma^2_{e_m}$ de $\sigma^2_{e_m} | \boldsymbol{\theta}, \boldsymbol{\Sigma}, \sigma^2_{e_o}, \boldsymbol{y} \sim \tilde{\upsilon}_{e_m}\tilde{S}^2_{e_m}\chi^{-2}_{\tilde{\upsilon}_{e_m}}$.
8. Calcular la matriz de las formas cuadráticas, $\boldsymbol{Q}$.
9. Muestrear $\boldsymbol{\Sigma}$ de $\boldsymbol{\Sigma} | \boldsymbol{\theta}, \sigma^2_{e_m}, \sigma^2_{e_o}, \boldsymbol{y} \sim IW\left(\upsilon + q, \boldsymbol{Q} + \boldsymbol{S}\right)$.
10. Repetir *m* veces los pasos 2–9.

En cada ciclo del procedimiento, los valores muestreados de los parámetros de interés (por ejemplo, los CVC) deben almacenarse con el objeto de computar luego estadísticos descriptivos posteriores o alguna otra característica de las distribuciones marginales. Con los valores almacenados es posible, incluso, calcular estadísticos descriptivos posteriores de ciertas funciones de estos parámetros (por ejemplo, heredabilidades) simplemente aplicando la función en cada ciclo de muestreo. En general, los primeros muestreos dependerán de los valores iniciales asignados arbitrariamente. Desde el punto de vista práctico, en consecuencia, estos ciclos iniciales de iteración son descartados para evitar esta dependencia y garantizar así la convergencia del algoritmo (García-Cortés *et al*., 1998). El número de ciclos que se descartan recibe el nombre de 'período de calentamiento' (*burn-in*).

2.2.3. Implementación del análisis con datos de campo

En esta sección se ilustra la implementación del análisis bayesiano jerárquico con el objeto de estimar CVC vía el muestreo de Gibbs para un archivo de datos de peso al destete en bovinos de carne. Los datos pertenecen a la empresa "Estancias y Cabaña Las Lilas" (Las Lilas), y estuvieron disponibles en virtud del convenio marco vigente entre la empresa y la Facultad de Agronomía de la Universidad de Buenos Aires (FAUBA).

2.2.3.1. Descripción del archivo de datos

El archivo de datos analizado corresponde al rodeo Angus de Las Lilas, e incluye 7229 registros de pesos al destete tomados sobre individuos nacidos entre 1972 y 2008, con 194 días de edad en promedio. El archivo se complementa, además, con una genealogía de 9936 animales. En la Tabla 2.1 se presenta una descripción detallada de los datos, con especial énfasis en características asociadas a la calidad de estimación de CVC bajo modelos con efectos maternos: el número promedio de crías por madre y el número de madres con registro fenotípico (*cf*. Gerstmayr, 1992; Maniatis y Pollott, 2003). En particular, es importante mencionar que todas las madres están identificadas y que no se han incluido datos de crías por transferencia embrionaria.

Tabla 2.1. Descripción de la base de datos del rodeo Angus de Las Lilas.

		ANGUS	
BASE de pedigree	**Individuos**	**Padres**	**Madres**
	9936	747	3404
BASE de datos	**Nº**	**Promedio, kg**	**DS, kg**
Registros de PD	7229	205,31	40,28
	Padres	**Madres**	**TOTAL**
Progenitores	264	2444	2708
(c/ registro de PD)	54	1386	1440
%	20,5	56,7	53,2
Nº prom. crías por progenitor	27,33	2,96	
% de progenitores c/:			
1 cría	13,64	30,20	
2 crías	7,95	21,77	
3 crías	4,92	15,47	
>3 crías	73,49	32,56	

PD = peso al destete; DS = desvío estándar.

2.2.3.2. Descripción de los análisis

Con el objetivo de estimar los CVC relevantes a esta población se implementó un análisis bayesiano jerárquico vía el algoritmo GS. El modelo de análisis fue el MAM clásico (ecuación [2.3]). El modelo incluyó los efectos fijos de sexo, edad de la madre, grupo de contemporáneos y la covariable edad al destete. Los CVC fueron estimados mediante un algoritmo GS como el descripto en la Sección 2.2.2.4.

Específicamente, se escribió un programa en Fortran 90 inspirado en las notas de clase de Misztal (2006) y el trabajo de Groeneveld y Kovac (1990). El programa presenta una estructura modular con dos subrutinas internas principales. La primera de ellas computa los elementos de la inversa de la matriz de relaciones aditivas según el algoritmo para el cálculo de los coeficientes de consanguinidad de Meuwissen y Luo (1992), y genera luego las correspondientes contribuciones de los efectos aleatorios a las MME. La segunda subrutina, por su parte, realiza un ciclo completo de muestreo de las distribuciones condicionales posteriores de todos los parámetros del modelo y actualiza, en consecuencia, sus valores. Los códigos están basados en programas del paquete BLUPF90 (Misztal *et al*., 2002) y programas F77 del grupo de investigación (Cantet y

Birchmeier, comunicación personal). Dada su extensión, no serán incluidos en el presente documento. Sin embargo, pueden ser solicitados al autor.

Una característica muy importante del programa merece un párrafo aparte. La demanda computacional del algoritmo de GS hasta aquí descripto está fuertemente determinada por el tiempo de cómputo asociado al muestreo del vector de parámetros de posición. En particular, de la expresión [2.18] se deduce que dicho muestreo involucra construir las MME en cada ciclo del GS, dado que los parámetros de la distribución condicional posterior dependen de elementos de la matriz de los coeficientes que, a su vez, se actualizan en la medida que se actualizan los CVC. Sin embargo, nótese que las MME [2.8] pueden rescribirse sucintamente según

$$\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}+\boldsymbol{E}^{-1}\right)\boldsymbol{\theta}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{y}, \qquad [2.34]$$

con

$$\boldsymbol{W}=\left(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{Z}_{p}\right) \qquad [2.35]$$

y

$$\boldsymbol{E}^{-1}=\begin{bmatrix}\boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \left(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{A}^{-1}\right)\sigma^{2}_{e_o} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_{d}\sigma^{2}_{e_o}\sigma^{-2}_{e_m}\end{bmatrix}. \qquad [2.36]$$

La descomposición de la matriz de los coeficientes en [2.34] sugiere una forma de reducir el tiempo de cómputo en este paso del muestreo de Gibbs: dado que no es necesario actualizar la matriz $\boldsymbol{W}$ en cada ciclo de muestreo, $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}$ puede construirse durante la primera iteración, y luego almacenarse en memoria para referencia en los ciclos subsiguientes. La misma estrategia puede seguirse luego con la inversa de la matriz de relaciones aditivas. De este modo, fue posible acelerar considerablemente la performance del programa por iteración.

Antes de proceder con la implementación del GS es necesario definir los hiperparámetros de las distribuciones a priori de los CVC. En general, se definieron los hiperparámetros de escala utilizando estimaciones REML de los CVC bajo el MAM como una sentencia sobre la media de dichas distribuciones, de acuerdo a la parameterización descripta en la Sección 2.2.1. En particular, las estimaciones REML y sus correspondientes errores estándares se obtuvieron mediante el paquete ASReml (Gilmour *et al*., 2006). Por su parte, los grados de credibilidad se definieron en valores más bien bajos, de modo de reflejar incertidumbre en torno a las medias a priori de los CVC. Los valores de los hiperparámetros especificados para los diferentes análisis del archivo de datos del rodeo Angus pueden consultarse en las tablas 2.2 y 2.3.

Entre otros aspectos importantes, la implementación del GS involucra decidir el número de cadenas a generar, por un lado, y determinar la longitud del período de calentamiento y el número de ciclos necesarios para asegurar una muestra representativa de la distribución marginal de interés, por otro (Gilks *et al*., 1996, cap. 1). En general, diferentes estrategias respecto a la implementación del algoritmo GS pueden afectar las correlaciones entre muestras de un mismo parámetro (‘autocorrelaciones’) y, por tanto, las tasas de convergencia (Sorensen y Gianola, 2002). Cuando las autocorrelaciones son

muy altas, las cadenas MCMC recorren muy lentamente el soporte de las distribuciones marginales de interés y, en consecuencia, es necesario generar un número importante de ciclos de muestreo para asegurar una muestra representativa. Desafortunadamente, no existe una regla infalible para determinar exactamente cuántos ciclos son necesarios. En la práctica existen básicamente dos alternativas para tomar una decisión al respecto a la hora de implementar el GS: o bien ejecutar una cadena muy larga (*e.g*. Geyer, 1992) o bien ejecutar varias cadenas más cortas (*e.g*. Gelman y Rubin, 1992). A modo de ilustración, en este trabajo se aplicaron ambas estrategias.

En primer lugar, y de acuerdo a las recomendaciones de Geyer (1992), se obtuvo una única cadena de medio millón de ciclos, luego de descartar 10.000 ciclos iniciales (2%) como período de calentamiento. La convergencia de la cadena se estudió mediante los diagnósticos de convergencia de cadena simple que provee el paquete BOA (Smith, 2007), ejecutado bajo entorno R (http://www.r-project.org/). Estadísticos descriptivos posteriores, 'tamaños efectivos de muestra' (ESS, según sus siglas en inglés) y autocorrelaciones de todos los CVC fueron, por último, obtenidos mediante el programa POSTGIBBSF90 del paquete BLUPF90 (Misztal *et al*., 2002).

En segundo lugar, y siguiendo ahora la estrategia de Gelman y Rubin (*cf*. Gelman, 1996), se obtuvieron tres cadenas de 100.000 ciclos. Cada una de las cadenas fue inicializada con diferentes combinaciones de valores para los CVC, que representaban puntos dispersos del soporte de las correspondientes distribuciones marginales. Más precisamente, los valores iniciales especificados correspondieron a las estimaciones REML ± 2×EE. El período de calentamiento se determinó por inspección visual de las gráficas de los muestreos en función del número de ciclos (*trace plots*). Luego, se determinó la convergencia mediante el test de Gelman y Rubin (1992), ejecutado a través del paquete BOA (Smith, 2007), bajo entorno R (http://www.r-project.org/). Finalmente, se computaron estadísticos descriptivos posteriores, ESS y correlaciones entre muestras de todos los CVC para la colección de muestras resultantes de las tres cadenas (*i.e*., luego de descartar los ciclos de calentamiento) mediante el programa POSTGIBBSF90 del paquete BLUPF90 (Misztal *et al*., 2002).

2.3. RESULTADOS

Los aspectos más relevantes de la implementación del análisis bayesiano jerárquico al conjunto de datos del rodeo Angus de Las Lilas se describen a continuación. Las MME incluyeron 22.418 ecuaciones. El tiempo de cómputo del análisis fue de alrededor de 10 ciclos por segundo en una computadora personal con procesador Pentium® 4 (CPU 3.6GHz, 3.11 GB de RAM).

En la Tabla 2.2 se presentan los estadísticos descriptivos posteriores, los ESS y las autocorrelaciones de todos los CVC bajo el MAM, obtenidos a partir de la cadena larga de 500.000 ciclos. Los tamaños efectivos de muestra (ESS) indican que el número de ciclos fue suficiente para obtener estadísticos posteriores precisos de las distribuciones marginales, incluso para aquellos CVC que presentaron autocorrelaciones muy altas para lapsos de hasta 200 muestreos. Por otro lado, los valores de las medias y los modos marginales posteriores fueron muy similares a las medias a priori (*i.e*., las estimaciones REML), aún cuando se especificó a priori una alta incertidumbre para estos últimos valores ($\upsilon = 5$ para todos los CVC). Cabe destacar, por último, que las secuencias de muestreos de todos los CVC pasaron todos los tests de convergencia de cadena simple que ofrece el paquete BOA (Smith, 2007), si bien para algunos de los CVC genéticos,

en particular para la covarianza genética directa-materna, el valor del estadístico *Z* de Geweke (1992) estuvo en el límite de los valores sugeridos.

Tabla 2.2. Parámetros a priori y estadísticos descriptivos posteriores de los CVC obtenidos para una cadena MCMC de 500.000 ciclos.

	CVC[1]				
	$\sigma^2_{e_o}$	$\sigma^2_{e_m}$	$\sigma^2_{a_o}$	$\sigma_{a_o a_m}$	$\sigma^2_{a_m}$
υ	5	5	5	5	5
S	450	93	190	-103	116
ESS	694	1251	483	518	551
Media	453,35	95,47	185,42	-101,12	110,95
Modo	454,06	92,98	177,34	-94,51	106,11
DS	20,60	13,64	33,88	23,11	21,72
IADP95	(413, 493)	(69, 122)	(123, 253)	(-147, -58)	(69, 153)
Autocorr. (Lapso)					
1	0,866	0,959	0,994	0,994	0,995
5	0,767	0,837	0,980	0,983	0,982
10	0,734	0,727	0,966	0,971	0,969
50	0,639	0,431	0,872	0,892	0,882
100	0,567	0,354	0,778	0,810	0,795
200	0,465	0,275	0,637	0,676	0,652

Refs.: υ = grados de credibilidad a priori; *S* = parámetro de escala a priori; ESS = tamaño efectivo de muestra; DS = desvío estándar; IADP95 = intervalo de alta densidad posterior del 95%.

[1] Componentes de (co)varianza: $\sigma^2_{e_o}$ = varianza del error; $\sigma^2_{e_m}$ = varianza de los efectos ambientales maternos permanentes; $\sigma^2_{a_o}$ = varianza aditiva directa; $\sigma^2_{a_m}$ = varianza aditiva materna; $\sigma_{a_o a_m}$ = covarianza genética directa-materna.

En lo que respecta ahora a la implementación del GS de acuerdo a la estrategia de Gelman y Rubin (1992), en la Figura 2.1 se presentan las gráficas de muestreos en función del número de iteraciones para la varianza del error y la varianza aditiva directa, y se señalan sobre ellas los puntos en los que las cadenas aparecen notoriamente superpuestas. A partir de la inspección de esta última gráfica se determinó un período de calentamiento de 7.500 ciclos. Sobre las gráficas se presentan, además, los 'factores potenciales de reducción de escala' (PSRF, *potential scale reduction factor*) (*cf*. Gelman, 1996), estadísticos que se utilizan para monitorear la convergencia de múltiples cadenas MCMC (*cf*. Gelman y Rubin, 1992). Los PSRF se calcularon mediante el paquete BOA (Smith, 2007) a partir del 90% de los ciclos resultantes de descartar los muestreos iniciales definidos como período de calentamiento. Nótese que los PSRF resultaron cercanos a la unidad, lo cual indica que el algoritmo está muestreando de las distribuciones marginales deseadas. Así, las 90.000 muestras finales de cada una las tres cadenas MCMC independientes se combinaron en una única cadena de 270.000 iteraciones, a partir de la cual se obtuvieron los estadísticos descriptivos posteriores de los CVC (Tabla 2.3).

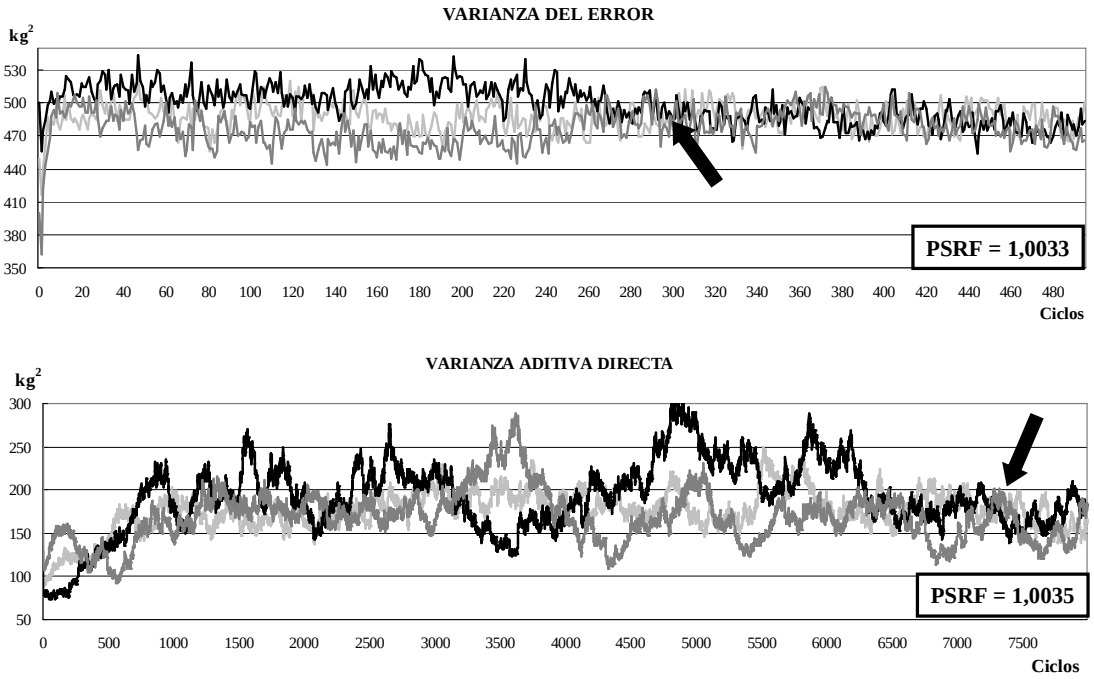


Figura 2.1. Gráficas de muestreos en función del número de ciclos para la varianza del error y la varianza aditiva directa. Las gráficas corresponden a tres cadenas MCMC inicializadas en puntos dispersos del soporte de las distribuciones marginales. Por inspección visual se puede determinar la longitud del período de calentamiento como el punto a partir del cual las cadenas se superponen completamente (flechas negras). Con el resto de los ciclos se computa el estadístico PSRF (Gelman y Rubin, 1992) que se utiliza para monitorear la convergencia de las cadenas.

En la Tabla 2.3, entonces, se presentan los estadísticos descriptivos posteriores de las distribuciones marginales todos los CVC bajo la implementación del GS sugerida por Gelman y Rubin (1992). En este caso los tamaños efectivos de muestra (ESS) fueron aproximadamente la mitad que los obtenidos para el análisis de una única cadena. Nótese que los resultados fueron obtenidos con la mitad de los ciclos, pero las autocorrelaciones aún mantuvieron los altos valores observados previamente. Aún así, los valores de los estadísticos descriptivos posteriores fueron muy similares.

Tabla 2.3. Parámetros a priori y estadísticos descriptivos posteriores de los CVC obtenidos a partir de ciclos de tres cadenas MCMC independientes.

		CVC[1]				
		$\sigma^2_{e_o}$	$\sigma^2_{e_m}$	$\sigma^2_{a_o}$	$\sigma_{a_o a_m}$	$\sigma^2_{a_m}$
υ		(5; 5; 5)	(5; 5; 5)	(5; 5; 5)	(5; 5; 5)	(5; 5; 5)
	Cad. 1	400	70	130	-50	70
S	Cad. 2	450	93	190	-103	116
	Cad. 3	500	120	250	-150	150
ESS		376	714	284	237	282
Media		454,71	96,55	182,85	-98,70	107,96
Modo		459,75	95,38	169,25	-88,62	106,29
DS		21,20	13,68	35,16	24,04	22,72
IADP95		(411, 495)	(70, 123)	(117, 254)	(-147, -54)	(66, 153)
Autocorr. (Lapso)						
	1	0,872	0,959	0,994	0,995	0,995
	5	0,779	0,834	0,982	0,985	0,985
	10	0,748	0,723	0,969	0,974	0,973
	50	0,659	0,432	0,883	0,904	0,897
	100	0,588	0,352	0,794	0,828	0,817
	200	0,491	0,277	0,655	0,697	0,676

Refs.: υ = grados de credibilidad a priori; S = parámetros de escala a priori; ESS = tamaño efectivo de muestra; DS = desvío estándar; IADP95 = intervalo de alta densidad posterior del 95%.

[1] Componentes de (co)varianza: $\sigma^2_{e_o}$ = varianza del error; $\sigma^2_{e_m}$ = varianza de los efectos ambientales maternos permanentes; $\sigma^2_{a_o}$ = varianza aditiva directa; $\sigma^2_{a_m}$ = varianza aditiva materna; $\sigma_{a_o a_m}$ = covarianza genética directa-materna.

Finalmente, en la Tabla 2.4 se presentan medias y desvíos estándares posteriores de los parámetros genéticos (*i.e.*, heradabilidad directa, h^2_o, heradabilidad materna, h^2_m, y correlación genética directa-materna, r_G) obtenidos bajo las dos implementaciones del GS. Estos estadísticos descriptivos posteriores se interpretan como las estimaciones puntuales y los errores estándares de estimación, respectivamente. Los resultados fueron exactamente iguales en ambos casos: las medias posteriores de las heredabilidades directa y materna fueron 0,25 y 0,15, respectivamente, mientras que la media posterior de la correlación directa-materna fue –0,70. En la Figura 2.2, por su parte, se presentan las densidades marginales posteriores de los tres parámetros genéticos, estimadas a través de un método no paramétrico basado en un núcleo Gaussiano (Silverman, 1986). La figura ilustra la riqueza del análisis bayesiano: el analista no dispone sólo de una estimación puntual del parámetro de interés, sino que cuenta con toda una distribución de probabilidad para realizar inferencias. Nótese, en particular, que la distribución marginal posterior de la heredabilidad materna muestra una mayor dispersión que la distribución posterior de la heredabilidad directa.

Tabla 2.4. Estimaciones y errores estándares para los parámetros genéticos bajo las dos implementaciones del muestreo de Gibbs.

	Parámetros genéticos[1]		
	h_o^2	h_m^2	r_G
Análisis	**Estim (EE)**	**Estim (EE)**	**Estim (EE)**
REML	0,26 (0,04)	0,16 (0,03)	-0,69 (0,07)
GS (Geyer)	0,25 (0,04)	0,15 (0,03)	-0,70 (0,08)
GS (Gelman y Rubin)	0,25 (0,04)	0,15 (0,03)	-0,70 (0,08)

Refs.: Estim = estimación REML o media posterior bayesiana; EE = error estándar aproximado (REML) o desvío estándar posterior bayesiano.

[1] Parámetros genéticos: h_o^2 = heredabilidad directa; h_m^2 = heredabilidad materna; r_G = correlación genética directa-materna.

2.4. DISCUSIÓN

En este capítulo se introdujo el MAM 'clásico' (Willham, 1963, Quaas y Pollak, 1980) y se describió en detalle un análisis bayesiano jerárquico con el objetivo de estimar los CVC del modelo. Es importante destacar que ni la formulación del modelo ni los resultados aquí descriptos con relación al análisis bayesiano son originales. De hecho, existe una cobertura muy extensa de estos temas en libros clásicos de la disciplina, como los de Henderson (1984, cap. 31), Mrode (2005, cap. 6) y, en particular, el de Sorensen y Gianola (2002, cap. 13.3). En cambio, el objetivo principal de este capítulo fue definir el marco de referencia sobre el que desarrollará el resto de la tesis. No obstante eso, la implementación del GS aquí descripta para estimar los CVC y los parámetros genéticos del rodeo Angus de Las Lilas no fue necesariamente estándar.

En particular, en la mayoría de las aplicaciones del GS bajo el MAM publicadas en la literatura se han definido distribuciones a priori uniformes para los CVC (*e.g*. Jensen *et al*., 1994, Quintanilla *et al*., 1999), generalmente con el objeto de representar ignorancia total respecto a los valores posibles de los parámetros a priori. Si bien este enfoque es generalmente aceptado, no está exento de críticas (*e.g*. Blasco, 2001). En contraste, aquí se han descripto distribuciones a priori informativas, particularmente distribuciones Gamma invertidas, parameterizadas en términos de ciertos valores 'razonables' para los CVC, por un lado, y en términos de la incertidumbre asociada a estos valores, por otro. La principal dificultad con este enfoque radica en que los resultados serán más o menos sensibles a la especificación de las distribuciones a priori según cuán informativos sean los datos. En general, si los datos son lo suficientemente informativos entonces la distribución a priori tendrá poca influencia en los resultados (Blasco, 2001). En este sentido, la robustez de los resultados obtenidos bajo las dos implementaciones del algoritmo de estimación aquí descriptas indicaría que la información contenida en el archivo de pesos al destete analizado resultó adecuada para estimar todos los CVC inherentes al MAM.

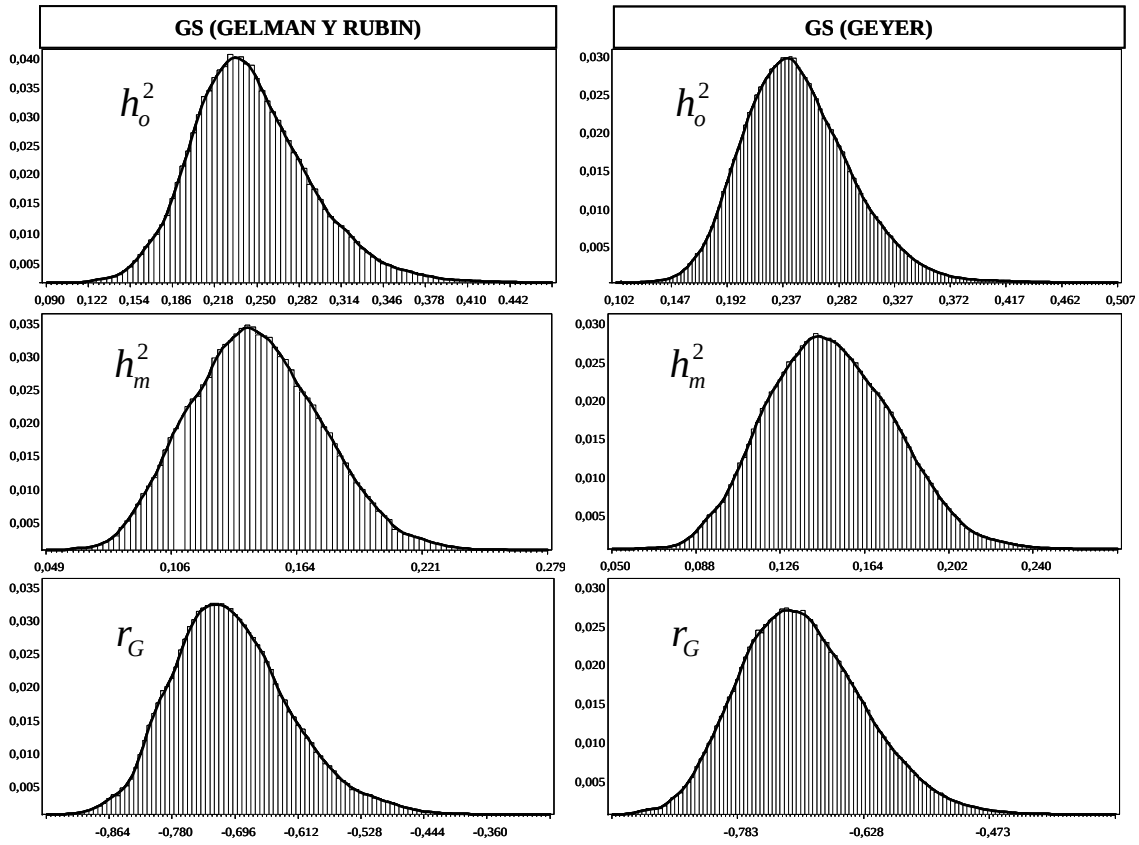


Figura 2.2. Distribuciones marginales posteriores de los parámetros genéticos. A la izquierda de la figura se presentan las densidades obtenidas a partir de 270.000 ciclos provenientes de combinar tres cadenas independientes (Gelman y Rubin, 1992). A la derecha, en cambio, se presentan las densidades obtenidas a partir de una única cadena de 500.000 ciclos (Geyer, 1992). Las curvas fueron obtenidas mediante de un método no paramétrico basado en un núcleo Gaussiano (Silverman, 1986).

En tal caso, la elección de una u otra implementación es indiferente. En términos generales, sin embargo, la principal determinante a la hora de tomar decisiones respecto a la implementación del GS es su factibilidad computacional. Al respecto, deben distinguirse dos aspectos bien diferentes con relación al tiempo de cómputo: 1. el número de operaciones aritméticas necesarias para completar un ciclo del algoritmo; y 2. el número de ciclos necesario para asegurar la convergencia del procedimiento. El número de operaciones por ciclo será una función lineal del número de individuos en el archivo de pedigree, mientras que el número de ciclos para asegurar la convergencia dependerá de las correlaciones entre muestras observadas. En todo caso, si el tiempo de cómputo por iteración no es limitante, entonces ejecutar una única cadena muy larga y descartar un número importante de muestreos iniciales para calcular los estadísticos descriptivos posteriores resulta en una estrategia de inferencia válida (Geyer, 1992). En cambio, cuando el tiempo de cómputo por ciclo es limitante es necesario tener mayor certeza del momento en el que el algoritmo converge; *i.e.*, comienza a muestrear de las distribuciones marginales de interés. En ese caso, es útil ejecutar una serie de cadenas inicializadas en puntos dispersos del soporte de las distribuciones de los parámetros de interés, y establecer luego la convergencia por inspección visual de las gráficas de muestreos en función del número de ciclos (Gelman y Rubin, 1992). Incluso, es posible ejecutar los algoritmos en

diferentes procesadores simultáneamente y, tras descartar los ciclos de calentamiento, colectar los resultados obtenidos para aumentar el número de muestreos sobre los que se basará la inferencia.

En este trabajo se aplicaron ambas implementaciones a modo de ilustración y, como era de esperar considerando que el tiempo de ejecución no fue limitante para obtener un número suficiente de muestreos, se obtuvieron los mismos resultados en términos de las estimaciones de los parámetros genéticos. Así, las medias posteriores de las heredabilidades directa y materna, y la correlación genética directa-materna, tomadas como estimaciones puntuales, fueron razonables y acordes a la literatura (*cf*. CSIRO, 2010), e incluso coincidentes con las estimaciones REML.

Por último, otros dos resultados obtenidos serán examinados con un mayor grado de detalle a la luz de las restricciones que impone el muestreo de los CVC genéticos de una distribución IW: 1. la dispersión diferencial que existe entre las distribuciones posteriores de la heredabilidad directa y de la heredabilidad materna; 2. las altísimas autocorrelaciones entre muestras para los CVC genéticos. Ambos resultados son frecuentes al estimar los parámetros genéticos del MAM vía el GS. Con respecto al primer punto, hay que señalar que en los archivos de datos utilizados comúnmente en los programas de mejoramiento genético animal en general existe menos información (o más incertidumbre) en torno al valor de la heredabilidad materna que en torno a la heredabilidad directa, básicamente porque existen menos relaciones de parentesco que provean contrastes informativos para estimar este parámetro. Como fuera discutido en el Capítulo 1, los efectos maternos se expresan con una generación de retraso respecto a los efectos directos y, además, están limitados a un sexo (Willham, 1980). Sin embargo, el analista no está en condiciones de modelar esta incertidumbre diferencial porque la distribución IW es función de un único hiperparámetro escalar, que restringe la matriz de covarianza genética en su conjunto. Por su parte, las altas correlaciones de muestreo observadas para los CVC genéticos son una consecuencia directa del muestreo conjunto de estos parámetros de una distribución multivariada IW. En el próximo capítulo se abordarán estos dos problemas, y se considerará y evaluará el uso alternativo de la distribución Wishart invertida generalizada (GIW).

3
La distribución Wishart invertida generalizada y su aplicación en el contexto de la estimación de parámetros genéticos en un modelo animal con efectos maternos[1]

[1] Munilla, S. y R. J. C. Cantet. 2011. Bayesian conjugate analysis using a generalized inverted Wishart distribution accounts for differential uncertainty among the genetic parameters – an application to the maternal animal model. *J. Anim. Breed. Genet.* (En prensa).

3.1. INTRODUCCIÓN

En este capítulo se considera el uso de la distribución Wishart invertida generalizada (GIW) para abordar el problema de la estimación de CVC genéticos en el marco de un análisis bayesiano jerárquico como el presentado en el capítulo anterior. La distribución GIW fue introducida originalmente por Brown *et al*. (1994) en el contexto de estudios sobre la evaluación del riesgo de contaminación del aire (*cf*. Le y Zidek, 2006), y constituye esencialmente una extensión de la distribución Wishart invertida (IW) con un mayor número de parámetros, un atributo que le confiere gran flexibilidad. En particular, la distribución GIW surge como una alternativa natural para especificar la estructura de covarianza a priori de observaciones que siguen una distribución normal multivariada con un patrón monótono de datos faltantes (Garthwaite y Al-Awadhi, 2001).

Se argumenta aquí que la distribución también puede ser utilizada para especificar una estructura de covarianza a priori más flexible para los CVC genéticos en el contexto de los modelos estadísticos utilizados por los mejoradores animales, en particular cuando existe información diferencial para estimar los diferentes componentes escalares. Considérese, por ejemplo, el MAM. Como fuera oportunamente discutido (véase Cap. 2.), en este caso existe en general menos información (más incertidumbre) en torno a la heredabilidad materna que en torno a la heredabilidad directa. En tal situación, el analista razonablemente tenderá a favorecer una especificación a priori que permita representar esta incertidumbre diferencial.

El capítulo está organizado del siguiente modo. En primer lugar, se introduce la distribución GIW en toda su generalidad. Luego, se presentan resultados teóricos con respecto a la especificación de la GIW como la distribución a priori de la matriz de covarianza genética del MAM en el contexto de un análisis bayesiano jerárquico. Finalmente, se describe un método de 'actualización bayesiana' (*bayesian updating*) para determinar los hiperparámetros de la distribución a priori de los CVC genéticos, basado en las propiedades de la distribución GIW, y se presentan luego estimaciones de los parámetros obtenidas ajustando datos de campo y datos simulados de peso al destete. Los resultados son comparados contra especificaciones a priori más estándares en términos de precisión de las estimaciones, errores estándares y comportamiento de convergencia de las cadenas de Markov.

3.2. MÉTODOS

3.2.1. Distribución Wishart invertida generalizada

Sea $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ una colección de vectores de orden $g \times 1$. Defínase luego $\boldsymbol{Y}$ ($n \times g$), tal que $\boldsymbol{Y}^{\mathrm{T}} = \left(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \right)$, y considérese finalmente una distribución de probabilidad normal multivariada, tal que

$$vec\left(\boldsymbol{Y}\right) | \boldsymbol{\Sigma} \sim NMV\left(\boldsymbol{0}, \boldsymbol{A} \otimes \boldsymbol{\Sigma}\right), \qquad [3.1]$$

Donde $vec\left(\cdot\right)$ representa al operador matricial 'vec' (Searle, 1982, cap. 12.9), $\boldsymbol{A}$ es una matriz ($n \times n$) simétrica y conocida, y $\boldsymbol{\Sigma}$ es una matriz de covarianza aleatoria ($g \times g$). En el contexto de un análisis bayesiano jerárquico, generalmente se asume que $\boldsymbol{\Sigma}$ sigue a priori una distribución $IW\left(\delta, \boldsymbol{\Psi}\right)$. En tal caso, nótese que mientras los elementos de la matriz simétrica y positiva definida $\boldsymbol{\Psi}$ constituyen un conjunto complementario de

hiperparámetros para modelar la esperanza a priori de $\boldsymbol{\Sigma}$, la incertidumbre respecto a estos valores está gobernada por un único parámetro escalar, δ (Brown, 2002). Alternativamente, es posible obtener una especificación más flexible a partir de la 'descomposición de Bartlett' (Bartlett, 1933) de $\boldsymbol{\Sigma}$.

Considérese, en primer lugar, una partición de la matriz de covarianza en 2 × 2 bloques,

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}. \qquad [3.2]$$

La descomposición de Bartlett de $\boldsymbol{\Sigma}$ es tal que $\boldsymbol{\Sigma} = \boldsymbol{T}\boldsymbol{\Delta}\boldsymbol{T}^{\mathrm{T}}$, con

$$\boldsymbol{\Delta} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \end{bmatrix} \text{ y } \boldsymbol{T} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} & \boldsymbol{I} \end{bmatrix}. \qquad [3.3]$$

Denótese $\boldsymbol{\tau} \equiv \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}$ y $\boldsymbol{\Gamma} \equiv \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$. Luego, es posible definir una transformación biyectiva $\boldsymbol{\Sigma} \to (\boldsymbol{\Sigma}_{11}, \tau, \Gamma)$ de la matriz de covarianza en los 'parámetros de Bartlett' del siguiente modo:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{11}\boldsymbol{\tau}^{\mathrm{T}} \\ \boldsymbol{\tau}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Gamma} + \boldsymbol{\tau}\boldsymbol{\Sigma}_{11}\boldsymbol{\tau}^{\mathrm{T}} \end{bmatrix}. \qquad [3.4]$$

En el caso más general de bloques múltiples, la descomposición puede aplicarse en forma recursiva. Antes de proceder, sin embargo, es necesario establecer cierta notación. A lo largo de este capítulo adoptaremos aquella de Le *et al*. (1999).

Defínase la partición de $\boldsymbol{\Sigma}$ en $k \times k$ bloques según

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{1,1} & \cdots & \boldsymbol{\Sigma}_{1,k} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{k,1} & \cdots & \boldsymbol{\Sigma}_{k,k} \end{bmatrix}, \qquad [3.5]$$

con $\boldsymbol{\Sigma}_{i,l}$ de orden $g_i \times g_l$, tal que $g_1 + \ldots + g_k = g$. Denótese luego la submatriz principal hasta el j-ésimo bloque como $\boldsymbol{\Sigma}^{[1,\ldots,j]}$. Es decir,

$$\boldsymbol{\Sigma}^{[1,\ldots,j]} = \begin{bmatrix} \boldsymbol{\Sigma}_{1,1} & \cdots & \boldsymbol{\Sigma}_{1,j} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{j,1} & \cdots & \boldsymbol{\Sigma}_{j,j} \end{bmatrix}. \qquad [3.6]$$

Sean además $\boldsymbol{\Sigma}^{[(j+1)j]} = \left(\boldsymbol{\Sigma}_{(j+1),1}, \ldots, \boldsymbol{\Sigma}_{(j+1),j}\right)$ y $\boldsymbol{\Sigma}^{[j(j+1)]} = \left(\boldsymbol{\Sigma}^{[(j+1)j]}\right)^{\mathrm{T}}$, esta última igualdad de acuerdo a la simetría de $\boldsymbol{\Sigma}$. Entonces, para $j = k - 1, \ldots, 1$

$$\boldsymbol{\Sigma}^{[1,\ldots,j+1]} = \begin{bmatrix} \boldsymbol{\Sigma}^{[1,\ldots,j]} & \boldsymbol{\Sigma}^{[1,\ldots,j]}\boldsymbol{\tau}_j^{\mathrm{T}} \\ \boldsymbol{\tau}_j\boldsymbol{\Sigma}^{[1,\ldots,j]} & \boldsymbol{\Gamma}_j + \boldsymbol{\tau}_j\boldsymbol{\Sigma}^{[1,\ldots,j]}\boldsymbol{\tau}_j^{\mathrm{T}} \end{bmatrix}, \qquad [3.7]$$

con $\boldsymbol{\tau}_j = \boldsymbol{\Sigma}^{[(j+1)j]}\left(\boldsymbol{\Sigma}^{[1,\ldots,j]}\right)^{-1}$ y $\boldsymbol{\Gamma}_j = \boldsymbol{\Sigma}_{(j+1),(j+1)} - \boldsymbol{\Sigma}^{[(j+1)j]}\left(\boldsymbol{\Sigma}^{[1,\ldots,j]}\right)^{-1}\boldsymbol{\Sigma}^{[j(j+1)]}$.

Considérese ahora una partición conformable de la matriz de observaciones $\boldsymbol{Y}$ en k bloques,

$$\boldsymbol{Y} = \left(\boldsymbol{Y}^{[1]}, \ldots, \boldsymbol{Y}^{[k]}\right) \qquad [3.8]$$

con $\boldsymbol{Y}^{[i]} = \left(\boldsymbol{y}_1^{[i]}, \ldots, \boldsymbol{y}_n^{[i]}\right)^{\mathrm{T}}$ para el i-ésimo bloque. Los bloques son de orden $n \times g_l$, tal que $g_1 + \ldots + g_k = g$. Esta notación pone énfasis en el hecho de que no existe necesariamente una correspondencia biyectiva entre un bloque de observaciones y cada coordenada de los vectores $\boldsymbol{y}$, aunque por conveniencia asumiremos que sí en desarrollos futuros, tal que $g_l = 1$ para todo l.

Ahora bien, usando las propiedades de la distribución normal (*cf.* Bauwens *et al.*, 1999, sección A.2.3) y la notación de la descomposición de Bartlett en bloques múltiples se pude expresar la distribución conjunta de $\boldsymbol{Y}$ como el producto de la siguiente secuencia de distribuciones condicionales (Brown, 2002):

$$\begin{aligned}
\boldsymbol{Y}^{[1]} &\sim N\left(\boldsymbol{0}, \boldsymbol{A} \otimes \boldsymbol{\Sigma}_{1,1}\right) \\
\boldsymbol{Y}^{[2]} \mid \boldsymbol{Y}^{[1]} &\sim N\left(\boldsymbol{Y}^{[1]}\boldsymbol{\tau}_1, \boldsymbol{A} \otimes \boldsymbol{\Gamma}_1\right) \\
&\vdots \\
\boldsymbol{Y}^{[j+1]} \mid \boldsymbol{Y}^{[1]}, \ldots, \boldsymbol{Y}^{[j]} &\sim N\left(\boldsymbol{Y}^{[1,\ldots,j]}\boldsymbol{\tau}_j, \boldsymbol{A} \otimes \boldsymbol{\Gamma}_j\right),
\end{aligned} \qquad [3.9]$$

para $j = 2, \ldots k-1$, con $\boldsymbol{Y}^{[1,\ldots,j]} = \left(\boldsymbol{Y}^{[1]}, \ldots, \boldsymbol{Y}^{[j]}\right)$.

Nótese que la expresión [3.9] sugiere un modo de especificar una distribución a priori con un mayor número de parámetros para la matriz de covarianza $\boldsymbol{\Sigma}$. De acuerdo a la independencia mutua entre $\boldsymbol{\Sigma}_{1,1}$ y los pares $\left(\boldsymbol{\tau}_j, \boldsymbol{\Gamma}_j\right)$, $j = 1, \ldots k-1$, una propiedad que descansa en la descomposición de Bartlett, asúmase que a priori

$$\begin{aligned}
\boldsymbol{\Sigma}_{1,1} &\sim IW\left(\delta_0, \boldsymbol{Q}_0\right) \\
\boldsymbol{\tau}_j \mid \boldsymbol{\Gamma}_j &\sim N\left(\boldsymbol{\tau}_{0j}, \boldsymbol{H}_j \otimes \boldsymbol{\Gamma}_j\right) \\
\boldsymbol{\Gamma}_j &\sim IW\left(\delta_j + g^{[1,\ldots,j]}, \boldsymbol{Q}_j\right),
\end{aligned} \qquad [3.10]$$

con $g^{[1,\ldots,j]} = g_1 + \ldots + g_j$. En [3.10], $\mathcal{H} = \left\{\delta_0, \boldsymbol{Q}_0; \delta_j, \boldsymbol{\tau}_{0j}, \boldsymbol{Q}_j, \boldsymbol{H}_j, j = 1, \ldots, k-1\right\}$ constituye el conjunto de hiperparámetros. En este contexto, se dice que la matriz de covarianza $\boldsymbol{\Sigma}$ sigue una distribución Wishart invertida generalizada (Brown, 2002) y se denota $\boldsymbol{\Sigma} \sim GIW\left(\mathcal{H}\right)$.

La distribución GIW se caracteriza esencialmente por el mayor número de parámetros que involucra, una característica que ofrece gran flexibilidad al momento de especificar el conocimiento a priori o una opinión experta respecto a estructura de covarianza de las observaciones. Adicionalmente, posee la ventaja de la simplicidad computacional, dado que los parámetros de Bartlett siguen distribuciones fáciles de muestrear;

específicamente, distribuciones normales y distribuciones IW de menor orden. Por último, nótese que la distribución GIW puede definirse en forma recursiva, dado que la submatriz principal $\boldsymbol{\Sigma}^{[1,\ldots,\ j]}$ de $\boldsymbol{\Sigma}$ pude considerarse

$$\boldsymbol{\Sigma}^{[1,\ldots,\ j]} \sim GIW\left(\delta_i, \boldsymbol{Q}_i, \boldsymbol{\tau}_{0i}, \boldsymbol{H}_i, i = 1, \ldots,\ j-1\right) \qquad [3.11]$$

sucesivamente para $j = k - 1, \ldots, 2$, si se define $\boldsymbol{\Sigma}^{[1,1]} \equiv \boldsymbol{\Sigma}_{1,1} \sim IW\left(\delta_0, \boldsymbol{Q}_0\right)$ para $j = 1$.

3.2.2. Especificación a priori usando la GIW: resultados teóricos

El objetivo a continuación es derivar resultados teóricos con respecto a la especificación de la distribución GIW como la función de densidad de probabilidad a priori de la matriz de covarianza genética, en el contexto de un análisis bayesiano jerárquico. En particular, considérese el problema de estimar CVC bajo el MAM (*cf.* Sorensen y Gianola, 2002, cap. 13.3). Como ya se ha comentado, en este caso es estándar asumir una distribución IW a priori para $\boldsymbol{\Sigma}$, principalmente porque esta distribución es conjugada (véase Cap. 2) y, en consecuencia, facilita la implementación del algoritmo GS. Como alternativa, se demostrará que asumir una distribución GIW extiende considerablemente el abanico de posibles especificaciones a priori, sin perder esta importante ventaja. En este punto se hará constante referencia al análisis bayesiano jerárquico para el MAM presentado en el capítulo precedente.

3.2.2.1. Partición del vector de valores de cría

Sea la distribución a priori del vector de valores de cría del MAM (ecuación [2.3])

$$\boldsymbol{a} = \begin{bmatrix} \boldsymbol{a}_o \\ \boldsymbol{a}_m \end{bmatrix} \sim N\left(\boldsymbol{0},\ \boldsymbol{\Sigma} \otimes \boldsymbol{A}\right). \qquad [3.12]$$

De acuerdo a los resultados presentados en la sección precedente, considérese expresar esta distribución conjunta como el producto de las siguientes distribuciones:

$$\begin{aligned} \boldsymbol{a}_o &\sim N\left(\boldsymbol{0}, \Sigma_{11} \times \boldsymbol{A}\right) \\ \boldsymbol{a}_m \mid \boldsymbol{a}_o &\sim N\left(\boldsymbol{a}_o \tau, \Gamma \times \boldsymbol{A}\right), \end{aligned} \qquad [3.13]$$

con $\tau = \Sigma_{12}\Sigma_{11}^{-1}$ y $\Gamma = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. Asúmase ahora que la distribución GIW se utiliza para representar la incertidumbre a priori respecto a la matriz de covarianza $\boldsymbol{\Sigma}$, de modo que los parámetros de Bartlett (Σ_{11}, τ y Γ) se distribuyen

$$\begin{aligned} \Sigma_{11} &\sim S_0 \chi^{-2}_{\upsilon_0} \\ \tau \mid \Gamma &\sim N\left(\tau_0, \Gamma \times H\right) \\ \Gamma &\sim S_1 \chi^{-2}_{\upsilon_1 + 1}, \end{aligned} \qquad [3.14]$$

donde $S\chi^{-2}_{\upsilon}$ representa a una distribución Chi-cuadrada escalada invertida con parámetros (υ, S), un caso especial de la distribución IW con una matriz de escala escalar. El conjunto de hiperparámetros en [3.14] es $\mathcal{H} = \left\{\upsilon_0, \upsilon_1, S_0, S_1, \tau_0, H\right\}$. Todos estos parámetros deben ser definidos por el analista.

3.2.2.2. Distribución condicional posterior de los parámetros de Bartlett

Como es estándar, la siguiente etapa del análisis bayesiano jerárquico involucra formar la distribución posterior conjunta de todas las incógnitas que comprende el modelo, multiplicando la función de verosimilitud por cada una de las distribuciones a priori. Luego, la distribución condicional posterior de cualquier parámetro de interés se deriva dejando al resto de ellos constante.

En particular, la distribución condicional posterior de la matriz de covarianza genética $\mathbf{\Sigma}$ bajo el modelo [2.3] será proporcional a

$$\begin{aligned} & p\left(\mathbf{\Sigma} \mid \mathcal{H}, \mathcal{D}\right) \propto \\ & \quad \propto p\left(\boldsymbol{a}_o \mid \Sigma_{11}\right) \times p\left(\Sigma_{11} \mid S_0, \upsilon_0\right) \times \\ & \quad \times p\left(\boldsymbol{a}_m \mid \boldsymbol{a}_o, \tau, \Gamma\right) \times p\left(\tau \mid \Gamma, \tau_0, H\right) \times p\left(\Gamma \mid S_1, \upsilon_1\right), \end{aligned} \qquad [3.15]$$

donde $\mathcal{D} = \left\{\boldsymbol{b}, \boldsymbol{a}_o, \boldsymbol{a}_m, \boldsymbol{e}_m, \sigma^2_{e_m}, \sigma^2_{e_o}, \boldsymbol{y}\right\}$. Explícitamente, y luego de algunos arreglos algebraicos,

$$\begin{aligned} & p\left(\mathbf{\Sigma} \mid \mathcal{H}, \mathcal{D}\right) \propto \\ & \quad \propto \left(\Sigma_{11}\right)^{-\frac{1}{2}(q+\upsilon_0+2)} \times exp\left\{-\frac{Q_{11}+S_0}{2\Sigma_{11}}\right\} \times \left(\Gamma\right)^{-\frac{1}{2}\left[(q+\upsilon_1+1)+2\right]} \times \\ & \quad \times exp\left\{-\frac{\left(\tau^2 Q_{11} - 2\tau Q_{12} + Q_{22}\right) + H^{-1}\left(\tau - \tau_0\right)^2 + S_1}{2\Gamma}\right\}, \end{aligned} \qquad [3.16]$$

donde se ha hecho uso de la siguiente notación para la matriz simétrica de sumas de cuadrados y productos cruzados

$$\boldsymbol{Q} = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \equiv \begin{bmatrix} \boldsymbol{a}_o^{\mathrm{T}} \boldsymbol{A}^{-1} \boldsymbol{a}_o & \boldsymbol{a}_o^{\mathrm{T}} \boldsymbol{A}^{-1} \boldsymbol{a}_m \\ \boldsymbol{a}_m^{\mathrm{T}} \boldsymbol{A}^{-1} \boldsymbol{a}_o & \boldsymbol{a}_m^{\mathrm{T}} \boldsymbol{A}^{-1} \boldsymbol{a}_m \end{bmatrix}. \qquad [3.17]$$

De la expresión [3.16] se deduce que la distribución condicional posterior de la matriz de covarianza genética, $\mathbf{\Sigma}$, puede considerarse proporcional al producto de tres distribuciones de probabilidad asociadas a los parámetros de Bartlett; *i.e.*,

$$\begin{aligned} & p\left(\mathbf{\Sigma} \mid \mathcal{H}, \mathcal{D}\right) \propto \\ & \quad \propto p\left(\Sigma_{11} \mid \mathcal{H}, \mathcal{D}\right) \times p\left(\tau, \Gamma \mid \mathcal{H}, \mathcal{D}\right) = \\ & \quad = p\left(\Sigma_{11} \mid \mathcal{H}, \mathcal{D}\right) \times p\left(\tau \mid \Gamma, \mathcal{H}, \mathcal{D}\right) \times p\left(\Gamma \mid \mathcal{H}, \mathcal{D}\right). \end{aligned} \qquad [3.18]$$

De hecho, se demostrará a continuación que las tres distribuciones coinciden con las correspondientes a los parámetros de Bartlett, de acuerdo a su definición en la ecuación [3.10]. La demostración será bosquejada aquí, de modo de poner énfasis en los resultados principales. Una derivación detallada de algunos pasos importantes se difiere al Apéndice A de este trabajo.

Nótese, en primer lugar, que la independencia entre Σ_{11} y el par (τ, Γ) se desprende directamente de [3.16]. Reteniendo los factores que sólo dependen de Σ_{11}, es posible escribir

$$p(\Sigma_{11} \mid \mathcal{H}, \mathcal{D}) \propto (\Sigma_{11})^{-\frac{1}{2}(q+\upsilon_0+2)} \times exp\left\{-\frac{Q_{11}+S_0}{2\Sigma_{11}}\right\}. \qquad [3.19]$$

Luego, tras definir $\tilde{S}_0 = Q_{11} + S_0$ y $\tilde{\upsilon}_0 = q + \upsilon_0$, la expresión [3.19] puede reconocerse como el núcleo de una distribución Chi-cuadrada escalada invertida; *i.e.*,

$$\Sigma_{11} \mid \mathcal{H}, \mathcal{D} \sim \tilde{S}_0 \chi^{-2}_{\tilde{\upsilon}_0}. \qquad [3.20]$$

En segundo lugar, ignorando todos los términos que no dependen de τ del argumento de la función exponencial en [3.16] se verifica que

$$p(\tau \mid \Gamma, \mathcal{H}, \mathcal{D}) \propto$$
$$\propto exp\left\{-\frac{\left(\tau^2 Q_{11} - 2\tau Q_{12}\right) + H^{-1}(\tau - \tau_0)^2}{2\Gamma}\right\}. \qquad [3.21]$$

Luego de algunas manipulaciones algebraicas sobre esta última expresión, en el Apéndice A se demuestra que

$$p(\tau \mid \Gamma, \mathcal{H}, \mathcal{D}) \propto$$
$$\propto exp\left\{-\frac{\left(Q_{11}+H^{-1}\right)\left[\tau - \left(Q_{12}+\tau_0 H^{-1}\right)\left(Q_{11}+H^{-1}\right)^{-1}\right]^2}{2\Gamma}\right\}, \qquad [3.22]$$

de donde se deduce inmediatamente que

$$\tau \mid \Gamma, \mathcal{H}, \mathcal{D} \sim N\left(\frac{Q_{12}+\tau_0 H^{-1}}{Q_{11}+H^{-1}}, \frac{\Gamma}{Q_{11}+H^{-1}}\right). \qquad [3.23]$$

Una representación más intuitiva de este último resultado puede obtenerse utilizando las siguientes identidades (Brown, 2002):

$$\tilde{H} \equiv \left(Q_{11}+H^{-1}\right)^{-1},\ W \equiv \tilde{H} \times H^{-1} \text{ y } \hat{\tau} \equiv Q_{11}^{-1} Q_{12}. \qquad [3.24]$$

A partir de ellas, los parámetros en [3.23] pueden rescribirse según

$$\tau \mid \Gamma, \mathcal{H}, \mathcal{D} \sim N\left(\tilde{\tau}_0,\ \Gamma \times \tilde{H}\right), \qquad [3.25]$$

donde $\tilde{\tau}_0 = W\tau_0 + (1-W)\hat{\tau}$. Esta representación indica que la esperanza condicional posterior es un promedio ponderado de la esperanza de la distribución a priori y de la información que proveen los datos a través del cociente de formas cuadráticas. Nótese además que las ponderaciones dependerán de la definición del hiperparámetro *H*. En particular, una elección estándar para retener la misma estructura media que con la dis-

tribución IW consiste en especificar $H = S_0^{-1}$ (Brown, 2002). En tal caso, se verifica que $W = \left(Q_{11} S_0^{-1} + 1\right)^{-1}$, de donde se desprende que si la información a priori y la información que proveen los datos respecto al valor de la varianza aditiva directa es la misma (*i.e.*, $S_0^{-1} = Q_{11}$), también lo serán las ponderaciones de ambos términos en la media posterior de τ. En cambio, si la información que proveen los datos es mayor, entonces el término asociado a las formas cuadráticas tendrá una ponderación más alta.

Regresando ahora al argumento principal, aún queda deducir la distribución condicional posterior de Γ. De [3.16],

$$\begin{aligned} & p\left(\Gamma \mid \mathcal{H}, \mathcal{D}\right) \propto \\ & \quad \propto \left(\Gamma\right)^{-\frac{1}{2}\left[\left(q+\upsilon_1+1\right)+2\right]} \times exp\left\{-\frac{S_1^* + S_1}{2\Gamma}\right\}, \end{aligned} \qquad [3.26]$$

donde S_1^* resulta de recolectar todos los términos de la función exponencial que no dependen de τ. La expresión [3.26] puede reconocerse como el núcleo de la siguiente distribución Chi-cuadrada escalada invertida

$$\Gamma \mid \mathcal{H}, \mathcal{D} \sim \tilde{S}_1 \chi^{-2}_{\tilde{\upsilon}_1+1}, \qquad [3.27]$$

con $\tilde{S}_1 = S_1^* + S_1$ y $\tilde{\upsilon}_1 = q + \upsilon_1$. Adicionalmente, en el Apéndice A se demuestra que

$$S_1^* = \left(\boldsymbol{a}_m - \boldsymbol{a}_o \hat{\tau}\right)^{\mathrm{T}} \boldsymbol{A}^{-1} \left(\boldsymbol{a}_m - \boldsymbol{a}_o \hat{\tau}\right) + W Q_{11} \left(\hat{\tau} - \tau_0\right)^2. \qquad [3.28]$$

De estos resultados se desprende que existen tres fuentes de información que contribuyen al valor que toma el parámetro de escala de la distribución condicional posterior de Γ. Primero, existe información a priori que contribuye S_1. Segundo, existe información que contribuyen los datos a través de la forma cuadrática en los valores de cría maternos ajustados, como puede apreciarse en el primer término a la derecha de la igualdad en [3.28]. De hecho, puede verificarse que este término es la expresión del estimador máximo-verosímil de Γ bajo la distribución de $\boldsymbol{a}_m \mid \boldsymbol{a}_o$ si $\hat{\tau}$ es interpretado como el valor verdadero de τ. La tercera fuente de información surgiría de esta última substitución, teniendo en cuenta que τ es estimado por $\hat{\tau}$.

3.2.2.3. Recuperando la matriz $\boldsymbol{\Sigma}$

Los resultados [3.20], [3.25] y [3.27] implican que la matriz de covarianza genética, $\boldsymbol{\Sigma}$, sigue una distribución condicional conjugada $GIW\left(\tilde{\upsilon}_0, \tilde{\upsilon}_1, \tilde{S}_0, \tilde{S}_1, \tilde{\tau}_0, \tilde{H}\right)$ a posteriori y, en consecuencia, la estimación de CVC puede llevarse a cabo mediante un algoritmo de GS. De hecho, en la etapa correspondiente del algoritmo sólo será necesario muestrear secuencialmente de las distribuciones condicionales de los parámetros de Bartlett y luego recuperar la matriz $\boldsymbol{\Sigma}$ aplicando la descomposición de Bartlett en sentido inverso; *i.e.*, calculando

$$\boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{11} & \Sigma_{11}\tau \\ \tau\Sigma_{11} & \Gamma + \tau^2 \Sigma_{11} \end{bmatrix}. \qquad [3.29]$$

3.2.2.4. Diferentes especificaciones a priori

Antes de proceder con la implementación del GS el analista debe definir el conjunto completo de hiperparámetros $\mathcal{H} = \{\upsilon_0, \upsilon_1, S_0, S_1, \tau_0, H\}$, de modo de describir su conocimiento a priori. En esta sección se discutirán tres especificaciones a priori diferentes, todas ellas basadas en las propiedades de la distribución GIW. En primer lugar, se definirá una distribución a priori no informativa, que permita reflejar completa incertidumbre respecto a la matriz de covarianza. Después, se presentará el conjunto particular de hiperparámetros que resultará en un muestreo equivalente de una distribución IW a posteriori. Finalmente, y con base en este último conjunto, se sugerirá modelar la incertidumbre diferencial entre los CVC genéticos asignando valores distintos a los parámetros υ_0 y υ_1.

Asúmase, en primer lugar, que la distribución a priori de $\mathbf{\Sigma}$ es proporcional a $|\mathbf{\Sigma}|^{-\frac{1}{2}\upsilon} = (\Sigma_{11})^{-\frac{1}{2}\upsilon} \times \Gamma^{-\frac{1}{2}\upsilon}$. En particular, si se define $\upsilon = 3$, la densidad corresponde a la distribución no informativa e invariante de Jeffreys (*cf*. Brown, 2002). Luego, la distribución condicional posterior de la matriz de covarianza genética puede escribirse explícitamente según

$$\begin{aligned} p(\mathbf{\Sigma} | \mathcal{H}, \mathcal{D}) &\propto \\ &\propto (\Sigma_{11})^{-\frac{1}{2}(q+\upsilon)} \times exp\left\{-\frac{Q_{11}}{2\Sigma_{11}}\right\} \times \\ &\times (\Gamma)^{-\frac{1}{2}(q+\upsilon)} \times exp\left\{-\frac{\left(\tau^2 Q_{11} - 2\tau Q_{12} + Q_{22}\right)}{2\Gamma}\right\}. \end{aligned} \qquad [3.30]$$

Entonces, recurriendo a los mismos argumentos que se han utilizado para derivar la distribución condicional posterior de los parámetros de Bartlett se puede verificar que $\mathbf{\Sigma}$ sigue condicionalmente una distribución $GIW\left(\tilde{\upsilon}_0, \tilde{\upsilon}_1, \tilde{S}_0, \tilde{S}_1, \tilde{\tau}_0, \tilde{H}\right)$ a posteriori con $\tilde{\upsilon}_0 = q+1$, $\tilde{\upsilon}_1 = q$, y

$$\begin{aligned} \tilde{S}_0 &= Q_{11}, \\ \tilde{S}_1 &= Q_{22} - Q_{11}^{-1} Q_{12}^2, \\ \tilde{\tau}_0 &= Q_{11}^{-1} Q_{12}, \\ \tilde{H} &= Q_{11}^{-1}, \end{aligned} \qquad [3.31]$$

donde Q_{ij} simboliza el elemento (i, j) de la matriz simétrica de sumas de cuadrados y productos cruzados definida en [3.17].

Alternativamente, asúmase una distribución IW a priori para $\mathbf{\Sigma}$ bajo un enfoque conjugado (*e.g*. Jensen *et al*., 1994), pero considere la posibilidad de muestrear secuencialmente de las distribuciones condicionales posteriores de los parámetros de Bartlett. Tal estrategia de muestreo sería ventajosa desde un punto de vista algorítmico, dado que sólo requiere muestrear tres variables normales estándares contra las que requeriría un muestreo directo de la distribución IW (Smith y Hocking, 1972). De hecho, algunas subrutinas disponibles para muestrear de la distribución IW se basan en la descomposición de Bartlett (*e.g*. la subrutina 'WSHRT' escrita en F77 por Smith y Hocking, 1972).

En el Apéndice B se demuestra que la equivalencia se basa en definir el siguiente conjunto de hiperparámetros

$$\begin{aligned} S_0 &= \Sigma_{11}^{*}, \\ S_1 &= \Sigma_{22}^{*} - \Sigma_{12}^{*2}\Sigma_{11}^{*-1}, \\ \tau_0 &= \Sigma_{12}^{*}\Sigma_{11}^{*-1}, \\ H &= \Sigma_{11}^{*-1}, \end{aligned} \qquad [3.32]$$

donde $\mathbf{\Sigma}^{*} = \left\{\Sigma_{ij}^{*}\right\}$ representa la matriz de escala de la distribución a priori de $\mathbf{\Sigma}$, y definiendo luego $\upsilon_0 = \upsilon + 1$ y $\upsilon_1 = \upsilon$, donde υ son los grados de credibilidad en común.

Es más, asignando valores distintos a los parámetros υ_0 y υ_1 es posible modelar la incertidumbre diferencial entre las estimaciones de las varianzas aditivas directa y materna. Esta última estrategia será explorada en la próxima sección. En el Apéndice B se presenta un algoritmo de muestreo fácil de acomodar dentro del código de un algoritmo de GS.

3.2.3. Especificación a priori usando la GIW: una aplicación

La gran mayoría de las asociaciones de criadores de ganado bovino de carne llevan adelante evaluaciones genéticas como parte de sus programas de control de producción (BIF, 2002, cap. 5). Los resultados principales de estas evaluaciones son las predicciones de los valores de cría de todos los individuos pertenecientes a la población bajo estudio. Como fuera oportunamente comentado (véase cap. 1), las predicciones se obtienen resolviendo las MME (*cf.* Henderson, 1984) resultantes del modelo utilizado para ajustar los datos, condicionalmente a los CVC estimados. La estimación de los CVC, por su parte, debe realizarse previamente a cada ejecución de la evaluación genética. En la práctica, sin embargo, una nueva estimación de CVC se lleva adelante sólo una vez que los archivos de datos se han incrementado lo suficiente. En todo caso, asúmase que la estimación de CVC se lleva a cabo mediante un algoritmo de GS en el marco de un análisis bayesiano como el descripto en el Capítulo 2. En este contexto, parece razonable utilizar los estadísticos posteriores obtenidos tras la ejecución anterior para especificar los correspondientes hiperparámetros en la subsiguiente, en el espíritu de un esquema de 'actualización bayesiana' (*Bayesian updating*). En esta sección se describe la aplicación de tal estrategia para estimar CVC aprovechando la flexibilidad que ofrece la distribución GIW para especificar el conocimiento a priori del analista. En particular, la estrategia fue utilizada para especificar la distribución a priori de los CVC del rodeo Angus de Las Lilas, y luego comparada contra otras especificaciones a priori más estándares. Además, fue puesta a prueba mediante un estudio de simulación estocástica.

3.2.3.1. Datos de campo

El archivo de datos de campo corresponde al archivo de pesos al destete del rodeo Angus de Las Lilas descripto en el capítulo precedente. La empresa lleva adelante una evaluación genética anual como una estrategia de comercialización de sus reproductores. Aún así, los CVC no son estimados con la misma frecuencia, si bien varias estimaciones se han llevado a cabo en la medida en que se acumulaban más datos. Imitando este esquema de trabajo se crearon dos subarchivos a partir de este archivo madre. El primero de ellos incluye los 4480 registros de los individuos nacidos hasta el año 1986. El se-

gundo, contiene los registros de 6290 animales tomados sobre individuos nacidos hasta el año 2000. Una descripción detallada de estos subarchivos se presenta en la Tabla 3.1.

Tabla 3.1. Características de los archivos de datos de peso al destete analizados.

	ANGUS			Simulado*
BASES de datos	**Archivo1**	**Archivo2**	**Total**	
Registros[1]	4480	6290	7229	4492
Pedigree	6080	8553	9936	5012
Conexiones del pedigree[2]	5,9	14,7	21,9	-
Padres				
Nº	119	199	264	65
% de padres con registro	7	16	20	69
Nº promedio de crías	38	32	27	69
Madres				
Nº	1608	2127	2444	1376
% de madres con registro	45	55	57	64
Nº promedio de crías	3	3	3	3

* Valores promedio sobre 39 réplicas. La variabilidad surge de asumir una tasa de fertilidad de 0,9.

[1] Pesos al destete tomados sobre individuos de unos 200 días de edad en promedio.

[2] Nº de elementos distintos a cero en la matriz $\boldsymbol{A}$ (en millones). No calculado para los datos simulados.

El objetivo fue estimar CVC vía el GS bajo diferentes especificaciones a priori. En general, todas las distribuciones fueron parameterizadas tal como se describiera en el Capítulo 2; *i.e.*, en términos de ciertos valores 'razonables' para los CVC, por un lado, y en términos de la incertidumbre asociada a estos valores, por otro. En particular, se utilizaron las estimaciones REML obtenidas mediante el paquete ASReml (Gilmour *et al*., 2006) para definir valores a priori verosímiles para los diferentes CVC [**REML**]. Luego, se llevó adelante una serie de análisis bayesianos vía el GS con diferentes estrategias respecto al grado de incertidumbre impuesto sobre estos valores.

En el primer análisis, se asumió que la incertidumbre a priori sobre los valores de los CVC era total [**NoINF**] y, en consecuencia, se ejecutó el algoritmo de estimación basado en la parameterización descripta en [3.31]. Recuérdese que en este caso no es necesario definir ningún hiperparámetro. En cambio, el algoritmo muestrea de las distribuciones condicionales posteriores de los parámetros de Bartlett que dependen únicamente de funciones de las formas cuadráticas en los datos. En tal escenario, las cadenas MCMC son completamente indiferentes a los valores iniciales una vez que han convergido y, por consiguiente, el 'método del acoplamiento de cadenas' (*cf*. García-Cortés *et al*., 1998) puede utilizarse en forma directa para establecer convergencia.

En segundo lugar, se consideró que las estimaciones REML constituyen un conocimiento a priori significativo y, en concordancia, se especificaron distribuciones Chi-cuadradas invertidas e IW a priori para todos los CVC. Más específicamente, los parámetros de escala fueron derivados luego de igualar las estimaciones REML de los CVC a las correspondientes medias a priori, de acuerdo a la parameterización descripta en el Capítulo 2. Luego, con el objetivo de evaluar la influencia de tal especificación a priori sobre los resultados, se definieron otros dos conjuntos de valores para las medias a priori: estos valores se correspondieron con las estimaciones REML $\pm$ 2 $\times$ EE. Por otro lado, dos valores diferentes de grados de credibilidad fueron asignados en cada análisis: 20

[**IW20**] y 100 [**IW100**]. Estos valores fueron particularmente escogidos para representar una incertidumbre moderada y baja, respectivamente, respecto a los conjuntos de valores especificados para las medias a priori. En la Tabla 3.2 se presentan las medias a priori y los grados de credibilidad utilizados en cada uno de los análisis que se llevó a cabo.

En una tercera estrategia, se examinó una especificación a priori 'experta' utilizando los estadísticos descriptivos posteriores de los CVC, obtenidos luego de ajustar los dos subarchivos de datos, para determinar los hiperparámetros de las distribuciones a priori en el análisis definitivo ([**GIW_S1**] y [**GIW_S2**], respectivamente). En particular, las estimaciones de los CVC para los subconjuntos de datos se obtuvieron mediante un análisis bayesiano basado en distribuciones a priori no informativas. Con estas estimaciones, luego, se derivaron los grados de credibilidad, υ_0 y υ_1, igualando las medias y varianzas marginales estimadas de Σ_{11} y Γ con las correspondientes medias y varianzas teóricas de distribuciones Chi-cuadradas invertidas; *i.e.*,

$$\begin{aligned}\upsilon_0 &= \frac{2\times\left[\hat{m}_i\left(\Sigma_{11}\right)\right]^2}{\hat{v}_i\left(\Sigma_{11}\right)} + 4, \\ \upsilon_1 &= \frac{2\times\left[\hat{m}_i\left(\Gamma\right)\right]^2}{\hat{v}_i\left(\Gamma\right)} + 4,\end{aligned} \qquad [3.33]$$

donde $\hat{m}_i(\cdot)$ y $\hat{v}_i(\cdot)$ denotan, respectivamente, la media y varianza marginal posterior de las correspondientes distribuciones de los parámetros de Bartlett, obtenidas tras ajustar el *i*-ésimo subarchivo de datos (i = 1, 2). A continuación, se utilizó la especificación a priori descripta en la ecuación [3.32] para definir los hiperparámetros S_0, S_1, τ_0 y H. Específicamente, los elementos de la matriz de escala $\mathbf{\Sigma}^*$ fueron calculados según

$$\begin{aligned}\Sigma^*_{11} &= \left(\upsilon_0 + 2\right)\times\hat{M}_i\left(\Sigma_{11}\right), \\ \Sigma^*_{12} &= \Sigma^*_{21} = \Sigma^*_{11}\times\hat{M}_i\left(\tau\right), \\ \Sigma^*_{22} &= \left(\Sigma^{*2}_{12}\Sigma^{*-1}_{11}\right) + \left(\upsilon_1 + 3\right)\times\hat{M}_i\left(\Gamma\right),\end{aligned} \qquad [3.34]$$

donde $\hat{M}_i(\cdot)$ representa los valores modales de las correspondientes distribuciones marginales posteriores de los parámetros de Bartlett, obtenidos tras ajustar el *i*-ésimo subarchivo de datos (i = 1, 2). Nótese que esta parameterización a priori implica interpretar los modos marginales posteriores como valores 'razonables' para los CVC genéticos.

Un último análisis fue ejecutado aplicando esta estrategia en forma recursiva; *i.e.*, definiendo las distribuciones a priori de los parámetros de Bartlett para el segundo subarchivo con los resultados del primero, y luego repitiendo el procedimiento para el archivo de datos completo [**GIW_S1|S2**]. Las medias a priori y los grados de credibilidad especificados para cada uno de estos análisis se presentan en la Tabla 3.2.

Tabla 3.2. Archivo Angus. Valores iniciales y grados de credibilidad utilizados para especificar la estructura de información a priori de los diferentes análisis.

	Parms. genéticos: medias a priori[1]			Grados de credibilidad[2]		
Análisis	h_o^2	h_m^2	r_G	υ	υ_0	υ_1
REML	-	-	-	-	-	-
NoINF	-	-	-	-	-	-
IW20_1	0,25	0,16	-0,69	20	-	-
IW20_2	0,20	0,12	-0,66	20	-	-
IW20_3	0,30	0,19	-0,75	20	-	-
IW100_1	0,25	0,16	-0,69	100	-	-
IW100_2	0,20	0,12	-0,66	100	-	-
IW100_3	0,30	0,19	-0,75	100	-	-
GIW_S1	0,25	0,18	-0,71	-	32	34
GIW_S2	0,21	0,14	-0,62	-	62	44
GIW_S2|S1	0,25	0,16	-0,67	-	105	85

[1] Las cifras en esta tabla fueron calculadas como funciones de las medias a priori especificadas para los diferentes CVC. Los tres conjuntos de valores en los análisis IW20 e IW100 corresponden a las estimaciones REML, REML – 2*EE y REML + 2*EE, respectivamente. Para los diferentes análisis basados en la distribución GIW, por su parte, las medias a priori fueron definidas como las modas marginales posteriores de los CVC obtenidos luego de ajustar los correspondientes subarchivos de datos, tal como se describe en el apartado 3.2.3.1.

[2] υ = grados de credibilidad de una distribución IW; υ_0 y υ_1 son los grados de credibilidad de las distribuciones Chi-cuadradas invertidas de los parámetros de Bartlett. Fueron derivados como se explica en el apartado 3.2.3.1.

Las líneas punteadas subdividen los diferentes análisis con relación al grado de incertidumbre supuesto para las medias a priori de los CVC. Las categorías son: "incertidumbre completa", "incertidumbre moderada", "baja incertidumbre" y "opinión a priori experta" en orden ascendente.

Se describen a continuación algunos detalles técnicos respecto a la implementación de los análisis. El GS utilizado en este trabajo fue similar al presentado en el Capítulo 2, con la particularidad de que el muestreo de la matriz de covarianza genética fue programado de acuerdo al algoritmo presentado en el Apéndice B de este trabajo. Para cada análisis, se ejecutó el programa y se generó una cadena de 100.000 ciclos. Luego, para el análisis **NoINF** se estableció la convergencia mediante el método del acoplamiento de cadenas (*cf*. García-Cortés *et al*., 1998), considerando una diferencia máxima entre cadenas de 10^{-3} para cada CVC, lo cual ocurrió en la iteración 7.606. Tomando una postura conservadora, entonces, se descartaron las 10.000 muestras iniciales como período de calentamiento. Los estadísticos descriptivos posteriores para todos los CVC y los correspondientes parámetros genéticos fueron calculados utilizando el programa POSTGIBBSF90 del paquete BLUPF90 (Mistzal *et al*., 2002). En particular, las medias posteriores, los desvíos estándares posteriores y las autocorrelaciones entre muestras para la heredabilidad directa, h_o^2, la heredabilidad materna, h_m^2, y la correlación genética directa-materna, r_G, fueron utilizados como criterio de comparación.

3.2.3.2. Estudio de simulación estocástica

Con el objeto de evaluar los procedimientos de estimación descriptos se llevó a cabo un estudio de simulación estocástica. A tal efecto, se simularon poblaciones cerradas con generaciones superpuestas bajo apareamiento aleatorio. En cada réplica, una población base no registrada de 500 vacas fue apareada al azar con 20 toros y produjeron la primera generación de progenie. Los valores fenotípicos de estos individuos fueron luego

muestreados de acuerdo al MAM clásico (ecuación [2.3]). En el paso siguiente, los toros y vacas más viejos fueron descartados de la población, de acuerdo a una tasa de reemplazo de 0,25 para los machos y 0,20 para las hembras. Sus reemplazos fueron seleccionados de la última generación utilizando las predicciones de los valores de cría directos como criterio de selección. Una segunda generación fue creada entonces apareando al azar toros y vacas de la generación parental recién creada con la única condición de que se evitaran apareamientos padre-hija y madre-hijo. Todo el procedimiento, por último, fue repetido hasta la décima generación. Cincuenta réplicas con esta estructura poblacional fueron generadas y analizadas en este estudio. Las principales características de la estructura poblacional, promediadas sobre réplicas, se incluyen en la Tabla 3.1, mientras que los valores de los parámetros genéticos utilizados en el proceso de generación de datos se presentan en la Tabla 3.4.

Todas las poblaciones simuladas fueron analizadas utilizando las mismas estrategias respecto a la estructura de información a priori de los CVC que las utilizadas para analizar el archivo de pesos al destete del rodeo Angus. Para cada réplica, el MAM fue ajustado y los CVC fueron estimados mediante los análisis **REML**, **NoINF** e **IW100**, como fuera descripto en el apartado anterior. Adicionalmente, se llevó a cabo un análisis **GIW** en dos etapas. Primero, los CVC fueron estimados para el subconjunto de datos hasta la octava generación a través de un análisis bayesiano jerárquico basado en una distribución a priori no informativa. Luego, los estadísticos posteriores fueron calculados y utilizados para determinar los hiperparámetros de las distribuciones a priori de los parámetros de Bartlett en un análisis que incluía ahora todos los datos simulados dentro de réplica. Para cada procedimiento de estimación bayesiano se obtuvieron 30.000 ciclos: los primeros 10.000 fueron descartados como período de calentamiento y el resto de ellos fue utilizado para calcular los estadísticos descriptivos posteriores. De la misma manera que con los datos de campo, se utilizaron las medias posteriores, los desvíos estándares posteriores y las autocorrelaciones para la heredabilidad directa, h_o^2, la heredabilidad materna, h_m^2, y la correlación genética directa-materna, r_G, promediados entre réplicas, para comparar los resultados.

Una digresión es necesaria en este punto. Once de las cincuenta réplicas exhibieron problemas de convergencia mientras se analizaban los correspondientes subconjuntos de datos. Básicamente, los valores de las modas marginales posteriores de los CVC genéticos, utilizados para inicializar los análisis con el conjunto completo de los datos, produjeron matrices de covarianza singulares y, en consecuencia, el GS abortó. Un análisis más profundo del problema, mostró que en tales casos el método del acoplamiento de cadenas fallaba en establecer la convergencia antes de los 10.000 ciclos, lo cual sugiere que hubiera sido necesario un mayor número de ciclos. Para no oscurecer las conclusiones, entonces, y dado que el número de ciclos por análisis estaba limitado al tiempo de ejecución, se decidió no incluir tales réplicas. En consecuencia, los resultados presentados en este trabajo fueron obtenidos promediando las 39 réplicas restantes.

3.3. RESULTADOS

En la Tabla 3.3 se presentan las estimaciones y errores estándares de los parámetros genéticos para los datos de peso al destete del rodeo Angus de Las Lilas. Con respecto a las estimaciones, nótese que si bien los resultados no fueron exactamente iguales entre todos los análisis, las cifras son bastante similares: las heredabilidades directa y materna estuvieron en el orden de 0,25 y 0,15, respectivamente, mientras que la estimación de la

correlación genética directa-materna fue de alrededor de –0,69. La mayor variabilidad respecto a los resultados se observó para los análisis **IW100**, principalmente debido a la fuerte influencia que ejerció la estructura de información a priori especificada. Por el contrario, las estimaciones de los análisis **IW20** exhibieron menor dispersión entre sí y, de hecho, fueron más cercanas a los valores obtenidos bajos los análisis **REML** y **NoINF**, independientemente de las medias a priori especificadas. Ahora, con respecto a los errores estándares, nótese que se ha puesto en evidencia un patrón consistente: en aquellos análisis en los que se ha supuesto a priori un menor grado de incertidumbre con respecto a las medias de los CVC especificados, los errores estándares obtenidos fueron menores en comparación con enfoques menos informativos. En particular, el análisis recursivo **GIW_S2|S1** mostró los menores errores estándares.

Tabla 3.3. Archivo Angus. Estimaciones y errores estándares de h_o^2, h_m^2 y r_G bajo las diferentes estrategias con respecto a la especificación a priori de los CVC.

	Parámetros genéticos					
	h_o^2		h_m^2		r_G	
Análisis*	**Estim.**	**EE**	**Estim.**	**EE**	**Estim.**	**EE**
REML	0,26	0,04	0,16	0,03	-0,69	0,07
NoINF	0,25	0,04	0,15	0,03	-0,69	0,08
IW20_1	0,25	0,04	0,15	0,02	-0,69	0,07
IW20_2	0,23	0,04	0,14	0,02	-0,69	0,07
IW20_3	0,26	0,04	0,16	0,02	-0,71	0,06
IW100_1	0,25	0,03	0,15	0,02	-0,69	0,04
IW100_2	0,21	0,03	0,12	0,02	-0,68	0,05
IW100_3	0,29	0,03	0,18	0,02	-0,73	0,04
GIW_S1	0,28	0,04	0,17	0,02	-0,71	0,05
GIW_S2	0,24	0,03	0,15	0,02	-0,66	0,05
GIW_S2|S1	0,27	0,02	0,16	0,02	-0,69	0,04

Referencias: Estim. = estimación REML o media posterior bayesiana; EE = error estándar aproximado (REML) o desvío estándar posterior bayesiano. Las líneas punteadas subdividen los diferentes análisis con relación al grado de incertidumbre supuesto para las medias a priori de los CVC. Las categorías son: "incertidumbre completa", "incertidumbre moderada", "baja incertidumbre" y "opinión a priori experta" en orden ascendente.

* Referir a la sección 3.2.3.1 y a la Tabla 3.2 para una descripción detallada de los análisis.

Por su parte, en la Tabla 3.4 se presentan las estimaciones y los correspondientes errores estándares para los parámetros genéticos obtenidos tras ajustar los datos simulados, promediados entre réplicas. En general, los resultados siguieron la misma tendencia que aquella descripta para los datos de campo. Tras ajustar el mismo modelo que se usara para generar los datos, los análisis **REML** y **NoINF** resultaron, en promedio, casi insesgados respecto a los verdaderos valores simulados. Es más, cuando las medias a priori de los CVC bajo el análisis **IW100** fueron especificadas con las estimaciones REML, la performance del procedimiento fue superior en términos del sesgo respecto a los verdaderos valores y en términos de los errores estándares. Por el contrario, cuando fueron especificadas con valores sobre-dispersos, las estimaciones resultaron en promedio desviadas de los verdaderos parámetros genéticos. En cambio, con el análisis **GIW** se obtuvieron estimaciones precisas, en promedio, mientras que los errores estándares se mantuvieron en valores relativamente pequeños.

Tabla 3.4. Datos simulados. Estimaciones y errores estándares de h_o^2, h_m^2 y r_G bajo las diferentes estrategias con respecto a la especificación a priori de los CVC.

	Parámetros genéticos					
	h_o^2 (Valor verdadero = 0,25)		h_m^2 (Valor verdadero = 0,15)		r_G (Valor verdadero = -0,70)	
Análisis*	**Estim.**	**EE**	**Estim.**	**EE**	**Estim.**	**EE**
REML	0,24 ± 0,04	0,05 ± 0,01	0,15 ± 0,03	0,04 ± 0,00	-0,69 ± 0,09	0,09 ± 0,02
NoINF	0,24 ± 0,05	0,04 ± 0,01	0,14 ± 0,04	0,03 ± 0,01	-0,72 ± 0,13	0,09 ± 0,03
IW100_1	0,24 ± 0,04	0,03 ± 0,00	0,15 ± 0,03	0,02 ± 0,00	-0,69 ± 0,09	0,04 ± 0,01
IW100_2	0,29 ± 0,05	0,03 ± 0,00	0,20 ± 0,03	0,02 ± 0,00	-0,72 ± 0,06	0,04 ± 0,01
IW100_3	0,17 ± 0,04	0,02 ± 0,00	0,08 ± 0,04	0,01 ± 0,00	-0,60 ± 0,16	0,06 ± 0,02
GIW	0,23 ± 0,04	0,03 ± 0,00	0,14 ± 0,04	0,02 ± 0,01	-0,70 ± 0,12	0,06 ± 0,02

Referencias: Estim. = estimación REML o media posterior bayesiana (promedio sobre 39 réplicas ± desvío estándar); EE = error estándar aproximado (REML) o desvío estándar posterior bayesiano (promedio sobre 39 réplicas ± desvío estándar). Las líneas punteadas subdividen los diferentes análisis con relación al grado de incertidumbre supuesto para las medias a priori de los CVC. Las categorías son: "incertidumbre completa", "incertidumbre moderada", "baja incertidumbre" y "opinión a priori experta" en orden ascendente.

* Los valores iniciales para cada réplica bajo los análisis IW100 corresponden a los estimaciones REML ± 2*EE. Referir a la sección 3.2.3.2 para una descripción detallada de los análisis GIW.

En la Figura 3.1 se presentan los gráficos de autocorrelaciones de los parámetros genéticos para los datos del rodeo Angus de Las Lilas. Para una mejor representación, se graficó sólo uno de los tres correlogramas correspondientes a los análisis **IW20** e **IW100**, dado que las curvas dentro de análisis fueron muy similares. Del mismo modo, sólo se presenta el correlograma del análisis **GIW_S2|S1**. Los análisis **IW100** y **GIW** mostraron mejores tasas de convergencia comparadas con los análisis **NoINF** e **IW20**.

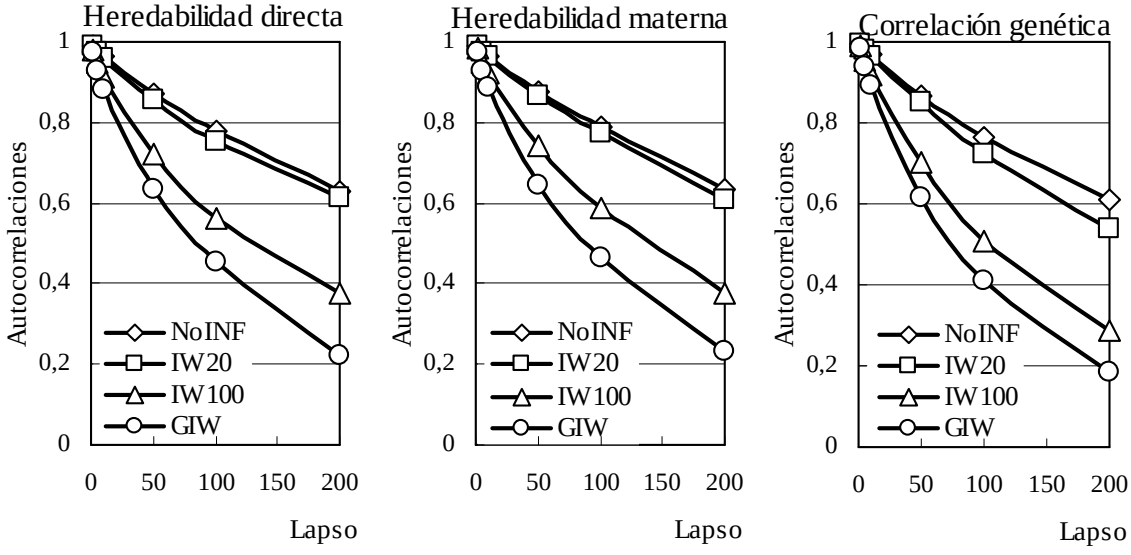


Figura 3.1. Archivo Angus. Correlogramas de los parámetros genéticos. El correlograma describe las autocorrelaciones entre muestras de un mismo parámetro en función del lapso entre muestras. Referir al texto para una descripción detallada de los análisis.

Por último, las autocorrelaciones lapso 10 y lapso 200 de los parámetros genéticos para los archivos de datos simulados se presentan en la Tabla 3.5. De nuevo, se observó

un mejor comportamiento de la cadena y, en consecuencia, una mejor tasa de convergencia cuando se asumió una menor incertidumbre en torno a las medias a priori especificadas para los CVC. De hecho, los análisis **IW100** y **GIW** mostraron la mejor tasa de convergencia. Sin embargo, aparecieron algunas diferencias en las autocorrelaciones para un lapso de 200 muestras entre ambos análisis. Estas diferencias fueron más importantes para la heredabilidad materna y la correlación genética directa-materna que para la heredabilidad directa. Al aspecto, vale la pena recordar que bajo el análisis **IW100** la matriz de covarianza genética está restringida en su conjunto por un único grado de credibilidad ($\upsilon = 100$), mientras que bajo el análisis **GIW** dos parámetros diferentes están involucrados (promediando entre réplicas $\upsilon_0 = 64$ y $\upsilon_1 = 19$, respectivamente).

Tabla 3.5. Datos simulados. Autocorrelaciones entre muestras de un mismo parámetro para h_o^2, h_m^2 y r_G para lapsos entre muestras de 10 y 200.

	Parámetros genéticos					
	h_o^2		h_m^2		r_G	
Análisis*	**Lapso10**	**Lapso200**	**Lapso10**	**Lapso200**	**Lapso10**	**Lapso200**
NoINF	0,92 ±0,06	0,48 ±0,14	0,95 ±0,06	0,62 ±0,14	0,96 ±0,02	0,55 ±0,17
IW100_1	0,83 ±0,03	0,12 ±0,07	0,84 ±0,02	0,12 ±0,08	0,83 ±0,02	0,09 ±0,06
IW100_2	0,82 ±0,03	0,11 ±0,06	0,82 ±0,02	0,10 ±0,08	0,80 ±0,02	0,07 ±0,05
IW100_3	0,83 ±0,03	0,15 ±0,09	0,87 ±0,02	0,20 ±0,12	0,85 ±0,02	0,15 ±0,07
GIW	0,85 ±0,03	0,19 ±0,10	0,91 ±0,04	0,33 ±0,16	0,90 ±0,04	0,29 ±0,23

Referencias: Las autocorrelaciones están promediadas sobre 39 réplicas ± desvío estándar. Las líneas punteadas subdividen los diferentes análisis con relación al grado de incertidumbre supuesto para las medias a priori de los CVC. Las categorías son: "incertidumbre completa", "incertidumbre moderada", "baja incertidumbre" y "opinión a priori experta" en orden ascendente.

* Los valores iniciales para cada réplica bajo los análisis IW100 corresponden a los estimaciones REML ± 2*EE. Referir a la sección 3.2.3.2 para una descripción detallada de los análisis GIW.

3.4. DISCUSIÓN

En este capítulo se introdujo la distribución Wishart invertida generalizada en toda su generalidad. La descripción se basó extensivamente en los trabajos de Brown (2002) y Le *et al*. (1999). En particular, se ha puesto énfasis en la flexibilidad que ofrece para describir el conocimiento a priori del analista respecto a la distribución de la matriz de covarianza genética en el contexto de un análisis bayesiano jerárquico. De todas maneras, es posible hallar nuevas aplicaciones siguiendo los argumentos aquí descriptos. Por ejemplo, una aplicación directa implicaría utilizar la distribución GIW como una alternativa natural para especificar la estructura de covarianza a priori de observaciones que siguen una distribución normal multivariada con un patrón monótono de datos faltantes (Garthwaite y Al-Awadhi, 2001). Tal aplicación en el contexto de un modelo animal multicarácter es el núcleo del algoritmo '*full conjugate Gibbs sampler*', una alternativa a los algoritmos de '*data augmentation*' con mejores tasas de convergencia (Cantet *et al*., 2004).

Adicionalmente, se han derivado resultados teóricos con respecto al uso de la GIW como la distribución a priori de la matriz de covarianza genética bajo el MAM. En particular, se ha demostrado que especificar una distribución GIW a priori constituye una alternativa conjugada y, en consecuencia, facilita la estimación de CVC mediante un algoritmo GS. De hecho, se probó que la especificación de una distribución IW puede considerarse un caso especial de la GIW, con base en un conjunto particular de hi-

perparámetros. Luego, es posible adaptar la distribución GIW o bien para representar incertidumbre diferencial entre distintos componentes de la matriz de covarianza, o bien para especificar una distribución a priori no informativa.

Ahora bien, proponer especificaciones a priori que representen apropiadamente la incertidumbre del analista es un problema bien diferente. En esta investigación se presentó una estrategia que surge de una práctica habitual en el proceso de ejecución de los programas de evaluación genética. La idea consiste en utilizar recursivamente estimaciones previas de los CVC para determinar los hiperparámetros en la siguiente ejecución. En particular, se ha propuesto derivar los grados de credibilidad de las distribuciones a priori de los parámetros de Bartlett igualando las medias y varianzas marginales posteriores de un subconjunto de los datos a sus correspondientes medias y varianzas teóricos. Así, se ha seguido un enfoque intuitivo de actualización bayesiana, explotando la "propiedad de 'memoria' del teorema de Bayes" (Gianola y Fernando, 1986).

La estrategia fue evaluada implementando un análisis bayesiano jerárquico a datos de campo y datos simulados, y luego comparada contra otras especificaciones a priori más estándares. En términos generales, la estrategia recursiva ha devuelto estimaciones puntuales precisas de los parámetros genéticos y errores estándares menores en comparación con especificaciones a priori menos informativas, al tiempo que mejoró las tasas de convergencia de las cadenas MCMC. Es necesario, de todas maneras, interpretar estos resultados con cierta precaución, dado que estas ventajas aparecieron asociadas a la especificación de los grados de credibilidad. De hecho, análisis basados en la distribución IW con un alto valor del hiperparámetro también produjeron menores errores estándares de estimación y una mejor tasa de convergencia. Sin embargo, cuando la media a priori bajo estos análisis fue especificada en valores sobredispersos, las estimaciones resultaron sesgadas con respecto a los verdaderos valores simulados. El riesgo de obtener estimaciones sesgadas al utilizar especificaciones a priori informativas es tratado en la revisión de Mistzal (2008).

En conclusión, se ha demostrado aquí que la incertidumbre diferencial respecto a los CVC genéticos en el contexto del MAM puede tenerse en cuenta asumiendo una distribución a priori GIW para la matriz de covarianza genética. Es más, la estimación de parámetros puede llevarse adelante utilizando un algoritmo de GS, dado que la distribución condicional posterior de la matriz de covarianza genética también pertenecerá a la familia GIW. En este trabajo se ha procurado modelar la incertidumbre diferencial especificando valores distintos para los grados de credibilidad de los parámetros de Bartlett. Teniendo en cuenta el conjunto mayor de hiperparámetros disponible, este enfoque puede resultar algo conservador. Sin embargo, casi no existe literatura respecto a este último punto (*e.g*. Garthwaite y Al-Awadhi, 2001).

La motivación con respecto al uso de la distribución GIW será más evidente en el próximo capítulo, en el que se extiende la formulación de MAM clásico para considerar asociaciones negativas de naturaleza ambiental entre efectos maternos. Como oportunamente se verá, en tal caso se tiene mayor certeza respecto al valor de la varianza aditiva directa, pero, en general, se quisiera imponer mayor incertidumbre respecto a los valores de la varianza aditiva materna y, en particular, de la correlación genética directa materna. Tal especificación diferencial de la incertidumbre no es posible al asumir una distribución IW a priori para la matriz de covarianza genética. Además, se verá que el tiempo de ejecución impone una restricción fuerte a la factibilidad del enfoque y, en general, es deseable mejorar las tasas de convergencia.

4
Estimación de la correlación ambiental madre–progenie bajo un modelo animal con efectos maternos[1]

[1] Munilla Leguizamón, S. y R. J. C. Cantet. 2010. Estimation of residual dam-offspring correlation for a maternal animal model through a Griddy Gibbs Sampler. En *9th World Congress on Genetics Applied to Livestock Production*, Leipzig, Alemania.

4.1. INTRODUCCIÓN

Bajo el modelo animal con efectos maternos (MAM) hasta aquí descripto, estimaciones muy negativas de la correlación genética directa-materna conducen a predicciones contrapuestas de los valores de cría directos y maternos. Así, animales con alto mérito predicho para la componente directa del carácter tienden a presentar valores de cría por debajo de la media para la componente materna, lo cual genera suspicacia y descrédito entre los criadores que utilizan las predicciones como una herramienta de selección. Tales estimaciones, sin embargo, son frecuentes en la literatura, aunque son tomadas más bien con escepticismo por los investigadores (Meyer, 1997). En general, se acepta que las estimaciones están sesgadas por una asociación negativa de naturaleza ambiental entre efectos maternos en generaciones adyacentes (*cf*. Koch, 1972; Baker, 1980), que el MAM clásico no contempla.

Para remediar este problema se han propuesto varias formulaciones alternativas del MAM clásico. Estas formulaciones básicamente incluyen: 1. regresar el fenotipo de la madre en el dato del individuo (*e.g*. Robinson, 1996); 2. ajustar por el efecto de abuela materna (*e.g*. Dodenhoff *et al*., 1998); 3. modelar la estructura de covarianza de los efectos ambientales maternos permanentes (*e.g*. Quintanilla *et al*., 1999); y 4. incluir una interacción aleatoria, típicamente, padre × grupo de contemporáneos (*e.g*. Gutiérrez *et al*., 2006). Una última alternativa, poco explorada aún, consiste en ajustar una covarianza entre los efectos ambientales maternos permanentes y el error del modelo para pares de observaciones madre–progenie (*cf*. Cantet, 1990; Koerhuis y Thompson, 1997). Esta alternativa, sin embargo, induce una estructura de covarianza del error no lineal en un parámetro de correlación 'ambiental' madre–progenie, que dificulta la estimación de los CVC. Para evitar este problema, Bijma (2006) sugirió ajustar una serie de tiempo de medias móviles a esta estructura de covarianza. Sin embargo, este procedimiento sólo es válido cuando las madres con registro fenotípico tienen una única cría.

El objetivo del presente capítulo, entonces, es desarrollar un procedimiento de estimación que sea aplicable en un contexto más amplio. En particular, se presentará una formulación alternativa del MAM que incluye un parámetro de correlación ambiental entre pares de observaciones madre–progenie, y se describirá luego un algoritmo de inferencia bayesiano para el parámetro, basado en un muestreo por grilla (GGS, por el inglés '*Griddy Gibbs sampler*') (*cf*. Ritter y Tanner, 1992). Finalmente, y a modo de ilustración, se presentará una estimación para el parámetro de correlación, obtenida tras ajustar el modelo a los datos de peso al destete del rodeo Angus de Las Lilas e implementar el correspondiente algoritmo de inferencia.

4.2. MÉTODOS

4.2.1. Descripción del modelo

4.2.1.1. Desarrollo teórico

Una vez más, considérese un carácter bajo la influencia de efectos maternos y defínase, en consecuencia, un MAM. Con el objetivo de enmarcar los desarrollos y la discusión que siguen considérese, en particular, el carácter peso al destete en bovinos de carne. El punto de partida será la covarianza entre el valor fenotípico de una madre W con el de su cría X. Las ecuaciones escalares de ambos individuos bajo el MAM (véase el Capítulo 2) son las siguientes:

$$\begin{aligned} y_W &= \boldsymbol{x}_W^{\mathrm{T}}\boldsymbol{b} + a_{oW} + a_{mU} + e_{mU} + e_{oW}, \\ y_X &= \boldsymbol{x}_X^{\mathrm{T}}\boldsymbol{b} + a_{oX} + a_{mW} + e_{mW} + e_{oX}, \end{aligned} \qquad [4.1]$$

donde *U* representa a la madre de *W*. Bajo el supuesto de que valores de cría y desvíos ambientales no están correlacionados, la covarianza entre los valores fenotípicos de ambos individuos será igual a:

$$\begin{aligned} Cov(y_W, y_X) &= Cov(a_{oW} + a_{mU}, a_{oX} + a_{mW}) + \\ &+ Cov(e_{mU} + e_{oW}, e_{mW} + e_{oX}). \end{aligned} \qquad [4.2]$$

En particular, tras desarrollar el último término de [4.2] se obtiene:

$$\begin{aligned} Cov(e_{mU} + e_{oW}, e_{mW} + e_{oX}) &= Cov(e_{mU}, e_{mW}) + \\ &+ Cov(e_{mU}, e_{oX}) + Cov(e_{oW}, e_{mW}) + Cov(e_{oW}, e_{oX}). \end{aligned} \qquad [4.3]$$

En este punto, asúmase que

$$Cov(e_{mU}, e_{mW}) = Cov(e_{mU}, e_{oX}) = Cov(e_{oW}, e_{oX}) = 0. \qquad [4.4]$$

Finalmente, denótese

$$Cov(e_{oW}, e_{mW}) \equiv \sigma_{e_o e_m}. \qquad [4.5]$$

La covarianza $\sigma_{e_o e_m}$ en [4.5] modela una asociación entre el desvío 'ambiental' de la observación de la madre y el efecto ambiental materno permanente de esta última, evaluado en la expresión del carácter en su cría. En otras palabras, esta covarianza indicaría que el ambiente que experimentó una hembra durante su crecimiento predestete podría impactar en forma permanente en el ambiente que ella –cuando madre– proveerá a sus crías. Según señala Bijma (2006), no hay razón para restringir esta correlación a casos especiales, como el síndrome de la ubre engrasada, sino que podría tratarse de un fenómeno más general. De hecho, dado que bajo la denominación de 'ambiente' en estos modelos se incluyen efectos génicos no aditivos, una acción pleiotrópica de naturaleza no aditiva podría generar asociación estadística. Nótese que, después de todo, la covarianza en [4.5] involucra al mismo conjunto de genes: el genotipo de la madre *W*.

Incluir la covarianza descripta en [4.5] modifica la expresión de la matriz de covarianza del modelo (expresión [2.7] en el Capítulo 2), dado que errores y efectos ambientales maternos permanentes estarán correlacionados entre pares de observaciones madre–progenie. Específicamente,

$$\begin{aligned} Cov(\boldsymbol{y}) &= \boldsymbol{Z}_o \boldsymbol{A} \boldsymbol{Z}_o^{\mathrm{T}} \sigma_{a_o}^2 + \left(\boldsymbol{Z}_o \boldsymbol{A} \boldsymbol{Z}_m^{\mathrm{T}} + \boldsymbol{Z}_m \boldsymbol{A} \boldsymbol{Z}_o^{\mathrm{T}}\right) \sigma_{a_o a_m} + \\ &+ \boldsymbol{Z}_m \boldsymbol{A} \boldsymbol{Z}_m^{\mathrm{T}} \sigma_{a_m}^2 + Cov\left(\boldsymbol{Z}_p \boldsymbol{e}_m, \boldsymbol{e}_o^{\mathrm{T}}\right) + Cov\left(\boldsymbol{e}_o, \boldsymbol{e}_m^{\mathrm{T}} \boldsymbol{Z}_p^{\mathrm{T}}\right) + \\ &+ \boldsymbol{Z}_p \boldsymbol{Z}_p^{\mathrm{T}} \sigma_{e_m}^2 + \boldsymbol{I} \sigma_{e_o}^2. \end{aligned} \qquad [4.6]$$

En particular,

$$\begin{aligned} &Cov\left(\boldsymbol{Z}_p\boldsymbol{e}_m, \boldsymbol{e}_o^{\mathrm{T}}\right) + Cov\left(\boldsymbol{e}_o, \boldsymbol{e}_m^{\mathrm{T}}\boldsymbol{Z}_p^{\mathrm{T}}\right) = \\ &= \boldsymbol{Z}_p Cov\left(\boldsymbol{e}_m, \boldsymbol{e}_o^{\mathrm{T}}\right) + Cov\left(\boldsymbol{e}_o, \boldsymbol{e}_m^{\mathrm{T}}\right)\boldsymbol{Z}_p^{\mathrm{T}} \\ &= \left(\boldsymbol{Z}_p\boldsymbol{C}_{mo} + \boldsymbol{C}_{mo}^{\mathrm{T}}\boldsymbol{Z}_p^{\mathrm{T}}\right)\sigma_{e_o e_m}, \end{aligned} \qquad [4.7]$$

con $Cov\left(\boldsymbol{e}_m, \boldsymbol{e}_o^{\mathrm{T}}\right) = \boldsymbol{C}_{mo}\sigma_{e_o e_m}$. Ahora, dado que $\boldsymbol{Z}_p\boldsymbol{C}_{mo} + \boldsymbol{C}_{mo}^{\mathrm{T}}\boldsymbol{Z}_p^{\mathrm{T}}$ es la matriz de incidencia de $\sigma_{e_o e_m}$ en la matriz de covarianza del vector de observaciones, se esperaría verificar que

$$\boldsymbol{Z}_p\boldsymbol{C}_{mo} + \boldsymbol{C}_{mo}^{\mathrm{T}}\boldsymbol{Z}_p^{\mathrm{T}} = \boldsymbol{S}, \qquad [4.8]$$

con $\boldsymbol{S}$ representando a una matriz simétrica con 1's en las posiciones donde confluyen la fila de una madre y las columnas de sus crías. Sin embargo, la solución del sistema [4.8] en los elementos de la matriz $\boldsymbol{C}_{mo}$ no es única, lo cual implica que, en última instancia, no será posible identificar el parámetro $\sigma_{e_o e_m}$ en forma inequívoca. En consecuencia, $\sigma_{e_o e_m}$ no es estimable. Alternativamente, defínase (Bijma, 2006)

$$\boldsymbol{R} = \boldsymbol{S}\rho + \boldsymbol{I}, \qquad [4.9]$$

donde $\rho = \sigma_{e_o e_m}\sigma_{e_o}^{-2}$ representa el parámetro de correlación ambiental entre pares observaciones madre–progenie. Entonces, es posible rescribir [4.6] según

$$\begin{aligned} Cov\left(\boldsymbol{y}\right) = &\boldsymbol{Z}_o\boldsymbol{A}\boldsymbol{Z}_o^{\mathrm{T}}\sigma_{a_o}^2 + \left(\boldsymbol{Z}_o\boldsymbol{A}\boldsymbol{Z}_m^{\mathrm{T}} + \boldsymbol{Z}_m\boldsymbol{A}\boldsymbol{Z}_o^{\mathrm{T}}\right)\sigma_{a_o a_m} + \\ &+ \boldsymbol{Z}_m\boldsymbol{A}\boldsymbol{Z}_m^{\mathrm{T}}\sigma_{a_m}^2 + \boldsymbol{Z}_p\boldsymbol{Z}_p^{\mathrm{T}}\sigma_{e_m}^2 + \boldsymbol{R}\sigma_{e_o}^2. \end{aligned} \qquad [4.10]$$

Bajo esta formulación, el parámetro ρ es estimable. Nótese, sin embargo, que el modelo presenta ahora una estructura de covarianza del error no lineal en ρ. Tal formulación dificulta la estimación de los CVC. En el caso particular en el que las madres con registro fenotípico tengan una única cría, entonces la matriz $\boldsymbol{R}$ presentará una estructura de covarianza Toeplitz y, en consecuencia, la estimación de ρ puede llevarse adelante ajustando una 'serie de tiempo de medias móviles' (*moving average time series*) a esta estructura de covarianza del error (*cf*. Bijma, 2006). No ocurre lo mismo cuando las madres con datos tienen más de una cría. En tal caso, es posible abordar el problema de la estimación del parámetro ρ mediante un método de inferencia bayesiano.

Antes de seguir, sin embargo, conviene detallar las diferencias en la formulación del modelo con respecto al MAM definido en el Capítulo 2. En este caso, se asumirá que

$$\boldsymbol{e}_o \mid \rho, \sigma_{e_o}^2 \sim NMV\left(\boldsymbol{0}, \boldsymbol{R}\sigma_{e_o}^2\right), \qquad [4.11]$$

donde

$$\boldsymbol{R}_{n\times n} = \left\{r_{ij}\right\}, \quad \text{con } r_{ij} = \begin{cases} 1 & \text{si } i = j \\ \rho & \text{si } (i,\ j) \text{ es un par madre - progenie} \\ 0 & \text{en caso contrario} \end{cases} \qquad [4.12]$$

Por su parte, las MME resultantes del modelo pueden escribirse sucintamente

$$\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{W}+\boldsymbol{E}^{-1}\right)\boldsymbol{\theta}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{y}, \qquad [4.13]$$

con

$$\boldsymbol{W}=\left(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{Z}_{p}\right) \qquad [4.14]$$

y

$$\boldsymbol{E}^{-1}=\begin{bmatrix}\boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \left(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{A}^{-1}\right)\sigma^{2}_{e_o} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_{d}\sigma^{2}_{e_o}\sigma^{-2}_{e_m}\end{bmatrix}. \qquad [4.15]$$

Nótese que, bajo esta formulación, el esfuerzo computacional necesario para construir el sistema de ecuaciones es potencialmente enorme, especialmente en archivos de datos con numerosos registros, dado que es necesario invertir la matriz $\boldsymbol{R}$ y computar luego las expresiones $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{W}$ y $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{y}$. Sin embargo, como se describe a continuación, es posible aligerar este problema considerando la particular estructura de la inversa de esta matriz cuando los registros se agrupan por familias maternas.

4.2.1.2. Familias maternas

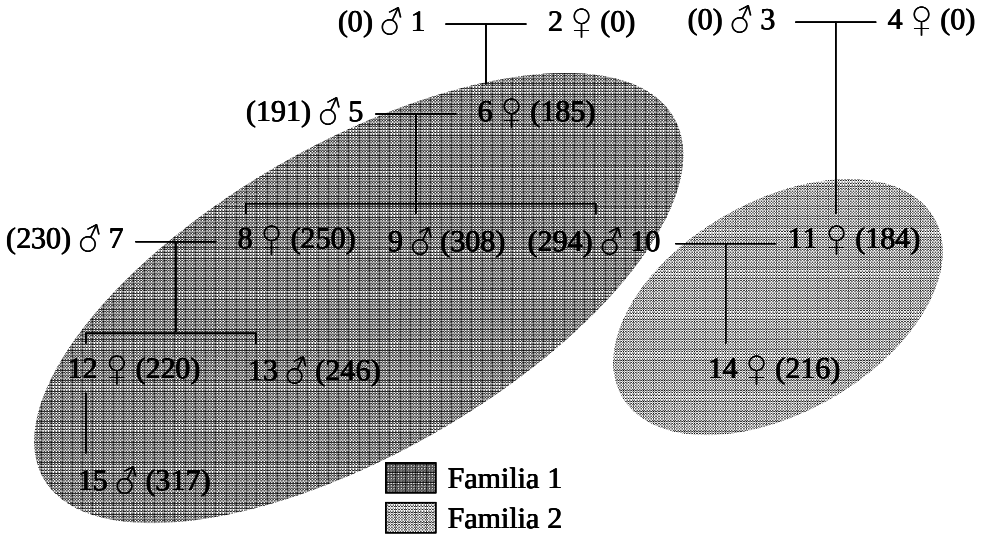
Una 'familia materna' es el conjunto de individuos definido por la primera madre con registro de performance dentro de una 'línea materna' (*female pathway*) junto toda su descendencia (Figura 4.1). Si se ordenan las observaciones por individuos dentro de familia materna, entonces es posible verificar que la matriz $\boldsymbol{R}^{-1}$ presenta una estructura diagonal en bloques (*i.e.*, $\boldsymbol{R}^{-1}=\oplus_{k=0}^{mf}\boldsymbol{R}_{k}^{-1}$), donde cada bloque $\boldsymbol{R}_{k}^{-1}$ (k = 1, …, mf) estará asociado a la k-ésima familia materna. El bloque $\boldsymbol{R}_{0}^{-1}$, por otro lado, reúne a aquellos individuos del conjunto de animales registrados que no pertenecen a ninguna familia materna; *i.e.*, todos los machos cuyas madres no hayan sido registradas y aquellas hembras sin descendencia cuyas madres no hayan sido registradas. Para estos últimos registros se asume una estructura de covarianza clásica; es decir, $\boldsymbol{R}_{0}^{-1}=\boldsymbol{I}$. En cambio, para una familia materna cualquiera, $\boldsymbol{R}_{k}^{-1}$ es completamente densa (Figura 4.1). Ahora bien, dado que en general las hembras tienen menos descendencia que los machos, las familias maternas no suelen ser muy grandes y, en consecuencia, invertir cada bloque en forma directa no implicaría una limitante computacional insalvable.

Por otro lado, para construir las MME en [4.13] es necesario computar también las expresiones $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{W}$ y $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{y}$. En este caso, la estructura en bloques de $\boldsymbol{R}^{-1}$ da lugar a un algoritmo de cómputo eficiente para estas expresiones. Dicho algoritmo se basa en definir submatrices $\boldsymbol{W}_{k}=\left\{w_{k(i,j)}\right\}$ (i = 1, …, f_k; j = 1, …, neq), donde f_k es el número de individuos con registro dentro de la k-ésima familia materna y neq representa el número de ecuaciones del sistema [4.13]. De acuerdo a esta definición, las matrices $\boldsymbol{W}_{k}$ son de orden ($f_k \times neq$) y contienen las filas de la matriz $\boldsymbol{W}$ asociadas a los registros de la k-ésima familia materna y al correspondiente bloque $\boldsymbol{R}_{k}^{-1}$. Esto implica que la ma-

triz $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{W}$, por ejemplo, puede computarse como la suma de (*mf* +1) matrices de orden *neq* × *neq*, de acuerdo a la expresión

$$\sum_{k=0}^{mf} \boldsymbol{W}_k^{\mathrm{T}}\boldsymbol{R}_k^{-1}\boldsymbol{W}_k. \qquad [4.16]$$

Esta formulación permite descomponer el problema de construir las MME en (*mf* +1) pasos, cada uno de los cuales involucra una de las familias maternas. En el Apéndice C se detallan los pasos necesarios para computar cada uno de los términos de [4.16], así como el vector $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{y}$, mediante contribuciones secuenciales.



ID	6	8	9	10	12	13	15	11	14	5	7
6	1,141	-0,249	-0,228	-0,228	0,052	0,050	-0,010	0,000	0,000	0,000	0,000
8	-0,249	1,143	0,050	0,050	-0,238	-0,229	0,048	0,000	0,000	0,000	0,000
9	-0,228	0,050	1,046	0,046	-0,010	-0,010	0,002	0,000	0,000	0,000	0,000
10	-0,228	0,050	0,046	1,046	-0,010	-0,010	0,002	0,000	0,000	0,000	0,000
12	0,052	-0,238	-0,010	-0,010	1,091	0,048	-0,218	0,000	0,000	0,000	0,000
13	0,050	-0,229	-0,010	-0,010	0,048	1,046	-0,010	0,000	0,000	0,000	0,000
15	-0,010	0,048	0,002	0,002	-0,218	-0,010	1,044	0,000	0,000	0,000	0,000
11	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,042	-0,208	0,000	0,000
14	0,000	0,000	0,000	0,000	0,000	0,000	0,000	-0,208	1,042	0,000	0,000
5	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000
7	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000

$\boldsymbol{R}^{-1}$ (*calculada con* $\rho = 0{,}2$ *y* $\sigma^2_{e_o} = 1{,}0$)

Figura 4.1. Ejemplo de familias maternas y estructura de la matriz $\boldsymbol{R}^{-1}$. (Arriba) El esquema representa una genealogía constituida por dos familias maternas (resaltadas en tono de sombras). Los individuos están numerados del 1 al 15, aunque sólo 11 de ellos presentan registro de performance, indicado entre paréntesis. (Debajo) Matriz $\boldsymbol{R}^{-1}$ calculada para $\rho = 0{,}2$ y varianza del error unitaria. Tras ordenar los individuos por familia materna, la matriz presenta una estructura diagonal en bloques. Nótese además que todas las submatrices dentro de bloque son densas, excepto para el último bloque, que corresponde a individuos que no pertenecen a ninguna familia materna.

4.2.1.3. Modelo 'operativo' alternativo

Antes de abordar el problema específico de la estimación del parámetro de correlación ρ, nótese que en el contexto de un método de inferencia MCMC, y bajo la formulación del modelo hasta aquí descripta, será necesario computar dentro de cada ciclo de muestreo $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{W}$ y $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{y}$, porque $\boldsymbol{R}^{-1}$ es función del ρ, cuyo valor se actualizará en cada iteración. La demanda en tiempo de cómputo de estas operaciones limitará severamente la utilidad del método cuando existen familias maternas con muchos individuos, dado que $\boldsymbol{R}_k^{-1}$ es una matriz completamente densa y genera muchas contribuciones a las MME. Considérese, en cambio, la siguiente formulación alternativa del modelo (Dr. Andrés Legarra, comunicación personal):

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{a} + \boldsymbol{Z}_p\boldsymbol{e}_m + \boldsymbol{e}_o + \boldsymbol{\varepsilon}, \qquad [4.17]$$

donde, a diferencia del MAM definido por la ecuación [2.3] (véase el Capítulo 2), $\boldsymbol{e}_o$ ($n \times 1$) es un vector aleatorio de desvíos ambientales directos, y ε es un vector de errores aleatorios, tal que $\boldsymbol{\varepsilon}\,|\,\sigma_\varepsilon^2 \sim NMV\left(\boldsymbol{0}, \boldsymbol{I}\sigma_\varepsilon^2\right)$, con σ_ε^2 fijo y de pequeña magnitud. Nótese que [4.17] da lugar al siguiente sistema de ecuaciones:

$$\left(\sigma_\varepsilon^{-2}\tilde{\boldsymbol{W}}^{\mathrm{T}}\tilde{\boldsymbol{W}} + \tilde{\boldsymbol{E}}^{-1}\right)\tilde{\boldsymbol{\theta}} = \sigma_\varepsilon^{-2}\tilde{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{y}, \qquad [4.18]$$

donde, ahora,

$$\tilde{\boldsymbol{W}} = \left[\begin{array}{c|c|c|c} \boldsymbol{X} & \boldsymbol{Z} & \boldsymbol{Z}_p & \boldsymbol{I}_n \end{array}\right],\ \tilde{\boldsymbol{\theta}} = \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{a} \\ \boldsymbol{e}_m \\ \boldsymbol{e}_o \end{bmatrix} \text{ y } \tilde{\boldsymbol{E}}^{-1} = \left[\begin{array}{c|c|c|c} \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \hline \boldsymbol{0} & \left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{A}^{-1}\right) & \boldsymbol{0} & \boldsymbol{0} \\ \hline \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_d\sigma_{e_m}^{-2} & \boldsymbol{0} \\ \hline \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{R}^{-1}\sigma_{e_o}^{-2} \end{array}\right]. \quad [4.19].$$

Bajo esta formulación alternativa, tanto la matriz $\sigma_\varepsilon^{-2}\tilde{\boldsymbol{W}}^{\mathrm{T}}\tilde{\boldsymbol{W}}$ como el vector $\sigma_\varepsilon^{-2}\tilde{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{y}$ son fijos de iteración a iteración y, en consecuencia, pueden computarse y almacenarse sólo durante el primer ciclo de un procedimiento MCMC. Esto permitirá acelerar considerablemente el tiempo de ejecución.

4.2.2. Estimación de ρ

4.2.2.1. Análisis bayesiano jerárquico

Considérese ahora abordar el problema de la estimación de parámetro ρ y otros CVC mediante un método de inferencia bayesiano, en el contexto de una construcción jerárquica del modelo [4.17]. Siguiendo los pasos descriptos en el Capítulo 2, es posible derivar todas las distribuciones condicionales posteriores de las incógnitas del modelo. Puede verificarse, de hecho, que todas las distribuciones condicionales pertenecen a familias conocidas, excepto por la distribución del parámetro de correlación. En ese caso, asumiendo para ρ una distribución a priori Uniforme acotada al intervalo [–1, +1], puede verificarse que la distribución condicional posterior será proporcional a:

$$p\left(\rho \,|\, \tilde{\boldsymbol{\theta}}, \boldsymbol{\Sigma}, \sigma_{e_m}^2, \sigma_{e_o}^2, \boldsymbol{y}\right) \sim \left|\boldsymbol{R}\right|^{-1/2} exp\left\{-\frac{\boldsymbol{e}_o^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{e}_o}{2\sigma_{e_o}^2}\right\}. \qquad [4.20]$$

En [4.20] el parámetro ρ está dentro de las matrices $\boldsymbol{R}$ y $\boldsymbol{R}^{-1}$. Como función de ρ, entonces, la densidad no pertenece a una familia conocida, lo cual imposibilita su muestreo directo y, en consecuencia, dificulta la implementación de un algoritmo GS. En este punto se introducirá un algoritmo muestreo para la distribución [4.20], basado en evaluar la función para una grilla de valores y conocido en la literatura como '*Griddy Gibbs sampler*' (GGS) (*cf.* Ritter y Tanner, 1992). El GGS se incorporará luego a la secuencia de etapas del GS al momento de muestrear ρ.

4.2.2.2. El algoritmo GGS

De acuerdo a su concepción original, el GGS aplica cuando es difícil muestrear de una distribución condicional posterior univariada en el contexto de la implementación de un algoritmo de GS (Ritter y Tanner, 1992). El GGS se basa en formar una aproximación de la inversa de la función de densidad acumulada (*cdf*, por el inglés '*cumulative density function*') mediante la evaluación de la distribución condicional posterior del parámetro en una grilla de valores. Una de las ventajas del algoritmo es que no es necesario conocer la forma analítica exacta de la distribución condicional posterior, sino sólo el 'núcleo' (*kernel*, en inglés) correspondiente. Una vez que la *cdf* inversa fue aproximada, el GGS procede tomando una muestra de una distribución Uniforme y, finalmente, interpolando el valor muestreado entre los valores de la grilla más cercanos.

En el caso específico del parámetro de correlación ambiental madre–progenie, ρ, el algoritmo GGS involucra los siguientes pasos:

1. Para el *j*-ésimo punto de la grilla, 'grid(*j*)', evaluar el núcleo de la distribución condicional posterior [4.20].
2. Computar la sumatoria de los valores calculados. El valor obtenido representará la constante de integración.
3. Para el *j*-ésimo punto de la grilla, normalizar el valor grid(*j*).
4. Aproximar la distribución posterior acumulada *cdf*(*j*) de ρ sumando progresivamente los valores normalizados.
5. Muestrear una uniforme [mín(*cdf*), máx (*cdf*)], '*uni*'.
6. Obtener un índice '*idx*' tal que $cdf(idx) < uni < cdf(idx + 1)$.
7. Actualizar finalmente el valor de ρ mediante una interpolación lineal entre grid(*idx*) y grid(*idx* + 1).

Los puntos en la grilla deben cubrir todo el espacio paramétrico. Sin embargo, el número total de valores de la grilla y los intervalos entre puntos dependerán de la factibilidad de implementar el algoritmo, como se ilustra a continuación.

4.2.3. Implementación del algoritmo de inferencia a datos de peso al destete

En esta sección se ilustra la implementación del algoritmo GGS con el objeto de estimar el parámetro de correlación ambiental madre–progenie para el archivo de datos de peso al destete del rodeo Angus de Las Lilas. En primer lugar, se presentan detalles técnicos relacionados con la programación y ejecución exitosa del programa de estimación. En segundo lugar, se describen las dos implementaciones diferentes del muestreo de Gibbs que se llevaron adelante. El archivo de datos utilizado fue descripto en el Capítulo 2 de

este trabajo, y puede referirse a la Tabla 2.1 para más detalles. El archivo presenta un total de 533 familias maternas.

4.2.3.1. Programación del GGS

Para llevar a cabo la estimación se adaptó el código del GS descripto en el Capítulo 2. Básicamente, se introdujeron dos modificaciones importantes. En primer lugar, se programó una serie de sentencias específicas para generar las contribuciones a las MME definidas en [4.18]. En particular, se incorporó una subrutina interna para reordenar los registros por familia materna antes de su lectura, un paso necesario para computar la matriz $\boldsymbol{R}^{-1}$ aprovechando la estructura diagonal en bloques (*i.e.*, $\boldsymbol{R}^{-1} = \oplus_{k=0}^{mf} \boldsymbol{R}_k^{-1}$). La inversión de las matrices $\boldsymbol{R}_k$, por su parte, se obtiene en forma directa a través de una subrutina de Cholesky. En segundo lugar, se incorporó dentro de la subrutina interna que realiza un ciclo completo del muestreo de Gibbs el bloque de sentencias para muestrear ρ, de acuerdo al algoritmo GGS presentado en la sección precedente.

Durante las primeras ejecuciones de prueba, sin embargo, se presentaron algunos inconvenientes asociados a la representación de los números en la computadora, que llevaron a la implementación una serie de ajustes importantes al programa. El primer problema apareció al momento de evaluar el núcleo de la distribución condicional posterior de ρ en los valores de la grilla. De [4.20] se desprende que la función a evaluar involucra el exponencial de la forma cuadrática $\boldsymbol{e}_o^{\mathrm{T}} \boldsymbol{R}^{-1} \boldsymbol{e}_o$, una operación que devuelve un número tan pequeño que no tiene representación en la máquina. Para salvar este inconveniente, entonces, se evaluó alternativamente el logaritmo de la función. Nótese, sin embargo, que para obtener la constante de integración en el segundo paso del GGS es necesario calcular una suma de logaritmos de números muy pequeños, una operación nada trivial. Para conseguirlo, se utilizó el algoritmo descripto por Primeaux (2005). Aún así, el procedimiento fallaba porque algunos puntos del soporte de ρ tenían valores muy pequeños de probabilidad. Para resolver este inconveniente, entonces, se adoptó un 'grilla adaptativa' (*adaptive grid*) (*cf.* Ritter y Tanner, 1992), un procedimiento que involucra los siguientes pasos. Primero, se define una grilla extendida de puntos sobre el espacio paramétrico de ρ, espaciada en intervalos de 0,05. Luego, se evalúa el logaritmo de la función para todos los puntos de esta grilla extendida, y se selecciona aquel valor que maximiza la función. Centrada en este valor, finalmente, se expande una grilla local espaciada en intervalos de 0,01. Con estos ajustes, entonces, fue posible ejecutar la versión definitiva del programa exitosamente. Los códigos completos del programa no serán incluidos en el documento, pero pueden ser solicitados al autor.

La performance del algoritmo GGS, por otro lado, se evaluó del siguiente modo. Como fuera descripto al presentar los pasos del algoritmo, el muestreo de la distribución condicional posterior de ρ requiere que se aproxime la correspondiente función de densidad acumulada. Para el éxito del procedimiento se esperaría que los puntos de la grilla local cubran la mayor parte de la densidad de la *cdf* en cada ciclo de muestreo, de modo que la interpolación sea lo más precisa posible. Como una medida de performance del procedimiento, entonces, se evaluó el porcentaje de grillas locales en los primeros 1.000 ciclos del algoritmo que cubrieron al menos el 0,95 de la densidad de probabilidad de la *cdf*.

4.2.3.2. Implementación del muestreo de Gibbs

Como fuera oportunamente discutido, construir las MME en cada ciclo del GS de acuerdo al procedimiento presentado en la Sección 4.2.1.2 de este capítulo y descripto en detalle en el Apéndice C involucra un esfuerzo computacional importante, que limita sensiblemente el número de ciclos de muestreo posibles en un intervalo razonable de tiempo. Por tal motivo, en un primer análisis se utilizó la implementación del GS sugerida por Gelman y Rubin (1992) (véase el Capítulo 2) para estimar el parámetro de correlación ambiental madre–progenie. Así, se obtuvieron tres cadenas de 5.000 ciclos cada una, inicializadas en valores dispersos del soporte de la distribución. Este número de ciclos, en particular, se decidió con base en los resultados del test de Raftery y Lewis (1992) ejecutado a través del paquete BOA (Smith, 2007) sobre una prueba piloto de 10.000 iteraciones. Este estudio preliminar sugirió un período de calentamiento de 1400 ciclos. Luego, se obtuvo una única cadena 35.000 ciclos y, a partir de los muestreos obtenidos, se llevaron adelante tests de convergencia a través del paquete BOA (Smith, 2007) y se computaron los estadísticos descriptivos posteriores de todos los CVC mediante el programa POSTGIBBSF90 del paquete BLUPF90 (Misztal *et al*., 2002).

Posteriormente se llevó a cabo un segundo análisis, luego de programar el modelo ‘operativo’ alternativo definido en [4.17]. En este segundo caso, el tiempo de cómputo sólo está limitado por la implementación del paso GGS dentro del muestro de Gibbs y, en consecuencia, es posible obtener un mayor número de ciclos por unidad de tiempo. Por otro lado, y de acuerdo al marco teórico de trabajo, se asumió que la incorporación del parámetro de correlación ρ al modelo afectaría principalmente el valor de los CVC maternos, pero no el de la varianza aditiva directa. En consecuencia, y aprovechando la flexibilidad que ofrece la distribución GIW para especificar en forma diferencial la incertidumbre a priori (véase el Capítulo 3), se decidió asignar un alto grado de credibilidad a la media a priori de la varianza aditiva directa, definida con base en la correspondiente estimación REML, mientras que, por el contrario, se especificó mayor incertidumbre respecto a las medias a priori de los CVC maternos. Luego, se generó una cadena MCMC de 100.000 ciclos y se descartaron los primeros 10.000 como período de calentamiento. Finalmente, se computaron los estadísticos descriptivos posteriores de ρ y de todos los CVC mediante el programa POSTGIBBSF90 del paquete BLUPF90 (Misztal *et al*., 2002). Con los resultados de este segundo muestreo, por último, se estimó la distribución marginal posterior de ρ a través de un método no paramétrico basado en un núcleo Gaussiano (Silverman, 1986).

4.3. RESULTADOS

La ejecución de los programas con el archivo de datos del rodeo Angus de Las Lilas fue llevada a cabo con éxito, si bien la demanda computacional constituyó una restricción importante. Por ejemplo, el tiempo de cómputo del programa que construye las MME en cada ciclo de muestreo fue de alrededor de un minuto por ciclo en una computadora personal con procesador Pentium® 4 (CPU 3.6GHz, 3.11 GB de RAM). En cambio, la versión definitiva del programa, basada en el modelo ‘operativo’ alternativo, demandó 7 segundos por ciclo en la misma computadora, es decir, 10 veces menos tiempo. En este punto, es importante mencionar que ambos programas devolvieron resultados muy similares cuando fueron inicializados exactamente con las mismas especificaciones. De todas maneras, la performance del procedimiento en lo que respecta al tiempo de cómputo aún está lejos de los 10 ciclos por segundo que demanda el GS bajo el MAM clásico.

Por otro lado, la performance del procedimiento GGS en lo que respecta a la precisión con la que se muestrearon los valores del parámetro de correlación fue muy buena: 92% de las grillas locales construidas durantes los primeros 1.000 ciclos del GS cubrieron más del 95% de la densidad de probabilidad de la correspondiente *cdf*, asegurando así una interpolación bastante precisa. Es importante mencionar que el espacio paramétrico cubierto por la grilla extendida no abarcó todo el soporte teórico de la distribución del parámetro, sino que éste fue restringido al intervalo (–0,3; +0,3). Durante las ejecuciones de puesta a punto de los programas, se verificó que con valores por fuera de este rango las matrices $\boldsymbol{R}_k$ para algunas de las familias maternas son singulares y, en consecuencia, es imposible obtener sus inversas.

En la Tabla 4.1 se presentan los estadísticos descriptivos posteriores, los ESS y las autocorrelaciones de todos los CVC, incluyendo el parámetro de correlación ambiental madre–progenie, ρ, obtenidos a partir de una cadena MCMC de 35.000 ciclos. Las medias y modos posteriores de todos los CVC fueron similares a los obtenidos bajo el MAM clásico, en tanto que la media posterior de ρ fue de 0,05, un resultado consistente bajo cualquier valor de inicialización (Figura 4.2). Por otro lado, los tamaños efectivos de muestra (ESS) para los CVC genéticos indican que el número de ciclos obtenido fue algo acotado. Nótese que las autocorrelaciones entre muestras fueron muy altas para lapsos de hasta 200 muestreos. De hecho, si bien las secuencias de muestreos de todos los CVC pasaron los tests de convergencia de cadena simple que ofrece el paquete BOA (Smith, 2007), los resultados fueron poco convincentes.

Tabla 4.1. Estadísticos descriptivos posteriores de los CVC obtenidos para una cadena MCMC de 35.000 ciclos*.

	CVC[1]					
	$\sigma^2_{e_o}$	ρ	$\sigma^2_{e_m}$	$\sigma^2_{a_o}$	$\sigma_{a_o a_m}$	$\sigma^2_{a_m}$
υ	20	-	100	20	20	20
S	450	-0,20	93	190	-103	116
Media	453,02	0,05	94,88	186,48	-107,13	117,01
Modo	451,51	0,05	95,22	191,28	-98,93	116,65
DS	18,31	0,03	9,11	28,96	19,47	18,35
IADP95	(417, 488)	(-0,01, 0,12)	(77, 113)	(134, 244)	(-145, -71)	(84, 157)
ESS	102	318	330	71	57	66
Autocorr.						
1	0,833	0,781	0,915	0,992	0,991	0,992
5	0,710	0,408	0,685	0,974	0,975	0,974
10	0,674	0,285	0,512	0,955	0,958	0,954
50	0,571	0,163	0,206	0,834	0,852	0,826
100	0,488	0,125	0,151	0,707	0,747	0,699
200	0,354	0,079	0,093	0,513	0,592	0,525

*35.000 ciclos. 1400 ciclos descartados como período de calentamiento.

Refs.: υ = grados de credibilidad a priori; *S* = parámetro de escala a priori; DS = desvío estándar; IADP95 = intervalo de alta densidad posterior del 95%; ESS = tamaño efectivo de muestra; Autocorr. = autocorrelaciones entre muestras (lapso).

[1] Componentes de (co)varianza: $\sigma^2_{e_o}$ = varianza del error; ρ = correlación ambiental madre–progenie; $\sigma^2_{e_m}$ = varianza de los efectos ambientales maternos permanentes; $\sigma^2_{a_o}$ = varianza aditiva directa; $\sigma^2_{a_m}$ = varianza aditiva materna; $\sigma_{a_o a_m}$ = covarianza genética directa-materna.

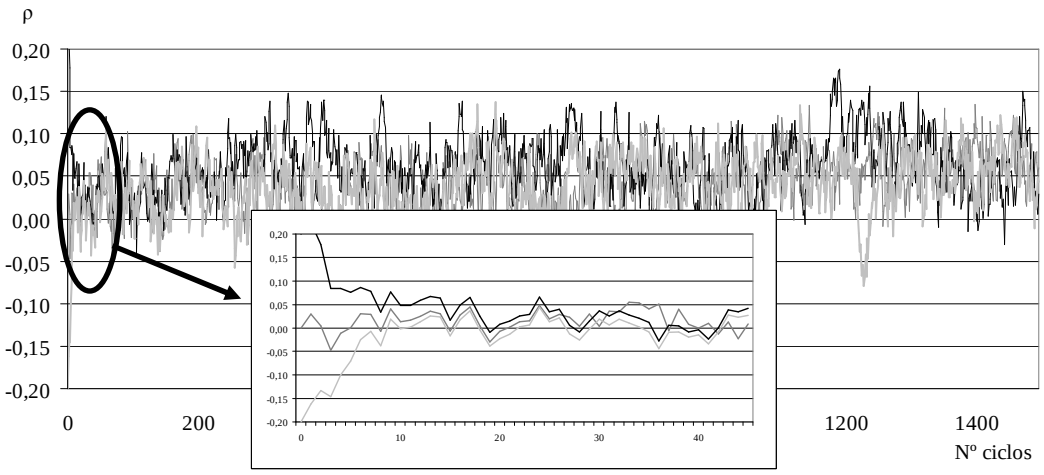


Figura 4.2. Gráfica de muestreos en función del número de ciclos para el parámetro de correlación. La gráfica corresponde a tres cadenas MCMC inicializadas en puntos dispersos del soporte de la distribución marginal. Nótese cómo ya en los primeros ciclos de muestreo el valor de parámetro se sitúa en torno al cero.

En la Tabla 4.2, por su parte, se presentan los estadísticos descriptivos posteriores, los ESS y las autocorrelaciones para todos los CVC, obtenidos ahora a partir de la cadena MCMC de 90.000 ciclos y bajo una especificación de la incertidumbre a priori diferente para los CVC genéticos. Las medias posteriores de todos los CVC fueron muy similares a las obtenidas bajo el primer análisis, mientras que los desvíos estándares posteriores resultaron, en términos generales, bastante menores. Por otro lado, los ESS fueron mayores con respecto a los obtenidos en el análisis previo, un resultado que responde al menor grado de incertidumbre especificado para la varianza aditiva directa y al mayor número de muestreos. Por último, nótese que las autocorrelaciones entre muestras fueron diferentes para los distintos CVC genéticos en conexión con el grado de incertidumbre a priori especificado a través de la distribución GIW.

Finalmente, en la Figura 4.3 se presenta la distribución marginal posterior estimada de ρ. La distribución resultó claramente unimodal y simétrica con una media de 0,05 (± 0,03) y un intervalo de alta densidad posterior del 95% entre (–0,01; +0,11), lo cual refleja que, para estos datos, el modelo pone una enorme masa sobre valores positivos, aunque pequeños, del parámetro.

4.4. DISCUSIÓN

En este capítulo se presentó en detalle una formulación alternativa del MAM 'clásico' (Willham, 1963, Quaas y Pollak, 1980), que incluye un parámetro de correlación ambiental entre pares de observaciones madre–progenie. El modelo se construye sobre la idea de que el ambiente que experimentó una hembra durante etapas tempranas de su desarrollo podría impactar en forma permanente en el ambiente que ella proveerá sus crías a futuro. Si bien esta idea es de larga trayectoria (*e.g.* Koch, 1972; Cantet *et al.*, 1988; Koerhuis y Thompson, 1997), en rigor, la formalización de este modelo se debe a Bijma (2006). En particular, en este trabajo se demostró que la formulación del modelo deriva de plantear una covarianza entre el desvío 'ambiental' de la observación de una madre y su efecto ambiental materno permanente, evaluado en la ecuación de su cría.

Tabla 4.2. Estadísticos descriptivos posteriores de los CVC obtenidos para una cadena MCMC de 100.000 ciclos bajo una especificación a priori diferencial para los CVC genéticos*.

	CVC[1]					
	$\sigma^2_{e_o}$	ρ	$\sigma^2_{e_m}$	$\sigma^2_{a_o}$	$\sigma_{a_o a_m}$	$\sigma^2_{a_m}$
υ	100	-	20	1000	20	20
S	450	0,05	93	190	-103	116
Media	444,69	0,05	93,93	191,24	-104,14	116,81
Modo	444,62	0,06	93,90	190,45	-103,39	117,84
DS	11,63	0,03	10,51	8,16	5,41	10,65
IADP95	(422, 468)	(-0,01, 0,11)	(73, 114)	(175, 207)	(-115, -94)	(96, 138)
ESS	514	404	150	373	473	67
Autocorr.						
1	0,602	0,755	0,934	0,899	0,896	0,977
5	0,567	0,731	0,921	0,839	0,802	0,941
10	0,537	0,703	0,905	0,796	0,737	0,918
50	0,372	0,536	0,807	0,577	0,488	0,827
100	0,267	0,396	0,711	0,409	0,349	0,749
200	0,159	0,226	0,550	0,215	0,192	0,612

*100.000 ciclos obtenidos bajo el modelo 'operativo' alternativo, con σ^2_ε = 5. Los 10.000 ciclos iniciales fueron descartados como período de calentamiento.

Refs.: υ = grados de credibilidad a priori; *S* = parámetro de escala a priori; DS = desvío estándar; IADP95 = intervalo de alta densidad posterior del 95%; ESS = tamaño efectivo de muestra; Autocorr. = autocorrelaciones entre muestras (lapso).

[1] Componentes de (co)varianza: $\sigma^2_{e_o}$ = varianza del error; ρ = correlación ambiental madre–progenie; $\sigma^2_{e_m}$ = varianza de los efectos ambientales maternos permanentes; $\sigma^2_{a_o}$ = varianza aditiva directa; $\sigma^2_{a_m}$ = varianza aditiva materna; $\sigma_{a_o a_m}$ = covarianza genética directa-materna.

La formulación del modelo da lugar a una estructura de covarianza del error no lineal en el parámetro de correlación, ρ, lo cual, a priori, sugiere un esfuerzo computacional enorme para construir las correspondientes MME, dado que ahora será necesario invertir la matriz de covarianza del error. Sin embargo, aquí se mostró que el problema puede aliviarse considerablemente en virtud de la estructura diagonal en bloques de la inversa de dicha matriz, que surge cuando los registros se ordenan por familias maternas; *i.e.*, por grupos de individuos emparentados por línea materna junto a toda su descendencia. Adicionalmente, se presentó un algoritmo de cómputo del sistema basado en una serie de contribuciones secuenciales al sistema, siguiendo los trabajos de Groeneveld y Kovac (1990) y Misztal (2006). De todas maneras, si existen familias maternas numerosas en el archivo de datos, la factibilidad del modelo aún podría verse comprometida, porque la inversa de cada bloque familiar es completamente densa y, en consecuencia, genera demasiadas contribuciones a la matriz de coeficientes.

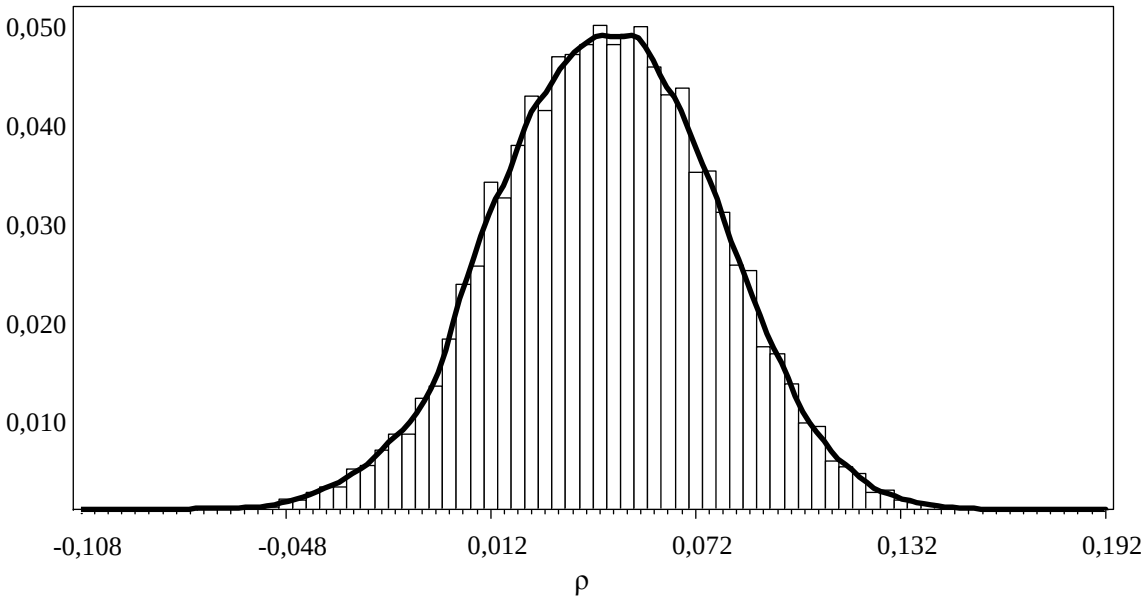


Figura 4.3. Distribución marginal estimada de ρ. El histograma se obtuvo a partir de los 90.000 ciclos remanentes luego de descartar los muestreos correspondientes al período de calentamiento. Por su parte, la curva superpuesta se obtuvo mediante de un método no paramétrico basado en un núcleo Gaussiano (Silverman, 1986).

Este último problema limita severamente la posibilidad de implementar un método MCMC con el objetivo de estimar los CVC para archivos con numerosos datos. En ese caso, dado que las matrices de covarianza del error de las familias maternas son función de ρ y este parámetro se actualiza en cada ciclo del procedimiento, las MME deberán construirse en cada iteración. Para salvar este inconveniente, en este trabajo se presentó un modelo operativo alternativo, basado en descomponer el vector de errores del MAM en un vector aleatorio de desvíos ambientales directos, por un lado, y un vector de errores aleatorios, con varianza fija y de pequeña magnitud, por otro. Bajo este modelo alternativo es posible construir y almacenar las MME durante el primer ciclo del procedimiento MCMC y, en consecuencia, acelerar considerablemente el tiempo de ejecución. Un ejemplo de esta estrategia, aplicado en el contexto de modelos de regresión aleatoria para curvas de crecimiento, puede consultarse en Nobre *et al*. (2003).

Por otro lado, la estructura de covarianza del error no lineal en ρ dificulta también la estimación de CVC, dado que la distribución de los errores pertenecerá ahora a una familia normal 'curva' (*cf*. Lehmann, 1983, pág. 45). Si bien Bijma (2006) ajustó una serie de tiempo de medias móviles a esta estructura de covarianza del error para obtener estimaciones del parámetro, este procedimiento sólo es válido cuando las madres tienen una única cría. Como alternativa, aquí se propuso abordar el problema de la estimación en un contexto más general, mediante un método de inferencia bayesiano bajo una construcción jerárquica del modelo. En este marco, fue posible derivar el núcleo de la distribución condicional posterior de ρ, una distribución que no pertenece a ninguna familia conocida. En este punto, entonces, se propuso muestrear de dicha distribución mediante el algoritmo GGS (Ritter y Tanner, 1992) ya dentro del muestreo de Gibbs. El GGS es un algoritmo de muestreo de una distribución univariada basado en formar una aproximación de la función de densidad acumulada mediante la evaluación del núcleo de la

distribución en una grilla de valores. Una vez que la *cdf* inversa fue aproximada, el algoritmo procede interpolando un desvío Uniforme entre los valores de la grilla más cercanos.

En esta investigación se implementó por primera vez un GGS para abordar un problema de estimación de CVC para datos de performance en especies ganaderas, en el contexto de un análisis bayesiano jerárquico. En particular, y con el objetivo de salvar ciertas limitaciones numéricas del procedimiento, se programó e implementó un GGS basado en una grilla adaptativa, que expande una grilla de puntos más estrecha en torno a aquel valor del espacio paramétrico que maximiza el logaritmo de la función de densidad. En rigor, aunque en un contexto y con una metodología bien diferentes, Iwaisaki *et al*. (2005) obtuvieron estimaciones REML del parámetro de correlación entre efectos ambientales maternos permanentes (*cf*. Quintanilla *et al*., 1999) mediante un método también basado en una búsqueda a partir de una grilla de valores (‘*grid search*’). En aquel estudio, sin embargo, los autores utilizaron una grilla mucho más acotada de valores, con sólo cuatro puntos distanciados a intervalos de 0,1. El GGS, en cambio, mostró una interpolación mucho más precisa, con puntos distanciados a 0,01. Por otro lado, la demanda computacional constituyó una restricción muy importante del procedimiento. En el contexto de la estimación del parámetro de correlación ambiental madre–progenie, el GGS mostró ser un algoritmo MCMC muy demandante en tiempo de cómputo, básicamente porque requiere invertir en cada ciclo de muestreo y en forma directa las matrices de covarianza del error de los bloques de familias maternas para cada punto de las grillas extendida y local. En consecuencia, es posible que el método sea demasiado restrictivo para un archivo de datos numeroso. En tal caso, será necesario recurrir a algún método MCMC alternativo para obtener muestras de la distribución condicional posterior de ρ. Algunos de estos métodos pueden consultarse en el libro de Gilks *et al*. (1996).

De todas maneras, el procedimiento fue implementado con éxito con los datos del rodeo Angus de Las Lilas, y se obtuvo por primera vez una estimación con datos de campo del parámetro de correlación ambiental madre–progenie para el carácter peso al destete, de acuerdo al modelo de Bijma (2006). La distribución marginal posterior estimada de ρ resultó unimodal, con una media cercana al cero, lo cual implica que el modelo puso una enorme masa sobre valores positivos y pequeños del parámetro. Consistentemente, las estimaciones de los otros CVC fueron similares a aquellas obtenidas bajo el MAM ‘clásico’ (véase el Capítulo 2). Para este conjunto de datos en particular, este resultado contradijo la expectativa original de que covarianzas de naturaleza ambiental entre pares de observaciones madre–progenie podrían sesgar la fuerte correlación genética estimada entre efectos directos y maternos. Eaglen y Bijma (2009) también estimaron una correlación ambiental madre–progenie cercana a cero al ajustar datos de facilidad de parto en ganado lechero mediante un modelo de serie de tiempo de medias móviles. Pero, hasta donde se sabe, no existen otras estimaciones del parámetro de correlación en la literatura.

5
Equivalencia entre modelos animales multirraciales y análisis bayesiano jerárquico para caracteres bajo la influencia de efectos maternos[1]

[1] Munilla Leguizamón, S. y R. J. C. Cantet. 2010. Equivalence of multibreed animal models and hierarchical Bayes analysis for maternally influenced traits. *Genet. Sel. Evol.*, 42(20): 1–12.

5.1. INTRODUCCIÓN

Los modelos animales multirraciales (MBAM) son modelos lineales mixtos que se utilizan para ajustar datos fenotípicos tomados sobre animales con diferente composición genética de origen racial. Bajo estos modelos, consideraciones tanto teóricas (Lo *et al*., 1993; Cantet y Fernando, 1995) como empíricas (Birchmeier *et al*., 2002; Cardoso y Tempelman, 2004) indican que la estructura de covarianza genética apropiada es heterogénea. Sin embargo, si bien la teoría ha sido establecida hace tiempo (Elzo y Famula, 1985; Elzo, 1990; Lo *et al*., 1993) y se han presentado métodos de inferencia clásica (*e.g*. Birchmeier *et al*., 2002; Elzo, 1994) y bayesiana (*e.g*. Cardoso y Tempelman, 2004), trabajos muy recientes de estimación de CVC en poblaciones multirraciales (*e.g*. Vergara *et al*., 2009a y 2009b) no tienen en cuenta esta particular estructura de dispersión, posiblemente porque no existe software de uso general disponible para ajustarla (García-Cortés y Toro, 2006).

En términos generales, la estimación de CVC en poblaciones multirraciales no es una tarea sencilla (*e.g*. Elzo y Wakeman, 1998; Birchmeier *et al*., 2002; Cardoso y Tempelman, 2004). Esencialmente, la dificultad radica en que los CVC escalares no pueden factorizarse de la inversa de la matriz de covarianza genética. En el marco de un análisis bayesiano jerárquico, por ejemplo, esto implica que la distribución condicional posterior de los CVC genéticos no tiene forma analítica estándar y, en consecuencia, no pueden emplearse algoritmos de fácil implementación como el GS en la estimación.

En este contexto, el enfoque basado en la descomposición de la matriz de covarianza genética en sus componentes por origen racial (García-Cortés y Toro, 2006) provee fórmulas más sencillas para la estimación de los CVC, fáciles de asimilar con la batería de técnicas de estimación disponibles en software de uso general. García-Cortés y Toro (2006) ilustraron la validez de su modelo empíricamente a través de un pequeño ejemplo numérico, pero no presentaron una derivación formal de la equivalencia con respecto al modelo formalizado por Cantet y Fernando (1995) utilizando los argumentos de la teoría genética cuantitativa de Lo *et al*. (1993), al menos en lo que respecta a la predicción de los valores de cría.

En este capítulo se abordará el problema. Básicamente, se derivará la equivalencia entre ambos modelos mediante una formulación algo distinta a la de García-Cortés y Toro (2006). Luego, se extenderá el modelo para incluir efectos maternos y se describirá la implementación de un análisis bayesiano jerárquico con el objetivo de estimar los CVC. Finalmente, se presentarán resultados de la implementación del análisis multirracial a datos de peso al destete de un cruzamiento experimental Angus × Hereford.

5.2. MÉTODOS

5.2.1. Equivalencia entre modelos animales multirraciales

Por simplicidad, considérese una población compuesta de dos razas *A* y *B*, que incluye individuos de las razas parentales y varios grupos raciales producto de diferentes cruzamientos entre sí. El carácter de interés está gobernado por un gran número de loci no ligados, que se encuentran en equilibrio gamético en los individuos de las razas parentales que dieron origen a la población. Asumiendo herencia aditiva, el valor genotípico del individuo *i* perteneciente a cualquier grupo racial en la población puede describirse según

$$G_i = \mu + \sum_{t=1}^{n} \left(\alpha_{S_i^t} + \alpha_{D_i^t} \right), \qquad [5.1]$$

donde μ es el valor genotípico esperado en el grupo racial de referencia, y $\alpha_{S_i^t}$, $\alpha_{D_i^t}$ representan los efectos aditivos de los alelos paterno y materno, respectivamente, que recibió el individuo *i* en el locus *t*. En este contexto, Lo *et al*. (1993) obtuvieron la expresión de la varianza genética aditiva como una función lineal de las varianzas aditivas de cada población contribuyente y un componente adicional asociado a la segregación de alelos con frecuencias génicas diferentes entre las poblaciones: la 'varianza de segregación' (*cf*. Wright, 1968; Lande, 1981). Para el caso de dos razas

$$Var(G_i) = f_A^i \sigma_{aA}^2 + f_B^i \sigma_{aB}^2 + 2\left(f_A^S f_B^S + f_A^D f_B^D \right) \sigma_{aS}^2 + \tfrac{1}{2} Cov(G_S, G_D), \qquad [5.2]$$

donde f_A^i y f_B^i representan, respectivamente, la proporción esperada de genes de las razas *A* y *B* en el individuo *i*, σ_{aA}^2 y σ_{aB}^2 las correspondientes varianzas genéticas aditivas de cada raza, y σ_{aS}^2 la varianza de segregación. El último término en [5.2] representa la covarianza entre los valores genotípicos de los padres del individuo, y puede desarrollarse expandiendo a la generación anterior. Bajo esta formulación, Lo *et al*. (1993) explicaron cómo calcular la matriz de covarianza genética mediante el método tabular (Emik y Terril, 1949) y su inversa en forma eficiente de acuerdo a los algoritmos propuestos por Henderson (1976) y Quaas (1988). Luego, Cantet y Fernando (1995) postularon un modelo animal que permite aplicar esta teoría para obtener predicciones BLUP de los valores de cría en el marco de una evaluación genética.

Alternativamente, García-Cortés y Toro (2006) desarrollaron la matriz de covarianza genética en sus componentes por origen racial. Para el caso de una población compuesta de dos razas puede verificarse que

$$\boldsymbol{G} = \boldsymbol{A}_A \sigma_{aA}^2 + \boldsymbol{A}_B \sigma_{aB}^2 + \boldsymbol{A}_S \sigma_{aS}^2, \qquad [5.3]$$

donde $\boldsymbol{A}_X$, $X = \{A, B, S\}$, son matrices de relaciones aditivas 'parciales' definidas de acuerdo a la fuente de variabilidad (García-Cortés y Toro, 2006). Estas matrices tienen dimensión $q \times q$ (donde q es el número de individuos) para asegurar que la suma sea conformable, pero presentan filas y columnas de ceros si existen individuos que no contribuyen a la componente correspondiente (por ejemplo, los individuos puros de la raza *A* no contribuyen a las componentes *B* y *S*) y, en consecuencia, son singulares. Esta formulación de la matriz de covarianza de los valores de cría se corresponde con un modelo animal convencional con varios factores aleatorios: los 'valores de cría por origen racial', $\boldsymbol{a}_X$, $X = \{A, B, S\}$. Es importante aclarar que bajo este modelo alternativo los valores de cría de individuos que no contribuyen a una componente en particular se definen fijos e iguales a cero, y se denominan 'nulos por origen racial'.

El modelo alternativo de García-Cortés y Toro (2006) facilita la estimación de CVC mediante el GS en el contexto de un análisis bayesiano jerárquico. Además, es equivalente al modelo de Cantet y Fernando (1995) en lo que respecta a su estructura de covarianza (véase la definición de modelos equivalentes de Henderson, 1985), dado que en su formulación ambos modelos son idénticos. Sin embargo, la equivalencia en lo que respecta a la predicción de los valores de cría no es obvia, porque la matriz de coeficientes de las ecuaciones del modelo mixto es singular y es necesario eliminar las ecuacio-

nes correspondientes a individuos con contribuciones nulas para resolverla y obtener resultados equivalentes (García-Cortés y Toro, 2006).

La propuesta, en cambio, es redefinir los vectores $\boldsymbol{a}_X$ de modo que sólo incluyan los q_X valores de cría no nulos por origen racial. Esto implica definir matrices de incidencia $\boldsymbol{Z}_X$ propias para cada componente, y rescribir la ecuación del modelo como

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Z}_A\boldsymbol{a}_A^* + \boldsymbol{Z}_B\boldsymbol{a}_B^* + \boldsymbol{Z}_S\boldsymbol{a}_S^* + \boldsymbol{e}, \qquad [5.4]$$

donde $\boldsymbol{Z}_X$ son matrices de incidencia de orden $n \times q_X$ de los q_X valores de cría no nulos por origen, $\boldsymbol{a}_X^*$, $X = \{A, B, S\}$. Nótese que esta formulación no incluye valores de cría forzados a cero cuando un individuo no contribuye a determinada componente, de modo que $\mathrm{Cov}(\boldsymbol{a}_X^*) = \boldsymbol{A}_X^*\sigma_{aX}^2$, donde $\boldsymbol{A}_X^*$, la submatriz de $\boldsymbol{A}_X$ sin las filas y columnas de ceros, es no singular. Defínase luego la matriz $\boldsymbol{M}_X$, de orden $q \times q_X$, tal que

$$\boldsymbol{Z}_X = \boldsymbol{Z}\boldsymbol{M}_X, \qquad [5.5]$$

donde $\boldsymbol{Z}$ es la matriz de incidencia de los efectos aleatorios en los modelos de García-Cortés y Toro (2006), y Cantet y Fernando (1995). Puede verificarse entonces que el producto matricial $\boldsymbol{M}_X\boldsymbol{A}_X^*$ recupera el patrón de filas nulas respecto a la matriz $\boldsymbol{A}_X$, y una subsiguiente posmultiplicación por $\boldsymbol{M}_X^{\mathrm{T}}$, por su parte, recupera el patrón de columnas nulas, de modo que

$$\boldsymbol{M}_X\boldsymbol{A}_X^*\boldsymbol{M}_X^{\mathrm{T}} = \boldsymbol{A}_X. \qquad [5.6]$$

Utilizando entonces [5.6] y [5.5] en [5.4] se obtiene

$$\begin{aligned} Cov(\boldsymbol{y}) &= \boldsymbol{Z}\left(\boldsymbol{M}_A\boldsymbol{A}_A^*\boldsymbol{M}_A^{\mathrm{T}}\sigma_{aA}^2 + \boldsymbol{M}_B\boldsymbol{A}_B^*\boldsymbol{M}_B^{\mathrm{T}}\sigma_{aB}^2 + \boldsymbol{M}_S\boldsymbol{A}_S^*\boldsymbol{M}_S^{\mathrm{T}}\sigma_{aS}^2\right)\boldsymbol{Z}^{\mathrm{T}} + \boldsymbol{R} \\ &= \boldsymbol{Z}\left(\boldsymbol{A}_A\sigma_{aA}^2 + \boldsymbol{A}_B\sigma_{aB}^2 + \boldsymbol{A}_S\sigma_{aS}^2\right)\boldsymbol{Z}^{\mathrm{T}} + \boldsymbol{R} \\ &= \boldsymbol{ZGZ}^{\mathrm{T}} + \boldsymbol{R} \\ &\equiv \boldsymbol{V}. \end{aligned} \qquad [5.7]$$

El modelo [5.4], en consecuencia, es equivalente al de Cantet y Fernando (1995) de acuerdo a la definición de Henderson (1985). Asimismo, dado que las componentes por origen racial del valor de cría no están correlacionadas entre sí, el BLUP de cada una de ellas puede escribirse (Henderson, 1950):

$$\begin{aligned} \mathrm{BLUP}(\boldsymbol{a}_X^*) &= E(\boldsymbol{a}_X^* \mid \boldsymbol{y}) \\ &= Cov(\boldsymbol{a}_X^*, \boldsymbol{y}^{\mathrm{T}})\left[Cov(\boldsymbol{y})\right]^{-1}\left[\boldsymbol{y} - E(\boldsymbol{y})\right] \\ &= Cov(\boldsymbol{a}_X^*, \boldsymbol{a}_X^{*\mathrm{T}})\boldsymbol{V}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{b}}\right) \\ &= \sigma_{aX}^2\,\boldsymbol{A}_X^*\,\boldsymbol{Z}_X^{\mathrm{T}}\boldsymbol{V}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{b}}\right). \end{aligned} \qquad [5.8]$$

La expresión [5.6] permite verificar que la suma de los $\mathrm{BLUP}(\boldsymbol{a}_X^*) = \hat{\boldsymbol{a}}_X^*$ sobre cada componente, premultiplicados por las correspondientes matrices $\boldsymbol{M}_X$ para asegurar que la suma sea conformable, es igual a

$$\begin{aligned}\sum_X \boldsymbol{M}_X \hat{\boldsymbol{a}}_X^* &= \sum_X \boldsymbol{M}_X \left[\left(\sigma_{aX}^2 \boldsymbol{A}_X^* \boldsymbol{Z}_X^{\mathrm{T}} \right) \boldsymbol{V}^{-1} \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{b}} \right) \right] \\ &= \sum_X \sigma_{aX}^2 \left(\boldsymbol{M}_X \boldsymbol{A}_X^* \boldsymbol{M}_X^{\mathrm{T}} \right) \boldsymbol{Z}^{\mathrm{T}} \left[\boldsymbol{V}^{-1} \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{b}} \right) \right] \\ &= \left(\sum_X \sigma_{aX}^2 \boldsymbol{A}_X \right) \left[\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{V}^{-1} \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{b}} \right) \right] \\ &= \boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{V}^{-1} \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{b}} \right) \\ &= \hat{\boldsymbol{a}},\end{aligned} \qquad [5.9]$$

donde $\hat{\boldsymbol{a}} = \mathrm{BLUP}(\boldsymbol{a})$, de acuerdo al modelo de Cantet y Fernando (1995). Nótese que si bien por simplicidad asumimos una población compuesta de dos razas en la derivación, la generalización de este argumento a *p* razas no presenta mayores inconvenientes.

5.2.2. Análisis bayesiano jerárquico para un MBAM con efectos maternos

Considérese ahora que el carácter de interés está bajo la influencia de efectos maternos y asúmase, en consecuencia, la estructura de covarianza descripta por Willham (1963). Adicionalmente, considérese la extensión de la teoría de Lo *et al*. (1993) a caracteres correlacionados presentada por Cantet y Fernando (1995). De acuerdo al enfoque alternativo presentado en la sección precedente, entonces, defínase el modelo

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \sum_{X=\{A,B,S\}} \left(\boldsymbol{Z}_{oX} \boldsymbol{a}_{oX}^* + \boldsymbol{Z}_{mX} \boldsymbol{a}_{mX}^* \right) + \boldsymbol{Z}_p \boldsymbol{e}_m + \boldsymbol{e}_o, \qquad [5.10]$$

donde $\boldsymbol{y}$ ($n \times 1$) es un vector de registros fenotípicos y $\boldsymbol{X}$ ($n \times p$) es la matriz de incidencia del vector de efectos fijos $\boldsymbol{b}$ ($p \times 1$), que incluye los efectos de raza. Sin perder generalidad, asúmase además que $\boldsymbol{X}$ es de rango completo. Por su parte, $\boldsymbol{a}_{oX}^*$ y $\boldsymbol{a}_{mX}^*$ son vectores aleatorios cuyos elementos corresponden a los q_X valores de cría directos y maternos no nulos por origen racial, respectivamente, y $\boldsymbol{e}_m$ ($d \times 1$) es el vector aleatorio que contiene los *d* efectos ambientales maternos permanentes. $\boldsymbol{Z}_{oX}$, $\boldsymbol{Z}_{mX}$ y $\boldsymbol{Z}_p$ representan las correspondientes matrices de incidencia. Finalmente, $\boldsymbol{e}_o$ ($n \times 1$) es el vector de errores. Para simplificar la notación, defínase $\boldsymbol{Z}_X = \left(\boldsymbol{Z}_{oX}, \boldsymbol{Z}_{mX} \right)$ y $\boldsymbol{a}_X^{*\mathrm{T}} = \left(\boldsymbol{a}_{oX}^{*\mathrm{T}}, \boldsymbol{a}_{mX}^{*\mathrm{T}} \right)$.

Considérese ahora la implementación de un análisis bayesiano jerárquico de este modelo, como el descripto en el Capítulo 2. En esta sección se extenderá aquella descripción para acomodar un MBAM con efectos maternos. En la primera etapa, entonces, es necesario especificar la distribución condicional conjunta de las observaciones. Asúmase un proceso normal multivariado

$$\boldsymbol{y} \mid \boldsymbol{b}, \boldsymbol{a}_A^*, \boldsymbol{a}_B^*, \boldsymbol{a}_S^*, \boldsymbol{e}_m, \sigma_{e_o}^2 \sim N\left(\boldsymbol{X}\boldsymbol{b} + \sum_{X=\{A,B,S\}} \boldsymbol{Z}_X \boldsymbol{a}_X^* + \boldsymbol{Z}_p \boldsymbol{e}_m, \boldsymbol{I}_n \sigma_{e_o}^2 \right). \qquad [5.11]$$

Luego, es necesario especificar la distribución a priori de los parámetros de posición; *i.e.*, $\boldsymbol{b}$, $\boldsymbol{a}_X^*$, $X = \{A,B,S\}$, y $\boldsymbol{e}_m$. En este punto se adoptarán las mismas distribuciones a priori que aquellas empleadas para el MAM, definidas por las expresiones [2.10] a [2.12] del Capítulo 2. Es decir, se asumirá todos los parámetros de posición siguen, a priori, distribuciones normales multivariadas. En particular, la distribución de los vectores de los valores de cría por origen racial, $\boldsymbol{a}_X^*$, será

$$\boldsymbol{a}_X^* \mid \boldsymbol{A}_X^*, \boldsymbol{\Sigma}_X \sim NMV\left(\boldsymbol{0}, \boldsymbol{\Sigma}_X \otimes \boldsymbol{A}_X^*\right) \text{con } \boldsymbol{\Sigma}_X = \begin{bmatrix} \sigma^2_{a_o X} & \sigma_{a_o a_m X} \\ \sigma_{a_o a_m X} & \sigma^2_{a_m X} \end{bmatrix}, \qquad [5.12]$$

para $X = \{A, B, S\}$, de acuerdo a la teoría genética clásica. En [5.12], $\boldsymbol{A}_X^*$ representa la matriz de relaciones aditivas parciales definida por García-Cortés y Toro (2006), pero sin las filas y columnas de ceros.

En el siguiente nivel de la jerarquía es necesario especificar las distribuciones a priori de los parámetros de dispersión. Bajo el modelo [5.10] éstos serán los escalares $\sigma^2_{e_o}$ y $\sigma^2_{e_m}$, y las matrices de covarianza genéticas $\boldsymbol{\Sigma}_X$, $X = \{A, B, S\}$. En este punto, se asumirán, al igual que en el caso del MAM, distribuciones conjugadas Gamma invertidas: Chi-cuadradas invertidas para los escalares y Wishart invertidas para las matrices.

Con estas especificaciones ya es posible obtener la distribución posterior conjunta. Asumiendo que $\boldsymbol{b}$, $\boldsymbol{a}_X^* \mid \boldsymbol{\Sigma}_X$, $\boldsymbol{\Sigma}_X$, $X = \{A, B, S\}$, $\boldsymbol{e}_m \mid \sigma^2_{e_m}$, $\sigma^2_{e_m}$ y $\sigma^2_{e_o}$ son todos independientes a priori,

$$\begin{aligned} & p\left(\boldsymbol{b}, \boldsymbol{a}_X^*, \boldsymbol{\Sigma}_X, \boldsymbol{e}_m, \sigma^2_{e_m}, \sigma^2_{e_o} \mid \boldsymbol{y}\right) \propto \\ & \quad \propto p\left(\boldsymbol{y} \mid \boldsymbol{b}, \boldsymbol{a}_X^*, \boldsymbol{e}_m, \sigma^2_{e_o}\right) \times p\left(\boldsymbol{b} \mid \boldsymbol{K}\right) \times \\ & \quad \times \prod_{X=\{A,B,S\}} \left[p\left(\boldsymbol{a}_X^* \mid \boldsymbol{A}_X^*, \boldsymbol{\Sigma}_X\right) \times p\left(\boldsymbol{\Sigma}_X \mid \upsilon_X, \boldsymbol{S}_X\right) \right] \times \\ & \quad \times p\left(\boldsymbol{e}_m \mid \sigma^2_{e_m}\right) \times p\left(\sigma^2_{e_m} \mid \upsilon_{e_m}, S^2_{e_m}\right) \times p\left(\sigma^2_{e_o} \mid \upsilon_{e_o}, S^2_{e_o}\right). \end{aligned} \qquad [5.13]$$

Explícitamente y tras agrupar factores afines

$$\begin{aligned} & p\left(\boldsymbol{b}, \boldsymbol{a}_X^*, \boldsymbol{e}_m, \boldsymbol{\Sigma}_X, \sigma^2_{e_m}, \sigma^2_{e_o} \mid \boldsymbol{y}\right) \propto \\ & \quad \propto exp\left\{-\left(\tfrac{1}{2}\right)\boldsymbol{b}^{\mathrm{T}} \boldsymbol{K}^{-1} \boldsymbol{b}\right\} \times \\ & \quad \times \left(\sigma^2_{e_o}\right)^{-\frac{1}{2}\left(\upsilon_{e_o}+n+2\right)} exp\left\{-\frac{\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e} - \upsilon_{e_o} S^2_{e_o}}{2\sigma^2_{e_o}}\right\} \times exp\left\{-\left(\tfrac{1}{2}\right)\boldsymbol{b}^{\mathrm{T}} \boldsymbol{K}^{-1} \boldsymbol{b}\right\} \times \\ & \quad \times \prod_{X=\{A,B,S\}} \left[\left|\boldsymbol{\Sigma}_X\right|^{-\frac{1}{2}\left(q_X+\upsilon_X+3\right)} exp\left\{-\left(\tfrac{1}{2}\right) tr\left[\boldsymbol{\Sigma}_X^{-1}\left(\boldsymbol{Q}_X + \boldsymbol{S}_X\right)\right]\right\} \right] \times \\ & \quad \times \left(\sigma^2_{e_m}\right)^{-\frac{1}{2}\left(\upsilon_{e_m}+d+2\right)} exp\left\{-\frac{\boldsymbol{e}_m^{\mathrm{T}}\boldsymbol{e}_m - \upsilon_{e_m} S^2_{e_m}}{2\sigma^2_{e_m}}\right\}, \end{aligned} \qquad [5.14]$$

donde $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b} - \sum_{X=\{A,B,S\}} \boldsymbol{Z}_X \boldsymbol{a}_X^* - \boldsymbol{Z}_p \boldsymbol{e}_p$ y $\boldsymbol{Q}_X = \begin{bmatrix} \boldsymbol{a}_{oX}^{*\mathrm{T}} \boldsymbol{A}_X^{*-1} \boldsymbol{a}_{oX}^* & \boldsymbol{a}_{oX}^{*\mathrm{T}} \boldsymbol{A}_X^{*-1} \boldsymbol{a}_{mX}^* \\ \boldsymbol{a}_{mX}^{*\mathrm{T}} \boldsymbol{A}_X^{*-1} \boldsymbol{a}_{oX}^* & \boldsymbol{a}_{mX}^{*\mathrm{T}} \boldsymbol{A}_X^{*-1} \boldsymbol{a}_{mX}^* \end{bmatrix}$.

A partir de la expresión analítica [5.14] es posible identificar la distribución condicional posterior de cualquier parámetro de interés, manteniendo el resto de ellos constante. Al igual que en el caso del análisis bayesiano jerárquico del MAM (véase el Capítulo 2), todas las distribuciones condicionales posteriores pertenecen a familias conocidas y, en consecuencia, pueden muestrearse con procedimientos estándares. Las expresiones de las distribuciones condicionales posteriores de todos los parámetros del MBAM con efectos maternos se presentan en el Apéndice D.

5.2.3. Implementación del análisis a datos experimentales

Finalmente, en esta sección se describe la implementación del análisis bayesiano jerárquico a datos de un cruzamiento experimental Angus × Hereford. Los datos provienen del 'Ruakura Agriculture Research Centre', perteneciente al 'AgResearch Crown Research Institute', Nueva Zelanda, y consisten, básicamente, en 3749 registros de peso al destete y su correspondiente genealogía (Tabla 5.1). Los registros fueron tomados entre 1973 y 1990 sobre individuos de las razas parentales y varios grupos raciales, incluyendo cruzamientos *inter-se*, retrocruzas y cruzamientos rotacionales (Tabla 5.2). Una descripción detallada del diseño de apareamientos y otros aspectos del experimento se puede consultar en Morris *et al*. (1994).

Tabla 5.1. Descripción del archivo de datos del rodeo experimental Angus × Hereford.

	ANGUS × HEREFORD		
BASE de pedigree	**Individuos**	**Toros**	**Vacas**
	4668	292	1698
BASE de datos	**Nº**	**Promedio, kg**	**DS, kg**
Registros de PD	3749	153,56	29,94
	Padres	**Madres**	**TOTAL**
Progenitores	216	1647	1863
(c/ registro de PD)	145	923	1068
%	67,13	56,04	57,33
Nº prom. crías x progenitor	16,05	2,28	
% de progenitores c/:			
1 cría	3,70	42,93	
2 crías	4,17	21,86	
3 crías	2,31	15,66	
>3 crías	89,81	19,55	

PD = peso al destete; DS = desvío estándar.

Con el objetivo de estimar parámetros relevantes a esta población experimental, se ajustó el modelo presentado en la sección precedente. El modelo incluyó los valores de cría directos y maternos por origen racial, y los efectos fijos de sexo, edad de la madre y la covariable días al nacimiento, de acuerdo a la descripción de Morris *et al*. (1994). Para contemplar diferencias en las medias fenotípicas por composición racial, por último, se incluyeron también efectos directos y maternos de raza y heterosis, de acuerdo a la parameterización propuesta por Hill (1982).

Los CVC fueron estimados mediante un algoritmo de GS similar al propuesto por García-Cortés y Toro (2006). En este trabajo, la estrategia de cómputo también se basó en construir las MME como si se tratara de un modelo animal con varios efectos aleatorios. Sin embargo, en lugar de descartar las ecuaciones correspondientes a individuos con contribuciones nulas, éstas nunca fueron generadas: el sistema, en cambio, se colapsó simplemente modificando las coordenadas de las contribuciones apropiadas; esto es, eliminando las filas y las columnas nulas. Nótese que esta alternativa tiene la ventaja de reducir el número de contribuciones generadas, aunque requiere que se identifiquen los animales con contribuciones nulas para cada componente por origen racial.

Tabla 5.2. Tipos de cruzamiento, genotipos y composiciones raciales representadas en el rodeo experimental Angus × Hereford.

Cruzamiento	Genotipos	N	f_A^i	f_A^S	f_A^D
Parental	ANGUS	711	1,00	1,00	1,00
Parental	HEREFORD	431	0,00	0,00	0,00
Inter-se	F1(H×A)	393	0,50	0,00	1,00
Inter-se	F1(A×H)	301	0,50	1,00	0,00
Inter-se	F2(HA×HA)	235	0,50	0,50	0,50
Inter-se	F2(AH×AH)	183	0,50	0,50	0,50
Inter-se	F3(F2×F2)	254	0,50	0,50	0,50
Inter-se	F4(F3×F3)	104	0,50	0,50	0,50
Retrocruza	B1(A×HA)	78	0,75	1,00	0,50
Retrocruza	B1(A×AH)	72	0,75	1,00	0,50
Retrocruza	B1(H×HA)	77	0,25	0,00	0,50
Retrocruza	B1(H×AH)	67	0,25	0,00	0,50
Retrocruza	B1(AH×A)	180	0,75	0,50	1,00
Retrocruza	B1(HAxH)	132	0,25	0,50	0,00
Rotacional	R3[A×B1(H×HA)]	77	0,63	1,00	0,25
Rotacional	R3[A×B1(H×AH)]	51	0,63	1,00	0,25
Rotacional	R3[H×B1(A×HA)]	96	0,38	0,00	0,75
Rotacional	R3[H×B1(AH×A)]	51	0,38	0,00	0,75
Rotacional	R4(A×R3)	67	0,69	1,00	0,38
Rotacional	R4(H×R3)	68	0,31	0,00	0,63
Avanzado	F3×F1(HA)	19	0,50	0,50	0,50
Avanzado	F3×F1(AH)	27	0,50	0,50	0,50
Avanzado	F3×F4	30	0,50	0,50	0,50
Avanzado	A×R4	21	0,66	1,00	0,31
Avanzado	H×R4	24	0,34	0,00	0,69
TOTAL		**3749**			

f_A^i, f_A^S, f_A^D: Fracción esperada de genes Angus ('composición racial') en individuos, padres y madres.

Para llevar a cabo la estimación se adaptó el código del GS descripto en el Capítulo 2. Básicamente, fue necesario acomodar la subrutina de Meuwissen y Luo (1992) para el cálculo de los coeficientes de consanguinidad con el objetivo de generar las contribuciones apropiadas a las matrices de relaciones aditivas parciales. Si bien García-Cortés y Toro (2006) adaptaron la subrutina de Quaas (1976) con este mismo propósito, la subrutina de Meuwissen y Luo (1992) presenta dos ventajas adicionales: 1. es un algoritmo más veloz; y 2. permite computar las contribuciones fila a fila, lo cual facilita la programación (*cf.* Mrode, 2005). Específicamente, entonces, se incluyó una subrutina interna dentro de la estructura modular del GS que genera las contribuciones de los efectos aleatorios y computa los elementos de las matrices de relaciones aditivas parciales de acuerdo a una versión modificada del algoritmo para el cálculo de los coeficientes de consanguinidad de Meuwissen y Luo (1992). La adaptación del algoritmo requirió redefinir la expresión para la varianza dentro de familia e inicializar la variable de trabajo 'FI' con los coeficientes de composición racial apropiados.

Por su parte, la implementación del análisis se llevó adelante en dos etapas. En la primera etapa se realizó un análisis exploratorio con el objetivo de definir valores 'razonables' para los parámetros de escala de las distribuciones a priori de los CVC. En una primera instancia, se ajustó el MAM clásico (véase Cap. 2) y se obtuvieron luego estimaciones REML de los CVC mediante el programa ASReml (Gilmour *et al*., 2006).

Los parámetros de escala de las distribuciones a priori de la varianza de los efectos ambientales maternos permanentes y de la varianza del error, entonces, se definieron de acuerdo a dichas estimaciones. Por su parte, los valores de las estimaciones de los CVC genéticos fueron arbitrariamente distribuidas entre las tres fuentes de variabilidad por origen racial. Una vez que los hiperparámetros de escala fueron especificados, se ejecutó el programa y se obtuvieron varias cadenas de entre un millón y dos millones de iteraciones, que variaban de acuerdo al signo de la covarianza genética directa-materna, a los grados de credibilidad asignados a los parámetros o bien al número de muestras descartados como período de calentamiento. Los estadísticos descriptivos posteriores y los diagnósticos de convergencia fueron razonablemente consistentes para todas las cadenas evaluadas en esta etapa del análisis y, en consecuencia, los resultados se utilizaron para definir los parámetros de escala de las distribuciones a priori de los CVC del análisis definitivo.

Con base en la experiencia recolectada en este análisis preliminar, entonces, en la segunda etapa del análisis se obtuvo una cadena larga de 3.500.000 iteraciones, de acuerdo a las recomendaciones de Geyer (1992). Se descartaron, luego, las primeras 100.000 iteraciones como período de calentamiento, y se utilizaron las 3.400.000 restantes para estudiar la convergencia de la cadena mediante los diagnósticos de convergencia de cadena simple que provee el paquete BOA (Smith, 2007), ejecutado bajo entorno R (http://www.r-project.org/). Medias, modos, medianas y desvíos estándares posteriores, así como intervalos de alta densidad posterior del 95%, de todos los CVC fueron, por último, obtenidos mediante el programa POSTGIBBSF90 del paquete BLUPF90 (Misztal *et al*., 2002).

5.3. RESULTADOS

Se describen a continuación algunos aspectos relevantes de la implementación del análisis multirracial al conjunto de datos Angus × Hereford. El análisis final llevó alrededor de cinco días de ejecución en una computadora personal con procesador Pentium® 4 (CPU 3.6GHz, 3.11 GB de RAM), a razón de 0,11 segundos por iteración. Los valores utilizados para inicializar los parámetros de escala y los grados de credibilidad de las distribuciones a priori de los componentes de varianza se presentan en la Tabla 5.3. En general, las autocorrelaciones entre muestras de un mismo parámetro fueron muy altas para todos los CVC, pero especialmente para aquellos asociados a los términos de segregación. Sin embargo, al tomar submuestras de las cadenas (‘*thinning*’) para un lapso adecuado las autocorrelaciones disminuyeron a valores razonables sin afectar los estadísticos descriptivos posteriores y, en consecuencia, la convergencia se estudió sobre la cadena larga de 3.400.000 iteraciones. Cabe destacar, por último, que las secuencias de muestreos de todos los componentes de varianza pasaron todos los tests de convergencia de cadena simple que ofrece el paquete BOA (Smith, 2007).

En la Tabla 5.3 se presentan los estadísticos descriptivos de las distribuciones marginales posteriores de los once componentes de varianza del modelo ajustado. Adicionalmente, en la Figura 5.1 se presentan las formas de estas densidades, estimadas a través de un método no paramétrico basado en un núcleo Gaussiano (Silverman, 1986). En general, los componentes de varianza genéticos mostraron distribuciones marginales posteriores con un alto grado de simetría, excepto por los componentes asociados a la segregación entre razas. Así, mientras que los valores medios de las varianzas de segregación para las componentes directa y materna del carácter fueron de $\bar{\sigma}^2_{a_o S} = 9,62\,\mathrm{kg}^2$ y

$\bar{\sigma}^2_{a_m S} = 13,37\,\text{kg}^2$, respectivamente, los valores modales rondaron los $3\,\text{kg}^2$ en ambos casos.

Tabla 5.3. Parámetros a priori y estadísticos descriptivos de las distribuciones marginales posteriores de los CVC.

							IADP95%	
CVC[1]	**υ**	***S***	**Media**	**DS**	**Mediana**	**Modo**	**Inferior**	**Superior**
$\sigma^2_{e_o}$	100	170	187,34	10,21	187,35	187,09	167,17	207,22
$\sigma^2_{e_m}$	100	80	95,53	9,91	95,24	98,75	76,47	115,17
$\sigma^2_{a_o A}$	20	85	120,74	20,43	119,54	115,82	82,22	161,46
$\sigma_{a_o a_m A}$	20	-25	-27,00	13,26	-26,11	-23,89	-53,70	-2,15
$\sigma^2_{a_m A}$	20	35	37,63	11,35	35,94	32,35	18,25	60,38
$\sigma^2_{a_o H}$	20	76	100,24	20,12	98,86	98,42	62,38	140,33
$\sigma_{a_o a_m H}$	20	-50	-56,31	19,64	-55,12	-56,55	-95,65	-19,13
$\sigma^2_{a_m H}$	20	70	95,18	24,61	92,96	88,29	50,29	144,21
$\sigma^2_{a_o S}$	5	10	9,62	6,24	8,10	3,68	1,28	21,96
$\sigma_{a_o a_m S}$	5	8	9,55	7,01	7,82	3,20	0,36	24,18
$\sigma^2_{a_m S}$	5	9	13,37	12,55	9,48	3,65	1,03	37,93

υ = grados de credibilidad a priori; *S* = parámetro de escala a priori.

[1] Componentes de (co)varianza: $\sigma^2_{e_o}$ = varianza del error; $\sigma^2_{e_m}$ = varianza de los efectos ambientales maternos permanentes; $\sigma^2_{a_o X}$ = varianza aditiva directa por origen racial; $\sigma^2_{a_m X}$ = varianza aditiva materna por origen racial; $\sigma_{a_o a_m X}$ = covarianza directa-materna por origen racial; *X* = {Angus, Hereford, segregación}; DS = desvío estándar; IADP95% = intervalo de alta densidad posterior del 95%.

Por otro lado, también se observaron algunas diferencias en los estadísticos descriptivos posteriores de los componentes de varianza genéticos según la fuente de variabilidad de origen racial. En el caso de las varianzas aditivas directas, por ejemplo, se obtuvo una pequeña diferencia de escala en los valores medios de acuerdo al origen Angus vs. Hereford ($\bar{\sigma}^2_{a_o A} = 120,74\,\text{kg}^2$ vs. $\bar{\sigma}^2_{a_o H} = 100,24\,\text{kg}^2$, respectivamente), con una dispersión similar. En cambio, las medias de las varianzas aditivas maternas mostraron diferencias importantes a favor del origen Hereford ($\bar{\sigma}^2_{a_m A} = 37,63\,\text{kg}^2$ vs. $\bar{\sigma}^2_{a_m H} = 95,18\,\text{kg}^2$), con una dispersión mucho mayor respecto al Angus. Por último, en lo que respecta a la covarianza genética directa-materna, la media de la distribución posterior fue negativa para ambos orígenes raciales, con aproximadamente la mitad de magnitud en el origen Angus respecto al origen Hereford ($\bar{\sigma}_{a_o a_m A} = -27,00\,\text{kg}$ vs. $\bar{\sigma}_{a_o a_m H} = -56,31\,\text{kg}$). La covarianza directa-materna de segregación, en cambio, fue positiva dentro del intervalo de alta densidad posterior del 95%, con un valor medio de $\bar{\sigma}_{a_o a_m S} = 9,55\,\text{kg}$ y un valor modal de 3,20 kg.

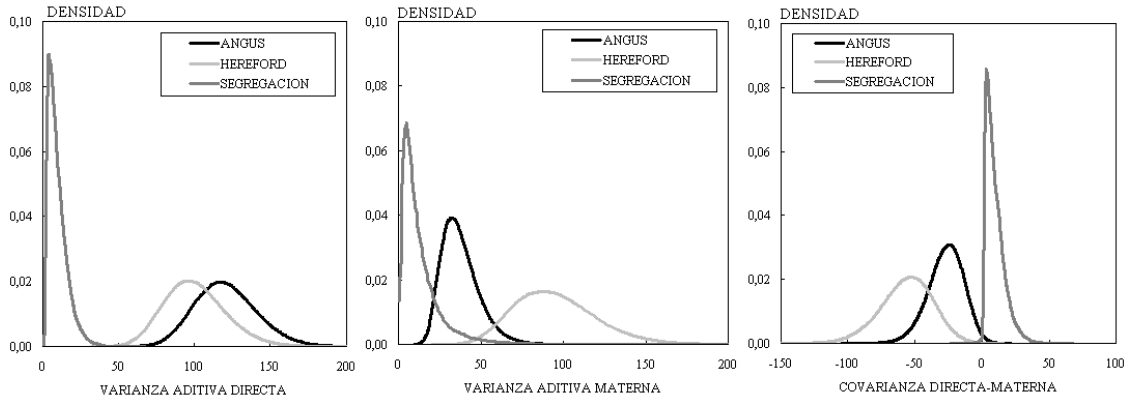


Figura 5.1. Distribuciones marginales posteriores estimadas de los CVC genéticos. Las gráficas fueron estimadas mediante un método no paramétrico basado en un núcleo Gaussiano (Silverman, 1986). Las curvas están desagregadas por fuente de variabilidad de origen racial.

Estadísticos descriptivos de las distribuciones posteriores de la heredabilidad directa, la heredabilidad materna y la correlación genética directa-materna para la población de referencia del modelo (*i.e.*, los individuos F_2) se presentan en la Tabla 5.4. Las heredabilidades, en este caso, fueron definidas como un cociente entre la varianza aditiva del carácter, expresada como una suma ponderada de los componentes por origen racial, y la varianza fenotípica de individuos pertenecientes al grupo de referencia. Las medias de las heredabilidades directa y materna fueron de 0,27 y 0,18, respectivamente, con algún pequeño desvío respecto a los valores modales en este último caso. La media de la correlación directa-materna, por su parte, fue de –0.33. Todos los cocientes de varianza fueron significativos al 95%, como se desprende de los correspondientes intervalos de alta densidad posterior.

Tabla 5.4. Estadísticos descriptivos para la heredabilidad directa, la heredabilidad materna y la correlación genética directa-materna.

	Media (DS)		Modo (IADP95i; IADP95s)	
Carácter[1]	**PD**	**AM**	**PD**	**AM**
PD	0,27 (0,03)	–0,33 (0,13)	0,26 (0,20; 0,33)	–0,35 (–0,57; –0,07)
AM		0,18 (0,03)		0,24 (0,11; 0,24)

DS = desvío estándar; IADP95i, IADP95s = límites inferior y superior del intervalo de alta densidad posterior del 95%.

[1]PD = peso al destete; AM = aptitud materna.

En la Tabla 5.5, finalmente, se presenta la contribución relativa de cada origen racial a las varianzas aditivas totales de las componentes directa y materna del carácter en individuos F_2. La contribución a la varianza aditiva del origen Angus fue superior a la contribución del origen Hereford para la componente directa (50,26% vs. 41,73%), mientras que la contribución del origen Hereford fue netamente dominante para la componente materna (23,59% vs. 59,65%). La contribución de la componente de segregación, por su parte, fue poco significativa para la componente directa del carácter (< 10%), pero más importante para la componente materna (≈ 17%), al menos cuando fue calculada utilizando las medias posteriores de las varianzas aditivas. En cambio, cuando se calcularon utilizando los valores modales de las distribuciones posteriores, en ambos

casos las contribuciones de las componentes de segregación fueron poco significativas: 3.32% y 5.71% para las componentes directa y materna, respectivamente.

Tabla 5.5. Varianzas aditivas directa y materna en individuos F_2 de acuerdo a la fuente de variabilidad de origen racial.

		% por origen racial			Total[1]
Varianzas aditivas en individuos F_2		Angus	Hereford	Segregación	kg[2]
Directa:	$\frac{1}{2}\sigma^2_{a_oA} + \frac{1}{2}\sigma^2_{a_oH} + \sigma^2_{a_oS}$	50,26%	41,73%	8,01%	120,11
Materna:	$\frac{1}{2}\sigma^2_{a_mA} + \frac{1}{2}\sigma^2_{a_mH} + \sigma^2_{a_mS}$	23,59%	59,65%	16,76%	79,78

[1] Calculado utilizando las medias posteriores.

5.4. DISCUSIÓN

En este capítulo se formalizó la equivalencia entre el modelo de análisis multirracial con varianzas aditivas heterogéneas propuesto por García-Cortés y Toro (2006) y aquel que deriva de la teoría genética cuantitativa (Lo *et al*., 1993; Cantet y Fernando, 1995). La derivación se basó en una formulación algo distinta a la de estos autores, en la que se redefinieron los vectores de valores de cría por origen racial sin incluir los individuos con contribuciones nulas. Luego, se definió una matriz que permite recuperar el patrón de filas y columnas nulas respecto a la matriz de incidencia de los valores de cría y, en consecuencia, respecto a las matrices de relaciones aditivas parciales. Finalmente, operando algebraicamente con estas matrices, se demostró la equivalencia entre los modelos en términos de su estructura de covarianza y predicción de los valores de cría. Si bien para simplificar la notación se asumió una población compuesta de dos razas en la derivación, la generalización a *p* razas sólo requiere redefinir apropiadamente los correspondientes vectores de valores de cría por origen racial.

Adicionalmente, se extendió la formulación del MBAM para incluir efectos maternos y se describió la implementación de un análisis bayesiano jerárquico con el objetivo de estimar los CVC. Como fuera discutido en el capítulo introductorio de esta tesis, el enfoque bayesiano es, en general, más intuitivo, más flexible y sus resultados son más informativos que otros métodos de inferencia. En el marco del MBAM, por ejemplo, el análisis bayesiano permite incorporar información a priori sobre los CVC asociados a las diferentes fuentes de variabilidad genética de origen racial (Cardoso y Tempelman, 2006). Cuando la incertidumbre a priori sobre estos parámetros es grande, sin embargo, una alternativa es utilizar especificar distribuciones a priori Gamma invertidas, parameterizadas de modo de poder reflejar dicha incertidumbre a través de la definición de los grados de credibilidad (Sorensen y Gianola, 2002), tal como se describió en esta investigación. En tal caso, las distribuciones condicionales resultantes son conjugadas y, en consecuencia, es posible implementar un algoritmo de GS como método de inferencia. De hecho, programar un algoritmo de GS con el objetivo de estimar CVC en el marco de un análisis multirracial con varianzas aditivas heterogéneas sólo requiere adaptar una subrutina para el cálculo de coeficientes de consanguinidad para generar las contribuciones apropiadas a las matrices de relaciones aditivas parciales (García-Cortés y Toro, 2006).

Como fuera discutido en el Capítulo 2 de esta tesis, algunos aspectos importantes al implementar el muestreo de Gibbs involucran decidir el número de cadenas a generar, elegir los valores iniciales y, por último, determinar el período de calentamiento y el

número de iteraciones necesarias para asegurar una muestra representativa de la distribución marginal de interés (Gilks *et al*., 1996). Con base en los altos niveles de autocorrelación entre muestras observados para todos los CVC, particularmente para aquellos asociados a la segregación entre razas, y dado que el tiempo de cómputo por ciclo no fue limitante, en este capítulo se adoptó el enfoque de Geyer (1992) para cubrir varios de los aspectos de la implementación. Así, los resultados aquí presentados se basaron en una única cadena muy larga, obtenida luego de descartar un número considerable de muestreos. Si bien utilizar submuestras de la cadena reduce la auto-correlación a niveles razonables, esta práctica no es obligatoria (Raftery y Lewis, 1996) ni definitivamente necesaria para obtener estadísticos posteriores precisos (Geyer, 1992).

Ahora bien, aunque el tiempo de cómputo no fue limitante en la presente implementación del GS, cabe preguntarse si el método es factible computacionalmente para archivos de datos de mayor dimensión. Al respecto, y como se discutiera en el Capítulo 2, deben distinguirse dos aspectos bien diferentes que afectan el tiempo de cómputo: 1. el número de operaciones aritméticas necesarias para completar un ciclo del algoritmo en función del número de individuos en el archivo de pedigree; y 2. el número de ciclos necesario para asegurar la convergencia del procedimiento. Las tareas que mayor tiempo de cómputo consumen son el muestreo del vector de parámetros de posición y el cómputo de las formas cuadráticas durante el muestreo de las matrices de covarianza. Estas tareas involucran operaciones aritméticas sobre matrices de considerable dimensión: la matriz de coeficientes de las MME y las matrices de relaciones aditivas parciales. Sin embargo, dado que estas matrices se almacenan en forma rala y que las operaciones sólo involucran a elementos distintos de cero, en última instancia el tiempo de cómputo por ciclo será lineal en el número de individuos en el pedigree (Misztal, 2006). Debe considerarse que en el caso de MBAM aquí descripto, la dimensión del sistema crece en forma cuadrática con el número de razas involucradas (García-Cortés y Toro, 2006). Sin embargo, el aumento en el número de ecuaciones será compensado por la existencia de ecuaciones nulas, y esto último dependerá de la composición racial de los individuos en el pedigree. Ahora, asegurar la convergencia de las cadenas es otro problema. En la presente implementación, en la que se asumió una enorme incertidumbre respecto los valores a priori de los CVC, los tests formales de convergencia resultaron poco concluyentes para cadenas de menos de un millón de ciclos. En consecuencia, para archivos de datos de mayor tamaño será necesario llevar a cabo alguna estrategia para mejorar la tasa de convergencia. Por ejemplo, si el analista cuenta con alguna información a priori respecto al valor de los CVC genéticos, podría implementar un muestreo basado en la distribución GIW (véase el Capítulo 3). Otras estrategias pueden consultarse en Gilks y Roberts (1996).

Por último, se ajustó el modelo de análisis multirracial aquí descripto a datos de peso al destete de un cruzamiento experimental Angus × Hereford y se obtuvieron, por primera vez, estimaciones para el juego completo de CVC descripto por Cantet y Fernando (1995) en el marco de un MBAM con efectos maternos. En rigor, Elzo y Wakeman (1998) reportaron previamente estimaciones REML de los CVC para un rodeo multirracial Angus × Brahman, aunque bajo un modelo bivariado padre–abuelo materno. En su trabajo, estos autores parameterizaron la variabilidad adicional que surge de la segregación entre razas en términos de la ‘varianza aditiva interracial’ (Elzo, 1994), un parámetro equivalente a dos veces la varianza de segregación definida por Lo *et al*. (1993). Los valores que obtuvieron para la varianza aditiva interracial materna y la covarianza aditiva interracial fueron, en términos absolutos, considerablemente mayores a los reportados aquí. Sin embargo, Elzo y Wakeman (1998) cuestionaron la validez de

estas estimaciones con base en el gran número de parámetros a estimar y la restringida información interracial en el archivo de datos. De hecho, muchos problemas de estimación asociados a conjuntos de datos poco informativos radican en la dificultad de cuantificar el error de estimación, particularmente en modelos con estructura jerárquica (O'Hara *et al*., 2008). Al incorporar incertidumbre en términos de distribuciones de probabilidad, el enfoque bayesiano permite sobrellevar este problema (Sorensen y Gianola, 2002; O'Hara *et al*., 2008).

Otros aspectos del análisis se discuten a continuación. Los resultados obtenidos en la presente investigación indicarían que en la población Angus × Hereford analizada las frecuencias génicas de las razas contribuyentes fueron originalmente similares entre sí. Esto se deduce de la contribución marginal de la varianza de segregación a la varianza aditiva total en los individuos F_2 (*e.g*. Lo *et al*., 1993; Birchmeier *et al*., 2002), al menos si se toman los valores modales posteriores como estimaciones puntuales de los CVC. En este punto, cabe mencionar que las distribuciones marginales posteriores de los componentes de segregación resultaron fuertemente asimétricas en todos los casos, un patrón muy similar al reportado por Cardoso y Tempelman (2004) al analizar datos de crecimiento postdestete de un cruzamiento Nelore × Hereford. Por otro lado, las medias posteriores de las heredabilidades directa y materna, y de la covarianza genética directa-materna para individuos de la población de referencia fueron razonables y acordes a la literatura (*cf*. CSIRO, 2010). Es importante destacar, sin embargo, que bajo el MBAM la varianza fenotípica individual es propia de cada composición racial y, en consecuencia, heredabilidades y correlaciones sólo tienen sentido dentro de cada grupo racial.

De hecho, las composiciones raciales constituyen un elemento clave del análisis: son necesarias tanto para el cómputo de las matrices de relaciones aditivas parciales, como para definir las variables regresoras de los coeficientes utilizados en el ajuste por raza y heterosis (*cf*. Lynch y Walsh, 1998, Cap. 9). Esto implica que es necesario conocer la composición racial de cada uno de los individuos de la población para ajustar el modelo aquí descripto. Sin embargo, archivos de datos con información precisa sobre las composiciones raciales de los individuos son difíciles de encontrar. Por otro lado, también es necesario contar con una adecuada estructura de información para obtener estimaciones precisas de los componentes de varianza; por ejemplo, la información para estimar los componentes de segregación está contenida en la performance de la progenie de padres cruza (Elzo y Wakeman, 1998). En este sentido, el archivo de datos utilizado en esta investigación reúne características excepcionales. Por un lado, cuenta con abundante información interracial, con mediciones tomadas sobre individuos de diversos grupos raciales y con muchas relaciones de parentesco conectando los grupos entre sí. Además, presenta una buena estructura de información en lo que respecta a la estimación de parámetros en modelos animales con efectos maternos (*cf*. Gerstmayr, 1992; Maniatis y Pollott, 2003), con un alto porcentaje de madres con información fenotípica y una elevada proporción de ellas con más de una cría. Sería interesante evaluar la performance del análisis aquí descripto con datos de campo, en especial para archivos con información interracial más restringida.

En conclusión, consideraciones teóricas y empíricas justifican la especificación de una estructura de covarianza genética heterogénea al ajustar modelos animales para la predicción de valores de cría en poblaciones multirraciales. En este contexto, el MBAM alternativo basado en la descomposición de la matriz de covarianza genética en diferentes fuentes de variabilidad por origen racial (García-Cortés y Toro, 2006) simplifica enormemente la estimación de CVC genéticos mediante la ejecución de un algoritmo de

GS. En este capítulo se ha demostrado que el modelo resultante es equivalente a aquél derivado utilizando de la teoría genética cuantitativa (Lo *et al*., 1993), y descripto por Cantet y Fernando (1995). Luego, la extensión del modelo para incluir efectos maternos y la implementación del correspondiente análisis bayesiano jerárquico con el objetivo de estimar CVC es directa.

6
Conclusiones generales

6.1. INTRODUCCIÓN

En los programas de mejoramiento genético animal frecuentemente se evalúan datos de performance para caracteres bajo la influencia de efectos maternos. En virtud de su naturaleza hereditaria, los efectos maternos pueden alterar la tasa y dirección de la respuesta a la selección en un programa de selección artificial (Kirkpatrick y Lande, 1989). En consecuencia, es importante que ambas componentes del carácter, es decir, tanto los efectos directos como los efectos maternos, sean puestos en consideración al tomar decisiones de selección (Heydarpour *et al*., 2008). Actualmente, y en forma masiva, se ajustan registros fenotípicos para caracteres de esta naturaleza empleando el modelo animal con efectos maternos que, en el contexto de este trabajo, ha sido denominado 'clásico' (MAM) (Willham, 1963, Quaas y Pollak, 1980). Bajo el MAM, las predicciones BLUP de los valores de cría para ambos caracteres componentes se obtienen resolviendo las denominadas ecuaciones del modelo mixto (MME) (*cf*. Henderson, 1984).

Como fuera discutido oportunamente, las MME dependen del conjunto de parámetros de dispersión asociados a la estructura de covarianza de las observaciones, los componentes de (co)varianza (CVC). Estrictamente, las propiedades estadísticas del predictor BLUP se basan en asumir que los CVC son conocidos. En la práctica, sin embargo, los CVC deben estimarse previamente a partir de los mismos datos. De hecho, la estimación de CVC constituye uno de los desafíos más importantes en el marco de los MAM. En la actualidad existen múltiples paquetes estadísticos de uso general que permiten obtener estimaciones de CVC bajo el MAM, bien mediante algoritmos REML, como, por ejemplo, el ASReml (Gilmour *et al*., 2006) o el WOMBAT (Meyer, 2007), o bien, en el contexto de un análisis bayesiano jerárquico, mediante el algoritmo del muestreo de Gibbs, como, por ejemplo, el MTGSAM (Van Tassell y Van Vleck, 1996) o la colección de programas GIBBSF90 (Misztal, 2002).

En la larga trayectoria de los MAM se han investigado y propuesto nuevas formulaciones del modelo, generalmente basadas en alguna extensión de la estructura de covarianza de los efectos genéticos y ambientales, con el objetivo de contemplar nuevas fuentes de variabilidad fenotípica. En términos generales, la difusión de estos modelos ha sido más bien limitada, fundamentalmente por tres motivos. En primer lugar, porque muchas veces no existen archivos de datos con la estructura de información apropiada para estimar los nuevos parámetros. En segundo lugar, porque, aunque estos archivos existieran, no existen métodos de inferencia apropiados para abordar el problema de la estimación. Y, finalmente, porque suele no existir software de uso general para embarcarse en dicha tarea.

En este trabajo se abordó el segundo de estos problemas. Específicamente, y desde un enfoque bayesiano, se presentaron contribuciones teóricas y metodológicas, así como aplicaciones a datos de campo, experimentales y simulados, con relación a novedosas estructuras de covarianza para los MAM. Dos formulaciones alternativas del MAM han sido particularmente tratadas: 1. La inclusión de un parámetro de correlación ambiental entre pares de observaciones madre–progenie (Bijma, 2006); y 2. la extensión de la estructura de covarianza genética en poblaciones multirraciales (Cantet y Fernando, 1995). En este capítulo se sintetizan y jerarquizan los desarrollos metodológicos descriptos a lo largo de este trabajo. La discusión está organizada del siguiente modo. En primer lugar, se discuten los problemas específicos que se abordaron. Luego, se sintetizan los principales resultados obtenidos en conexión con los problemas planteados. Finalmente, se presentan las líneas de investigación futura que quedaron abiertas en conexión con los resultados obtenidos.

6.2. PROBLEMAS ESPECÍFICOS QUE SE ABORDARON EN ESTE TRABAJO

En este trabajo se adhirió al enfoque bayesiano para abordar el problema de la estimación de CVC bajo el MAM. Como fuera discutido en el capítulo introductorio, el enfoque bayesiano es más intuitivo, sus resultados son más informativos y, al menos en lo que respecta a métodos de inferencia para modelos jerárquicos, su implementación es más sencilla, si bien con relación a este último punto es necesario indicar que los métodos son altamente dependientes de la capacidad de cómputo. En este contexto, entonces, el punto de partida de la presente investigación fue la implementación del análisis bayesiano jerárquico vía el algoritmo del muestreo de Gibbs (GS), descripta en detalle en el Capítulo 2. Este capítulo constituyó, de hecho, el marco de referencia de toda la tesis. Allí se ilustró también la implementación del análisis a un archivo de datos de peso al destete en bovinos de carne, y se discutieron en detalle algunos resultados obtenidos a la luz de las restricciones que impone el muestreo de los CVC genéticos de una distribución Wishart invertida (IW). Este fue el primero de los problemas atacados.

6.2.1. Muestreo de la matriz covarianza genética a partir de una distribución IW

En el marco de un análisis bayesiano jerárquico, la IW es la alternativa natural para modelar la distribución a priori de los CVC genéticos. Sin embargo, presenta una limitante importante: mientras que la distribución es función de un conjunto completo de parámetros que permiten modelar la media a priori de la distribución de la matriz de covarianza genética, la incertidumbre respecto a estos valores está gobernada por un único parámetro escalar (Brown, 2002). Esta limitante genera básicamente dos problemas. En primer lugar, no permite modelar la incertidumbre diferencial que existe entre los parámetros de dispersión directos y maternos. Al respecto, en el Capítulo 2 se discutió que en general existe más incertidumbre en torno al valor de la heredabilidad materna que en torno al valor de la heredabilidad directa, básicamente porque existen menos relaciones de parentesco que provean contrastes informativos para estimar este parámetro en los archivos de datos; recuérdese que los efectos maternos se expresan con una generación de retraso respecto a los efectos directos y, además, están limitados a un sexo (Willham, 1980). En segundo lugar, al aplicar el GS se suelen observar altísimas correlaciones de muestreo entre los CVC genéticos, lo cual aumenta considerablemente la demanda en tiempo de cómputo del algoritmo para asegurar su convergencia.

Por otro lado, otro aspecto de la implementación del GS con el objetivo de estimar CVC bajo un MAM está sujeto a controversia: no existe consenso general con respecto a la forma de parameterizar las distribuciones a priori de los parámetros de dispersión. Una alternativa al respecto es utilizar distribuciones a priori Uniformes con el supuesto objetivo de representar ignorancia total respecto a los valores posibles de los parámetros a priori. Sin embargo, este enfoque ha sido fuertemente cuestionado (*cf.* Blasco, 2001). En cambio, si el analista decide utilizar distribuciones a priori informativas, entonces los resultados de los análisis serán más o menos sensibles a las especificaciones a priori.

6.2.2. Estimaciones negativas de la covarianza genética directa-materna

Además de los aspectos relacionados con la implementación del algoritmo GS, en este trabajo también se abordó el problema de la identificación y estimación de CVC asociados a fuentes de variabilidad fenotípica no contempladas en la formulación del MAM clásico. Al respecto, uno de los problemas sobre los que más se ha escrito desde los trabajos seminales de Falconer (1965) y Koch (1972), es el de la posible existencia de una correlación de naturaleza ambiental entre los efectos maternos en generaciones adyacen-

tes, y que generalmente se utiliza como argumento para justificar las altísimas correlaciones genéticas entre efectos directos y maternos frecuentemente reportadas en la literatura. En el Capítulo 1 se presentó una revisión bibliográfica de este problema, mientras que en el Capítulo 4 se enumeraron las diferentes formulaciones alternativas del MAM que se sucedieron para contemplar este efecto. En aquel capítulo se hizo especial énfasis en el modelo propuesto por Bijma (2006), que incluye un parámetro de correlación entre pares de observaciones madre–progenie. Si bien la formulación que propuso Bijma (2006) dificulta la estimación de los CVC, en el caso particular en el que las madres con registro fenotípico tengan una única cría, entonces la matriz de covarianza de error presentará una estructura Toeplitz y, en consecuencia, la estimación del parámetro puede llevarse adelante ajustando una serie de tiempo de medias móviles (*cf*. Bijma, 2006). Sin embargo, cuando las madres con datos tienen más de una cría este enfoque no es válido.

6.2.3. Equivalencia entre modelos de análisis multirracial

Por último, otro de los problemas que se trató en este trabajo con relación a la identificación y estimación de CVC asociados a nueves fuentes de variabilidad fenotípica, es la necesidad de extender la estructura de covarianza de los valores de cría para contemplar nuevas fuentes de variabilidad genética en poblaciones multirraciales. Al respecto, si bien la teoría ha sido establecida hace tiempo (Elzo y Famula, 1985; Elzo, 1990; Lo *et al*., 1993), el modelo resultante no se ha difundido posiblemente porque los métodos de inferencia propuestos (*e.g*. Birchmeier *et al*., 2002; Elzo, 1994; Cardoso y Tempelman, 2004) son difíciles de implementar y no existe software de uso general para encarar la estimación. Recientemente, García-Cortés y Toro (2006) desarrollaron un modelo alternativo basado en la descomposición de la matriz de covarianza genética en sus diferentes fuentes de variabilidad por origen racial, que facilita enormemente la estimación mediante la implementación de un algoritmo GS. Si bien estos autores ilustraron la validez de su modelo con respecto al que se deriva de la teoría genética cuantitativa mediante un pequeño ejemplo numérico, no presentaron una derivación formal de la equivalencia. Por otro lado, tampoco existía una formulación del modelo que incluya efectos maternos ni estimaciones disponibles del correspondiente juego de CVC genéticos descriptos por Cantet y Fernando (1995).

6.3. CONTRIBUCIONES DE LA TESIS

A lo largo de este trabajo se han presentado contribuciones teóricas y metodológicas con relación a la estimación de parámetros de dispersión en MAM sujetos a diferentes estructuras de covarianza. En términos generales, los resultados reportados representan aportes novedosos a la problemática específica descripta en la sección precedente. El objetivo de esta sección es sintetizar los resultados obtenidos, rescatando en particular aquellos resultados que han constituido un aporte original a la disciplina, jerarquizando las conclusiones y ubicando los desarrollos descriptos con relación a los problemas planteados. Para una discusión más organizada, se han clasificado las contribuciones de acuerdo a su naturaleza en teóricas, metodológicas y de aplicación.

6.3.1. Contribuciones teóricas

Las contribuciones teóricas, a su vez, se han clasificado en ‘derivaciones’, por un lado, y en ‘modelos alternativos propuestos’, por otro. Los párrafos dentro de las subsecciones están organizados por capítulos.

6.3.1.1. Derivaciones

En el Capítulo 3, se introdujo la distribución Wishart invertida generalizada (GIW) en la disciplina del mejoramiento genético animal. La distribución GIW fue desarrollada originalmente por Brown *et al*. (1994) en el contexto de estudios sobre la evaluación del riesgo de contaminación del aire, y constituye esencialmente una extensión de la distribución IW con un mayor número de parámetros. Si bien la descripción se basó extensamente en el trabajo de Brown (2002), aquí se realizó un esfuerzo importante por adaptar la notación algo críptica, al menos desde el punto de vista de un mejorador animal, de este autor. En este aspecto, el trabajo de Le *et al*. (1999), de quienes se tomó la notación, constituyó un aporte de gran valor. Por otro lado, la distribución GIW fue presentada en toda su generalidad con la expectativa de motivar el hallazgo de nuevas aplicaciones en investigaciones futuras, aprovechando particularmente la flexibilidad que ofrece para describir el conocimiento a priori respecto a la distribución de una matriz de covarianza en el contexto de un análisis bayesiano jerárquico.

Luego, se derivaron resultados teóricos con respecto a la especificación de la GIW como la distribución a priori de la matriz de covarianza genética del MAM. Se demostró además que esta especificación constituye una alternativa conjugada y, en consecuencia, facilita la estimación de CVC mediante un algoritmo GS. Por otro lado, se discutieron tres especificaciones a priori diferentes para la matriz de covarianza genética, todas ellas basadas en la distribución GIW. Estas incluyeron: 1. una especificación no informativa a priori; 2. Una especificación que devuelve un muestreo posterior de una distribución IW; y 3. la extensión de esta última especificación para modelar la incertidumbre diferencial entre distintos componentes escalares de la matriz de covarianza genética.

Por su parte, en el Capítulo 4 se demostró que la formulación del MAM de Bijma (2006) deriva de plantear una covarianza entre el desvío ‘ambiental’ de la observación de una madre y su efecto ambiental materno permanente, evaluado en la ecuación de su cría. De este modo, se conciliό el modelo que incluye un parámetro de correlación entre pares de observaciones madre–progenie con una serie de ideas de larga trayectoria en la disciplina.

Finalmente, en el Capítulo 5 se formalizó la equivalencia entre el modelo de análisis multirracial propuesto por García-Cortés y Toro (2006) y aquel que deriva de la teoría genética cuantitativa (Lo *et al*., 1993). La derivación se basó en una formulación algo distinta del modelo, en la que se redefinieron los vectores de valores de cría por origen racial sin incluir los individuos con contribuciones nulas. Luego, se definió una matriz que permite recuperar el patrón de filas y columnas nulas, y operando algebraicamente con estas matrices se demostró la equivalencia entre los modelos en términos de su estructura de covarianza y predicción de los valores de cría.

6.3.1.2. Modelos alternativos propuestos

Por otro lado, en algunos capítulos de la tesis se han presentado y elaborado modelos alternativos que facilitan la implementación de los métodos de inferencia propuestos. En el Capítulo 4, en primer lugar, se presentó un modelo operativo alternativo al propuesto por Bijma (2006), que permitió acelerar considerablemente el tiempo de ejecución del algoritmo GS. Este modelo operativo alternativo, basado en descomponer el vector de errores del MAM en un vector aleatorio de desvíos ambientales directos, por un lado, y un vector de errores aleatorios, con varianza fija y de pequeña magnitud, por otro, no es estrictamente equivalente al modelo de Bijma (2006) (*cf*. Henderson, 1985). Sin embar-

go, al comparar resultados obtenidos bajo los dos modelos con exactamente las mismas especificaciones casi no hubieron diferencias. En cambio, el MBAM presentado en el Capítulo 5, basado en redefinir los vectores de valores de cría por origen racial sin incluir los individuos con contribuciones nulas, y extendido luego en su formulación para acomodar efectos maternos, es estrictamente equivalente al modelo de Cantet y Fernando (1995). Este último modelo facilita conceptualmente el algoritmo GS, porque no requiere que se generen filas y columnas de ceros en la matriz de los coeficientes de las MME y, en consecuencia, esta matriz es invertible.

6.3.2. Contribuciones metodológicas

Las principales contribuciones metodológicas de esta investigación fueron la elaboración y descripción de métodos de inferencia bayesianos para los CVC inherentes a las diferentes formulaciones del MAM, bajo una construcción jerárquica del modelo. En todos los casos, las estimaciones se llevaron a cabo mediante métodos MCMC. En particular, el algoritmo de muestreo utilizado fue siempre el GS, aunque con ciertas particularidades según el caso. A modo de referencia, en el Capítulo 2 se describió en detalle la implementación del análisis bayesiano jerárquico para el MAM clásico y se presentó, luego, paso a paso el algoritmo de muestreo. En esta sección se discute cómo se adaptó este algoritmo de muestreo para encarar la estimación de CVC bajo las formulaciones alternativas del MAM. Por último, se discuten algunos detalles referidos a la programación de estos algoritmos de muestreo.

6.3.2.1. Algoritmos de muestreo

Como fuera discutido en la sección precedente, en el Capítulo 3 se derivaron resultados teóricos respecto a la especificación de la GIW como la distribución a priori de la matriz de covarianza genética. Si bien es estándar asumir una distribución IW a priori para esta matriz, se demostró que asumir una distribución GIW, en cambio, extiende el abanico de posibles especificaciones a priori, sin perder las importantes ventajas de la distribución IW. En este caso, el algoritmo de muestreo debe ser adaptado tal como se describe en detalle en el Apéndice B. El algoritmo es muy sencillo: en lugar de muestrear de una distribución IW, sólo requiere dos muestreos de distribuciones Chi-cuadradas invertidas y un muestreo de una distribución Normal. Luego, se recuperan los elementos de la matriz de covarianza genética aplicando la descomposición de Bartlett en sentido inverso.

Por otro lado, la adaptación del algoritmo de muestreo para estimar el parámetro de correlación ambiental madre–progenie fue bastante más complicada. Como se describe en el Capítulo 4, la incorporación de este parámetro de correlación induce una estructura no diagonal en la matriz de covarianza del error, que conlleva cambios importantes en las sentencias que se utilizan para contribuir a la matriz de coeficientes de las MME. Además, los registros deben reordenarse para aprovechar la estructura diagonal en bloques de la inversa de la matriz de covarianza del error, aunque este paso puede obviarse si se ordenan los datos previamente a la ejecución. Finalmente, es necesario incorporar el bloque de sentencias para muestrear el parámetro de correlación de acuerdo al algoritmo GGS (Ritter y Tanner, 1992). En resumen, la adaptación del GS para estimar los CVC bajo el modelo de Bijma (2006) no es nada trivial. Peor aún si suman las modificaciones que fue necesario introducir para evitar los problemas de la representación de números muy pequeños en la computadora; *i.e.*, la evaluación del logaritmo de la distribución condicional posterior del parámetro de correlación y la implementación de la grilla adaptativa. De todas maneras, el algoritmo de muestreo aquí descripto per-

mite abordar el problema de la estimación del parámetro de correlación en un contexto bien general.

En cambio, la implementación del análisis bayesiano jerárquico vía el GS para estimar el conjunto de CVC inherentes al MBAM con efectos maternos bajo el modelo multirracial descripto en el Capítulo 5 fue considerablemente más sencilla. En este caso, la estrategia de muestreo del parámetro de posición se basó en construir las MME como si se tratara de un modelo animal con varios efectos aleatorios, tal como describieran García-Cortés y Toro (2006). Luego, fue necesario acomodar la subrutina de Meuwissen y Luo (1992) para el cálculo de los coeficientes de consanguinidad con el objetivo de generar las contribuciones apropiadas a las matrices de relaciones aditivas parciales.

6.3.2.2. Programación de los algoritmos de inferencia

Algunos detalles generales respecto a la programación de los algoritmos de inferencia se describen a continuación. Todos los algoritmos de muestreo descriptos en la sección anterior fueron programados en el lenguaje Fortran 90. El programa madre de todos ellos fue un código escrito específicamente para construir y resolver las MME, e inspirado en las notas de clase de Misztal (2006) y el trabajo de Groeneveld y Kovac (1990). Más adelante, se incorporó al programa una subrutina interna para muestrear las distribuciones condicionales posteriores de acuerdo al algoritmo descripto en el Capítulo 2 de este trabajo. En general, la programación y puesta a punto de los algoritmos de inferencia llevó la mayor parte del tiempo y esfuerzo total dedicados a la ejecución de esta tesis. Otros programas accesorios, como rutinas para el manejo de las bases de datos y pedigree, también fueron programados durante el período de trabajo. Entre ellos, se destaca el simulador empleado para poner a prueba la estrategia recursiva con base en la distribución GIW presentada en el Capítulo 3. Una vez más, es importante aclarar que si bien los códigos de todos estos programas no fueron anexados al presente documento debido a su extensión, pueden ser solicitados al autor. Una de las tareas a futuro con respecto a estos programas será ponerlos a disponibilidad en algún sitio de Internet.

6.3.3. Implementación de los métodos de inferencia

En general, los métodos de inferencia y algoritmos de muestreo desarrollados en este trabajo fueron aplicados a conjuntos de datos de diferente naturaleza; *i.e.*, se utilizaron datos de campo, experimentales y simulados. En todos los casos, el carácter evaluado fue el peso al destete en bovinos de carne, por la sencilla razón de su disponibilidad inmediata. Es importante destacar, sin embargo, que la discusión y metodologías que se describieron en este trabajo aplican directamente a otras especies y a otros caracteres bajo la influencia de efectos maternos. En esta sección se discuten, en primer lugar, ciertas características de los diferentes archivos de datos empleados. Luego, se discuten en términos generales detalles de la implementación de los algoritmos de muestreo y de las estimaciones que se obtuvieron.

6.3.3.1. Archivos de datos

En general, se procuró que los archivos de datos que se utilizaron para implementar los algoritmos de inferencia fueran lo suficientemente informativos como para obtener estimaciones precisas de los diferentes CVC. Específicamente, se procuró que los datos contaran con características favorables respecto a la calidad de estimación de CVC bajo modelos con efectos maternos: fundamentalmente, y de acuerdo a los resultados de Gerstmayr (1992) y Maniatis y Pollott (2003), se aseguró un número promedio de crías

por madre no menor a dos, por un lado, y un porcentaje de madres con registro fenotípico mayor al 40%, por otro. El archivo de datos del rodeo Angus de Las Lilas, utilizado en tres capítulos diferentes de esta tesis, reunió características excepcionales en este sentido. Por su parte, el archivo experimental del rodeo Angus × Hereford neozelandés no sólo contó con una excelente estructura de información para la estimación de CVC bajo el MAM, sino que además disponía de información sumamente rica respecto a las composiciones raciales de los individuos en la población, un elemento clave del análisis multirracial. Finalmente, resta mencionar que la estructura poblacional generada en el estudio de simulación estocástica respondió estrechamente a estudios de esta naturaleza descriptos en la literatura, como el de Quintanilla *et al*. (1999).

Una pequeña digresión es importante en este punto. En todos los archivos de datos analizados se procuró que las madres de todos los individuos con registro fenotípico estuvieran identificadas. Además, no se incluyeron en los archivos datos de crías por transferencia embrionaria. Con respecto al primer punto, y como discutiera Henderson (1988), si existen madres sin identificar en el archivo de datos se viola el supuesto de homoscedasticidad en la varianza de error. Para salvar este inconveniente, entonces, sería necesario extender el vector de valores de cría para incorporar a las madres no identificadas como 'madres fantasmas' (*cf*. Cantet *et al*., 1992b). Con respecto al segundo punto, merece destacarse que la incorporación de registros por transferencia embrionaria requiere un conocimiento, como mínimo, de la raza y edad de la madre receptora para una correcta imputación de los efectos fijos del modelo. Por otro lado, la incorporación de registros de transferencia embrionaria enriquecería notablemente la estructura de información del archivo de datos (*e.g*. Meyer, 1992).

6.3.3.2. Estimaciones obtenidas

A continuación se discuten detalles de la implementación de los diferentes algoritmos de muestreo. En función de los objetivos planteados, en cada capítulo se ajustó el MAM correspondiente a los datos, y se obtuvieron luego estimaciones para todos los CVC utilizando los métodos de inferencia propuestos. En primer lugar, se obtuvieron estimaciones REML y sus correspondientes errores estándares mediante el paquete ASReml (Gilmour *et al*., 2006). En general, las estimaciones REML fueron utilizadas como una sentencia sobre la media de la distribución a priori de los CVC en los diferentes análisis bayesianos llevados a cabo. En algunos casos, se utilizaron también las estimaciones REML desviadas de sus respectivos errores estándares para especificar medias a priori sobredispersas.

En el Capítulo 2, y a modo de ilustración, se presentaron resultados obtenidos mediante dos estrategias diferentes con respecto a la implementación del GS. Como se discutiera en aquel capítulo, en la práctica existen básicamente dos alternativas para tomar una decisión sobre cómo implementar el algoritmo de muestreo: o bien ejecutar una cadena muy larga (*e.g*. Geyer, 1992) o bien ejecutar varias cadenas más cortas (*e.g*. Gelman y Rubin, 1992). En virtud de que los resultados expresados en términos de los parámetros genéticos fueron exactamente iguales, se concluyó que la información contenida en el archivo de pesos al destete del rodeo Angus de Las Lilas resultó adecuada para estimar todos los CVC inherentes al MAM. En general, si los datos proveen suficiente información, tanto la implementación del algoritmo como la distribución a priori tendrán poca influencia en los resultados. En modelos jerárquicos, sin embargo, mayor cantidad de datos no implica necesariamente mayor información (Blasco, 2001). Esto último puede ser particularmente importante al momento de estimar CVC en MAM so-

breparametrizados. En tal caso, se vuelve esencial contar con la información que proveen los contrastes entre diferentes parientes para identificar los parámetros de manera inequívoca.

En el Capítulo 3, por su parte, se presentó una estrategia para determinar los hiperparámetros de las distribuciones a priori de los CVC, basada en las propiedades de la distribución GIW y que surge de una práctica habitual en el proceso de ejecución de los programas de evaluación genética. La idea consiste básicamente en utilizar recursivamente estimaciones previas de los CVC para determinar los hiperparámetros en la siguiente ejecución, explotando así la "propiedad de 'memoria' del teorema de Bayes" (Gianola y Fernando, 1986). Comparada contra especificaciones a priori más estándares mediante un estudio de simulación estocástica, la estrategia produjo estimaciones precisas de los parámetros genéticos, con menores errores estándares y mejor tasa de convergencia. Sin embargo, aún es necesario encuadrar esta estrategia en un marco teórico más formal.

Por otro lado, en esta investigación se implementó por primera vez un algoritmo GGS para abordar un problema de estimación de CVC para datos de performance en especies ganaderas. El procedimiento fue ejecutado con éxito utilizando los datos del rodeo Angus de Las Lilas, y se obtuvo por primera vez una estimación con datos de campo del parámetro de correlación ambiental madre–progenie para el carácter peso al destete. En el Capítulo 4, se graficó la distribución marginal posterior estimada del parámetro, que resultó unimodal con una media cercana al cero, a partir de lo cual se dedujo que el modelo puso una enorme masa sobre valores positivos aunque pequeños del parámetro. Este resultado contradijo la expectativa original de que covarianzas de naturaleza ambiental entre pares de observaciones madre–progenie podrían sesgar la fuerte correlación genética estimada para este conjunto de datos en particular.

Finalmente, y tras implementar el MBAM con efectos maternos a los datos del cruzamiento experimental Angus × Hereford, en el Capítulo 5 se presentó la primera estimación del juego completo de nueve CVC genéticos que surge del modelo descripto por Cantet y Fernando (1995) siguiendo los argumentos de la teoría genética cuantitativa desarrollados por Lo *et al*. (1993). En general, las distribuciones marginales posteriores de los componentes de segregación resultaron fuertemente asimétricas en todos los casos. En particular, se obtuvo una estimación positiva de la covarianza genética directa-materna de segregación.

6.4. LÍNEAS DE INVESTIGACIÓN FUTURA

Los resultados que se obtuvieron a lo largo de estos años de trabajo han dejado abiertas nuevas líneas de investigación que podrían llevarse adelante a futuro. En esta sección se presentan algunos posibles caminos de acción, siempre en conexión con los resultados obtenidos en esta investigación.

6.4.1. Método GIW

En este trabajo se presentó la distribución GIW en toda su generalidad con la expectativa de motivar nuevas aplicaciones en investigaciones futuras, en particular aprovechando la flexibilidad que ofrece el mayor número de parámetros para describir el conocimiento a priori respecto a la distribución de una matriz de covarianza. Como fuera mencionado en el Capítulo 3, una aplicación directa implicaría utilizar la distribución GIW como una alternativa natural para especificar la estructura de covarianza a priori de ob-

servaciones que siguen una distribución normal multivariada con un patrón monótono de datos faltantes (Garthwaite y Al-Awadhi, 2001). Otra alternativa, sugerida por uno de los referís del artículo científico que derivara del capítulo, sería su aplicación en matrices residuales restringidas, en el contexto de modelos lineales (binarios) de umbral.

Por otro lado, también quedó abierta la posibilidad de aprovechar aún más la flexibilidad que ofrece la distribución GIW a través de la especificación de algún otro hiperparámetro. Al respecto, en este trabajo se modeló cierta incertidumbre diferencial sólo a través de los grados de credibilidad de los parámetros de Bartlett. Sin embargo, teniendo en cuenta el conjunto mayor de hiperparámetros disponible, este enfoque parece algo conservador. Ahora bien, esta última posibilidad debe distinguirse del problema de determinar valores para esos hiperparámetros. Por ejemplo, utilizar los grados de credibilidad para reflejar incertidumbre respecto a la media de una distribución a priori es un problema bien diferente a asignar valores numéricos a estos parámetros. Con relación a esto último, en el Capítulo 3 se presentó un método recursivo para determinar valores de los hiperparámetros de las distribuciones a priori de los CVC genéticos. Sin embargo, aún no se formalizó esta estrategia. Básicamente, existen dos enfoques que se pueden tomar al respecto: 1. tratar el problema desde el punto de vista de un método 'bayesiano empírico' (*empirical Bayes*) (*cf.* Casella, 2001); y 2. abordar el problema a través de algún algoritmo de actualización bayesiana, como, por ejemplo, el 'filtro de Kalman' (*Kalman filter*, *cf.* Meinhold y Singpurwalla, 1983). La evaluación de estos enfoques constituye todavía una tarea pendiente.

6.4.2. El algoritmo GGS

Aún más desafíos se plantean luego de analizar los resultados obtenidos tras estimar el parámetro de correlación ambiental madre–progenie mediante el algoritmo GGS. En primer lugar, la gran demanda computacional del algoritmo podría constituir una limitante insalvable en archivos de datos numerosos. En tal caso, una estrategia alternativa sería recurrir a algún otro método MCMC para obtener muestras de la distribución condicional posterior del parámetro. Por otro lado, el valor cercano a cero que se obtuvo en este trabajo para la media posterior del parámetro no es consistente con la expectativa de reducir la correlación genética directa-materna, que para los datos del rodeo Angus de Las Lilas es altísima. En contraste, algunos análisis preliminares que se están llevando a cabo actualmente con otro archivo de datos de peso al destete, correspondiente al rodeo Hereford de Las Lilas, devuelven consistentemente una estimación negativa, en torno a –0,2, para el parámetro de correlación. Sin embargo, las estimaciones de los otros CVC aparecen muy sensibles a las correspondientes medias de las distribuciones a priori especificadas, lo cual sugiere que existe algún problema de identificabilidad de los CVC.

6.4.3. MBAM con efectos maternos

Con respecto al MBAM con efectos maternos, por último, sería interesante evaluar la performance del algoritmo de muestreo con datos de campo, dado que éstos suelen ser más restrictivos en muchos aspectos. Sin embargo, por la forma en la que se suelen registrar los diferentes cruzamientos en una población multirracial, parece difícil a priori hallar archivos de datos con buena información respecto a las composiciones raciales. Actualmente, existen técnicas moleculares basadas en los microsatélites que se utilizan para los tests de paternidad que, en conjunto con algoritmos de recomposición (*peeling*) basados en la información genealógica, quizás puedan asistir en la determinación de las

composiciones raciales de todos los individuos pertenecientes a una población compuesta.

6.5. CONCLUSIÓN

En conclusión, en este trabajo se presentaron métodos novedosos de inferencia de CVC en modelos jerárquicos asociados al análisis de datos de performance para caracteres bajo efectos maternos. Basados en la construcción de modelos bayesianos jerárquicos, estos métodos son plausibles de ser implementados a través de algunas modificaciones más o menos sencillas del algoritmo del muestreo de Gibbs.

APÉNDICES

APÉNDICE A.

Resultados sobre las distribuciones condicionales posteriores de los parámetros de Bartlett

En la sección 3.2.2.2 del Capítulo 3 se expusieron una serie de resultados en conexión con las distribuciones condicionales posteriores de los parámetros de Bartlett que resultan de la descomposición de la matriz de covarianza genética $\mathbf{\Sigma}$ bajo el MAM. En este apéndice se presentan derivaciones detalladas de aquellos resultados.

Se comenzará con la distribución condicional posterior de τ tal como resulta de la expresión [3.21] tras ignorar todos los términos que no dependen de τ del argumento de la función exponencial en [3.16]:

$$p(\tau \mid \Gamma, \mathcal{H}, \mathcal{D}) \propto \\ \propto exp\left\{-\frac{\left(\tau^2 Q_{11} - 2\tau Q_{12}\right) + H^{-1}\left(\tau - \tau_0\right)^2}{2\Gamma}\right\}. \qquad [A.1]$$

Completando ahora el binomio cuadrado en el primer término de la función exponencial

$$\begin{aligned} \tau^2 Q_{11} - 2\tau Q_{12} &= Q_{11}\left(\tau^2 - 2\tau Q_{11}^{-1} Q_{12}\right) \\ &= Q_{11}\left[\tau^2 - 2\tau Q_{11}^{-1} Q_{12} + \left(Q_{11}^{-1} Q_{12}\right)^2 - \left(Q_{11}^{-1} Q_{12}\right)^2\right] \\ &= Q_{11}\left(\tau - Q_{11}^{-1} Q_{12}\right)^2 - Q_{11}^{-1} Q_{12}^2. \end{aligned} \qquad [A.2]$$

El ultimo término en [A.2] no depende de τ y, en consecuencia, es absorbido por la constante de integración. El siguiente paso involucra combinar las formas cuadráticas:

$$Q_{11}\left(\tau - Q_{11}^{-1} Q_{12}\right)^2 + H^{-1}\left(\tau - \tau_0\right)^2 \qquad [A.3]$$

Para lograrlo, se hará uso de la siguiente identidad (Sorensen y Gianola, 2002, pág. 227)

$$M\left(z-m\right)^2 + B\left(z-b\right)^2 = \left(M+B\right)\left(z-c\right)^2 + \frac{MB}{M+B}\left(m-b\right)^2, \qquad [A.4]$$

con $c = \left(M+B\right)^{-1}\left(M\,m + B\,b\right)$. Nótese entonces que definiendo $M = Q_{11}$, $m = Q_{12}Q_{11}^{-1}$, $B = H^{-1}$, $b = \tau_0$ y $z = \tau$, y luego descartando el segundo término en [A.4], dado que no depende de τ, se llega a la expresión [3.22], de donde se deduce que τ se distribuye condicionalmente como una variable normal univariada a posteriori.

Luego, se derivará la expresión para el parámetro de escala de la distribución condicional posterior de Γ, $\tilde{S}_1$, como fuera presentada en [3.28]. Tal como se mencionó oportunamente, este paso involucra recolectar todos los términos que no dependen de τ del argumento de la función exponencial en [3.16]. Estos son: 1. el parámetro de escala de la distribución a priori de Γ, S_1; 2. la forma cuadrática en los valores de cría maternos, Q_{22}; 3. el término $-Q_{11}^{-1}Q_{12}^2$, descartado al completar el binomio cuadrado en [A.2]; y 4. el segundo término a la derecha de la ecuación en [A.4], que luego del reemplazo de variables correspondiente y utilizando luego el conjunto de identidades definido en

[3.24] resulta igual a $WQ_{11}\left(\hat{\tau}-\tau_0\right)^2$. Ahora bien, operando con las formas cuadráticas, y tras sumar y restar $Q_{11}^{-1}Q_{12}^2$,

$$\begin{aligned} Q_{22}-2Q_{11}^{-1}Q_{12}^2+Q_{11}^{-1}Q_{12}^2 &= Q_{22}-2Q_{12}\hat{\tau}+Q_{12}\hat{\tau} \\ &= Q_{22}-2Q_{12}\hat{\tau}+Q_{11}\hat{\tau}^2 \\ &= \boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{a}_m-2\left(\boldsymbol{a}'_o\boldsymbol{A}^{-1}\boldsymbol{a}_m\right)\hat{\tau}+\boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{a}_o\hat{\tau}^2 \\ &= \left(\boldsymbol{a}_m-\boldsymbol{a}_o\hat{\tau}\right)^{\mathrm{T}}\boldsymbol{A}^{-1}\left(\boldsymbol{a}_m-\boldsymbol{a}_o\hat{\tau}\right), \end{aligned} \qquad [A.5]$$

y, en consecuencia,

$$\tilde{S}_1=\left(\boldsymbol{a}_m-\boldsymbol{a}_o\hat{\tau}\right)^{\mathrm{T}}\boldsymbol{A}^{-1}\left(\boldsymbol{a}_m-\boldsymbol{a}_o\hat{\tau}\right)+WQ_{11}\left(\hat{\tau}-\tau_0\right)^2+S_1. \qquad [A.6]$$

Fórmulas para la distribución condicional posterior de la matriz de covarianza en el caso más general de la descomposición de Bartlett en bloques múltiples pueden hallarse en los trabajos de Brown (2002) y Le *et al*. (1999).

APÉNDICE B.

Algoritmo de muestreo para la distribución GIW

Asúmase que el analista define una distribución IW a priori para $\boldsymbol{\Sigma}$ bajo un enfoque condicional conjugado. Como fuera mencionado en la sección 3.2.2.4 del Capítulo 3, la expresión [3.32] define el conjunto particular de hiperparámetros de la distribución GIW que equivalentemente recupera una muestra de la correspondiente distribución condicional posterior IW. En este apéndice se demuestra tal equivalencia y se presenta un algoritmo de muestreo para la distribución GIW.

Más específicamente, considérese un análisis bayesiano jerárquico del MAM vía el GS como fuera descripto en la sección 2.2.2 del Capítulo 2. En ese caso, la matriz de covarianza genética se muestrea de la distribución condicional posterior IW definida por $IW\left(\upsilon+q,\ \boldsymbol{Q}^{*}\right)$, con $\boldsymbol{Q}^{*}=\boldsymbol{Q}+\boldsymbol{S}$. Aquí, $\boldsymbol{Q}$ representa la matriz simétrica de sumas de cuadrados y productos cruzados definida en [3.17] mientras que $\boldsymbol{S}$ representa la matriz de escala de la distribución a priori de $\boldsymbol{\Sigma}$, en general parameterizada según $\boldsymbol{S}=\upsilon\boldsymbol{S}^{*}$, donde $\boldsymbol{S}^{*}$ representa una matriz a priori de valores 'razonables' para los CVC genéticos y υ son los grados de credibilidad en esos valores a priori.

Por otro lado, utilizando el conjunto de hiperparámetros definido en [3.32] en las ecuaciones [3.20], [3.25] y [3.27] se obtienen las siguientes distribuciones condicionales posteriores de los parámetros de Bartlett

$$\begin{aligned}
&\Sigma_{11}\,|\,\mathcal{H},\mathcal{D}\sim Q_{11}^{*}\chi_{\tilde{\upsilon}_0}^{-2},\\
&\tau\,|\,\Gamma,\mathcal{H},\mathcal{D}\sim N\left(Q_{11}^{*-1}Q_{12}^{*},\,Q_{11}^{*-1}\Gamma\right),\\
&\Gamma\,|\,\mathcal{H},\mathcal{D}\sim\left(Q_{22}^{*}-Q_{11}^{*-1}Q_{12}^{*2}\right)\chi_{\tilde{\upsilon}_1+1}^{-2},
\end{aligned}\qquad [B.1]$$

con $\tilde{\upsilon}_0=\upsilon_0+q=\upsilon+q+1$ y $\tilde{\upsilon}_1=\upsilon_1+q=\upsilon+q$.

Ahora bien, para probar la equivalencia entre ambas estrategias de muestreo se mostrará a continuación que el producto de los 'núcleos' (*kernels*, en inglés) de las tres distribuciones en [B.1] corresponde al núcleo de la distribución IW apropiada. Explícitamente, y luego de reemplazar los parámetros de Bartlett con los correspondientes elementos de la matriz de covarianza $\boldsymbol{\Sigma}$, la multiplicación resulta

$$\begin{aligned}
&p\left(\Sigma_{11}\,|\,\mathcal{H},\mathcal{D}\right)\times p\left(\tau\,|\,\Gamma,\mathcal{H},\mathcal{D}\right)\times p\left(\Gamma\,|\,\mathcal{H},\mathcal{D}\right)\propto\\
&\propto\left(\Sigma_{11}\right)^{-\frac{1}{2}\left[(\upsilon+q+1)+2\right]}\times\left(\Sigma_{22}-\Sigma_{11}^{-1}\Sigma_{12}^{2}\right)^{-\frac{1}{2}\left[(\upsilon+q)+1+2\right]}\times\\
&\times exp\left\{-\frac{1}{2}\left[\frac{Q_{11}^{*}}{\Sigma_{11}}+\frac{Q_{11}^{*}\left(\Sigma_{11}^{-1}\Sigma_{12}-Q_{11}^{*-1}Q_{12}^{*}\right)^{2}}{\left(\Sigma_{22}-\Sigma_{11}^{-1}\Sigma_{12}^{2}\right)}+\frac{\left(Q_{22}^{*}-Q_{11}^{*-1}Q_{12}^{*2}\right)}{\left(\Sigma_{22}-\Sigma_{11}^{-1}\Sigma_{12}^{2}\right)}\right]\right\}.
\end{aligned}\qquad [B.2]$$

Nótese primero que

$$\begin{aligned}
\left(\Sigma_{11}\right)^{-\frac{1}{2}\left[(\upsilon+q+1)+2\right]}\left(\Sigma_{22}-\Sigma_{11}^{-1}\Sigma_{12}^{2}\right)^{-\frac{1}{2}\left[(\upsilon+q)+1+2\right]}&=\left(\Sigma_{11}\Sigma_{22}-\Sigma_{21}\Sigma_{12}\right)^{-\frac{1}{2}(\upsilon+q+3)}\\
&=\left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}(\upsilon+q+3)}.
\end{aligned}\qquad [B.3]$$

Luego, trabajando las funciones exponenciales, se llega a

$$
\begin{aligned}
&exp\left\{-\frac{1}{2}\left[\frac{Q_{11}^{*}}{\Sigma_{11}}+\frac{Q_{11}^{*}\left(\Sigma_{11}^{-1}\Sigma_{12}-Q_{11}^{*-1}Q_{12}^{*}\right)^{2}+\left(Q_{22}^{*}-Q_{11}^{*-1}Q_{12}^{*2}\right)}{\Sigma_{22}-\Sigma_{11}^{-1}\Sigma_{12}^{2}}\right]\right\}= \\
&=exp\left\{-\frac{1}{2}\left[\frac{Q_{11}^{*}}{\Sigma_{11}}+\frac{Q_{11}^{*}\Sigma_{11}^{-2}\Sigma_{12}^{2}-2Q_{12}^{*}\Sigma_{11}^{-1}\Sigma_{12}+Q_{22}^{*}}{\Sigma_{22}-\Sigma_{11}^{-1}\Sigma_{12}^{2}}\right]\right\}= \\
&=exp\left\{-\frac{1}{2}\left[\left(\frac{Q_{11}^{*}\Sigma_{22}-Q_{12}^{*}\Sigma_{12}}{|\boldsymbol{\Sigma}|}\right)+\left(\frac{Q_{22}^{*}\Sigma_{11}-Q_{12}^{*}\Sigma_{12}}{|\boldsymbol{\Sigma}|}\right)\right]\right\}= \\
&=exp\left\{-\frac{1}{2}\left[\left(Q_{11}^{*}\Sigma^{11}+Q_{12}^{*}\Sigma^{12}\right)+\left(Q_{22}^{*}\Sigma^{22}-Q_{12}^{*}\Sigma^{12}\right)\right]\right\}= \\
&=exp\left\{-\frac{1}{2}tr\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}^{*}\right)\right\},
\end{aligned}
\qquad \text{[B.4]}
$$

donde Σ^{ij} simboliza el elemento (i, j) de la matriz $\boldsymbol{\Sigma}^{-1}$.

Los resultados [B.3] y [B.4] muestran que la expresión [B.2] puede reconocerse como el núcleo de una distribución $IW\left(\upsilon+q,\ \boldsymbol{Q}^{*}\right)$. Entonces, la densidad IW a priori para la matriz de covarianza genética bajo el MAM puede considerarse un caso especial de una distribución GIW tras definir el conjunto específico de hiperparámetros en [3.32]. Es más, nótese que la ecuación [B.1] sugiere un algoritmo de muestreo. Asúmase que el analista desea especificar valores diferentes para los parámetros υ_0 y υ_1 de modo de reflejar incertidumbre diferencial a priori. Entonces, el siguiente algoritmo tomará muestras de la distribución condicional posterior de la matriz de covarianza genética $\boldsymbol{\Sigma}$:

1*a*. Definir υ_0 y υ_1, y formar luego la matriz $\boldsymbol{\Sigma}^{*}$ en forma secuencial con los siguientes elementos:

$$\Sigma_{11}^{*}=\left(\upsilon_0+2\right)S_{11}^{*}.$$

$$\Sigma_{12}^{*}=\Sigma_{21}^{*}=\left(S_{12}^{*}S_{11}^{*-1}\right)\Sigma_{11}^{*}.$$

$$\Sigma_{22}^{*}=\left(\upsilon_1+3\right)\left(S_{22}^{*}-S_{12}^{*2}S_{11}^{*-1}\right)+\left(\Sigma_{12}^{*2}\Sigma_{11}^{*-1}\right).$$

1*b*. Computar $\tilde{\upsilon}_0=q+\upsilon_0$ y $\tilde{\upsilon}_1=q+\upsilon_1$.

2. Formar la matriz $\boldsymbol{Q}^{*}=\boldsymbol{Q}+\boldsymbol{\Sigma}^{*}$.

3. Muestrear Σ_{11} de $\Sigma_{11}\,|\,\mathcal{H},\mathcal{D}\sim Q_{11}^{*}\chi_{\tilde{\upsilon}_0}^{-2}$.

4. Muestrear Γ de $\Gamma\,|\,\mathcal{H},\mathcal{D}\sim\left(Q_{22}^{*}-Q_{11}^{*-1}Q_{12}^{*2}\right)\chi_{\tilde{\upsilon}_1+1}^{-2}$.

5 Muestrear τ de $\tau\,|\,\Gamma,\mathcal{H},\mathcal{D}\sim N\left(Q_{11}^{*-1}Q_{12}^{*},\,Q_{11}^{*-1}\Gamma\right)$.

6. Recuperar la matriz $\boldsymbol{\Sigma}$ calculando $\boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{11} & \Sigma_{11}\tau \\ \tau\Sigma_{11} & \Gamma + \tau^2\Sigma_{11} \end{bmatrix}$.

7. Repetir los pasos 2–6 dentro de cada ciclo del algoritmo GS.

La definición de la matriz de escala a priori en el paso 1*a* del algoritmo de muestreo se basa en utilizar valores 'razonables' para los CVC a priori (los elementos de la matriz $\boldsymbol{S}^*$) como una sentencia sobre el modo de la distribución de los parámetros de Bartlett, y luego resolver para cada elemento de la matriz $\boldsymbol{\Sigma}^*$. En particular, definiendo $\upsilon_0 = \upsilon + 1$ y $\upsilon_1 = \upsilon$ el algoritmo tomará una muestra de una distribución IW, con matriz de escala a priori igual a $\boldsymbol{\Sigma}^* = (\upsilon + 3)\boldsymbol{S}^*$.

APÉNDICE C.

Cómputo de las MME para observaciones ordenadas por familias maternas

En el Capítulo 4 se presentó una formulación alternativa del MAM que incluye un parámetro de correlación en la estructura de covarianza del error. Bajo esta formulación alternativa, la construcción de las MME demanda un esfuerzo computacional mucho mayor, dado que involucra invertir la matriz de covarianza del error, $\boldsymbol{R}$, y computar luego expresiones en los elementos de dicha matriz. Sin embargo, como se describe en la Sección 4.2.2.1, cuando los registros se agrupan por familias maternas, $\boldsymbol{R}^{-1}$ presenta una estructura diagonal en bloques que permite reducir considerablemente el tiempo de cómputo.

En conexión con esto último, el objetivo particular de este apéndice es describir en detalle una forma eficiente de calcular las expresiones $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{W}$ y $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{y}$. Se comenzará por la primera de ellas. Nótese que, de acuerdo a [4.16], los cálculos pueden descomponerse en (mf +1) pasos, que involucran contribuciones asociadas a cada una de las familias maternas. En particular, se puede calcular un elemento arbitrario (s, t) de $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{W}$ empleando la siguiente fórmula:

$$\sum_{k=1}^{mf+1}\left[\sum_{j=1}^{f_k} w_{k\left(\sum_{m=1}^{k-1} f_m + j,\,t\right)}\left(\sum_{i=1}^{f_k} w_{k\left(\sum_{m=1}^{k-1} f_m + i,\,s\right)} r_k^{ij}\right)\right], \qquad [C.1]$$

donde $w_{k(i,j)}$ es el elemento (i, j) de la matriz $\boldsymbol{W}_k$ y r_k^{ij} es el elemento (i, j) de la matriz $\boldsymbol{R}_k^{-1}$. Ya dentro del k-ésimo paso, se puede rescribir la expresión dentro del corchete en [C.1] según

$$\sum_{j=1}^{f_k}\sum_{i=1}^{f_k} w_{\left(\sum_{m=1}^{k-1} f_m + j\right),t}\, w_{\left(\sum_{m=1}^{k-1} f_m + i\right),s}\, r_k^{ij}. \qquad [C.2]$$

De [C.2] se deduce que cualquier elemento de $\boldsymbol{W}_k^{\mathrm{T}}\boldsymbol{R}_k^{-1}\boldsymbol{W}_k$ puede computarse como una suma ponderada de todos los elementos de la matriz $\boldsymbol{R}_k^{-1}$. Cada uno de los factores de ponderación, por su parte, será el producto de dos elementos de la matriz $\boldsymbol{W}$. Este resultado sugiere un algoritmo de cómputo del sistema basado en contribuciones secuenciales (véanse los algoritmos de Groeneveld y Kovac, 1990, y las notas de clase de Misztal, 2006). Esto resultará más claro cambiando la óptica del problema. En este punto, se seguirán los desarrollos de Misztal (2006). Primero, se trabajará la matriz de de incidencia de los efectos fijos, $\boldsymbol{X}$, de orden $n \times p$. Luego, se extenderán los resultados a las matrices de incidencia de los efectos aleatorios. Finalmente, se describirá el cómputo secuencial del vector $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{y}$.

C.1. Matrices $\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{R}_k^{-1}\boldsymbol{X}_k$

Considérese el k-ésimo paso en la construcción de la matriz de coeficientes del sistema de ecuaciones. Defínase, en este contexto, la matriz $\boldsymbol{X}_k$ como la submatriz de orden $f_k \times p$ de la matriz de incidencia de los efectos fijos (obsérvese que $\boldsymbol{X}_k$ es el bloque de $\boldsymbol{W}_k$ correspondiente a los p efectos fijos). Luego, considérese una descomposición de esta

submatriz en f_k términos, cada uno de los cuales corresponde a una matriz nula, excepto por la *i*-ésima fila de $\boldsymbol{X}_k$, que denotaremos $\boldsymbol{x}_{k,i}^{\mathrm{T}}$. De modo que

$$\boldsymbol{X}_k = \sum_{i=1}^{f_k} \boldsymbol{X}_{k,i}. \qquad [C.3]$$

De [C.3] se desprende que

$$\begin{aligned}
\boldsymbol{X}_k^{\mathrm{T}} \boldsymbol{R}_k^{-1} \boldsymbol{X}_k &= \left(\sum_{i=1}^{f_k} \boldsymbol{X}_{k,i}^{\mathrm{T}} \right) \boldsymbol{R}_k^{-1} \left(\sum_{j=1}^{f_k} \boldsymbol{X}_{k,j} \right) \\
&= \sum_{i=1}^{f_k} \sum_{j=1}^{f_k} \boldsymbol{X}_{k,i}^{\mathrm{T}} \boldsymbol{R}_k^{-1} \boldsymbol{X}_{k,j} \\
&= \sum_{i=1}^{f_k} \sum_{j=1}^{f_k} \boldsymbol{x}_{k,i} \boldsymbol{x}_{k,j}^{\mathrm{T}} r_k^{ij} \\
&= \sum_{i=1}^{f_k} \boldsymbol{x}_{k,i} \boldsymbol{x}_{k,i}^{\mathrm{T}} r_k^{ii} + \sum_{j \neq i} \boldsymbol{x}_{k,i} \boldsymbol{x}_{k,j}^{\mathrm{T}} r_k^{ij}.
\end{aligned} \qquad [C.4]$$

En [C.4] puede observarse más claramente cómo es el patrón de contribuciones. En primer lugar, nótese que para construir $\boldsymbol{X}_k^{\mathrm{T}} \boldsymbol{R}_k^{-1} \boldsymbol{X}_k$ es necesario leer y almacenar las líneas de datos de todos los individuos dentro de la familia materna, dado que algunas contribuciones involucran las líneas de datos de individuos diferentes, pero siempre dentro de la familia. Luego, el problema se reduce a direccionar apropiadamente elementos de $\boldsymbol{R}_k^{-1}$ multiplicados por un coeficiente que dependerá de la naturaleza del efecto fijo involucrado (*i.e.*, categórico o covariable). En total, para *t* efectos fijos, cada bloque $\boldsymbol{X}_k^{\mathrm{T}} \boldsymbol{R}_k^{-1} \boldsymbol{X}_k$ será construido a través de $t^2 \times f_k^2$ contribuciones secuenciales, o más bien a través de $\frac{1}{2}(t \times f_k)\left[(t \times f_k)+1\right]$ contribuciones si sólo se almacena la porción triangular superior de la matriz de coeficientes.

C.2. Matrices de incidencia de los efectos aleatorios: $\boldsymbol{Z}_{o,k}^{\mathrm{T}} \boldsymbol{R}_k^{-1} \boldsymbol{Z}_{o,k}$, $\boldsymbol{Z}_{m,k}^{\mathrm{T}} \boldsymbol{R}_k^{-1} \boldsymbol{Z}_{m,k}$ y $\boldsymbol{Z}_{p,k}^{\mathrm{T}} \boldsymbol{R}_k^{-1} \boldsymbol{Z}_{p,k}$

Dada la naturaleza de estas matrices de incidencia, donde cada fila es un vector nulo excepto por un "1" en la posición correspondiente a la identificación del individuo (recuérdese que los individuos deben estar identificados en forma única), este caso es una simplificación del caso anterior. En general, el individuo identificado con el número *s*, y cuya observación es la *i*-ésima dentro de la familia materna *k*, contribuirá con r_k^{ii} al elemento diagonal (*s*, *s*) y con r_k^{ij} al elemento (*s*, *t*), donde *t* es la columna correspondiente al individuo identificado con dicho número, y que se ubica en la *j*-ésima posición dentro de la familia materna. Nótese que no es necesario asumir que los individuos estén ordenados dentro de familia materna. De hecho, se asume que los individuos están ordenados de acuerdo a su identificación numérica, para facilitar la contribución de los efectos aleatorios a la matriz de coeficientes.

C.3. Vectores $\boldsymbol{X}_k^{\mathrm{T}} \boldsymbol{R}_k^{-1} \boldsymbol{y}_k$

De manera análoga al caso de $\boldsymbol{W}^{\mathrm{T}} \boldsymbol{R}^{-1} \boldsymbol{W}$, el cómputo del vector $\boldsymbol{W}^{\mathrm{T}} \boldsymbol{R}^{-1} \boldsymbol{y}$ puede descomponerse en (*mf* +1) pasos, que involucran contribuciones asociadas a cada una de las familias maternas. En este caso, las contribuciones de los efectos fijos en el *k*-ésimo paso serán:

$$\begin{aligned}\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{R}_k^{-1}\boldsymbol{y}_k &= \left(\sum_{i=1}^{f_k}\boldsymbol{X}_{k,i}^{\mathrm{T}}\right)\boldsymbol{R}_k^{-1}\boldsymbol{y}_k \\ &= \sum_{i=1}^{f_k}\boldsymbol{X}_{k,i}^{\mathrm{T}}\boldsymbol{R}_k^{-1}\boldsymbol{y}_k \\ &= \sum_{i=1}^{f_k}\boldsymbol{X}_{k,i}^{\mathrm{T}}\boldsymbol{r}_{k,i}^{\mathrm{T}}\boldsymbol{y}_k \\ &= \sum_{i=1}^{f_k}\boldsymbol{x}_{k,i}\sum_{j=1}^{f_k}r_k^{ij}y_{k,j}.\end{aligned} \qquad [C.5]$$

Es decir, por cada uno de los *t* efectos fijos, el *i*-ésimo individuo dentro de la familia materna contribuirá con el valor del producto cruzado de la *i*-ésima fila en $\boldsymbol{R}_k^{-1}$ por el subvector de las observaciones dentro de la familia, $\boldsymbol{y}_k$, multiplicado, a su vez, por el coeficiente del nivel asociado. En total, entonces, cada bloque $\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{R}_k^{-1}\boldsymbol{y}_k$ podrá ser construido a través de $t\times f_k^2$ contribuciones secuenciales. Las contribuciones asociadas a los efectos aleatorios, por su parte, constituyen un caso especial de [C.5].

APÉNDICE D.

Distribuciones condicionales posteriores bajo el MBAM con efectos maternos

En el Capítulo 5 se presentó una extensión del MBAM de García-Cortés y Toro (2006) para acomodar efectos maternos. En este apéndice se presentan las distribuciones condicionales posteriores de todos los parámetros del modelo en el contexto de un análisis bayesiano jerárquico. Los pasos necesarios para derivar estas distribuciones son exactamente los mismos que aquellos descriptos en el Capítulo 2, siguiendo de cerca los trabajos de Jensen *et al*. (1994) y el libro de Sorensen y Gianola (2002).

Se partirá de la expresión de la distribución posterior conjunta en [5.14]. Defínase, en primer lugar, el vector de parámetros de posición según $\boldsymbol{\theta}^{\mathrm{T}} \equiv \left(\boldsymbol{b}^{\mathrm{T}}, \boldsymbol{a}_A^{*\mathrm{T}}, \boldsymbol{a}_B^{*\mathrm{T}}, \boldsymbol{a}_S^{*\mathrm{T}}, \boldsymbol{e}_m^{\mathrm{T}}\right)$. Luego, la distribución condicional posterior de $\boldsymbol{\theta}$ es proporcional a

$$\begin{aligned} & p\left(\boldsymbol{\theta} \mid \boldsymbol{\Sigma}_{X=\{A,B,S\}}, \sigma_{e_m}^2, \sigma_{e_o}^2, \boldsymbol{y}\right) \propto \\ & \quad \propto p\left(\boldsymbol{y} \mid \boldsymbol{b}, \boldsymbol{a}_X^*, \boldsymbol{e}_m, \sigma_{e_o}^2\right) \times p\left(\boldsymbol{b} \mid \boldsymbol{K}\right) \times \\ & \quad \times p\left(\boldsymbol{e}_m \mid \sigma_{e_m}^2\right) \times \prod_{X=\{A,B,S\}} p\left(\boldsymbol{a}_X^* \mid \boldsymbol{A}_X^*, \boldsymbol{\Sigma}_X\right). \end{aligned} \qquad [D.1]$$

Explícitamente,

$$\begin{aligned} & p\left(\boldsymbol{\theta} \mid \boldsymbol{\Sigma}_{X=\{A,B,S\}}, \sigma_{e_m}^2, \sigma_{e_o}^2, \boldsymbol{y}\right) \propto \\ & \quad \propto exp\left\{-\frac{\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}}{2\sigma_{e_o}^2}\right\} \times exp\left\{-\left(\tfrac{1}{2}\right)\boldsymbol{b}^{\mathrm{T}}\boldsymbol{K}^{-1}\boldsymbol{b}\right\} \times \\ & \quad \times exp\left\{-\frac{\boldsymbol{e}_m^{\mathrm{T}}\boldsymbol{e}_m}{2\sigma_{e_m}^2}\right\} \times \prod_{X=\{A,B,S\}} \left[exp\left\{-\frac{\boldsymbol{a}_X^{*\mathrm{T}}\left(\boldsymbol{\Sigma}_X^{-1} \otimes \boldsymbol{A}_X^{*-1}\right)\boldsymbol{a}_X^*}{2\sigma_{e_o}^2}\right\}\right]. \end{aligned} \qquad [D.2]$$

Operando algebraicamente (*e.g*. Jensen *et al*., 1994) puede demostrarse que

$$p\left(\boldsymbol{\theta} \mid \boldsymbol{\Sigma}_{X=\{A,B,S\}}, \sigma_{e_m}^2, \sigma_{e_o}^2, \boldsymbol{y}\right) \sim NMV\left(\hat{\boldsymbol{\theta}}, \boldsymbol{C}^{-1}\sigma_{e_o}^2\right). \qquad [D.3]$$

En [D.3], $\hat{\boldsymbol{\theta}} = \boldsymbol{C}^{-1}\boldsymbol{r}$ es la solución a las MME del modelo [5.10], $\boldsymbol{C}^{-1}$ la inversa de la correspondiente matriz de coeficientes y $\boldsymbol{r}$ es el vector de 'términos a la derecha de las ecuaciones' (*right hand side*). A diferencia del modelo de García-Cortés y Toro (2006), el sistema de ecuaciones que deriva de [5.10] admite inversa. Nótese además que a la diagonal correspondiente a cada efecto fijo es necesario sumarle k_i^{-1}, donde k_i es la cantidad a través de la cual se refleja gran incertidumbre a priori respecto al valor de los efectos fijos.

Por su parte, la distribución condicional posterior de la varianza del error es proporcional a

$$\begin{aligned} & p\left(\sigma_{e_o}^2 \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}_{X=\{A,B,S\}}, \sigma_{e_m}^2, \boldsymbol{y}\right) \propto \\ & \quad \propto p\left(\boldsymbol{y} \mid \boldsymbol{b}, \boldsymbol{a}_X^*, \boldsymbol{e}_m, \sigma_{e_o}^2\right) \times p\left(\sigma_{e_o}^2 \mid \upsilon_{e_o}, S_{e_o}^2\right). \end{aligned} \qquad [D.4]$$

Explícitamente,

$$p\left(\sigma^2_{e_o} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}_{X=\{A,B,S\}}, \sigma^2_{e_m}, \boldsymbol{y}\right) \propto$$
$$\propto \left(\sigma^2_{e_o}\right)^{-\frac{1}{2}\left(\upsilon_{e_o}+n+2\right)} exp\left\{-\frac{\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}+\upsilon_{e_o}S^2_{e_o}}{2\sigma^2_{e_o}}\right\}, \qquad [D.5]$$

con $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{Xb} - \sum_{X=\{A,B,S\}} \boldsymbol{Z}_X \boldsymbol{a}^*_X - \boldsymbol{Z}_p \boldsymbol{e}_p$. Defínase luego

$$\tilde{S}^2_{e_o} = \frac{\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}+\upsilon_{e_o}S^2_{e_o}}{\tilde{\upsilon}_{e_o}}, \quad \text{con } \tilde{\upsilon}_{e_o} = \upsilon_{e_o} + n. \qquad [D.6]$$

Entonces

$$p\left(\sigma^2_{e_o} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}_{X=\{A,B,S\}}, \sigma^2_{e_m}, \boldsymbol{y}\right) \propto$$
$$\propto \left(\sigma^2_{e_o}\right)^{-\frac{1}{2}\left(\tilde{\upsilon}_{e_o}+2\right)} \exp\left\{-\frac{\tilde{\upsilon}_{e_o}\tilde{S}^2_{e_o}}{2\sigma^2_{e_o}}\right\}. \qquad [D.7]$$

Por inspección, la expresión [D.7] corresponde al núcleo de una distribución Chi-cuadrada invertida con parámetros $\tilde{\upsilon}_{e_o}$ y $\tilde{\upsilon}_{e_o}\tilde{S}^2_{e_o}$. En consecuencia

$$\sigma^2_{e_o} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}_{X=\{A,B,S\}}, \sigma^2_{e_m}, \boldsymbol{y} \sim \tilde{\upsilon}_{e_o}\tilde{S}^2_{e_o}\chi^{-2}_{\tilde{\upsilon}_{e_o}}. \qquad [D.8]$$

Argumentos similares permiten derivar la distribución condicional posterior de la varianza de los efectos ambientales maternos permanentes:

$$p\left(\sigma^2_{e_m} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}_{X=\{A,B,S\}}, \sigma^2_{e_o}, \boldsymbol{y}\right) \propto$$
$$\propto p\left(\boldsymbol{e}_m \mid \sigma^2_{e_m}\right) \times p\left(\sigma^2_{e_m} \mid \upsilon_{e_m}, S^2_{e_m}\right). \qquad [D.9]$$

Explícitamente,

$$p\left(\sigma^2_{e_m} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}_{X=\{A,B,S\}}, \sigma^2_{e_o}, \boldsymbol{y}\right) \propto$$
$$\propto \left(\sigma^2_{e_m}\right)^{-\frac{1}{2}\left(\upsilon_{e_p}+d+2\right)} exp\left\{-\frac{\boldsymbol{e}^{\mathrm{T}}_m\boldsymbol{e}_m+\upsilon_{e_m}S^2_{e_m}}{2\sigma^2_{e_m}}\right\}. \qquad [D.10]$$

Definiendo

$$\tilde{S}^2_{e_m} \equiv \frac{\boldsymbol{e}^{\mathrm{T}}_m\boldsymbol{e}_m+\upsilon_{e_m}S^2_{e_m}}{\tilde{\upsilon}_{e_m}}, \text{con } \tilde{\upsilon}_{e_m} \equiv \upsilon_{e_m} + d, \qquad [D.11]$$

entonces

$$p\left(\sigma^2_{e_m} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}_{X=\{A,B,S\}}, \sigma^2_{e_o}, \boldsymbol{y}\right) \propto$$
$$\propto \left(\sigma^2_{e_m}\right)^{-\frac{1}{2}\left(\tilde{\upsilon}_{e_m}+2\right)} exp\left\{-\frac{\tilde{\upsilon}_{e_m}\tilde{S}^2_{e_m}}{2\sigma^2_{e_m}}\right\}. \qquad [D.12]$$

Por inspección, la expresión [D.12] corresponde al núcleo de una distribución Chi-cuadrada invertida con parámetros $\tilde{\upsilon}_{e_m}$ y $\tilde{\upsilon}_{e_m}\tilde{S}^2_{e_m}$. En consecuencia,

$$\sigma^2_{e_m} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}_{X=\{A,B,S\}}, \sigma^2_{e_o}, \boldsymbol{y} \sim \tilde{\upsilon}_{e_m}\tilde{S}^2_{e_m}\chi^{-2}_{\tilde{\upsilon}_{e_m}}. \qquad [D.13]$$

Finalmente, resta obtener la distribución condicional posterior de las matrices de covarianza genética de los valores de cría por origen racial. Por ejemplo, para la componente correspondiente a la raza *A*, esta distribución es proporcional a

$$p\left(\boldsymbol{\Sigma}_A \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}_{R=\{B,S\}}, \sigma^2_{e_m}, \sigma^2_{e_o}, \boldsymbol{y}\right) \propto$$
$$\propto p\left(\boldsymbol{a}^*_A \mid \boldsymbol{A}^*_A, \boldsymbol{\Sigma}_A\right) \times p\left(\boldsymbol{\Sigma}_A \mid \upsilon_A, \boldsymbol{S}_A\right). \qquad [D.14]$$

En [D.14], $\boldsymbol{\Sigma}_R$ representa a las matrices de covarianza genética de las otras fuentes de variabilidad genética. Bajo la distribución condicional, éstas matrices se tratan como constantes. Explícitamente, entonces,

$$p\left(\boldsymbol{\Sigma}_A \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}_{R=\{B,S\}}, \sigma^2_{e_m}, \sigma^2_{e_o}, \boldsymbol{y}\right) \propto$$
$$\propto \left|\boldsymbol{\Sigma}_A\right|^{-\frac{1}{2}\left(q_A+\upsilon_A+3\right)} exp\left\{-\left(\tfrac{1}{2}\right) tr\left[\boldsymbol{\Sigma}^{-1}_A\left(\boldsymbol{Q}_A+\boldsymbol{S}_A\right)\right]\right\}. \qquad [D.15]$$

Esta expresión corresponde al núcleo de una distribución Wishart invertida; *i.e.*,

$$\boldsymbol{\Sigma}_A \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}_{R=\{B,S\}}, \sigma^2_{e_m}, \sigma^2_{e_o}, \boldsymbol{y} \sim IW\left(\upsilon_A+q_A, \boldsymbol{Q}_A+\boldsymbol{S}_A\right). \qquad [D.16]$$

Nótese que este resultado se extiende a las matrices de covarianza genética de las otras fuentes de variabilidad genética.

BIBLIOGRAFÍA

Baker, R. L. 1980. The role of maternal effects on the efficiency of selection in beef cattle: a review. *Proc. N.Z. Soc. Anim. Prod.*, 40: 285–303.

Bartlett, M. S. 1933. On the theory of statistical regression. *Proc. Roy. Soc. Edinburgh*, 53: 260–283.

Bauwens, L., M. Lubrano y J. F. Richard. 1999. *Bayesian inference in dynamic econometric models*. Oxford University Press, New York, USA.

Beef Improvement Federation (BIF). 2002. *Guidelines for Uniform Beef Improvement Programs*, 8th Ed. University of Georgia, Athens, USA.

Bijma, P. 2006. Estimating maternal genetic effects in livestock. *J. Anim. Sci*., 84: 800–806.

Birchmeier, A. N., R. J. C. Cantet, R. L. Fernando, C. A. Morris, F. Holgado, A. Jara y M. Santos Cristal. 2002. Estimation of segregation variance for birth weight in beef cattle. *Livest. Prod. Sci*., 76: 27–35.

Blasco, A. 2001. The Bayesian controversy in animal breeding. *J. Anim. Sci*., 79: 2023–2046.

Bondari, K., R. L. Willham y A. E. Freeman. 1978. Estimates of direct and maternal genetic correlations for pupa weight and family size of *Tribolium*. *J. Anim. Sci*., 47: 358–365.

Box, G. E. P. y G. C. Tiao. 1973. *Bayesian inference in statistical analysis*. Addison-Wesley Publishing Co, Reading, MA, USA.

Brown, P. J. 2002. The generalized inverted Wishart distribution. Pp. 1079–1083 en El-Shaarawi, A. H. y W. W. Piegorsch (Eds.) *Encyclopaedia of environmetrics*. Wiley, UK.

Brown, P. J., N. D. Le y J. V. Zidek. 1994. Inference for a covariance matrix. Pp. 77–92 en Freeman, P. R. y A. F. M. Smith (Eds.) *Aspects of uncertainty: a tribute to D. V. Lindley*, Wiley, New York, USA.

Bulmer, M. G. 1980. *The mathematical theory of quantitative genetics*. Clarendon Press, Oxford, UK.

Cantet, R. J. C. 1990. Estimation and prediction problems in mixed linear models for maternal genetics effects. *Ph.D Thesis*. University of Illinois, Urbana.

Cantet, R. J. C. y R. L. Fernando. 1995. Prediction of breeding values with additive animal models for crosses from two populations. *Genet. Sel. Evol*., 27: 323–334.

Cantet, R. J. C., D. D. Kress, D. C. Anderson, D. E. Doornbos, P. J. Burfening y R. L. Blackwell. 1988. Direct and maternal variances and covariances and maternal phenotypic effects on preweaning growth of beef cattle. *J. Anim. Sci*., 66: 648–660.

Cantet, R. J. C., R. L. Fernando y D. Gianola. 1992a. Bayesian inference about dispersion parameters of univariate mixed models with maternal effects: theoretical considerations. *Genet. Sel. Evol*., 24: 107–135.

Cantet, R. J. C., L. R. Schaeffer y C. Smith. 1992b. Reduced animal model with differential genetic grouping for direct and maternal effects. *J. Anim. Sci*., 70: 1730–1741.

Cantet, R. J. C., A. N. Birchmeier y J. P. Steibel. 2004. Full conjugate analysis of normal multiple traits with missing records using a generalized inverted Wishart. *Genet. Sel. Evol*., 36: 49–64.

Cardoso, F. F. y R. J. Tempelman. 2004. Hierarchical Bayes multiple-breed inference with an application to genetic evaluation of a Nelore-Hereford population. *J. Anim. Sci*., 82: 1589–1601.

Casella, G. 2001. Empirical Bayes Gibbs sampling. *Biostatistics*, 2: 485–500.

Casella, G. y E. I. George. 1992. Explaining the Gibbs Sampler. *The Am. Stat*., 46: 167–174.

Cockerham, C. C. 1954. An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics*, 39: 859–882.

CSIRO, 2010. *AAABG Genetic Parameters*. En http://www.gparm.csiro.au/index.html. Último acceso: enero de 2010.

Daniels, M. J. y M. Pourahmadi. 2002. Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika*, 89: 553–566.

Dodenhoff, J., L. D. Van Vleck, S. D. Kachman y R. M. Koch. 1998. Parameter estimates for direct, maternal, and grandmaternal genetic effects for birth weight and weaning weight in Hereford cattle. *J. Anim. Sci*., 76: 2521–2527.

Eaglen, S. A. E. y P. Bijma. 2009. Genetic parameters of direct and maternal effects for calving ease in Dutch Holstein-Friesian cattle. *J. Dairy Sci*., 92: 2229–2237.

Eisen, E. J. 1967. Mating designs for estimating direct and maternal genetic variances and direct-maternal covariances. *Can. J. Genet. Cytol*., 9: 13–22.

Elzo, M. A. 1990. Recursive procedures to compute the inverse of multiple trait additive genetic covariance matrix in inbreed and noninbreed multibreed populations. *J. Anim. Sci*., 68: 1215–1228.

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.

Elzo, M. A. y T. R. Famula. 1985. Multibred sire evaluation procedures within a country. *J. Anim. Sci*., 60: 942–952.

Elzo, M. A. y D. L. Wakeman. 1998. Covariance components and prediction for additive and nonadditive preweaning growth genetic effects in an Angus-Brahman multibreed herd. *J. Anim. Sci*., 76: 1290–1302.

Emik, L. O. y C. E. Terril. 1949. Systematic procedures for calculating inbreeding coefficients. *J. Hered.*, 40: 51–55.

Falconer, D. S. 1965. Maternal effects and genetic response. Genetics Today. En: *Proc. XI Int. Cong. of Genetics*, Hague, Netherlands, 3: 763.

Falconer, D. S. y T. F. C. Mackay. 1996. *Introduction to quantitative genetics. 4th Edition*. Longman, UK.

Fisher, R. A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Tras. Roy. Soc. Edinb.*, 52: 399–433.

García-Cortés, L. A. y M. A. Toro. 2006. Multibreed analysis by splitting the breeding values. *Genet. Sel. Evol.*, 38: 601–615.

García-Cortés, L. A., M. Rico y E. Groeneveld. 1998. Using coupling with the Gibbs sampler to assess convergence in Animal models. *J. Anim. Sci.*, 76: 441–447.

Garthwaite, P. H. y S. A. Al-Awadhi. 2001. Non-conjugate prior distribution assessment for multivariate normal sampling. *J. R. Statist. Soc. Ser. B*, 63: 95–110.

Gelfand, A. E. y A. F. M. Smith. 1990. Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.*, 85: 398–409.

Gelman, A. 1996. Inference and monitoring convergence. Pp. 131–144 en Gilks, W. R., S. Richardson y D. J. Spiegelhalter (Eds.) *Markov Chain Monte Carlo in practice*. Chapman and Hall, Boca Raton, US.

Gelman, A. y D. B. Rubin. 1992. Inference from iterative simulation using multiple sequences. *Stat. Sci.*, 7: 457–511.

Geman, S. y D. Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattn. Anal. Mach. Intel.*, 6: 721-741.

Gerstmayr, S. 1992. Impact of data structure on the reliability of the estimated genetic parameters in an animal model with maternal effects. *J. Anim. Breed. Genet.*, 109: 321–336.

Geweke, J. 1992. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. Pp. 169–193 en Bernardo, J. M., J. O. Berger, A. P. Dawid y A. F. M. Smith (Eds.) *Bayesian Statistics 4*, Oxford University Press, Oxford, UK.

Geyer, C. J. 1992. Practical Markov Chain Montecarlo. *Stat. Sci.*, 7: 473–511.

Gianola, D. y J. L. Foulley. 1982. Non linear prediction of latent genetic liability with binary expression: an empirical Bayes approach. En: *Proc. 2nd World Congress on Genetics Applied to Livestock Production*, Madrid, España, 7: 293–303.

Gianola, D. y R. L. Fernando. 1986. Bayesian methods in animal breeding theory. *J. Anim. Sci.*, 63: 217–244.

Gilks, W. R. y G. O. Roberts. 1996. Strategies for improving MCMC. Pp. 89–114 en Gilks, W. R., S. Richardson y D. J. Spiegelhalter (Eds.) *Markov Chain Monte Carlo in practice*. Chapman and Hall, Boca Raton, US.

Gilks, W. R., S. Richardson y D. J. Spiegelhalter. 1996. *Markov Chain Monte Carlo in practice*. Chapman and Hall, Boca Raton, US.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, S. J. Welham y R. Thompson. 2006. *ASReml User Guide Release 2.0.* VSN International Ltd. Hemel Hempstead, HP1 1ES, UK.

Graybill, F. A. 1961. *An introduction to linear statistical models. Vol. 1*. McGraw-Hill, NY, USA.

Groeneveld, E. y M. Kovac. 1990. A generalized computing procedure for setting up and solving mixed linear models. *J. Dairy Sci*., 73: 513–531.

Gutiérrez, J. P., I. Fernández, I. Alvarez, L. J. Royo y F. Goyache. 2006. Sire × contemporary group interactions for birth weight and preweaning growth traits in the Asturiana de los Valles beef cattle breed. *Livest. Sci*., 99: 61–68.

Henderson, C. R. 1950. Estimation of genetic parameters (abstract). *Ann. Math. Statist*., 21: 309–310.

Henderson, C. R. 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics*, 32: 69–83.

Henderson, C. R. 1984. *Applications of linear models in animal breeding*. University of Guelf, Guelf, Ontario, Canada.

Henderson, C. R. 1985. Equivalent linear models to reduce computations. *J. Dairy Sci*., 68: 2267–2277.

Henderson, C. R. 1988. Theoretical basis and computational methods for a number of different animal models. *J. Dairy Sci*., 71 (suppl 2): 35–53.

Henderson, C. R., O. Kempthorne, S. R. Searle y C. M. von Krosigk. 1959. Estimation of environmental and genetic trends from records subject to culling. *Biometrics*, 15: 192–218.

Heydarpour, M., L. R. Schaeffer y M. H. Yazdi. 2008. Influence of population structure on estimates of direct and maternal parameters. *J. Anim. Breed. Genet*., 125: 89–99.

Hill, W. G. 1982. Dominance and epistasis as components of heterosis. *J. Anim. Breed. Genet*., 99: 161–168.

Hill, W. G., M. E. Goddard, y P. M. Visscher. 2008. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet*., 4: e1000008.

Hobert, J. P. y G. Casella. 1996. The effects of improper priors on Gibbs sampling in hierarchical linear models. *J. Amer. Statist*., 91: 1461–1473.

Iwaisaki, H., S. Tsuruta, I. Misztal y J. K. Bertrand. 2005. Estimation of correlation between maternal permanent environmental effects of related dams in beef cattle. *J. Anim. Sci*., 83: 537–542.

Jensen, J., C. S. Wang, D. A. Sorensen y D. Gianola. 1994. Bayesian inference on variance and covariance components for traits influenced by maternal and direct genetic effects, using the Gibbs sampler. *Acta Agric. Scand*., 44: 193–201.

Kempthorne, O. 1954. The correlations between relatives in a random mating population. *Proc. Royal Stat. Soc*., 143: 103–113.

Kirkpatrick, M. y R. Lande. 1989. The evolution of maternal characters. *Evolution*, 43: 485–503.

Koch, R. M. 1972. The role of maternal effects in animal breeding: VI. Maternal effects in beef cattle. *J. Anim. Sci*., 35: 1316–1323.

Koerhuis, A. N. M. y Thompson, R. 1997. Models to estimate maternal effects for juvenile body weight in broiler chickens. *Genet. Sel. Evol*., 29: 225–249.

Lande, R. 1981. The minimum number of genes contributing to quantitative variation between and within populations. *Genetics*, 99: 541–553.

Le, N. D. y J. V. Zidek. 2006. *Statistical analysis of environmental space-time processes*. Springer Science+Business Media, Inc., NY, USA.

Le, N. D., L. Sun y J. V. Zidek. 1999. *Bayesian spatial interpolation and backcasting using Gaussian – generalized inverted Wishart model*. Technical Report 185. Statistics Dept., UBC, Vancouver, Canada.

Lehmann, E. L. 1983. *Theory of point estimation*. John Wiley y Sons, NY, USA.

Lo, L. L., R. L. Fernando y M. Grossman. 1993. Covariance between relatives in multibreed populations: additive model. *Theor. Appl. Genet*., 87: 423–430.

Lynch, M. y B. Walsh. 1998. *Genetics and analysis of quantitative traits*. Sinauer Associates, Inc., Suderland, MA, USA.

Maniatis, N. y G. E. Pollott. 2003. The impact of data structure on genetic (co)variance components of early growth in sheep, estimated using an animal model with maternal effects. *J. Anim. Sci*., 81: 101–108.

Meinhold, R. J. y N. D. Singpurwalla. 1983. Understanding the Kalman Filter. *The Amer. Stat*., 37: 123–127.

Meuwissen, T. H. E. y Z. Luo. 1992. Computing inbreeding coefficients in large populations. *Genet. Sel. Evol*., 24: 305–313.

Meyer, K. 1989. Restricted maximum likelihood to estimate variance components for animal models with several random effects using a derivative-free algorithm. *Genet. Sel. Evol*., 21: 317–340.

Meyer, K. 1992. Bias and sampling covariances of estimates of variance components due to maternal effects. *Genet. Sel. Evol*., 24: 487–509.

Meyer, K. 1997. Estimates of genetic parameters for weaning weight of beef cattle accounting for direct-maternal environmental covariances. *Livest. Prod. Sci*., 52: 187–199.

Meyer, K. 2007. WOMBAT – A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). *J. Zhejiang Univ. Sci. B*, 8:815–821.

Misztal, I. 2006. *Computational techniques in animal breeding. Course notes*. Disponible en: http://nce.ads.uga.edu/~ignacy. Último acceso: octubre de 2011.

Misztal, I. 2008. Reliable computing in estimation of variance components. *J. Anim. Breed. Genet*., 125: 363–370.

Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet y D. H. Lee. 2002. BLUPF90 and related programs (BGF90). En: *Proc. 7th World Congress on Genetics Applied to Livestock Production*, Montpellier, Francia.

Morris, C. A., R. L. Baker, N. G. Cullen y D. L. Johnson. 1994. Rotation crosses and inter se matings with Angus and Hereford cattle for five generations. *Livest. Prod. Sci*., 39: 157–172.

Mrode, R. A. 2005. *Linear models for the prediction of animal breeding values*. CAB International Wallingford, Oxfordshire, UK.

Nobre, P. R. C., I. Misztal, S. Tsuruta, J. K. Bertrand, L. O. C. Silva y P. S. Lopes. 2003. Analyses of growth curves of Nellore cattle by multiple-trait and random regression models. *J. Anim. Sci*., 81: 918–926.

O'Hara, R. B., J. M. Cano, O. Ovaskainen, C. Teplitsky y J.S. Alho. 2008. Bayesian approaches in evolutionary quantitative genetics. *J. Evol. Biol*., 21: 949–957.

Pearle, J. 2000. *Causality: models, reasoning, and inference*. Cambridge University Press, Cambridge, UK.

Primeaux, D. 2005. Programming with Gaussian logarithms to compute the approximate addition and subtraction of very small (or very large) positive numbers. Pp. 129–132 en: *Proc. 6th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing and 1st ACIS International Workshop on Self-Assembling Wireless Networks*, Maryland, USA.

Quaas, R. L. 1976. Computing the diagonal elements and inverse of a large numerator relationship matrix. *Biometrics*, 32: 949–956.

Quaas, R. L. 1988. Additive genetic model with groups and relationships. *J. Dairy Sci*., 71: 1338–1345.

Quaas, R. L. y E. J. Pollak. 1980. Mixed model methodology for farm and ranch beef cattle testing programs. *J. Anim. Sci*., 51: 1277–1287.

Quintanilla, R., L. Varona, M. R. Pujol y J. Piedrafita. 1999. Maternal animal model with correlation between maternal environmental effects of related dams. *J. Anim. Sci*., 77: 2904–2917.

Raftery, A. E. y S. M. Lewis. 1992. How many iterations in the Gibbs sampler? Pp. 765–776 en Bernardo, J. M., J. O. Berger, A. P. Dawid y A. F. M. Smith (Eds.) *Bayesian Statistics 4*, Oxford University Press, Oxford, UK.

Raftery, A. E. y S. M. Lewis. 1996. Implementing MCMC. Pp 115–130 en Gilks, W. R., S. Richardson y D. J. Spiegelhalter (Eds.) *Markov Chain Monte Carlo in practice*. Chapman and Hall, Boca Raton, US.

Ritter, C. y M. A. Tanner. 1992. Facilitating the Gibbs sampler: the Gibbs Stopper and the Griddy-Gibbs sampler. *J. Am. Stat. Assoc*., 87: 861–868.

Robinson, D. L. 1996. Models which might explain negative correlations between direct and maternal genetic effects. *Livest. Prod. Sci*., 45: 111–122.

Searle, S. R. 1982. *Matrix algebra useful for statistics*. John Wiley & Sons, Inc., NY, USA.

Searle, S. R., G. Casella y C. E. McCulloch. 1992. *Variance components*. John Wiley & Sons, Inc., NY, USA.

Silverman, B.W. 1986. *Density estimation for statistics and data analysis*. Chapman and Hall, London, UK.

Smith, B. 2007. BOA: An R package for MCMC output convergence assessment and posterior inference. *J. Stat. Soft*., 21: 1–37.

Smith, W. B. y R. R. Hocking. 1972. Algorithm AS 53: Wishart variate generator. *Appl. Statist*., 21: 341–345.

Sorensen, D. A. y D. Gianola. 2002. *Likelihood, Bayesian, and MCMC methods in quantitative genetics*. Springer-Verlag, NY, USA.

Thompson, R. 1976. The estimation of maternal genetic variances. *Biometrics*, 32: 903–917.

Van Tassell, C. P. y L. D. Van Vleck. 1996. Multiple-trait Gibbs sampler for animal models: flexible programs for Bayesian and likelihood-based (co)variance component inference. *J. Anim. Sci*., 74: 2586–2597.

Vergara, O. D., M. F. Ceron-Muñoz, E. M. Arboleda, Y. Orozco y G. A. Ossa. 2009a. Direct genetic, maternal genetic, and heterozygosity effects on weaning weight in a Colombian multibreed beef cattle population. *J. Anim. Sci*., 87: 516–521.

Vergara, O. D., M. A. Elzo, M. F. Ceron-Muñoz y E.M. Arboleda. 2009b. Weaning weight and post-weaning gain genetic parameters and genetic trends in a Blanco Orejinegro–Romosinuano–Angus–Zebu multibreed cattle population in Colombia. *Livest. Sci*., 124: 156–162.

Visscher, P. M., W. G. Hill y N. R. Wray. 2008. Heritability in the genomics era – concepts and misconceptions. *Nat. Rev. Genet.*, 9: 255–266.

Willham, R. L. 1963. The covariance between relatives for characters composed of components contributed by related individuals. *Biometrics*, 19: 18–27.

Willham, R. L. 1972. The role of maternal effects in animal breeding: III. Biometrical aspects of maternal effects in animals. *J. Anim. Sci.*, 35: 1288–1293.

Willham, R. L. 1980. Problems in estimating maternal effects. *Livest. Prod. Sci.*, 7: 405–418.

Wright, S. 1921a. Systems of mating. Parts I–V. *Genetics*, 6: 111–178.

Wright, S. 1921b. Correlation and causation. *J. Agric. Res*., 20: 557–585.

Wright, S. 1922. Coefficients of inbreeding and relationship. *Am. Nat*., 56: 330–338.

Wright, S. 1968. *Evolution and the genetics of populations. Vol. 1. Genetics and biometrical foundations*. University of Chicago Press, Chicago, USA.