COLEGIO DE POSTGRADUADOS

INSTITUCIÓN DE ENSEÑANZA E INVESTIGACIÓN
EN CIENCIAS AGRÍCOLAS

CAMPUS MONTECILLO

SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA
ESTADÍSTICA

MODELOS DE SELECCIÓN GENÓMICA PARA PREDICCIÓN DE CARACTERES COMPLEJOS EN HÍBRIDOS DE MAÍZ

ROCÍO GUADALUPE ACOSTA PECH

T E S I S

PRESENTADA COMO REQUISITO PARCIAL PARA
OBTENER EL GRADO DE:

DOCTORA EN CIENCIAS

MONTECILLO, TEXCOCO, EDO. DE MÉXICO
2017

CARTA DE CONSENTIMIENTO DE USO DE LOS DERECHOS DE AUTOR Y DE LAS REGALIAS COMERCIALES DE PRODUCTOS DE INVESTIGACION

En adición al beneficio ético, moral y académico que he obtenido durante mis estudios en el Colegio de Postgraduados, el que suscribe ROCIO GUADALUPE ACOSTA PECH, Alumno (a) de esta Institución, estoy de acuerdo en ser participe de las regalías económicas y/o académicas, de procedencia nacional e internacional, que se deriven del trabajo de investigación que realicé en esta institución, bajo la dirección del Profesor DR. PAULINO PÉREZ RODRÍGUEZ, por lo que otorgo los derechos de autor de mi tesis MODELOS DE SELECCIÓN GENÓMICA PARA PREDICCIÓN DE CARACTERES COMPLEJOS EN HÍBRIDOS DE MAÍZ

y de los producto de dicha investigación al Colegio de Postgraduados. Las patentes y secretos industriales que se puedan derivar serán registrados a nombre el colegio de Postgraduados y las regalías económicas que se deriven serán distribuidas entre la Institución, El Consejero o Director de Tesis y el que suscribe, de acuerdo a las negociaciones entre las tres partes, por ello me comprometo a no realizar ninguna acción que dañe el proceso de explotación comercial de dichos productos a favor de esta Institución.

Montecillo, Mpio. de Texcoco, Edo. de México, a 24 de JULIO de 2017

Firma del
Alumno (a)

Pérez Rdz.

DR. PAULINO PÉREZ RODRÍGUEZ

Vo. Bo. del Consejero o Director de Tesis

La presente tesis titulada: **Modelos de Selección Genómica para Predicción de Caracteres complejos en híbridos de Maíz.**, realizada por la alumna: **Rocío Guadalupe Acosta Pech**, bajo la dirección del Consejo Particular indicado ha sido aprobada por el mismo y aceptada como requisito parcial para obtener el grado de:

DOCTORA EN CIENCIAS

SOCIOECONOMÍA, ESTADÍSTICA E INFORMATICA ESTADÍSTICA

CONSEJO PARTICULAR

CONSEJERO Pérez Rdz.
Dr. Paulino Pérez Rodríguez.

ASESOR Elizabeth González E.
Dra. Elizabeth González Estrada.

ASESOR Luis A. Rodríguez C.
Dr. Luis Alfonso Rodríguez Carvajal.

ASESOR
Dr. Ciro Velaseo Cruz.

ASESOR
Dr. Javier Suárez Espinosa.

Montecillo, Texcoco, Estado de México, 2017.

Modelos de Selección Genómica para Predicción de Caracteres Complejos en Híbridos de Maíz

Rocío Guadalupe Acosta Pech, Dra.

Colegio de Postgraduados, 2017

RESUMEN

La predicción del rendimiento híbrido es muy importante en los programas de mejoramiento agrícola. En el fitomejoramiento, los ensayos multi-ambientales desempeñan un papel importante en la selección de rasgos importantes, tales como estabilidad a través de ambientes, rendimiento de grano y resistencia a plagas. Las condiciones ambientales modulan la expresión génica causando interacción genotipo× ambiente (G×A), de tal manera que las correlaciones genéticas estimadas del rendimiento de líneas individuales a través de ambientes resumen la acción conjunta de los genes y las condiciones ambientales. Este trabajo propone un modelo estadístico genético que incorpora G×A junto con la aptitud combinatoria general y específica para predecir el rendimiento de híbridos en ambientes. El modelo propuesto también se puede aplicar a cualquier otra especie híbrida con diferentes grupos parentales. En este estudio se evaluó el poder predictivo de dos modelos de predicción de rendimiento de híbridos utilizando la técnica de validación cruzada aplicada en datos híbridos de maíz, que comprenden 2,724 híbridos derivados de 507 líneas Dent y 24 líneas Flint, que fueron evaluadas para tres rasgos en 58 ambientes durante 12 años; los análisis se realizaron para cada año. En promedio, los modelos genómicos que incluyen la interacción de la aptitud combinatoria general y específica con los ambientes tienen mayor poder predictivo que los modelos genómicos sin interacción con ambientes (van del 12 al 22 %, dependiendo del rasgo). Se concluye que la inclusión de G×A en la predicción de rendimiento de híbridos de maíz no probados aumenta la precisión de los modelos genómicos.

Por otra parte, se sabe que la selección genómica se ha convertido en una herramienta muy conocida para la selección de candidatos en programas de fitomejoramiento de plantas y animales. En el caso de rasgos cuantitativos, es usual suponer que la distribución de la variable respuesta se puede aproximar con una distribución normal. Sin embargo, es bien sabido que el proceso de selección conduce a distribuciones que son asimétricas. Existe una amplia literatura estadística sobre distribuciones asimétricas; una distribución particularmente interesante es la normal asimétrica, que incluye un parámetro de forma que le permite adoptar formas asimétricas. Por lo tanto, la distribución normal asimétrica puede considerarse una generalización de la distribución normal. Se propone el uso de la distribución normal asimétrica en el contexto de regresión con aplicaciones en Selección Genómica, donde por lo general el número de predictores excede ampliamente el tamaño de la muestra. Con el fin de hacer los cálculos factibles, se utilizó una representación

estocástica de una variable aleatoria normal asimétrica que permite el uso de técnicas de cadenas de Markov Monte Carlo (MCMC) para ajustar el modelo propuesto. Se evaluó el poder predictivo y la bondad de ajuste del modelo y se comparó con el modelo de regresión Ridge Bayesiano utilizando simulaciones de datos. También se evaluó el poder predictivo del modelo propuesto utilizando un conjunto de datos reales de maíz. Los resultados muestran que la bondad de ajuste (utilizando el criterio de información de la devianza y el número efectivo de parámetros) favoreció el modelo propuesto; sin embargo, el poder predictivo evaluado a través de la validación cruzada entre la regresión Ridge Bayesiana y el modelo propuesto fue aproximadamente el mismo.

Palabras clave: Selección Genómica, Poder Predictivo, Validación cruzada, Normal-Asimétrica (SNB), Regresión Ridge Bayesiana (RRB).

Genomic Selection Models for Prediction of Complex Traits in Maize Hybrids

Rocío Guadalupe Acosta Pech, Dra.

Colegio de Postgraduados, 2017

ABSTRACT

The prediction of hybrid performance (HP) is very important in agricultural breeding programs. In plant breeding, multi-environment trials play an important role in the selection of important traits, such as stability across environments, grain yield and pest resistence. Environmental conditions modulate gene expression causing genotype×environment interaction (G×E), such that the estimated genetic correlations of the performance of individual line across environments summarize the joint action of genes and environmental conditions. This work proposes a genomical statistical model that incorporates G×E for general and specific combining ability for predicting the performance of hybrids in environments. The proposed model can also be applied to any other hybrid species with distinct parental pools. In this study, we evaluated the predictive ability of two HP prediction models using a cross-validation approach applied in extensive maize hybrid data, comprising 2724 hybrids derived from 507 dent lines and 24 flint lines, which were evaluated for three traits in 58 environments over 12 years; analyses were performed for each year. On average, genomic models that include the environments have greater predictive ability than genomic models without interaction with environments (ranging from 12 to 22 %, depending on the trait). We concluded that including G×E in the prediction of untested maize hybrids increases the accuracy of genomic models.

Genomic selection (GS) has become as a very well-known tool for selecting candidates in plant and animal breeding programs. In the case of quantitative traits, it is usual to assume that the distribution of the response variable can be approximated with a normal distribution. However, it is well known that the selection process leads to distributions that are skewed. There is vast statistical literature on skewed distributions; a particularly interesting distribution is the skew normal, which includes a shape parameter that allows it to adopt skewed forms. Therefore, the skew normal distribution can be considered a generalization of normal distribution. Here we propose using the skew normal distribution in the regression context with GS application where usually the number of predictors vastly exceeds the sample size. In order to make the computations feasible, we used a stochastic representation of a skew normal random variable which allows using standard Markov Chain Monte Carlo (MCMC) techniques to fit the proposed model. We evaluated the predictive power and goodness of fit of the proposed model and compared it to the Bayesian Ridge Regression model using simulated datasets. We also evaluated the predictive power of the proposed model using a real maize dataset. Results show that

the goodness of fit (deviance information criterion and effective number of parameters) favored the proposed model; nevertheless, the predictive power evaluated through cross-validation of the Bayesian Ridge Regression and the proposed model was about the same.

Key words: Genomic Selection, Predictive power, Cross-validation, Bayesian Skew-Normal (BSN), Bayesian Ridge Regression (BRR).

AGRADECIMIENTOS

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por el apoyo económico brindado durante la estancia y realización de mis estudios de postgrado.

Al Colegio de Postgraduados, por haberme dado la oportunidad de continuar mi formación académica en sus aulas.

A los integrantes de mi Consejo Particular:

Dr. Paulino Pérez Rodríguez por su paciencia, dedicación e interés mostrado en orientarme para poder llevar a buen término este trabajo. Muchas gracias Doc por el conocimiento compartido y haber permitido que trabajara con usted.

Dra. Elizabeth González Estrada por su apoyo, sugerencias y comentarios realizados a la presente tesis, tiempo dedicado a compartir su conocimiento, charlas interminables y su incondicional amistad.

Dr. Luis Alfonso Rodríguez Carvajal, por los comentarios y observaciones realizadas a este escrito, pero sobretodo por su valiosa amistad a lo largo de mi vida académica, ser un gran amigo y un excelente compañero. Gracias.

Doctores:

Dr. Ciro Velasco Cruz,

Dr. Javier Suárez Espinosa,

por el tiempo empleado y las sugerencias hechas en la revisión de este trabajo final.

Dr. José Crossa por todas las observaciones, comentarios y sugerencias hechas al presente escrito en pro de un mejor desarrollo del mismo.

A todos los profesores, personal administrativo: Laurita, Geno, Isa, Toñita, Jackie, Mary y todas aquellas personas que de alguna manera me apoyaron en esta gran aventura, muchas gracias a todos ellos.

DEDICADO A:

Mis padres:

Francisco y ***Matilde.***
... mis primeros maestros, gracias por darme la vida.
Todo lo que soy se lo debo a ustedes. Los amo!

Mi adorada hermana, amiga, compañera, cómplice, colega y todo aquello que con palabras no se expresa:

Nayeli.
Gracias por el amor inmenso que me tienes y estar siempre a mi lado,
te quiero hermana!!!

Mi grandiosa amiga:

Dulce.
Gracias por el apoyo incondicional en todo este tiempo y tu cariño sincero.

Mis hermanos:

Óscar, Grissel, Dianela, Anabel.
Mis cómplices de viaje en esta aventura llamada vida!!!

Mi amiga de aventuras:

Elizabeth.
Gracias, por tu maravillosa compañía, entrañable amistad,
tiempo compartido y sincero afecto!
...when hope hardly seems worth having...

"Es, pues, la fe la certeza de lo que se espera, la convicción de lo que no se ve." Hebreos 11:1-3.

◇

CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

Capítulo 1

Introducción

El mejoramiento genético vegetal o animal es la ciencia, el arte y el negocio de mejorar los organismos para el beneficio de los seres humanos (Bernardo, 2002). El maíz *(Zea mays L.)* es uno de los tres granos más ampliamente producidos a nivel mundial junto al trigo y el arroz, y es el de mayor consumo por el hombre (FAO, 2013), además de la infinidad de productos industriales derivados.

Desde 1977, Sprague y Eberhart señalaron que los rendimientos más altos de maíz se alcanzan con híbridos de cruza simple de líneas endogámicas de alta Aptitud Combinatoria General (ACG) y alto rendimiento derivados por autofecundación. Sin embargo, este tipo de líneas no siempre está disponible y es necesaria su formación para utilizarlas en cruzas simples y generar materiales híbridos superiores a los actuales. La necesidad de producir híbridos surge como una opción para los productores, y fitomejoradores, con la finalidad de generar mayor rendimiento de grano y por ende, compensar la demanda alimentaria debida a los aumentos de población.

Bajo esta perspectiva se hace necesario explorar o proponer nuevas alternativas estadísticas que mejoren el proceso de predicción de rendimiento de granos, en un esquema donde los modelos lineales juegan un rol importante al aplicarlos en espacios de alta dimensión con un enfoque Bayesiano.

Este capítulo tiene como finalidad introducir el problema de investigación y la metodología utilizada para alcanzar sus objetivos.

1.1. Selección Genómica: una herramienta para predicción

La selección genómica (SG) o predicción genómica se define como la selección de individuos basada en su valor genético (Meuwissen *et al.*, 2001). Los valores genéticos tienen como función explicar y/o predecir el comportamiento de rasgos fenotípicos tales como altura de la planta, rendimiento, etc. Estos valores se predicen en función del efecto de miles de marcadores moleculares.

Con los recientes avances tecnológicos para el estudio y mejoramiento de rasgos cuantitativos, la SG tiene como principal herramienta el uso de páneles de marcadores densos, SNP (polimorfismos de un solo nucleótido), distribuidos por todo el genoma que sirven para explorar una porción muy significativa de la variabilidad genética presente en el genoma (Vélez, 2015).

El modelo genético base para estudiar el valor *fenotípico* (P) está dado en componentes atribuibles a la influencia del *genotipo* (G) y del *ambiente* (A) (Lynch *et al.*, 1998):

$$P = G + A, \tag{1.1}$$

donde genotipo es el arreglo particular de genes que presenta el individuo, y el ambiente como todas las circunstancias no genéticas que afectan al valor fenotípico (Lynch *et al.*, 1998). El modelo 1.1 se puede escribir como

$$y_{ik} = \mu + g_i + a_k + e_{ik},\ i = 1, 2, \ldots, n \tag{1.2}$$

donde y_{ik} representa la respuesta o fenotipo del individuo i en el ambiente k, y depende del *i-ésimo* valor genético, g_i, del efecto del *k-ésimo* ambiente, a_k, y de un error aleatorio e_{ik}, que se supone sigue una distribución normal con media 0 y varianza σ_e^2, μ representa una media general.

Para incorporar los marcadores en un modelo de SG, los valores genéticos $g_i; (i = 1, 2, \ldots, n)$, se modelan con base a una regresión paramétrica donde las covariables, x_{ij}, son los marcadores; es decir, $g_i = \sum_{j=1}^{p} x_{ij}\beta_j$, por tanto el valor fenotípico expresado en 1.2 queda representado como

$$y_{ik} = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + a_k + e_{ik},\ j = 1, 2, \ldots, p, \tag{1.3}$$

siendo β_j el valor que representa el efecto del *j-ésimo* marcador en la respuesta y_{ik} (Crossa

et al., 2000). Matricialmente, el modelo expresado en 1.3 es:

$$\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{X\beta} + \boldsymbol{a} + \boldsymbol{e}.$$

Goddard y Hayes (2007); Hayes *et al.* (2001); Meuwissen *et al.* (2001) propusieron este enfoque que es el más usado en SG (Crossa *et al.*, 2011). Crossa *et al.* (2000) mencionan que las estimaciones por mínimos cuadrados de la media general, del modelo descrito en (1.2), están dadas por $\hat{\mu} = \bar{y}_{..}$, $\hat{g}_i = \bar{y}_{i.} - \bar{y}_{..}$ y $\hat{a}_k = \bar{y}_{.j} - \bar{y}_{..}$, siempre y cuando se cumpla que $n > p$; es decir, siempre que el número de individuos sea mayor que el número de genotipos.

Si el efecto ambiental no es considerado en el modelo 1.2 y si se supone, sin pérdida de generalidad $\mu = 0$, entonces éste se reduce a:

$$y_i = g_i + e_i,$$

con $e_i \sim N(0, \sigma_e^2)$, y por lo tanto el modelo 1.3, queda expresado como:

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + e_i, \text{ o matricialmente } \boldsymbol{y} = \mathbf{X\beta} + \mathbf{e}.$$

Bajo este esquema, la estimación de los parámetros $\boldsymbol{\beta}$ se basa en el *Método de Mínimos Cuadrados Ordinarios (MCO)*, que consiste en la minimización del cuadrado del error,

$$r_i = \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2,$$

o bien en notación matricial,

$$\mathbf{r} = (\boldsymbol{y} - \boldsymbol{X\beta})^t(\boldsymbol{y} - \boldsymbol{X\beta}),$$

entonces, el estimador del vector de parámetros $\boldsymbol{\beta}$ se obtiene de:

$$\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X\beta})^t(\boldsymbol{y} - \boldsymbol{X\beta}),$$

cuya solución es:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y},$$

siempre y cuando $\boldsymbol{X}$ sea de rango completo. Sin embargo, es común en las aplicaciones de los modelos de SG que el número p, de marcadores moleculares de los que se dispone siempre exceda al número de observaciones, $p \gg n$, generando que el uso de *MCO* no sea el método adecuado para hallar estimadores de los parámetros de interés. Otra desventaja que surge en el análisis de esta relación funcional y que es debido a la alta dimensionalidad de la matriz de SNP's, es la multicolinealidad aproximada existente, la cual puede ser ocasionada por las altas correlaciones entre las variables. Esto ocasiona que la matriz $\boldsymbol{X}^t\boldsymbol{X}$ sea singular, por lo que no es posible la estimación de $\boldsymbol{\beta}$ ya que las ecuaciones normales no se pueden resolver de manera única.

Piepho (2009) menciona que es posible resolver el problema de multicolinealidad, si en lugar de usar todos los marcadores, se selecciona un subconjunto de marcadores por alguno de los métodos conocidos de selección de variables, regresión hacia adelante, regresión hacia atrás, etc.; sin embargo, la eficiencia de estos métodos queda en entredicho debido a la alta correlación existente entre marcadores.

Crossa *et al.* (2010) comentan sobre el uso de métodos de regresión penalizados, para estimar el efecto conjunto de todos los marcadores disponibles empleados en el análisis, como la *regresión Ridge* (Hoerl y Kennard, 1970), *regresión LASSO* (*Least Absolute Shrinkage and Selection Operator*, por sus siglas en inglés.) (Tibshirani, 1996) o versiones Bayesianas de éstos, son más adecuadas.

Heffner *et al.* (2009) hacen énfasis en el objetivo principal de la Selección Genómica, la estimación de los valores de cría para individuos que solo tienen datos genotípicos usando un modelo *entrenado* con individuos de los cuales se tiene información genotípica y fenotípica. De hecho, la estimación de los valores de cría para rasgos cuantitativos ha sido a través del BLUP (*Best Linear Unbiased Prediction*, por sus siglas en inglés) utilizando datos fenotípicos y de sus familiares (Henderson, 1984).

En el enfoque bayesiano, se asigna una distribución a priori a los parámetros del modelo de regresión (Tibshirani, 1996), de hecho la distribución a priori es Gaussiana con dos hiperparámetros, la media y la varianza, sin embargo se supone a la media como cero y la varianza se convierte en el parámetro que balancea el ajuste y la complejidad del modelo.

En el esquema de la regresión Ridge, el parámetro de penalización o encogimiento se define como $\lambda = \dfrac{\sigma_{\boldsymbol{e}}^2}{\sigma_{\boldsymbol{\beta}}^2}$ y la distribución a posteriori de los paramétros $\boldsymbol{\beta}$ es:

$$p(\boldsymbol{\beta}|\boldsymbol{y}, \mathbf{X}) \propto \boldsymbol{N}(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{\beta}, \sigma_{\boldsymbol{e}}^2\boldsymbol{I})\boldsymbol{N}(\boldsymbol{\beta}|\mathbf{0}, \sigma_{\boldsymbol{\beta}}^2\boldsymbol{I}).$$

En el contexto *LASSO* Bayesiano de Park y Casella (2008), la distribución a priori asociada a los parámetros de regresión, $\boldsymbol{\beta}$, es doble exponencial, es decir,

$$p(\boldsymbol{\beta}|\sigma^2_{\boldsymbol{e}}) = \prod_{j=1}^{p} \frac{\lambda}{2\sigma^2_{\boldsymbol{e}}} e^{\frac{-\lambda|\beta_j|}{\sigma^2_{\boldsymbol{e}}}},$$

logrando con esto un mayor encogimiento de los coeficientes hacia el cero con mayor rapidez que la regresión Ridge. Ambos enfoques son utilizados en selección genómica (de los Campos *et al.*, 2010), y la función *BLR* de la biblioteca de `R` (R Core Team, 2017) con el mismo nombre (Pérez-Rodríguez *et al.*, 2010) permite hacer la estimación y predicción Bayesiana.

La regresión no paramétrica, *RKHS* (*Reproducing Kernel Hilbert Spaces*, por sus siglas en inglés) también ha sido una herramienta muy utilizada en *SG*, en este esquema los marcadores se utilizan para construir una estructura de covarianza entre los valores genéticos (de los Campos *et al.*, 2010),

$$Cov(g_i, g_j) \propto K(\mathbf{x}_i, \mathbf{x}_j),$$

donde $\mathbf{x}_i, \mathbf{x}_j$ denotan vectores de los marcadores genotípicos para los individuos i y j, $K(\cdot,\cdot)$ se conoce como *RK* (reproducing kernel), una función positiva definida. Una de las desventajas de este enfoque está en la interpretación de los parámetros, no es sencilla, sin embargo el modelo es muy útil ya que permite obtener predicciones más precisas (González-Recio *et al.*, 2008).

Como se menciona en de los Campos *et al.* (2009a), la regresión *RKHS* proporciona un marco de referencia para la evaluación genética de caracteres cuantitativos, que puedan ser utilizados para incorporar información sobre las genealogías, marcadores o cualquier otra forma de caracterizar los antecedentes genéticos de los individuos de tal manera que puedan estar disponibles en el futuro y ser útiles para las predicciones.

1.2. Objetivos

Se definen los siguientes objetivos a alcanzar con el presente trabajo.

- Desarrollar un modelo de Selección Genómica para predicción de rendimiento de híbridos.
- Aplicar los modelos propuestos a conjuntos de datos reales, agregando los siguientes términos de interacción: genotipo×ambiente, híbrido×ambiente y padres×ambiente.

- Comparar el poder predictivo del modelo propuesto con un modelo sin ningún término de interacción.
- Generalizar el modelo de regresión Ridge usado en SG, para variables respuestas cuya distribución es asimétrica.

1.3. Metodología

Para el desarrollo de este trabajo, se inicia con una revisión de la literatura acerca de los modelos más comunes, utilizados para predecir rendimiento de híbridos de maíz en multi-ambientes. Esto con la finalidad de conocer las herramientas actuales y disponibles con la que se está abordando este reto.

Se analiza el problema de multicolinealidad y dimensionalidad, generado por el gran número de marcadores que contiene un **QTL** (*loci* de rasgos cuantitativos o cuantificables) considerando las propuestas de Tikhonov y Arsenin (1977); Tikhonov (1963) y Hoerl y Kennard (1970).

Se revisan los métodos de regresión penalizada Bayesiana, las características principales de estos modelos con respecto a su habilidad predictiva basada en el promedio de las correlaciones obtenidas.

Se hace uso de la librería de funciones BGLR (de los Campos y Pérez-Rodríguez, 2015) del paquete estadístico R (R Core Team, 2017) para el ajuste de los modelos propuestos para predicción de rendimiento de híbridos, haciendo uso de un conjunto de datos de rendimiento de híbridos de maíz, del programa de mejoramiento de la compañía RAGT (`https://www.ragtsemences.com`).

Para el caso del modelo con errores cuya distribución es normal asimétrica, SN (skew-normal), se utilizan dos conjuntos de datos. El primero consiste de $n = 599$ líneas de trigo del Programa Global de Trigo del CIMMyT, donde el fenotipo evaluado fue el *rendimiento del grano* en cuatro mega-ambientes. Estos datos se encuentran disponibles libremente en internet y están incluidos en la biblioteca de funciones BGLR. El segundo conjunto de datos proviene del proyecto de Tolerancia del Maíz a la Sequía (DTMA, por sus siglas en inglés) del Programa Global de Maíz del CIMMyT.

Posteriormente, se evalúa el poder predictivo del modelo en base a dos medidas de bondad de ajuste, el pD (número efectivo de parámetros) y el DIC (Criterio de Información de Devianza) propuestos por Spiegelhalter *et al.* (2002) como una medida del ajuste de un modelo jerárquico que es penalizado por la complejidad. El primero con el número efectivo de parámetros y el segundo con el promedio de la devianza.

1.4. Organización del Trabajo

En el capítulo 2 se realiza una revisión de los modelos en Selección Genómica que más se han utilizado en la predicción del rendimiento de híbridos de maíz en multi-ambientes, tanto desde el punto de vista Clásico como Bayesiano. En este mismo capítulo, se revisan propiedades esenciales de la distribución normal asimétrica, en el contexto de regresión lineal.

En el capítulo 3 se presenta y analiza la aplicación de los modelos propuestos a un conjunto de datos reales, para predecir rendimiento de híbridos de maíz. Se describen los modelos ajustados y se evalúa su poder predictivo a través de la técnica de validación cruzada.

El capítulo 4 reporta el análisis de un modelo de regresión lineal cuando la variable respuesta tiene distribución normal asimétrica y se ejemplifica su aplicación en el contexto de predicción en Selección Genómica, a través de dos conjuntos de datos. Se analiza el modelo propuesto y se evalúa su poder predictivo en base a dos medidas de bondad ajuste, el pD y el DIC.

El capítulo 5 resume los resultados y conclusiones obtenidos de los análisis, con una breve discusión de los mismos.

Se anexan los programas en `R` (R Core Team, 2017) empleados para realizar todos los ajustes y simulaciones.

Capítulo 2

Revisión de literatura

En este capítulo se presenta un resumen del estado del arte relacionado con los temas que serán abordados en los siguientes capítulos. En específico, los modelos de selección genómica utilizados en predicción de híbridos, el modelo mixto y la distribución normal-asimétrica (SN).

2.1. La interacción genotipo × ambiente

El concepto de la interacción genotipo por ambiente (G×A) se define como el comportamiento genético diferencial, que muestran los genotipos cuando se les somete a diferentes ambientes (Sánchez, 1974). Este término indica que diferentes genotipos responden a cambios ambientales de diferentes maneras, implicando en casos extremos que la clasificación de los genotipos pueda ser alterada por un cambio en el ambiente (Lynch *et al.*, 1998).

El estudio de G×A, permite la clasificación de los genotipos en función de su rendimiento bajo dos esquemas diferentes: estables o adaptados a un ambiente particular (Kandus *et al.*, 2010). La estabilidad se refiere a la habilidad que tiene el genotipo para ser consistente, con rendimientos altos o bajos en varios ambientes mientras que la adaptabilidad está en función del ajuste que hace un organismo a las condiciones ambientales para sobrevivir (Balzarini *et al.*, 2005). La función que describe la expresión fenotípica (ej. rendimiento) de un genotipo en relación a cambios ambientales se denomina *norma de reacción* (Lynch *et al.*, 1998) y ésta es variable, es decir, para cada genotipo, rasgo fenotípico y variable ambiental pueden existir diferentes normas de reacción.

La comprensión de la base genética de la adaptación y de sus causas fisiológicas y ambientales es de fundamental importancia para entender la interacción genotipo ×

ambiente, para evaluar la asociación entre valores fenotípicos y genotípicos, y para mejorar la selección de genotipos superiores y estables (Crossa *et al.*, 1999).

La inclusión de G×A siempre ha sido una preocupación en el análisis de experimentos multiambientales en mejoramiento de plantas. Se han propuesto y utilizado varios modelos para describir la respuesta media de los genotipos sobre los ambientes y para estudiar e interpretar la interacción G×A en experimentos agrícolas (Eberhart y Russell, 1966; Finlay y Wilkinson, 1963; López-Cruz *et al.*, 2015; Yates y Cochran, 1938).

Históricamente, se han desarrollado una gran cantidad de modelos estadísticos para estudiar la interacción genotipo × ambiente. Los modelos lineales, modelos bilineales y modelos lineales-bilineales han sido utilizados para describir la respuesta media de genotipos en ambientes y para estudiar e interpretar la interacción G×A en experimentos agrícolas (Crossa, 2012).

En los últimos años, varios estudios han propuesto el uso de modelos de SG que incorporan el término G×A. Por ejemplo, Burgueño *et al.* (2012) extendieron el modelo de predicción (GBLUP) de un solo ambiente, a un contexto multi-ambientes y reportaron ganancias en la predicción con el modelo multi-ambientes. Del mismo modo, Jarquín *et al.* (2014) y Heslot *et al.* (2014) modelaron G×A utilizando marcadores moleculares así como covariables ambientales, y mostraron que la incorporación del término de interacción incrementa el poder predictivo de los modelos.

2.2. Modelos en Selección Genómica

El enfoque para el análisis de la interacción G×A, ha evolucionado a lo largo del tiempo, generando una gran cantidad de literatura sobre modelos y estrategias para su estudio (Malosetti *et al.*, 2014; Pérez-Rodríguez *et al.*, 2015), con el objetivo final de evaluar su impacto en la predicción.

Meuwissen *et al.* (2001) fueron los primeros en introducir las técnicas conocidas como métodos de *Selección Genómica* (SG), en el contexto de mejoramiento animal. Los autores proponen, de una manera simple, incorporar marcadores moleculares en los modelos estadísticos utilizados para estimar el valor genético de un individuo; por lo tanto, modelan los valores fenotípicos con base en miles de marcadores moleculares usando un modelo lineal. En la actualidad, la SG ha sido de gran utilidad en mejoramiento vegetal, se tiene registro de que el primer estudio de simulación en una especie en particular, maíz (*Zea mays* L.), fue hecha por Bernardo y Yu (2007) y consistió en la comparación de la SG y la selección asistida por marcadores (MAS), resultando que la primera es más efectiva. Posteriormente se han hecho estudios para otros cultivos, cebada (Lorenzana y Bernardo, 2009), trigo (Zhao *et al.*, 2013), arroz

(Xu *et al.*, 2014), etc. El modelo paramétrico básico para predicción en SG propuesto por Meuwissen *et al.* (2001) está dado por:

$$y_i = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + e_i,$$

o en términos matriciales

$$\boldsymbol{y} = \mu\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{2.1}$$

donde $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^t$ vector respuesta, $\mathbf{X}$ matriz de los marcadores, codificados como $0, 1, 2$ para homocigotos dominantes, recesivos y heterocigotos, respectivamente, y $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^t$ los parámetros desconocidos, que resumen el efecto de los marcadores y $\boldsymbol{e}$ error aleatorio, $\boldsymbol{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

Las estimaciones de los parámetros por los métodos convencionales, Máxima Verosimilitud (MV) y Mínimos Cuadrados Ordinarios (MCO) se vuelven técnicas no viables para obtener dichas estimaciones, esto ocurre, porque en los análisis que involucran datos genómicos, es común que $n << p$ (maldición de la dimensionalidad Bellman y Kalaba, 1961), lo que implica que la inversa de $\mathbf{X}^t\mathbf{X}$ no existe, por la alta dimensionalidad y la múltiple colinealidad entre sus columnas; para resolver tal inconveniente suelen usarse los métodos de regularización. Algunos de ellos se presentan a continuación.

2.2.1. Regresión Ridge

Considerado como uno de los Métodos de Regresión Penalizada, ya que minimiza la Suma de Cuadrados del Error (SCE) sujeta a ciertas restricciones en la estimación de los coeficientes de regresión. Aunque se obtienen estimadores sesgados, tienen la ventaja de ser de menor varianza logrando con esto una mayor precisión en la predicción.

Recordando que el estimador de $\boldsymbol{\beta}$ por mínimos cuadrados, en el modelo 2.1, bajo el supuesto sin pérdida de generalidad de que $\mu = 0$, está dado por:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{y},$$

la inestabilidad que presenta este estimador ante la multicolinealidad, se puede reducir agregando una constante $\lambda > 0$ a cada término de la diagonal de $\mathbf{X}^t\mathbf{X}$ antes de invertirla, generando con esto el estimador *ridge*. En la Regresión Ridge (Hoerl y Kennard, 1970), el objetivo es minimizar la suma de cuadrados penalizada con la norma L_2 :

$$\sum_{i=1}^{n}(y_i - \mathbf{x}_i^t\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}{\beta_j}^2 = \| \boldsymbol{y} - \boldsymbol{X\beta} \|^2 + \lambda \| \boldsymbol{\beta} \|^2 . \tag{2.2}$$

Minimizando la expresión dada en la ecuación (2.2), se obtiene el estimador Ridge con respecto a $\boldsymbol{\beta}$ y considerando $\lambda > 0$ fijo y conocido:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^t\boldsymbol{y}. \tag{2.3}$$

Este estimador penalizado involucra *'encogimiento'* y por consiguiente evita sobreajuste del modelo, además estabiliza las estimaciones en comparación a la estimación hecha por MCO (Hoerl y Kennard, 1970; Piepho, 2009). Sin embargo tiene algunas desventajas: al incluir todas las covariables al modelo, no es un método apropiado si el objetivo es la selección de variables; el valor de λ depende de $\boldsymbol{\beta}$ el cual es desconocido y penaliza a todos los ${\beta_j}'s$.

2.2.2. Regresión LASSO

Tibshirani (1996) propuso el método de Regresión LASSO (*Least Absolute Shrinkage and Selection Operator* por sus siglas en inglés) que combina la selección de variables y la penalización de los parámetros. El estimador LASSO se obtiene al minimizar

$$\min_{\boldsymbol{\beta}}\{(\boldsymbol{y} - \boldsymbol{X\beta})^t(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda \| \boldsymbol{\beta} \|\}$$

para algún $\lambda > 0$ y $\| \boldsymbol{\beta} \|$ denota la norma L_1. Este método encoge rápidamente los efectos de los marcadores cuya significancia es pequeña, haciéndolos cero. El algoritmo *LARS* (*Least Angle Regression Selection* por sus siglas en inglés) propuesto por Efron *et al.* (2004), produce resultados similares a los procedimientos de selección de variables hacia adelante y hacia atrás; lo cual es conveniente cuando se desea alguna característica de interés o cuando se trata de predictores altamente correlacionados. Esto último puede ser una desventaja, si todos los predictores están correlacionados la regresión LASSO es superada por la regresión Ridge (Tibshirani, 1996; Zou y Hastie, 2005).

2.2.3. Regresión Ridge Bayesiana

La regresión Ridge Bayesiana (BRR) se obtiene al asignar una distribución *a priori* a cada elemento del vector de parámetros $\boldsymbol{\beta}$,

$$p(\boldsymbol{\beta}|\sigma^2_\beta) = \left(\frac{1}{2\pi\sigma^2_\beta}\right)^{\frac{p}{2}} exp\left\{-\frac{1}{2\sigma^2_\beta}\boldsymbol{\beta}^t\boldsymbol{\beta}\right\}.$$

Suponiendo que, $\boldsymbol{e} \sim N(\mathbf{0}, \sigma^2_e\mathbf{I})$ y si se asigna una distribución a priori no informativa para σ^2_e, es decir,

$$p(\sigma^2_e) \propto \frac{1}{\sigma^2_e},$$

se obtiene que la distribución a posteriori de $\boldsymbol{\beta}$ está dada por:

$$p(\boldsymbol{\beta} \mid datos) \propto exp\left\{-\frac{1}{2\sigma^2_e}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^t\boldsymbol{\Sigma}_\beta(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\right\}. \tag{2.4}$$

Entonces, de (2.4) se desprende que $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}^*, \boldsymbol{\Sigma}_\beta)$ donde:

$$\boldsymbol{\beta}^* = \left(\mathbf{X}^t\mathbf{X} + \frac{\sigma^2_e}{\sigma^2_\beta}\mathbf{I}\right)^{-1}\mathbf{X}^t\boldsymbol{y}$$

y

$$\boldsymbol{\Sigma}_\beta = \left(\mathbf{X}^t\mathbf{X} + \frac{\sigma^2_e}{\sigma^2_\beta}\mathbf{I}\right)^{-1},$$

por lo tanto, el estimador de $\boldsymbol{\beta}$ si $\lambda = \dfrac{\sigma^2_e}{\sigma^2_\beta}$, está dado por:

$$\hat{\boldsymbol{\beta}} = E[\boldsymbol{\beta} \mid \boldsymbol{y}] = (\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^t\boldsymbol{y},$$

que es el estimador Ridge que se obtuvo en (2.3).

2.2.4. LASSO Bayesiano (BL)

Park y Casella (2008) proponen un enfoque Bayesiano para la regresión LASSO usando una a priori condicional *Laplace* (o Doble Exponencial) para $\boldsymbol{\beta}$;

$$p(\boldsymbol{\beta} \mid \lambda, \sigma^2_e) = \prod_{j=1}^{p}\left(\frac{\lambda}{2\sigma^2_e}\right) exp\left\{-\frac{\lambda|\beta_j|}{\sigma^2_e}\right\},$$

y una a priori no informativa para σ_e^2 dada por

$$p(\sigma_e^2) \propto \frac{1}{\sigma_e^2}.$$

Por lo tanto, la distribución final estará dada por:

$$p(\boldsymbol{\beta}|\lambda, datos) \propto \prod_{j=1}^{p} p(\beta_j|\sigma_e^2) \prod_{i=1}^{n} N(y_i|\mathbf{x}_i^t\boldsymbol{\beta}, \sigma_e^2)p(\sigma_e^2)$$

Como se puede observar, obtener la distribución a posteriori no es posible de forma analítica, por lo tanto, se utiliza un modelo jerárquico, para obtener las distribuciones condicionales que permiten obtener muestras de la distribución conjunta haciendo uso del muestreador de Gibbs (Geman y Geman, 1984). Pérez-Rodríguez *et al.* (2010) desarrollaron el paquete BLR (Bayesian Linear Regression) para R (R Core Team, 2017) el cual ajusta de forma eficiente los modelos BRR y BL.

Los modelos anteriores fueron desarrollados para hacer predicciones basadas en información genotípica y bajo supuestos de una única varianza asociada a los marcadores. Meuwissen *et al.* (2001) propusieron los modelos llamados *Bayes A* y *Bayes B* en los cuales se supone que cada marcador tiene su propia varianza.

2.2.5. Bayes A, Bayes B

Meuwissen *et al.* (2001) propusieron dos modelos, Bayes A y Bayes B. En Bayes A, la distribución a priori del efecto de un marcador β_j se supone normal con media cero y varianza $\sigma_{\beta_j}^2$ y la varianza asociada con el efecto de cada marcador, se supone con distribución a priori χ-*cuadrada escalada invertida* (Gianola *et al.*, 2009). En el primer nivel se tiene:

$$\beta_j \mid \sigma_{\beta_j}^2 \sim N(0, \sigma_{\beta_j}^2);$$

y en el segundo nivel

$$\sigma_{\beta_j}^2 \mid \nu, S \sim \chi^{-2}(\nu, S).$$

Por lo tanto, la distribución marginal a priori inducida para β_j se obtiene por integración,

$$\begin{aligned} p(\beta_j \mid \nu, S) &= \int_0^\infty N(\beta_j|0, \sigma^2_{\beta_j}) p(\sigma^2_{\beta_j} \mid \nu, S) d\sigma^2_{\beta_j} \\ &\propto \int_0^\infty (\sigma^2_{\beta_j})^{-((1+\nu+2)/2)} exp\left(-\frac{\beta_j^2 + S}{2\sigma^2_{\beta_j}}\right) d\sigma^2_{\beta_j} \\ &\propto \left(1 + \frac{\beta_j^2}{\nu S^2}\right)^{-\frac{(\nu+1)}{2}}, \end{aligned}$$

la última expresión corresponde al núcleo de la densidad de una distribución t (Gianola *et al.*, 2009). En Bayes B, Meuwissen *et al.* (2001) proponen:

$$\begin{aligned} \beta_j \mid \sigma^2_{\beta_j} &\sim \begin{cases} c & \sigma^2_{\beta_j} = 0 \\ N(0, \sigma^2_{\beta_j}) & \sigma^2_{\beta_j} > 0, \end{cases} \\ \sigma^2_{\beta_j} \mid p_0 &= \begin{cases} 0 & \text{con probabilidad } p_0 \\ \chi^{-2}(\nu, S) & \text{con probabilidad } 1 - p_0. \end{cases} \end{aligned}$$

Implicando que la distribución a priori conjunta de β_j y $\sigma^2_{\beta_j}$ dado un parámetro $p_0 \in (0, 1)$, sea la siguiente:

$$p(\beta_j, \sigma^2_{\beta_j} \mid p_0) = \begin{cases} \beta_j = c \ y \ \sigma^2_{\beta_j} = 0 & \text{con probabilidad } p_0 \\ N(0, \sigma^2_{\beta_j})\chi^{-2}(\nu, S) & \text{con probabilidad } 1 - p_0. \end{cases}$$

Marginalmente, después de integrar, la a priori para β_j toma la siguiente forma (Gianola *et al.*, 2009):

$$p(\beta_j \mid p_0) = \begin{cases} \beta_j = c & \text{con probabilidad } p_0 \\ t(0, \nu, S) & \text{con probabilidad } 1 - p_0. \end{cases}$$

Por lo tanto, Bayes B se reduce a Bayes A si $p_0 = 0$.

Posterior a esto se han desarrollado diferentes modelos, *Bayes C, Bayes*c$-\pi$, etc., con el objetivo final de hacer más precisas las predicciones basadas en información genotípica. Aunque en estos modelos la variable ambiente no es incluida de manera explícita, su efecto es incorporado en el error aleatorio.

2.2.6. Regresión RKHS

Un método semiparamétrico utilizado en selección genómica, para predicción de fenotipos, es el denominado regresión RKHS (*Reproducing Kernel Hilbert Spaces*, por sus siglas en inglés) (Gianola *et al.*, 2006; Gianola y van Kaam, 2008), que surge como una opción a los enfoques paramétricos para capturar las múltiples y complejas interacciones que potencialmentes pueden surgir en modelos de predicción genómica. En este método, se supone que los marcadores moleculares se usan para crear una estructura que modela la dependencia entre individuos; $K(\mathbf{x}_i, \mathbf{x}_j)$, donde $\mathbf{x}_i, \mathbf{x}_j$, son vectores de marcadores del *i-ésimo* y *j-ésimo* individuo, respectivamente (de los Campos *et al.*, 2009b).

Utiliza una función llamada *kernel* (normalmente, una densidad simétrica alrededor del cero), mediante la cual se transforma el conjunto de datos (marcadores moleculares), en un conjunto de distancias entre pares de observaciones, generando una matriz cuadrada que puede ser utilizada como matriz de varianzas y covarianzas para un efecto aleatorio en el modelo mixto.

Debido a que la regresión *RKHS* no asume linealidad, se considera que podría captar de mejor manera los efectos no aditivos. El modelo se representa como (Heslot *et al.*, 2012):

$$\boldsymbol{y} = \mathbf{W}\boldsymbol{\theta} + \mathbf{K}_h\boldsymbol{\alpha} + \mathbf{e},$$

donde $\boldsymbol{\theta}$ es un vector de efectos fijos, $\boldsymbol{W}$ matriz de incidencia para los efectos fijos, $\mathbf{e}$ representa el error aleatorio y se supone que $\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2\mathbf{I})$ y $\boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{K}_h\sigma_\alpha^2)$. $\mathbf{K}_h$ es una matriz que depende de la función RK(*Reproducing Kernel*, por sus siglas en inglés), con parámetro de suavizamiento h, el cual mide la "distancia genética" entre marcadores y controla la razón de decaimiento de la correlación entre marcadores.

Una función kernel muy conocida, es el *kernel Gaussiano* representado de la siguiente manera

$$K_h(\boldsymbol{x}_i, \boldsymbol{x}_j) = exp(-hd_{ij}),$$

donde d_{ij} es la distancia euclideana entre los marcadores del individuo i con el individuo j, es decir:

$$d_{ij} = \sum_{k=1}^{p} \frac{(x_{ik} - x_{jk})^2}{p},$$

siendo p igual al número de marcadores. En Crossa *et al.* (2010) y de los Campos *et al.* (2009a, 2010) se pueden consultar más detalles de este método.

2.3. El modelo mixto

Los modelos mixtos son una extensión de los modelos de regresión que permiten la incorporación de efectos aleatorios (Ruppert *et al.*, 2003). La ecuación que los representa está dada por:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

donde $\boldsymbol{y}$ es un vector de n observaciones, $\boldsymbol{\beta}$ es el vector que representa los efectos fijos, $\mathbf{u}$ el vector que representa los efectos aleatorios, $\mathbf{X}$ matriz que relaciona las observaciones con los efectos fijos y $\mathbf{Z}$ matriz que relaciona las observaciones con los efectos aleatorios, $\mathbf{e}$ es el vector de los errores aleatorios.

Resumiendo, los supuestos usuales sobre la esperanza y la varianza de los componentes aleatorios, se tiene que:

$$E\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}$$

y

$$Cov\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}\sigma^2,$$

con $\mathbf{G}$ y $\mathbf{R}$ matrices positivas definidas conocidas y σ^2 una constante positiva. Cuando se supone distribución normal para el vector de observaciones, la función de densidad queda completamente determinada por el vector de valores esperados y la matriz de varianzas y covarianzas. La matriz de varianzas y covarianzas de $\boldsymbol{y}$ está dada por:

$$\begin{aligned} Var(\boldsymbol{y}) = \mathbf{V} &= Var(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}) \\ &= \mathbf{Z}Var(\mathbf{u})\mathbf{Z}^t + Var(\mathbf{e}) \\ &= \mathbf{Z}\mathbf{G}\mathbf{Z}^t + \mathbf{R}. \end{aligned}$$

Las estimaciones por mínimos cuadrados generalizados, pueden usarse para estimar los efectos fijos del modelo mixto. Estas estimaciones se obtienen minimizando $(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^t\mathbf{V}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})$ con respecto a $\boldsymbol{\beta}$ y el estimador del vector de efectos fijos está dado por:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{V}^{-1}\boldsymbol{y}. \tag{2.5}$$

Si todas las componentes de varianza en $\mathbf{V}$ son conocidas, este estimador es el *mejor estimador lineal insesgado*(BLUE, *Best Linear Unbiased Estimator*, por sus siglas en inglés). Si se supone que $\mathbf{u}$ y $\mathbf{e}$ tienen distribución normal, la mejor estimación se logra con métodos basados en máxima verosimilitud (ML, Maximum Likelihood por sus siglas en inglés) y máxima verosimilitud restringida (REML, Restricted Maximum Likelihood por sus siglas en inglés). Para estimar el vector de efectos aleatorios, $\mathbf{u}$, Henderson (1963) demostró que el mejor predictor lineal insesgado, BLUP(*Best Linear Unbiased Predictor* por sus siglas en inglés) de $\mathbf{u}$ está dado por:

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}^t\mathbf{V}^{-1}(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{2.6}$$

El cálculo de los estimadores BLUE y BLUP requiere calcular la inversa de la matriz $\mathbf{V}$, cuya dimensión es del orden de la dimensión de $\boldsymbol{y}$. Aunque en algunos casos obtener la inversa puede ser fácil por las características de su estructura, en general la naturaleza de los datos y el elevado número de ellos, hace imposible o muy costosa la inversión de $\mathbf{V}$.

La solución a este problema, la encontró Henderson (1953), al desarrollar un conjunto de ecuaciones denominadas *Ecuaciones del Modelo Mixto* (MME, Mixed Model Equations por sus siglas en inglés), las cuales permiten obtener de manera conjunta las estimaciones de $\boldsymbol{\beta}$ y $\mathbf{u}$ bajo el supuesto de que las estructuras de covarianza son conocidas (Witkovský, 2012). Estas ecuaciones son derivadas por la maximización de la densidad conjunta de $\boldsymbol{y}$ y $\mathbf{u}$ la cual está dada, para $Var(\mathbf{e}) = \mathbf{R}$ y $Var(\mathbf{u}) = \mathbf{G}$ por:

$$f(\boldsymbol{y}, \boldsymbol{u}) = \frac{exp\{-[(\boldsymbol{y} - \boldsymbol{X\beta} - \mathbf{Zu})^t\mathbf{R}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta} - \mathbf{Zu}) + \mathbf{u}^t\mathbf{G}^{-1}\mathbf{u}]/2\}}{(2\pi)^{(n+df_u)/2}|\mathbf{R}|^{1/2}|\mathbf{G}|^{1/2}} \tag{2.7}$$

Igualando a cero las derivadas parciales de (2.7) con respecto a $\boldsymbol{\beta}$ y $\mathbf{u}$, se obtienen las denominadas, ecuaciones del modelo mixto,

$$\begin{bmatrix} \mathbf{X}^t\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^t\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^t\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^t\mathbf{R}^{-1}\boldsymbol{y} \\ \mathbf{Z}^t\mathbf{R}^{-1}\boldsymbol{y} \end{bmatrix},$$

cuya solución coincide con 2.5 y 2.6.

2.4. Distribución Normal-Asimétrica

En el contexto de los modelos lineales, existen muchas aplicaciones reales donde los supuestos de normalidad y homocedasticidad no se satisfacen. Ocurre que en muchos casos, la distribución de frecuencias asociada a la variable respuesta es marcadamente asimétrica. Azzalini (1985) introduce el concepto de la *Distribución Normal-Asimétrica*, que permite modelar la asimetría presente en un conjunto de datos. Su función de densidad de probabilidades está dada por:

$$f_Z(z) = 2\phi(z)\Phi(\lambda z)I_{(-\infty,\infty)}(z), \tag{2.8}$$

donde $\phi(\cdot)$ y $\Phi(\cdot)$ denotan la función de densidad y distribución de una variable normal estándar, respectivamente, y $\lambda \in \mathbb{R}$ regula la asimetría (forma).

Si Z tiene la función de densidad dada en (2.8), entonces se escribe $Z \sim SN(\lambda)$. Algunas propiedades de la distribución SN, que la hacen matemáticamente tratable son:

1) Si $\lambda = 0$ en (2.8) entonces $Z \sim N(0,1)$.

2) Si $Z \sim SN(\lambda)$ entonces $-Z \sim SN(-\lambda)$.

3) Si $\lambda \to \infty$ entonces (2.8) converge a la distribución media-normal, $2\phi(z)I_{(0,\infty)}(z)$.

4) Si $Z \sim SN(\lambda)$ entonces $Z^2 \sim \chi^2_1$.

Usando los resultados de Azzalini (1985), la función generadora de momentos de (2.8) está dada por

$$M_Z(t) = 2exp(t^2/2)\Phi\left(\frac{\lambda t}{\sqrt{1+\lambda^2}}\right),$$

de donde es posible obtener la media y la varianza, dados por:

$$E(Z) = \sqrt{\frac{2}{\pi}}\frac{\lambda}{\sqrt{1+\lambda^2}}, \quad Var(Z) = 1 - \frac{2}{\pi}\frac{\lambda^2}{1+\lambda^2}. \tag{2.9}$$

La función de distribución acumulada de (2.8) está dada por:

$$F_Z(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\pi} exp(-x^2/2) \int_{-\infty}^{\lambda x} exp(-t^2/2)dtdx$$

En la Figura 2.1 se puede observar la densidad de la Normal Asimétrica para distintos valores de λ. Algunas de las propiedades enunciadas se pueden identificar con facilidad en la Figura 2.1.

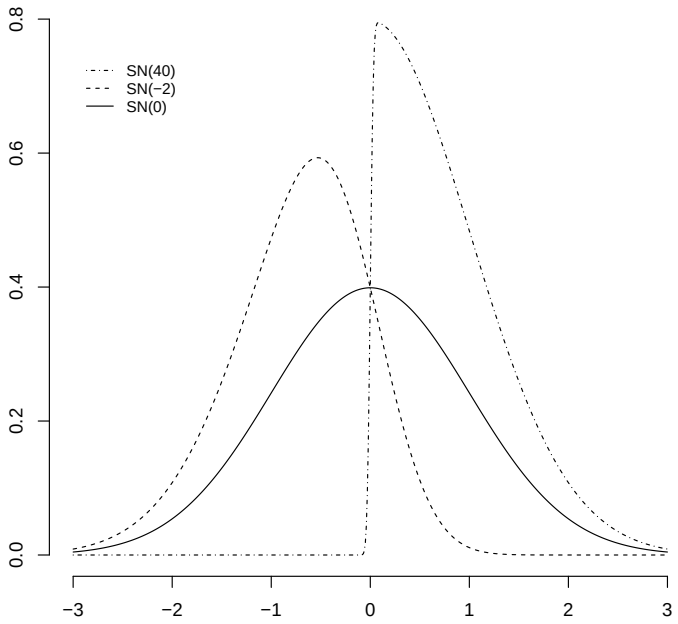

Figura 2.1: Densidad de la Normal Asimétrica con distintos parámetros de forma λ. Observe que si $\lambda = 0$ se obtiene la densidad $N(0, 1)$ y si $\lambda \to \infty$ se obtiene la distribución media normal.

Esta distribución tiene algunas propiedades de la distribución normal univariada, pero presenta problemas desde el punto de vista de la estimación de sus parámetros, en particular, el estimador de máxima verosimilitud para el parámetro de forma tiende a infinito con probabilidad positiva (Azzalini, 1985; Pérez-Rodríguez *et al.*, 2017).

2.5. Inferencia Bayesiana

La metodología Bayesiana está basada en la interpretación subjetiva de la probabilidad y tiene como punto central el Teorema de Bayes. Bajo el enfoque Bayesiano, el vector de parámetros de interés, $\boldsymbol{\theta}$, sobre el cual se desea hacer algún tipo de inferencia es desconocido, de tal modo que el objetivo es usar datos, $\boldsymbol{y}$, junto con un modelo paramétrico dado, para realizar análisis sobre los parámetros desconocidos. Lo único que se requiere para el proceso de inferencia Bayesiana es la especificación previa de

una distribución a priori, la cual representa el conocimiento acerca del parámetro antes de obtener cualquier información respecto a los datos.

Un modelo básico tiene un vector de parámetros de interés ($\boldsymbol{\theta}$), unos datos observados ($\boldsymbol{y}$) que se relacionan a través de la función de densidad de un modelo paramétrico $p(\boldsymbol{y}|\boldsymbol{\theta})$ (función de verosimilitud) y el objetivo es relacionar probabilísticamente $\boldsymbol{\theta}$ con los datos. El Teorema de Bayes permite esta relación, al obtenerse la densidad a posteriori de $\boldsymbol{\theta}|\boldsymbol{y}$, es decir:

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

y $p(\boldsymbol{y})$, se puede obtener de dos maneras:

$$p(\boldsymbol{y}) \quad = \quad \begin{cases} \sum_{\boldsymbol{\Theta}} p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) & \text{si } \boldsymbol{\theta} \text{ es discreto.} \\ \int_{\boldsymbol{\Theta}} p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} & \text{si } \boldsymbol{\theta} \text{ es continuo,} \end{cases}$$

donde $\boldsymbol{\Theta}$, denota el espacio paramétrico de $\boldsymbol{\theta}$ y $p(\boldsymbol{y})$ es la constante de normalización de $p(\boldsymbol{\theta}|\boldsymbol{y})$, y es llamada la distribución marginal de los datos o la distribución predictiva inicial (Carlin y Louis, 2000). Entonces, como $p(\boldsymbol{y})$ no depende de $\boldsymbol{\theta}$ se tiene que:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

Generalmente es difícil obtener $p(\boldsymbol{\theta}|\boldsymbol{y})$, de forma analítica, por lo cual es común utilizar algoritmos de integración numérica como Metropolis-Hastings (Hastings, 1970), muestreador de Gibbs (Geman y Geman, 1984), etc., los cuales se describen a continuación.

2.6. Muestreador de Gibbs

El muestreador de Gibbs (Geman y Geman, 1984) es una técnica de Cadenas de Markov Monte Carlo (MCMC), que permite muestrear de una distribución conjunta utilizando sus distribuciones condicionales.

Suponga que $p(\boldsymbol{\theta}), \boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)^t$ es una densidad conjunta y se desea:

$$p(\theta_1) = \int \ldots \int p(\theta_1, \ldots, \theta_k) d\theta_2, d\theta_3, \ldots, d\theta_k,$$

pero en muchas casos la integral anterior es muy compleja de obtener de forma analítica. En estos casos el Muestreador de Gibbs, el cual se describe brevemente a continuación, es un método numérico alternativo para resolver el problema.

Sea $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^t$, donde θ_i es univariado o multivariado y suponga que se puede muestrear de las densidades condicionales totales, $p(\boldsymbol{\theta}_i|\boldsymbol{y}, \boldsymbol{\theta}_j, j \neq i)$. El muestreador de Gibbs genera una cadena de Markov como sigue: Se comienza con un valor inicial $\boldsymbol{\theta}_0 = (\theta_1^0, \theta_2^0, \ldots, \theta_k^0)^t$ en el soporte de $p(\boldsymbol{\theta}|\boldsymbol{y})$. En el tiempo t, la realización $\boldsymbol{\theta}_t = (\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_k^{(t)})^t$ se obtiene de la siguiente manera:

1. Se hace $t = 1$.
2. Se genera $\theta_1^{(t)}$ de $p(\theta_1|\boldsymbol{y}, \theta_2^{(t-1)}, \ldots, \theta_k^{(t-1)})$.
3. Se genera $\theta_2^{(t)}$ de $p(\theta_2|\boldsymbol{y}, \theta_1^{(t)}, \theta_3^{(t-1)}, \ldots, \theta_k^{(t-1)})$.

   $\vdots$
4. Se genera $\theta_k^{(t)}$ de $p(\theta_k|\boldsymbol{y}, \theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_{k-1}^{(t)})$.
5. Repetir los pasos $2-4$ hasta tener un número B de muestras.

Las primeras m muestras se denominan, muestras de *calentamiento* (burn-in), así que se pueden desechar y solamente tomar en cuenta las últimas $B - m$ muestras.

2.7. Muestreador de Metropolis

Este algoritmo es un método de Monte Carlo vía Cadenas de Markov (MCMC, por sus siglas en inglés) el cual ha sido ampliamente utilizado en Física, restauración de imágenes y actualmente en Estadística Carlin y Louis (2000). El método en su versión original fue propuesto por Metropolis *et al.* (1953), y evoluciona al conocido algoritmo como Metropolis-Hastings (Hastings, 1970). A continuación se hace una breve descripción de su funcionamiento.

Suponga que se desea muestrear de una distribución $p(\boldsymbol{\theta})$ completamente conocida, entonces, para construir la cadena $\{\boldsymbol{\theta}^{(t)}\}$, se definen las probabilidades de transición de la siguiente manera.

Se genera un punto candidato $\boldsymbol{\theta}^*$ de una distribución propuesta (comúnmente llamada *generadora de candidatos*), $q(\cdot|\boldsymbol{\theta})$, la cual sólo depende del estado $\boldsymbol{\theta}$. El punto $\boldsymbol{\theta}^* = \boldsymbol{\theta}^{(t)}$ se acepta, con probabilidad de transición definida de la siguiente manera:

Sea $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$ una distribución de transición y se define

$$\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \text{mín}\left(\frac{p(\boldsymbol{\theta}^*|\boldsymbol{y})q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}|\boldsymbol{y})q(\boldsymbol{\theta}^*|\boldsymbol{\theta})}, 1\right),$$

la razón de aceptación, con la cual se decide si se acepta o se rechaza el punto candidato. Si el punto $\boldsymbol{\theta}^*$ no se acepta, entonces la cadena no se mueve y $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.

Algoritmo

1. Dar un valor inicial $\boldsymbol{\theta}^{(0)}$ en el soporte de $p(\cdot)$.
2. Generar una observación $\boldsymbol{\theta}^*$ de $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})$.
3. Generar una variable $u \sim U(0, 1)$.
4. Si $u < \alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(t-1)})$,

   $$\text{hacer} \quad \boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*,$$

   en caso contrario,

   $$\text{hacer} \quad \boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}.$$

Esto genera una cadena de Markov $(\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \cdots, \boldsymbol{\theta}^{(k)}, \cdots)$, donde la probabilidad de transición de $\boldsymbol{\theta}^{(t-1)}$ a $\boldsymbol{\theta}^{(t)}$, depende sólo de $\boldsymbol{\theta}^{(t-1)}$ y no de $\boldsymbol{\theta}^{(0)}, \cdots, \boldsymbol{\theta}^{(t-2)}$ y cuya distribución de transición es:

$$P(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t-1)}) = \alpha(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t-1)})q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t-1)}).$$

Si la distribución propuesta q es simétrica, es decir:

$$q(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = q(\boldsymbol{\theta}|\boldsymbol{\theta}^*),$$

entonces

$$\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \text{mín}\left(\frac{p(\boldsymbol{\theta}^*|\boldsymbol{y})}{p(\boldsymbol{\theta}|\boldsymbol{y})}, 1\right).$$

Existen dos casos particulares de $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$ que comúnmente se utilizan (Smith y Roberts, 1993), *caminata aleatoria e independiente*.

Caso Caminata Aleatoria: Sea $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = q_1(\boldsymbol{\theta}^* - \boldsymbol{\theta})$, donde $q_1(\cdot)$ es una densidad de probabilidades simétrica centrada en el origen. Entonces el punto candidato se toma como:

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}^{(t)} + \boldsymbol{\eta},$$

donde $\boldsymbol{\eta}$ es una variable aleatoria de incremento de la distribución propuesta q y con base en lo anterior:

$$\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \text{mín}\left(\frac{p(\boldsymbol{\theta}^*|\boldsymbol{y})}{p(\boldsymbol{\theta}|\boldsymbol{y})}, 1\right).$$

Caso Independiente: Sea $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = q_0(\boldsymbol{\theta}^*)$, donde $q_0(\cdot)$ es una densidad de probabilidad sobre $\boldsymbol{\theta}$, por lo tanto:

$$\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \text{mín}\left(\frac{\omega(\boldsymbol{\theta}^*)}{\omega(\boldsymbol{\theta}}, 1\right),$$

con $\omega(\boldsymbol{\theta}) = \dfrac{p(\boldsymbol{\theta}|\boldsymbol{y})}{q_0(\boldsymbol{\theta})}$ (Tierney, 1994).

Capítulo 3

Predicción del Rendimiento de Híbridos de Maíz usando Modelos G×A

La predicción del rendimiento de híbridos es muy importante en los programas agrícolas. En el fitomejoramiento, las interacciones multi-ambientales para evaluar experimentos de genotipado con el ambiente, desempeñan un papel importante en la selección de fenotipos con buenas características.

En este capítulo abordamos la predicción del rendimiento de híbridos en multiambientes usando dos modelos de selección genómica con la interacción *genotipo* × *ambiente*. El trabajo ya ha sido publicado, ver Acosta-Pech *et al.* (2017).

3.1. Introducción

Con el uso de la tecnología doble-haploide, el número de híbridos simples que pueden ser probados en campo ha aumentado significativamente, por lo tanto la predicción del rendimiento de ellos es de gran importancia en los programas de mejoramiento modernos.

van Eeuwijk *et al.* (2010) señalan que, aunque algunos investigadores sostienen que la Aptitud Combinatoria Específica (ACE) de líneas parentales es el factor principal que determina el rendimiento de híbridos, otros concluyen que la principal fuerza de éste es la acción genética aditiva (Bernardo, 1996a,b; Duvick, 2005). Sin embargo, cuando se estudia el rendimiento de híbridos, es importante considerar dos fuentes de variación: Aptitud Combinatoria General (ACG) o efectos aditivos entre líneas y Aptitud Combinatoria Específica o efectos no aditivos entre híbridos, tales como dominancia o

desviaciones epistáticas.

El uso de los modelos mixtos para calcular el Mejor Predictor Lineal Insesgado (BLUP, por sus siglas en inglés) de híbridos, utilizando únicamente datos de campo de individuos desarrollados de líneas relacionadas con datos de marcadores o pedigrí, fue propuesta y usada por Bernardo (1994). Los resultados mostraron que la relación entre líneas puede mejorar la precisión predictiva de híbridos no probados.

Posteriormente (Bernardo, 1996a,b, 1999), usando la matriz de relaciones de pedigrí (obtenida de los coeficientes de ancestría) obtuvo resultados prometedores en la predicción de rendimiento de híbridos no observados basados en algunos híbridos observados. Estos BLUP de los híbridos no observados, en función de la matriz de relaciones de pedigrí son análogos a la predicción de líneas no observadas con base a marcadores moleculares, como originalmente los propuso Meuwissen *et al.* (2001) en su trabajo de Selección Genómica usando todos los marcadores posibles.

Tomando como base el trabajo desarrollado por Bernardo (1994, 1996a, 1999) y los resultados de los estudios en predicción genómica llevados a cabo en los últimos 4-5 años, la predicción tipo BLUP por medio de la regresión Ridge, RR-BLUP o su modelo equivalente, con la matriz de relaciones genómicas (GBLUP) (VanRaden, 2008), se ha empleado ampliamente en predicción de rendimiento de híbridos (Massman *et al.*, 2013; Piepho, 2009; Schrag *et al.*, 2010; Technow y Melchinger, 2013; Technow *et al.*, 2012; Xu *et al.*, 2014; Zhao *et al.*, 2013).

Los modelos de selección genómica para predicción de rendimiento de híbridos, desarrollados por Massman *et al.* (2013), consideran la varianza de la aptitud combinatoria general así como también la varianza debida a la aptitud combinatoria específica entre líneas endogámicas; estos autores compararon la precisión en la predicción de rendimiento de híbridos del modelo genómico, RR-BLUP, con la precisión en la predicción del parentesco basado en marcadores e información acerca de líneas relacionadas, BLUP. Como resultado, los autores no encontraron ninguna mejora en la precisión de la predicción del rendimiento de híbridos con el RR-BLUP genómico en comparación con el BLUP estándar, ambos calculados en un gran número de ambientes.

En el mejoramiento de plantas, los experimentos multi-ambientales para evaluar interacciones *genotipo* × *ambiente* (G×A) juegan un papel importante en la selección de fenotipos con un buen rendimiento y estabilidad a través de ambientes. Las condiciones ambientales modulan la expresión genética, y esto induce la interacción G×A de tal modo que las correlaciones genéticas estimadas del desempeño de una línea individual a través de ambientes, resume la acción conjunta de los genes y las condiciones ambientales (López-Cruz *et al.*, 2015). Estudios recientes demostraron que los modelos lineales mixtos permiten considerar estructuras ambientales correlacionadas en el marco de GBLUP y por lo tanto, pueden predecir rendimiento de fenotipos no observados usando marcadores moleculares y pedigrí.

Burgueño *et al.* (2012) fueron los primeros en usar modelos GBLUP con marcadores y pedigrí para evaluar predicción genómica bajo G × A; los autores mostraron que modelar G × A usando marcadores y pedigrí aumenta considerablemente el poder predictivo de los modelos. Heslot *et al.* (2014) modelaron datos de cultivos para estudiar la interacción G × A, y de los resultados que se obtuvieron con un gran conjunto de datos de trigo mostraron que la precisión en la predicción se incrementa en promedio un 11 %. Jarquín *et al.* (2014) propusieron el modelo GBLUP con efectos aleatorios, donde los efectos principales y la interacción de los marcadores y las covariables ambientales se introducen usando estructuras de covarianza de alta dimensionalidad, entre los marcadores y las covariables ambientales. Los autores mencionan que el incremento en la precisión predictiva de los modelos ajustados cuando se incorpora el término G × A, se incrementó entre el 17 % y el 34 % en comparación con aquellos modelos donde no se incluye este término. Por lo tanto, dada la importancia del término G × A en mejoramiento vegetal y el incremento en el poder predictivo de los modelos que se ajustan para predecir rendimiento, la pregunta es cómo incorporar este término en modelos de predicción genómica para predecir rendimiento de híbridos.

Se supone que la incorporación de $G \times A$ en los modelos GBLUP, puede aumentar la precisión en la predicción de rendimiento de híbridos. Sin embargo, ninguno de los estudios anteriores sobre predicción de rendimiento de híbridos ha incorporado, de manera explícita, el término $G \times A$ en los modelos ajustados. El objetivo principal de este capítulo es, en base a los modelos desarrollados por Technow *et al.* (2012) y Massman *et al.* (2013), incluir el término de interacción $G \times A$ haciendo uso del modelo norma de reacción propuesto por Jarquín *et al.* (2014). Los modelos propuestos consideran la interacción de los efectos de Aptitud Combinatoria Específica × ambiente y Aptitud Combinatoria General × ambiente.

Se hace una aplicación del modelo propuesto a través de un conjunto de datos con 2,724 híbridos de maíz que se obtuvieron de la cruza de 531 líneas endogámicas, de las cuales 507 líneas fueron *Dent* y 24 líneas *Flint* que se usaron como probadores. Las líneas fueron genotipadas con Illumina 50k y se evaluaron durante 12 años ($2004 - 2015$) en 58 diferentes localidades que se usaron para ilustrar el desempeño de los modelos ajustados.

3.2. Modelos Estadísticos

Massman *et al.* (2013) propusieron un modelo para predecir rendimiento de híbridos usando información genotípica de padres. Esta información se utilizó para construir matrices de relación para los padres y para los híbridos. El modelo propuesto, es un modelo mixto lineal que incluye efectos aleatorios debido a la aptitud combinatoria general de los padres y aptitud combinatoria específica de los híbridos, cuyas matrices de varianza-covarianza se construyen sobre los marcadores. En Technow *et al.* (2014) se comparó la precisión en la predicción obtenida mediante los modelos de predicción

GBLUP (Bernardo, 1996a; Massman *et al.*, 2013) y BayesB, y se concluyó que el poder de predicción de ambos métodos era aproximadamente la misma.

En esta propuesta, se extienden los modelos tipo GBLUP con el fin de tener en cuenta el efecto del ambiente y el efecto de la interacción Genotipo× Ambiente. Se ajustan dos modelos:

- $GBLUP + Amb.$
- $GBLUP + Amb + Híbrido \times Amb + Padres \times Amb.$

El primer modelo es el que se discutió en Massman *et al.* (2013) y Technow *et al.* (2014). El segundo modelo, extiende el modelo 1 al incluir el efecto de la interacción. Una descripción de los modelos ajustados se presenta a continuación. En ambos modelos se supone varianzas homogéneas para los ambientes.

Modelo 1: GBLUP+Amb

El modelo lineal para el rendimiento de una cruza simple que incluye el efecto principal del ambiente ($\boldsymbol{\beta}_E$ año-localidad) está dado por (Technow *et al.*, 2014):

$$\boldsymbol{y} = \mathbf{Z}_E\boldsymbol{\beta}_E + \mathbf{Z}_D\mathbf{g}_D + \mathbf{Z}_F\mathbf{g}_F + \mathbf{Z}_H\mathbf{h} + \boldsymbol{e}, \tag{3.1}$$

donde $\boldsymbol{y}$ es el vector respuesta, es decir, la información fenotípica (rendimiento) de los híbridos ajustados, $\mathbf{Z}_E$ es la matriz diseño asociada a los ambientes, $\boldsymbol{\beta}_E$ representa el vector de efectos ambientales y se supone que $\boldsymbol{\beta}_E \sim N(\mathbf{0}, \sigma^2_E\mathbf{I})$, $\mathbf{g}_D$ es el vector de efectos aleatorios asociado a la aptitud combinatoria general (ACG) de las líneas Dent, $\mathbf{g}_F$ es el vector de efectos aleatorios asociado a ACG de marcadores para las líneas Flint y $\mathbf{h}$ es el vector de efectos aleatorios asociado a la aptitud combinatoria específica (ACE) para los híbridos. $\mathbf{Z}_D, \mathbf{Z}_F, \mathbf{Z}_H$ son matrices de incidencia que relacionan $\boldsymbol{y}$ con $\mathbf{g}_D, \mathbf{g}_F, \mathbf{h}$; además se supone que $\mathbf{g}_D \sim N(\mathbf{0}, \sigma^2_D\mathbf{G}_D)$, $\mathbf{g}_F \sim N(\mathbf{0}, \sigma^2_F\mathbf{G}_F)$ y $\mathbf{h} \sim N(\mathbf{0}, \sigma^2_H\mathbf{H})$, σ^2_D, σ^2_F y σ^2_H son los parámetros de varianza asociados con aptitud combinatoria general y específica, respectivamente y finalmente, $\boldsymbol{e} \sim N(\mathbf{0}, \sigma^2_e\mathbf{I})$, donde σ^2_e es la varianza asociada con los errores.

Las matrices de relaciones genómicas, $\mathbf{G}_D$ y $\mathbf{G}_F$, se construyen en base a los marcadores (VanRaden, 2008). Sea $\mathbf{W}_k, k \in \{Dent, Flint\}$ la matriz de marcadores cuyas entradas son 0 y 2 para denotar homocigotos recesivos y dominantes, respectivamente. Sea $\mathbf{Z}_k$ la matriz de marcadores centrada y estandarizada, esto es,

$$z_{ijk} = \frac{(x_{ijk} - 2p_{jk})}{\sqrt{4p_{jk}(1-p_{jk})}},$$

donde i indexa a los individuos y j a los marcadores, p_{jk} es la frecuencia alélica del alelo referente en la población de las líneas *Dent* o *Flint* (ver Technow *et al.*, 2014).

Entonces

$$\mathbf{G}_k = \frac{\mathbf{Z}_k \mathbf{Z}_k^t}{p}$$

(López-Cruz *et al.*, 2015; Technow *et al.*, 2014) donde p es el número de marcadores. Ésto genera un valor promedio de la diagonal de $\mathbf{G}_k$ cercano a uno; por lo tanto σ_k^2 se define en la misma escala que σ_e^2, es decir como un parámetro de varianza.

Los elementos de la matriz $\mathbf{H}$ se obtienen directamente de las matrices $\mathbf{G}_D$ y $\mathbf{G}_F$ (ver Bernardo, 2002, pags:231-232), y Technow *et al.* (2014). La derivación se hace usando el hecho de que la aptitud combinatoria específica de los híbridos se representa como una interacción de primer orden entre líneas maternas y paternas. Se incluye la derivación del resultado porque la técnica es usada para introducir la interacción G × A

Sea h_{ij} representando el efecto de la interacción de un híbrido obtenido a partir de una cruza simple del individuo i en la población Dent y el individuo j en la población Flint. Supongamos que $h_{ij} = \mathbf{g}_{D_i} \times \mathbf{g}_{F_j}$, entonces la media y la función de covarianza son las siguientes:

$$\mathbb{E}[h_{ij}] = \mathbb{E}[\mathbf{g}_{D_i} \times \mathbf{g}_{F_j}] = \mathbb{E}[\mathbf{g}_{D_i}] \times \mathbb{E}[\mathbf{g}_{F_j}] = 0,$$

y la covarianza está dada por:

$$\begin{aligned}
Cov(h_{ij}, h_{i'j'}) &= \mathbb{E}[h_{ij} \times h_{i'j'}] - \mathbb{E}[h_{ij}] \times \mathbb{E}[h_{i'j'}] \\
&= \mathbb{E}[h_{ij} \times h_{i'j'}] - 0 \times 0 \\
&= \mathbb{E}[(\mathbf{g}_{D_i} \times \mathbf{g}_{F_j}) \times (\mathbf{g}_{D_{i'}} \times \mathbf{g}_{F_{j'}})] \\
&= \mathbb{E}[(\mathbf{g}_{D_i} \times \mathbf{g}_{D_{i'}}) \times (\mathbf{g}_{F_j} \times \mathbf{g}_{F_{j'}})] \\
&= \mathbb{E}[\mathbf{g}_{D_i} \times \mathbf{g}_{D_{i'}}] \times \mathbb{E}[\mathbf{g}_{F_j} \times \mathbf{g}_{F_{j'}}] \\
&= Cov(\mathbf{g}_{D_i}, \mathbf{g}_{D_{i'}}) \times Cov(\mathbf{g}_{F_j}, \mathbf{g}_{F_{j'}}) \\
&\propto \mathbf{G}_{D_{ii'}} \times \mathbf{G}_{F_{jj'}}
\end{aligned}$$

donde $\mathbf{G}_{D_{ii'}}$ y $\mathbf{G}_{F_{jj'}}$ son las entradas respectivas de las matrices $\mathbf{G}_D$ y $\mathbf{G}_F$. En notación compacta, la matriz $\mathbf{H}$ para todos las posibles cruzas, se obtiene como el producto Kronecker de $\mathbf{G}_D$ y $\mathbf{G}_F$, esto es $\mathbf{H} = \mathbf{G}_D \otimes \mathbf{G}_F$ (Covarrubias-Pazaran, 2016).

Modelo 2: GBLUP+Amb+Híbrido × Amb + Padres × Amb

Jarquín *et al.* (2014) sugirieron modelar la interacción entre los marcadores y las covariables ambientales utilizando un proceso Gaussiano con una clase específica de funciones de covarianza que se genera en base a un modelo de norma de reacción. Estos autores mostraron, que si la función de covarianza generada por los términos de interacción, se obtiene utilizando un modelo multiplicativo de primer orden, la estructura de covarianza es el producto Haddamard(celda×celda) de dos estructuras de covarianza, una que describe la información genética y la otra el efecto ambiental. Usando este enfoque se extiende el modelo (3.1) al incluir el término de la interacción Genotipo(Híbrido) × Ambiente y la interacción Padres × Ambiente. El modelo propuesto es el siguiente:

$$\boldsymbol{y} = \mathbf{Z}_E\boldsymbol{\beta}_E + \mathbf{Z}_D\mathbf{g}_D + \mathbf{Z}_F\mathbf{g}_F + \mathbf{Z}_H\mathbf{h} + \mathbf{u}_H + \mathbf{u}_D + \mathbf{u}_F + \boldsymbol{e}, \tag{3.2}$$

donde $\mathbf{u}_H \sim N(\mathbf{0}, \sigma^2_{hA}\mathbf{V}_H)$, $\mathbf{u}_D \sim N(\mathbf{0}, \sigma^2_{DA}\mathbf{V}_D)$, $\mathbf{u}_F \sim N(\mathbf{0}, \sigma^2_{FA}\mathbf{V}_F)$, σ^2_{hA}, σ^2_{DA},σ^2_{FA} son componentes de varianza asociados con la interacción híbrido×ambiente, Dent× ambiente, Flint×ambiente, respectivamente y $\mathbf{V}_H, \mathbf{V}_D, \mathbf{V}_F$ son las matrices de varianzas-covarianzas asociadas. Los elementos de la matriz $\mathbf{V}_H$ se pueden obtener, siguiendo el enfoque de Jarquín *et al.* (2014). Suponiendo que la interacción entre híbridos y ambientes Eh_{ijk} se puede representar como $h_{ij} \times E_k$, donde $E_k = \beta_{E_k}, k = 1, \cdots, E$(ambientes), entonces la media y la función de covarianzas son las siguientes.

$$\mathbb{E}[h_{ij} \times E_k] = \mathbb{E}[h_{ij}] \times \mathbb{E}[E_k] = 0 \times 0 = 0,$$

y la covarianza, es como sigue

$$\begin{aligned} Cov(Eh_{ijk}, Eh_{i'j'k'}) &= \mathbb{E}[Eh_{ijk} \times Eh_{i'j'k'}] - \mathbb{E}[h_{ijk}] \times \mathbb{E}[h_{i'j'k'}] \\ &= \mathbb{E}[h_{ijk} \times E_k \times h_{i'j'k'} \times E_{k'}] - 0 \times 0 \\ &= \mathbb{E}[(h_{ij} \times h_{i'j'}) \times (E_k \times E_{k'})] \\ &= \mathbb{E}[h_{ij} \times h_{i'j'}] \times \mathbb{E}[E_k \times E_{k'}] \\ &= Cov(h_{ij}, h_{i'j'}) \times Cov(E_k, E_{k'}) \\ &\propto \mathbf{G}_{D_{ii'}} \times \mathbf{G}_{F_{jj'}} \times Cov(E_k, E_{k'}). \end{aligned}$$

Observe que la $Cov(E_k, E_{k'}) \neq 0$ si $k = k'$. Usando estos resultados, la matriz de varianzas-covarianzas está dada por

$$\mathbf{V}_H = \mathbf{Z}_h\mathbf{H}\mathbf{Z}_h^{'}\#\mathbf{Z}_E\mathbf{Z}_E^{'}$$

donde # representa el producto Hadamard. Note que si las observaciones están ordenadas por ambientes, entonces $\mathbf{V}$ es una matriz diagonal por bloques, cuya estructura es similar a la del modelo GBLUP marcador ×ambiente en López-Cruz *et al.* (2015). Sin embargo no es necesario ordenar las observaciones por ambiente para ajustar el modelo. De manera análoga, se pueden obtener $\mathbf{V}_D = \mathbf{Z}_D\mathbf{G}_D\mathbf{Z}_D'\#\mathbf{Z}_E\mathbf{Z}_E'$ y $\mathbf{V}_F = \mathbf{Z}_F\mathbf{G}_F\mathbf{Z}_F'\#\mathbf{Z}_E\mathbf{Z}_E'$ (ver Jarquín *et al.*, 2014, para más detalles).

Con respecto a la expresión asociada a la ecuación de la covarianza que se muestra arriba, se observa que Eh_{ijk} representa el efecto de la interacción de un híbrido, h_{ij} que se obtiene de una sola cruza entre el individuo i de la población Dent y el individuo j de la población Flint evaluado en el ambiente k. De la misma manera se define $Eh_{i'j'k'}$. Si $i = i'$ y $j = j'$ entonces $Cov(h_{ij}, h_{i'j'}) = Var(h_{ij})$ que corresponde a los elementos de la diagonal de $\mathbf{H}$, de otro modo, éstos representan los elementos fuera de la diagonal de $\mathbf{H}$.

3.2.1. Evaluación del modelo

Los modelos 3.1 y 3.2 se ajustaron usando los datos completos en cada año, y las estimaciones de los parámetros de varianza se obtuvieron de este análisis. Los hiper-parámetros de las distribuciones a priori de se fijaron de acuerdo a las reglas dadas en el material suplementario en de los Campos y Pérez-Rodríguez (2015). Es decir, para los parámetros asociados a la varianza se tiene:

$$\begin{aligned} \sigma^2_{\boldsymbol{\beta}_E} &\sim \chi^{-2}(df_{\boldsymbol{\beta}}, S_{\boldsymbol{\beta}}) \\ \sigma^2_{\mathbf{e}} &\sim \chi^{-2}(df_{\mathbf{e}}, S_{\mathbf{e}}) \end{aligned}$$

En el caso de $\sigma^2_D, \sigma^2_F, \sigma^2_H, \sigma^2_{hA}, \sigma^2_{DA}, \sigma^2_{FA}$, las distribuciones a priori, también fueron, $\chi^{-2}(df, S)$ y cada una con sus respectivos grados de libertad. Se obtuvieron las predicciones en base a la estimación de los BLUP's de los componentes aleatorios y se realizaron las inferencias con respecto a la media a posteriori obtenida de las interaciones realizadas en el muestreador de Gibbs, esto porque la función de pérdida que se usa en el paquete BGLR, es la función de pérdida cuadrática. Las predicciones se obtienen como sigue:

$$\hat{\boldsymbol{y}} = \mathbf{Z}_E\hat{\boldsymbol{\beta}}_E + \mathbf{Z}_D\hat{\mathbf{g}}_D + \mathbf{Z}_F\hat{\mathbf{g}}_F + \mathbf{Z}_H\hat{\mathbf{h}},$$

y

$$\hat{\boldsymbol{y}} = \mathbf{Z}_E\hat{\boldsymbol{\beta}}_E + \mathbf{Z}_D\hat{\mathbf{g}}_D + \mathbf{Z}_F\hat{\mathbf{g}}_F + \mathbf{Z}_H\hat{\mathbf{h}} + \hat{\mathbf{u}}_H + \hat{\mathbf{u}}_D + \hat{\mathbf{u}}_F.$$

3.2.2. Validación Cruzada

La validación cruzada es un método muy común que se usa para estimar el error de predicción y comparar diferentes modelos. Consiste en dividir los datos en subconjuntos disjuntos, comúnmente llamados grupos (en este estudio se usaron 5) y así los datos quedan divididos en conjuntos de tamaño aproximado $k = [n/5]$. Se utilizó esta técnica para estimar el poder predictivo de los modelos ajustados. Para este fin, se simuló un problema común al que se enfrentan los mejoradores cuando prueban nuevas líneas usando experimentos de campo incompletos: cómo predecir el rendimiento de híbridos desarrollados recientemente. Los híbridos son evaluados en algunos ambientes pero no en otros y su rendimiento tiene que ser predicho en ambientes en los que no se evaluaron. Para imitar este problema, se realizó un análisis de validación cruzada utilizando un esquema que se conoce como CV2 que considera algunas líneas que se observan en algunos ambientes, pero no en otros. El problema es predecir las líneas que no se observaron en esos ambientes (ver por ejemplo Burgueño *et al.*, 2012; Jarquín *et al.*, 2014). En CV2, los registros individuales de un híbrido se asignan a los grupos (Pérez-Rodríguez *et al.*, 2015). Se realizó una validación cruzada de *5-grupos* para cada año, esto implica que se ajustaron 120 modelos (12×5×2) usando técnicas de Cadenas de Markov Monte Carlo.

La validación cruzada de *5-grupos* se ha utilizado mucho en otros estudios, bajo este esquema 80 % de los registros se usan en el conjunto de entrenamiento y 20 % en el conjunto de prueba. Los modelos 1 y 2 se ajustaron utilizando los registros en el conjunto de entrenamiento y las predicciones se obtuvieron en el conjunto de prueba con la finalidad de obtener la precisión en la predicción de los modelos. Se obtuvo el coeficiente de correlación de Pearson entre las observaciones en el conjunto de prueba y los valores observados, para cada año y se calculó una correlación promedio mediante la ponderación de la correlación individual en cada sitio de acuerdo con el número de híbridos predichos en cada uno de ellos. Correlaciones altas entre valores observados y ajustados, es un indicador de que el modelo ajustado será mejor para predecir valores futuros. En el anexo A, se incluye el código R (R Core Team, 2017) utilizado para generar las particiones para este tipo de validación cruzada.

3.3. Software

Los modelos descritos arriba se ajustaron utilizando el paquete de R (R Core Team, 2017) Regresión Lineal Bayesiana Generalizada (*BGLR*, Bayesian Generalized Linear Regression). El software puede ser descargado libremente de `https://cran.r-project.org/web/packages/BGLR/index.html`. Para más detalles ver de los Campos y Pérez-Rodríguez (2015). En el Apéndice A se incluyen los códigos en R utilizados para ajustar los dos modelos. Las inferencias se hicieron con base a 30,000 iteraciones para el muestreador de Gibbs (Geman y Geman, 1984), 5,000 de las

cuales se tomaron como calentamiento.

3.4. Datos Experimentales

Fenotipos

El conjunto de datos fue proporcionado por el programa de mejoramiento de maíz en *RAGT* (https://www.ragtsemences.com). Los datos incluyen 2,724 híbridos de maíz provenientes de 531 líneas endogámicas de maíz, 507 son líneas Dent y 24 líneas Flint que se utilizaron como probadores. Se evaluaron los híbridos durante 12 años (2004-2005) en 58 localidades diferentes. Los rasgos analizados fueron: porcentaje de almidón ajustado (%SC), porcentaje de materia seca (%DMC) y rendimiento de ensilaje (YLD) en kg/ha (el ensilaje es un proceso de conservación del forraje basado en una fermentación láctica del pasto que produce ácido láctico y una disminución del pH por debajo de 5). Los fenotipos se ajustaron tomando en cuenta los efectos del diseño de campo.

Se utilizaron dos diseños de campo en los ensayos: *i*) el diseño de bloques aumentado estándar, cuando las líneas verificadas se hallan en un diseño de bloques completos al azar (Federer y Raghavarao, 1975); y *ii*) un bloque completo al azar con dos repeticiones para cada híbrido. La Figura 3.1 muestra una representación esquemática de los híbridos probados, obtenidos cruzando las líneas Dent y Flint. Los híbridos probados en el campo, están representados en cuadros negros, alrededor del 22 % de todos los híbridos posibles fueron probados en el campo.

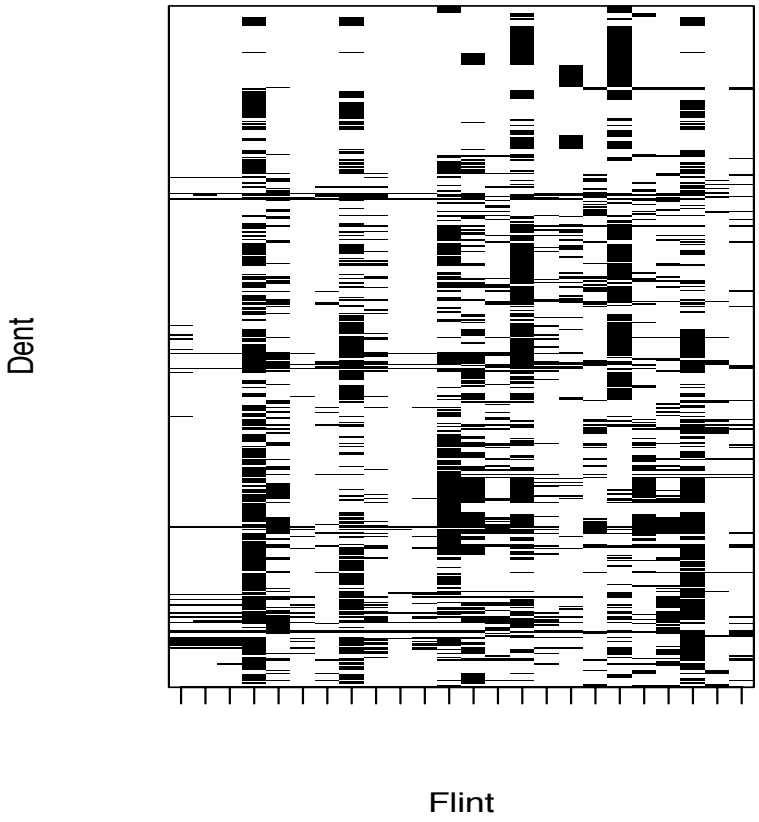

Figura 3.1: Representación esquemática de los híbridos probados en campo, obtenidos cruzando 507 líneas Dent con 24 líneas Flint. Los híbridos probados en campo están representados por cuadros en negro. Alrededor del 22 % de todos los posibles híbridos fueron probados.

Genotipos

Las líneas se genotiparon usando el chip *50k Illumina* para maíz (http://www.illumina.com), de las cuales se obtuvieron 49,013 SNP's. Se aplicaron controles de calidad estándar a los datos, removiendo todos los marcadores no bi-alélicos y marcadores no mapeados. Utilizando el software *Beagle v3.2* (Browning y Browning, 2009), https://faculty.washington.edu/browning/beagle/b3.html) se imputaron valores faltantes en los genotipos. Después de la edición, 22,690 marcadores estuvieron disponibles para hacer las predicciones.

La Figura 3.2 muestra un gráfico de los primeros dos vectores propios de la matriz de

relaciones genómicas para las líneas Dent y Flint, las cuales se distinguen claramente en él. La proporción de la varianza explicada por los dos primeros componentes principales fue del 20 % y sólo 90 componentes principales fueron necesarios para explicar el 80 % de la suma de los valores propios de la matriz de relaciones genómicas (los datos no se muestran).

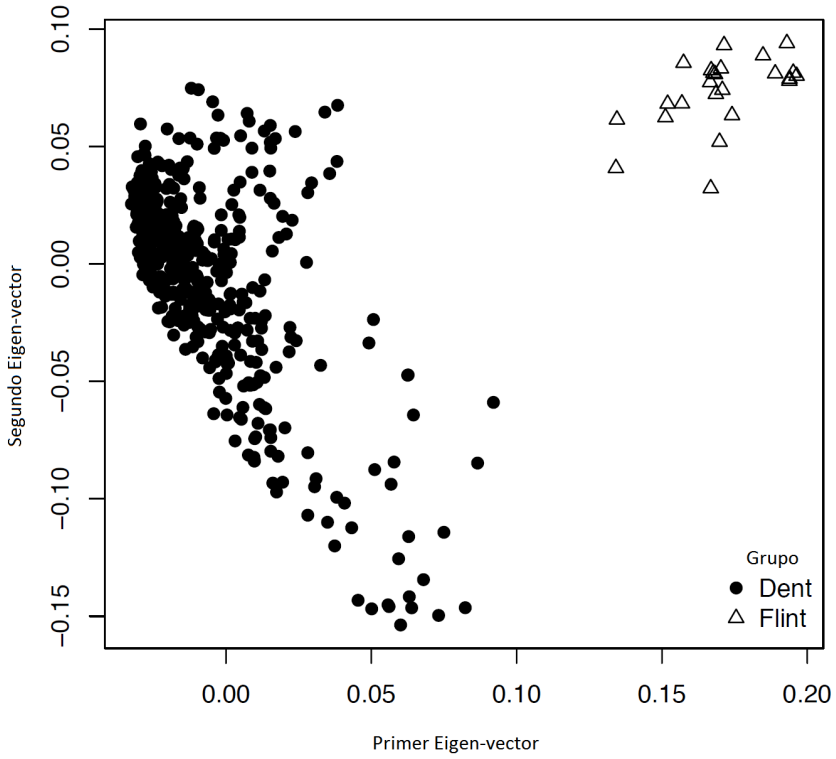

Figura 3.2: Gráfico de los dos primeros vectores propios de la matriz de relaciones Genómicas de las líneas Dent y Flint.

3.5. Estimación de los Parámetros de Varianza

Se describe la estimación de los parámetros de varianza de los modelos ajustados (3.1) y (3.2) para determinar la proporción de la varianza total que es explicada por cada componente. Los parámetros de varianza de los modelos ajustados son: ambientes (A), aptitud combinatoria general para las líneas Dent (D), aptitud combinatoria general para las líneas Flint (F), aptitud combinatoria específica de híbridos (H) y la interacción entre híbridos y ambientes (H×A), líneas Dent × ambiente (D×A), líneas Flint × ambiente (F×A) y residuales (Res), los cuales se obtuvieron a partir del análisis de datos completos cuando se ajustaron los modelos (3.1) y (3.2). Los resultados se muestran en las Tablas 3.1 y 3.2 para *YLD* y en el caso de *%SC* y *%DMC* los resultados se incluyen en el anexo A, Tablas A.1-A.4. En general, el parámetro asociado a la varianza del efecto principal,

ambientes (Localidades) representa una gran proporción de la varianza total, sin embargo, el componente A no suele ser considerado en la estimación de los parámetros de varianza, esto es debido al propósito de distinguir diferencias entre los otros parámetros, que son pequeños en comparación con la magnitud de A.

Cuando el modelo (3.2) se ajusta la varianza residual disminuye consistentemente, en comparación con el modelo (3.1). Dependiendo del rasgo, la disminución promedio en la varianza residual es de aproximadamente del 10 al 16 %.

Para YLD los resultados de los modelos ajustados en todos los años indican que la ACG del componente Flint (F) mostró una variabilidad mucho mayor que la explicada por la ACG del componente dent (D). En general la ACE, mostró la menor variabilidad de los tres componentes genéticos. En términos de la varianza total explicada por los tres componentes de interacción, al ajustar el modelo (3.2), F×A explicó la mayor variabilidad, seguido por los componentes D×A y H×A. El patrón de los resultados obtenidos en los parámetros de varianza para los rasgos %SC y %DMC para todos los años y con los dos modelos, fue similar al obtenido para YLD (ver Anexo A).

Tabla 3.1: Parámetros de varianza estimados (A=ambientes; D=Dent; F=Flint; H=Híbridos, Res=Residual), desviación estándar (en paréntesis) y porcentaje de varianza dentro de los ambientes explicada por cada efecto aleatorio para YLD estimado, ajustando el Modelo (3.1).

Modelo 1: GBLUP+Amb					
Año	A	D	F	H	Res
2004	349.5(125.5) --	61.2(17.9) 14.2	163.5(73.3) 36.1	42.3(11) 9.9	169.4(8.3) 39.8
2005	662.3(228.8) --	74.9(22.4) 15.5	97.8(43) 19.8	64.4(16.2) 13.4	243.8(10.2) 51.2
2006	449.2(132.6) --	77.4(23.6) 16.8	84.7(37.4) 18	55.3(13.6) 12.1	241.1(9.2) 53.1
2007	437.7(146.5) --	89.2(25.6) 17.6	126.1(55.1) 24.2	69(15.1) 13.8	222.1(8.3) 44.3
2008	619.2(181.3) --	67.8(17.5) 14.9	109.7(47.6) 23.3	49.8(10.3) 11	229.5(7.9) 50.8
2009	517.9(155) --	56.4(12.7) 12.4	162.3(68.1) 33.9	44.6(8.1) 9.8	199(6.3) 43.9
2010	464(132.2) --	68.2(15) 16.5	79.4(29.7) 18.9	52.7(9.7) 12.8	211.8(6.2) 51.7
2011	618(188.5) --	62.3(13.1) 14.8	133.1(52.6) 30.4	44.7(8) 10.7	185.1(5.7) 44.2
2012	588.2(162.7) --	36.3(7.2) 10	90.7(31.6) 24.4	46.5(7.4) 12.8	191(5) 52.8
2013	488.9(138.5) --	51.4(10) 12.2	77.1(27.9) 18	41.8(6.8) 10	250.5(6.4) 59.8
2014	404.5(116.3) --	67.8(13.4) 15.7	81.9(30.8) 18.6	47.4(7.8) 11	235.5(5.5) 54.7
2015	663.4(172.4) --	50.1(8.9) 12.7	82.5(29.4) 20.5	42.2(6.4) 10.8	219.5(5) 56

Tabla 3.2: Parámetros de varianza estimados (E=ambientes; D=Dent; F=Flint; H=Híbridos; H×E=Híbridos×Env; D×E=Dent×Env; F×E=Flint×Env; Res=Residual), desviación estándar (en paréntesis) y porcentaje de varianza dentro de los ambientes explicada por cada efecto aleatorio para YLD estimado, ajustando el Modelo (3.2).

Modelo2: GBLUP+Amb+Híbrido×Amb+Padres×Amb								
Año	A	D	F	H	H ×A	D × A	F × A	Res
2004	306.1(113.5)	43.5(14.3)	122.9(59.2)	30.5(8.7)	23(5.7)	27.6(6.9)	43.5(12.3)	123.3(7.1)
	–	10.6	28.5	7.5	5.6	6.8	10.6	30.3
2005	656.5(230.4)	56.2(19)	70.7(33.5)	62.3(17.5)	36.8(9.3)	34.9(8.9)	43.6(12.1)	176.9(8.8)
	–	11.6	14.4	12.9	7.7	7.3	9.1	37
2006	448.3(137.6)	61.3(20.9)	57(26.2)	44(12.9)	32.7(8)	29.6(7.8)	33.3(8.3)	186.7(8.3)
	–	13.7	12.6	9.9	7.4	6.7	7.5	42.2
2007	426.3(143.8)	73.4(23.2)	94.3(45.9)	62.1(15.4)	32.3(7.8)	29(6.6)	42.5(10)	167.2(7.2)
	–	14.6	18.3	12.5	6.5	5.8	8.5	33.7
2008	584(170.8)	54.4(15.5)	68.1(30.1)	34.2(8)	30.7(6.5)	41(8.3)	41.5(8.2)	157.6(6.4)
	–	12.7	15.6	8	7.2	9.6	9.7	37.1
2009	485.1(152.4)	45.3(11.1)	131(56.5)	36.3(7.9)	39.3(7.7)	30.2(6.3)	39.4(9.1)	145.8(5.8)
	–	9.8	27.2	7.9	8.5	6.5	8.5	31.6
2010	468.9(140.4)	52.3(12.6)	52.8(21.3)	49.5(9.9)	22.7(4.7)	37.1(6.3)	29.7(5.8)	158.8(5.6)
	–	13	12.9	12.3	5.7	9.2	7.4	39.6
2011	611.1(192.9)	57.3(13)	107.9(45.4)	34.8(7.3)	27.3(4.8)	22.4(4.2)	43.3(7.9)	135.8(4.9)
	–	13.4	24.5	8.2	6.4	5.3	10.2	32
2012	572.9(160.7)	28.1(6.1)	68.2(26.6)	41.6(7.3)	21.1(3.8)	21.3(3.7)	27.1(5.2)	153.9(4.5)
	–	7.8	18.5	11.5	5.9	5.9	7.5	42.8
2013	436.3(129.1)	40.5(8.9)	56.4(22.3)	32.1(6.1)	22.6(4.1)	23.6(4.5)	39.8(7.8)	206.5(6)
	–	9.6	13.2	7.6	5.4	5.6	9.4	49.2
2014	357(103.7)	65.5(13.1)	60.1(23.9)	44.7(8.2)	22.5(4.7)	17.7(3.4)	45.2(7.7)	180.8(5.1)
	–	15	13.6	10.3	5.2	4.1	10.4	41.6
2015	639(172.4)	40.1(8)	56.3(21.5)	35.5(6.1)	32.7(5.1)	19.5(3.5)	31.7(5.7)	174.9(4.8)
	–	10.3	14.2	9.1	8.4	5	8.1	44.9

3.6. Precisión en la Predicción de los modelos (3.1) y (3.2)

Las correlaciones promedio entre fenotipos y predicciones obtenidas en CV2 se reportan por modelo en la Tabla 3.3 para el rasgo *YLD* y para los rasgos *%SC* y *%DMC* en las Tablas A.5 y A.6 del anexo A, respectivamente. Los resultados se obtuvieron mediante el coeficiente de correlación de Pearson y oscilaron entre 0.42 a 0.50 para el modelo (3.1) y de 0.48 a 0.60 para el modelo (3.2) dependiendo del tratamiento analizado.

Al tener correlaciones mayores con el modelo (3.2), se muestra que tiene un mayor poder predictivo que con el modelo (3.1). La Tabla 3.3 también reporta el cambio en el porcentaje de la precisión predictiva del modelo (3.2) versus el modelo (3.1).

Tabla 3.3: [a] %Cambio M1 vs M2 = $(r_{M2} - r_{M1})/r_{M1} \times 100$

Año	M1	M2	$\%Cambio\ M1 \text{vs} M2^a$
2004	0.5380	0.56817	5.59
2005	0.3876	0.4667	20.41
2006	0.3436	0.3806	10.77
2007	.5138	0.5616	8.35
2008	0.3147	0.4756	51.13
2009	0.3999	0.4666	16.68
2010	0.4256	0.4736	11.28
2011	0.4766	0.5564	16.74
2012	0.4953	0.5339	7.79
2013	0.3692	0.4246	15.01
2014	0.4200	0.5089	21.17
2015	0.3830	0.4435	15.80
Promedio	0.4227	0.4883	16.73

Para *YLD*, el porcentaje de cambio (%) para el modelo (3.2) *v.s.* modelo (3.1) fue de 16.73 % (Tabla 3.3) En dos años (2005 y 2008), la superioridad en el poder predictivo del modelo (3.2) sobre (3.1) alcanzó 20 % y 50 %, respectivamente. Para el rasgo %SC, el cambio porcentual promedio para el modelo (3.2) frente al modelo (3.1) fue de 21.74 % (Tabla A.5). Finalmente para el rasgo %DMC, el porcentaje promedio de aumento en la precisión de la predicción fue de 21.74 % (Tabla A.6). En resumen, en la mayoría de los casos el cambio porcentual promedio en la precisión predictiva del modelo (3.2) sobre el modelo (3.1) es positivo, lo cual indica que el modelo (3.2) tiene mejor poder predictivo que el modelo (3.1).

Estos resultados también se representan en la Figura 3.3 que muestra la distribución del cambio porcentual para el modelo (3.2) *v.s* (3.1) para los rasgos %DMC, %SC y YLD. Los resultados indican la importancia de incorporar G× A para aumentar la precisión de la predicción híbrida en los tres rasgos analizados en este estudio.

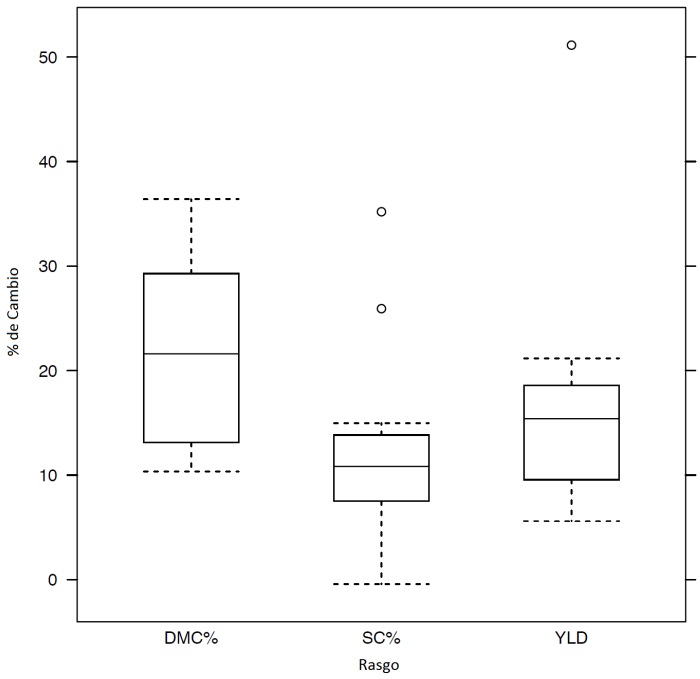

Figura 3.3: Comparación de Modelos. Gráfico de caja y bigote del porcentaje de cambio en la precisión de la predicción, calculado tomando como base al modelo (3.1). Porcentaje de cambio M1 vs M2 $= \frac{(r_{M1}-r_{M2})}{r_{M1}} \times 100$, donde r_{M1} y r_{M2} son los coeficientes de correlación de Pearson para los modelos (3.1) y (3.2), respectivamente.

3.7. Discusión de resultados

Diversos estudios han documentado los beneficios de usar modelos multi-ambientales para evaluar el rendimiento de genotipos en diferentes condiciones ambientales (Burgueño *et al.*, 2012; Dawson *et al.*, 2013; Jarquín *et al.*, 2014). Los análisis multi-ambientes pueden modelar la interacción $G \times A$ usando funciones de covarianza (Burgueño *et al.*, 2012) y marcadores. El objetivo principal en este capítulo fue demostrar que incluir el término $G \times A$, incrementa el poder predictivo de un modelo de Selección Genómica usado para predecir rendimiento de híbridos, cuando solo se tiene información genotípica de los padres. En este capítulo se usó la función de covarianza propuesta por Jarquín *et al.* (2014), quienes la definieron como el producto Haddamart de la matriz de genotipos y la matriz diseño asociada a los ambientes. Usando este enfoque, los autores muestran que un gran porcentaje de la varianza fenotípica es explicada por el efecto principal de los ambientes; esto concuerda con los resultados de los análisis donde se obtuvo los parámetros de varianza para los ambientes (Tablas 3.1, 3.2, A.1-A.6 en el anexo A).

En el caso de la aptitud combinatoria general de los padres, la aptitud combinatoria de los

híbridos y la interacción entre líneas parentales y los ambientes, varían de un tratamiento a otro, pero en general estos términos explican una proporción considerable de la varianza total y cuando se incluyen en el modelo completo, la precisión en la predicción de híbridos no observados se incrementó, como se observa en la Tabla 3.3. Los resultados de este estudio en el maíz muestran la importancia de la precisión en la predicción genómica del rendimiento híbrido, basado tanto en aptitud combinatoria general como en aptitud combinatoria específica. En el mismo sentido, los resultados concuerdan con los de otros investigadores que consideran que el rendimiento híbrido se determina, no solo por los efectos aditivos debidos a la aptitud combinatoria general femenina y masculina, sino también a la interacción de dominancia intra genética que se produce por los efectos de la aptitud combinatoria específica (van Eeuwijk *et al.*, 2010). Además, el efecto G×A, se considera un factor importante no genético que afecta a la heterosis. Este es el primer estudio donde se muestra como la interacción intra genética debida a la dominancia y su interacción con los ambientes se puede modelar para explotar los efectos de interacción genética con los ambientes.

El modelo propuesto es similar al modelo de Massman *et al.* (2013) y Technow *et al.* (2014) excepto que aquí se incluyen los términos de interacción, ACE×A y ACG×A. Estos efectos de interacción afectan de manera positiva la precisión en la predicción. El error residual disminuyó consistentemente al ajustar el modelo (3.2), lo que indica que el término de interacción G×A contribuye con una proporción considerable de la varianza entre los ambientes. Esto es consistente con los hallazgos de Jarquín *et al.* (2014), quienes reportaron que incluir el término de interacción reduce significativamente la varianza del error e incrementa la capacidad predictiva del modelo. También mencionan que la correlación predictiva puede ser afectada por la fuerza de las relaciones genéticas entre líneas, de modo que mediante el uso de un modelo sin G×A para las líneas que no están relacionadas, la predicción de las correlaciones puede ser baja.

La varianza genética no fue significativa en comparación con la predicción del rendimiento en los ambientes. Sin embargo, Bernardo (1994) señala que estimaciones precisas de las variaciones genéticas no son necesarias para predecir del rendimiento de cruzas simples, y que aproximaciones de sus valores son suficientes. En este estudio se supuso la varianza del error homogénea a través de los ambientes, pero el supuesto puede relajarse para considerar el hecho de que los mismos genotipos pueden responde de manera diferente en cada ambiente.

Evaluamos la precisión en la predicción utilizando un esquema de validación cruzada. Este método cuantifica la precisión en la predicción del rendimiento en una combinación particular de año-ambiente incluidas en el conjunto de datos (Pérez-Rodríguez *et al.*, 2015). Crossa *et al.* (2011) mencionan que un enfoque sencillo para evaluar la capacidad predictiva, consiste en dividir los datos en una *muestra de entrenamiento* y una *muestra de validación* o *conjunto de prueba*. Los modelos se ajustaron utilizando la *muestra de entrenamiento* y los modelos ajustados se usaron para predecir resultados en la *muestra de validación*. Este enfoque es apropiado para conjuntos de datos grandes pero no es recomendado en conjuntos de datos pequeños,

porque el tamaño de las muestra de entrenamiento y validación podrían también ser pequeñas (Hastie *et al.*, 2011). Las ganancias en la precisión de la predicción cuando G×A se incluye en el modelo, también son consistentes con los resultados presentados en otros estudios (Burgueño *et al.*, 2012; Jarquín *et al.*, 2014; López-Cruz *et al.*, 2015).

Los resultados confirman la superioridad del modelo (3.2) ($GBLUP + Amb + Híbrido \times Ambiente + Padres \times Ambiente$) con respecto a su capacidad de predicción. Sin embargo, como se menciona en López-Cruz *et al.* (2015), los modelos con interacción están sujetos a la estructura de la matriz de co-varianzas; por ejemplo, la covarianza entre ambientes debe ser positiva y constante entre ambientes. Así, el modelo de interacción se adapta mejor a los ambientes que están correlacionados positivamente.

En un estudio sobre la predicción genómica de RH para identificar cruzas únicas superiores, al principio de un programa de mejoramiento híbrido de maíz, Kadam *et al.* (2016) utlizaron un modelo inicial que incluía la aptitud combinatoria general de los padres, la aptitud combinatoria específica y sus interacciones con los ambientes. Aunque la predicción del rendimiento de cruzas simples se realizó utilizando la aptitud combinatoria general de los padres y la covarianza entre cruzas simples para rendimiento de grano con diferentes esquemas de cruzas simples, prueba/entrenamiento, los autores no modelaron la aptitud combinatoria general de los padres×ambiente y/o la aptitud combinatoria específica×ambiente, y por lo tanto no cuantificaron su impacto en la precisión de la predicción del rendimiento híbrido.

Los resultados obtenidos en este estudio, indican claramente el beneficio de no solo incluir los diversos términos de interacciones en el modelo, sino también modelarlos usando una estructura de varianza-covarianza apropiada, en este caso, dada por el producto Haddamard del modelo propuesto. El modelo utilizado en este estudio permite compartir información de ambientes correlacionados de tal manera que, los híbridos (o padres) observados en algunos ambientes pueden predecirse en otros ambientes en los que no se observaron.

En general, la incorporación de G×A en los modelos para predicción genómica del rendimiento de híbridos, se pueden hacer con cualquier cultivo y en la mayoría de los GBLUP que se han utilizado en estudios genómicos recientes para evaluar la predicción genómica de rendimiento de híbridos en diferentes ambientes.

Capítulo 4

Regresión con errores aleatorios normal asimétricos: Una Aplicación en Selección Genómica

En este capítulo se analiza el ajuste de un modelo de regresión con errores normal asimétricos. Ilustramos su desempeño en el contexto de Selección Genómica a través de un experimento simulado y posteriormente con una aplicación usando datos reales. Evaluamos el poder predictivo del modelo usando dos criterios de Información, el DIC y el pD.

4.1. Introducción

En estudios genéticos de plantas o animales, es común encontrar rasgos cuantitativos cuya distribución no es normal, esto sucede porque los datos se obtienen de múltiples fuentes o contienen observaciones aisladas (Li *et al.*, 2015). Landfors *et al.* (2011) señalan que a menudo es necesario normalizar los datos para eliminar la variación introducida en el desarrollo del experimento, con el riesgo de que tales técnicas de normalización no sean capaces de eliminar el sesgo porque una gran parte de las observaciones pueden ser afectadas de manera positiva o negativa por alguno de los tratamientos. También sucede que sea difícil encontrar una transformación adecuada para los datos, lo que provoca problemas de estimación e interpretación de los resultados obtenidos (Fernandes *et al.*, 2007). Para resolver esta última cuestión se han desarrollado diferentes métodos que son lo suficientemente flexibles para representar los datos y para reducir los supuestos pocos realistas (Arellano-Valle *et al.*, 2005).

En el marco de selección genómica (Meuwissen *et al.*, 2001) se utilizan información

fenotípica y genotípica (marcadores moleculares densos) para predecir los valores genéticos de los candidatos a la selección. La disponibilidad de marcadores moleculares de alta densidad de muchas especies agrícolas, junto con resultados prometedores de simulaciones (por ejemplo Meuwissen *et al.*, 2001) y estudios empíricos en plantas (Crossa *et al.*, 2010, 2011) y animales (por ejemplo, VanRaden, 2008; Weigel *et al.*, 2009), promueve el uso de la selección genómica en varios programas de mejoramiento genético.

En selección genómica, un modelo paramétrico que se utiliza para predecir los fenotipos en base a los marcadores disponibles es el de regresión lineal (Meuwissen *et al.*, 2001), $y_i = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + e_i$, donde x_{ij} representa a los marcadores genotípicos, $x_{ij} \in \{0, 1, 2\}$ el conjunto que representa el número de copias de un marcador dialélico (SNP), β_j es el efecto aditivo del alelo codificado como uno en el *j-ésimo* marcador y e_i, el efecto del error aleatorio. En notación matricial, el modelo es expresado como:

$$\boldsymbol{y} = \mu\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e},$$

siendo $\boldsymbol{y} = \{y_i\}$, $\boldsymbol{\beta} = \{\beta_j\}$ y $\boldsymbol{e} = \{e_i\}$ los valores de los fenotipos, efecto de los marcadores y errores del modelo respectivamente, $\mathbf{X} = \{x_{ij}\}$ es una matriz que contiene a los marcadores genotípicos de dimensión $n \times p$.

Sin embargo, la incorporación de un gran número de marcadores moleculares (p) en un modelo de regresión, generalmente ($p >> n$), afectando con esto, las estimaciones de los parámetros, un problema conocido como de *maldición de dimensionalidad*. Aunado a éste, existe otro problema, *el problema de multicolinealidad*, es decir la correlación lineal entre variables explicativas. Estos problemas causan sobreestimación de la varianza de los estimadores.

Por tanto, una manera de resolver estos problemas es utilizando algún método de regresión penalizada. La mayoría de éstos métodos son equivalentes a la moda posterior de los coeficientes de regresión en cierta clase de modelos Bayesianos. La literatura en selección genómica ofrece una amplia gama de métodos Bayesianos que permiten el análisis de diferentes tipos de distribución a priori asignada a los efectos de los marcadores $\beta_j, j = 1, \ldots, p$, lo cual lleva a lo que se conoce como Alfabeto Bayesiano (Gianola, 2013).

En la mayoría de los modelos mencionados anteriormente, se asume que la variable respuesta proviene de una distribución normal, sin embargo en el análisis de algunos rasgos de interés en SG, es posible que éstos presenten un comportamiento asimétrico, lo cual implica trabajar con distribuciones que son asimétricas, por ejemplo, el tiempo de floración masculina y femenina en plantas, el intervalo de tiempo entre floraciones, enfermedades, etc.

Una manera de resolver el problema de la asimetría en el conjunto de datos, es a través de algún tipo de transformación de la variable respuesta, por ejemplo, usando la transformación de Box-Cox (Box y Cox, 1964), logaritmo o raíz cuadrada o usando un modelo que permita ajustar respuestas asimétricas.

La asimetría puede aparecer como consecuencia de un muestreo no aleatorio, o bien, porque las variables observadas representan una muestra que se ha truncado con respecto a una variable oculta, por ejemplo, selección de sujetos a los cuales se les mide un rasgo (Y) condicionado a un segundo rasgo (O), que cumpla con algún límite superior o inferior, entonces la distribución condicional de $Y|O > o$, para un valor fijo de o conduce a una distribución asimétrica (Arnold y Beaver, 2000). Existe una amplia literatura para distribuciones asimétricas, una de particular interés es la Normal-Asimétrica (SN) (Azzalini, 1985). La distribución SN es una generalización de la distribución normal donde se agrega un parámetro de forma que controla la asimetría.

En este capítulo se propone el uso de la distribución Normal-asimétrica para los términos del error en un modelo de regresión de alta dimensionalidad, $p >> n$, en el contexto de Selección Genómica. El modelo utiliza la representación estocástica de la variable respuesta propuesta por Arnold y Beaver (2000), a fin de facilitar los cálculos.

4.2. Modelos Estadísticos

En esta sección se describe el modelo que será implementado para el análisis de los datos. La estimación desde el punto de vista Bayesiano de los parámetros y los resultados de la simulación son presentados en las siguientes secciones.

4.2.1. Modelo Normal-Asimétrico

La distribución Normal-asimétrica, introducida por Azzalini (1985), permite modelar conjuntos de datos que presentan cierta evidencia de asimetría a través de su función de distribución de probabilidad. Esta distribución incluye a la distribución normal como un caso particular. Su función de densidad está dada por:

$$f_U(u|\lambda) = 2\phi(u)\Phi(\lambda u), \quad u \in \mathbb{R} \tag{4.1}$$

donde $\phi(\cdot)$ y $\Phi(\cdot)$ denotan las funciones de densidad y distribución de una variable normal estándar $N(0,1)$. El parámetro $\lambda \in (-\infty, \infty)$, es el que regula la asimetría y cuando $\lambda = 0$, la ecuación (4.1) se transforma en la función de densidad de una variable aletoria

Normal estándar. La esperanza y varianza están dadas por las expresión en (2.9) y el coeficiente de asimetría de U está dado por la siguiente expresión:

$$\gamma_1 = \frac{[E(U)]^3}{\{1 - [E(U)]^2\}^{3/2}}.$$

Si se considera una transformación lineal de U, $Y = \xi + \omega U$, siendo ξ un parámetro de localidad y ω un parámetro de escala, se obtiene una distribución Normal-Asimétrica para Y, con parámetros (ξ, ω, λ) y se denota como $Y \sim SN_D(\xi, \omega, \lambda)$, donde D representa la parametrización directa de la SN. La función de densidad de Y está dada por:

$$f(y|\xi, \omega, \lambda) = 2\frac{1}{\omega}\phi\left(\frac{y-\xi}{\omega}\right)\Phi\left[\lambda\left(\frac{y-\xi}{\omega}\right)\right].$$

El problema es estimar los parámetros ξ, ω y λ.

4.2.2. Estimación de parámetros

Muchas investigaciones han surgido con respecto a la estimación de los parámetros indexados a la distribución $SN_D(\lambda, \xi, \omega)$. Desde el punto de vista clásico se pueden obtener vía estimadores de momentos, o bien estimadores de máxima verosimilitud, sin embargo, los estimadores de momentos no dan buenos resultados bajo el esquema de la parametrización directa (Pewsey, 2000), y aunque la función de verosimiltud se puede calcular, al momento de maximizarla surgen diferentes problemas. Azzalini (1985) obtuvo la matriz de información de Fisher de los estimadores de máxima verosimilitud (MLE) y en un trabajo posterior Azzalini y Capitanio (1999) mencionan que la forma de la verosimilitud y los MLE hacen complicado realizar inferencias sobre los parámetros empezando con la maximización numérica de la función de verosimilitud y que no es posible removerla con una reparametrización de la función de verosimilitud.

Por ejemplo, cuando $\xi = 0$ y $\omega = 1$, la función de verosimilitud es monótona, creciente o decreciente, implicando con ello que el estimador de máxima verosimilitud de λ no existe (Azzalini, 1985). Específicamente la verosimilitud perfil de λ tiene un punto estacionario en $\lambda = 0$ independientemente de la muestra observada (Liseo y Loperfido, 2006), y cuando ξ, ω y λ son desconocidos el problema es más complicado porque el Hessiano se vuelve singular (Azzalini y Genton, 2008; Chiogna, 1998; Rusell y González, 2002). Azzalini y Capitanio (1999) proponen utilizar una reparametrización de la función de densidad para resolver este problema, la cual es denominada "parametrización centrada".

El problema consiste en reparametrizar de (ξ, ω, λ) a $(\mu, \sigma_e^2, \gamma_1)$. Si $U \sim SN_D(\lambda)$ y $\lambda =$

$\frac{\rho}{\sqrt{1-\rho^2}}$ con $\rho \in (-1,1)$ entonces $E(U) = E_U = \sqrt{\frac{2}{\pi}}\rho$ y $Var(U) = 1 - \frac{2}{\pi}\rho^2$, , se puede observar que $\rho = \dfrac{\lambda}{\sqrt{1+\lambda^2}}$, y en algunos trabajos (Azzalini, 1985; Liseo y Loperfido, 2006), esta parametrización es llamada la parametrización δ, entonces, $S_U = \sqrt{Var(U)} = \sqrt{1 - \frac{2}{\pi}\rho^2}$, por lo tanto la función de densidad en (2.8) queda expresada de la siguiente forma:

$$Y = \xi + \omega U = \mu + \sigma_e Z_0, \quad Z_0 = \left(\frac{U - E(U)}{\sqrt{Var(U)}}\right),$$

donde Z_0 es una variable aleatoria estandarizada con media cero y varianza uno, $\mu \in \mathbb{R}$ y $\sigma_e > 0$ denotan la media y la desviación estándar de Y respectivamente y se escribe $Y \sim SN_C(\mu, \sigma_e, \gamma_1)$, donde γ_1 es el coeficiente de asimetría y está dado por:

$$\gamma_1 = \sqrt{\frac{2}{\pi}}\rho^3 \left(\frac{4}{\pi} - 1\right)\left(1 - \frac{2\rho^2}{\pi}\right)^{-3/2},$$

y su rango de variación es $(-0.99527, 0.99527)$. La $E(Y) = \mu$ y la $V(Y) = \sigma_e^2$.

Si se considera el siguiente modelo de regresión:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i,$$

bajo el supuesto de que los errores son independientes e idénticamente distribuidos, $e_i \sim SN_C(0, \sigma_e^2, \gamma_1)$, y_i variable respuesta, x_{ij} covariables de interés. Con $E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$ y $Var(y_i) = \sigma_e^2$ por lo tanto:

$$y_i \sim SN_C(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \sigma_e^2, \gamma_1).$$

Ahora el trabajo se enfoca en obtener estimaciones de los parámetros $\beta_0, \sigma_e^2, \gamma_1, \boldsymbol{\beta}$ y σ_β^2. Para abordar este problema, haremos uso de Métodos Bayesianos y la técnica de truncamiento oculto que se describen en las siguientes secciones.

Enfoque Bayesiano para estimar parámetros

Liseo y Loperfido (2006), abordan el problema de estimación de los parámetros de la distribución normal asimétrica, bajo el esquema de la parametrización directa, utilizando la distribución a priori no informativa de Jeffreys cuando $\xi = 0, \omega = 1$ y λ es desconocido y en el caso cuando los tres parámetros son desconocidos. La distribución a priori de referencia para el caso de los tres parámetros, obtenida por estos autores está dada por:

$$p(\xi,\omega,\lambda) = p(\xi,\omega)p(\lambda) \sim \frac{1}{\omega} g^{1/2}(\lambda),$$

donde

$$\begin{aligned}
g(\lambda) &= \frac{i_{11}i_{22}i_{33} + 2i_{12}i_{13}i_{23} - i_{11}i_{23}^2 - i_{33}i_{12}^2 - i_{22}i_{13}^2}{i_{33}i_{22} - i_{23}^2} \\
\{i_{jk}\} &= \begin{pmatrix} a_2 & -\lambda a_2/\omega & \left(\frac{b}{(1+\lambda^2)^{3/2}} - \lambda a_1\right)/\omega \\ -\lambda a_2/\omega & (2+\lambda^2 a_2)/\omega^2 & \left(b\delta\frac{1+2\lambda^2}{1+\lambda^2} + \lambda^2 a_1\right)/\omega^2 \\ \left(\frac{b}{(1+\lambda^2)^{3/2}} - \lambda a_1\right) & b\delta\frac{1+2\lambda^2}{1+\lambda^2} + \lambda^2 a_1 & (1+\lambda^2 a_0)/\omega^2 \end{pmatrix} \\
b &= \sqrt{\frac{2}{\pi}} \\
\delta &= \frac{\lambda}{\sqrt{1+\lambda^2}} \\
a_j = a_j(\lambda) &= \int_{\mathbb{R}} 2z^j \frac{\phi(\lambda z)\phi(z)}{\Phi(z)} dz, \qquad j = 0,1,2.
\end{aligned}$$

Por lo tanto, trabajar con la distribución a priori de referencia propuesta, es muy complicado como se puede observar. En este estudio, se propone trabajar con la parametrización centrada para obtener estimadores de los parámetros de interés, bajo las siguientes distribuciones a priori:

$$\begin{aligned}
\beta_0|\sigma^2_{\beta_0} &\sim N(0,\sigma^2_{\beta_0}) \\
\boldsymbol{\beta}|\sigma^2_{\boldsymbol{\beta}} &\sim N(\mathbf{0},\sigma^2_{\boldsymbol{\beta}}\mathbf{I}) \\
p(\sigma^2_e|S_e,df_e) &= \chi^{-2}(S_e,df_e) \\
p(\sigma^2_{\boldsymbol{\beta}}|S_{\boldsymbol{\beta}},df_{\boldsymbol{\beta}}) &= \chi^{-2}(S_{\boldsymbol{\beta}},df_{\boldsymbol{\beta}}) \\
p(\rho|a_0,b_0) &\propto \left(\frac{1-\rho}{2}\right)^{a_0-1}\left(1-\frac{1-\rho}{2}\right)^{b_0-1}.
\end{aligned}$$

Siendo λ una función de ρ, es decir $\lambda = \frac{\rho}{\sqrt{1-\rho^2}}$. Por lo tanto la distribución a priori conjunta es:

$$
\begin{aligned}
p(\beta_0, \sigma_e^2, \rho, \boldsymbol{\beta}, \sigma_{\boldsymbol{\beta}}^2 | \Omega) \quad \propto \quad & \frac{1}{\sqrt{2\pi\sigma_{\beta_0}^2}} exp\left\{ -\frac{1}{2\sigma_{\beta_0}^2}\beta_0^2 \right\} \\
\times \quad & \left(\sigma_e^2\right)^{-(df_e/2+1)} exp\left\{ -\frac{S_e}{2\sigma_e^2} \right\} \\
\times \quad & \left(\frac{1-\rho}{2}\right)^{a_0-1} \left(1 - \frac{1-\rho}{2}\right)^{b_0-1} \\
\times \quad & \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi\sigma_{\boldsymbol{\beta}}^2}} exp\left\{ -\frac{1}{2\sigma_{\boldsymbol{\beta}}^2}\beta_j^2 \right\} \\
\times \quad & \left(\sigma_{\boldsymbol{\beta}}^2\right)^{-(df_\beta/2+1)} exp\left\{ -\frac{S_\beta}{2\sigma_{\boldsymbol{\beta}}^2} \right\},
\end{aligned}
$$

donde Ω es el conjunto de hyper-parámetros de las distribuciones *a priori*.

4.2.3. Truncamiento Oculto

Una muestra truncada es aquella donde se excluyen totalmente ciertos valores de la población de acuerdo a algún tipo de restricción (Cohen, 2016), o bien cuando las variables observadas representan una muestra que ha sido truncada por alguna variable oculta (Arnold y Beaver, 2000). En ambos casos, esto implica truncar la distribución de la variable aleatoria de interés. Usando la propuesta de Arnold y Beaver (2000) y Kim (2005), truncamos una normal bivariada de la siguiente manera:

Sean V y W dos variables aleatorias cuya distribución conjunta está dada como sigue:

$$
\begin{pmatrix} V \\ W \end{pmatrix} = N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right),
$$

con $\rho \in (-1, 1)$ y definimos la variable aleatoria U como sigue

$$
U = \begin{cases} W & \text{si } V \geq 0 \\ 0 & \text{de otro modo} \end{cases}
$$

Entonces $U \sim SN(\lambda)$, con $\lambda = \dfrac{\rho}{\sqrt{1-\rho^2}}$ (Arnold y Beaver, 2000; Liseo y Parisi, 2013). La representación descrita anteriormente, es utilizada para escribir una función de verosimilitud aumentada (Kim, 2005; Liseo y Parisi, 2013), al incorporar la variable latente $Z = V > 0$ *"como si"* ésta hubiera sido observada, y permite simular de (4.1).

Ya se han mencionado los diferentes inconvenientes que existen con respecto a la estimación de los parámetros, vía método de momentos y máxima verosimilitud. En el contexto Bayesiano, el uso de la técnica de Cadenas de Markov Monte Carlo (MCMC), utilizando la representación estocástica anterior de la distribución, ayuda a resolver el problema de estimación.

Haciendo uso de resultados de la distribución Normal Bivariada, se obtiene que la distribución condicional de $U|Z = z$ es $N(\rho z, 1-\rho^2)$ y $Z = V > 0 \sim NT(0, 1, 0, \infty)$, es decir Z es una variable aleatoria Normal Truncada, con parámetro de localidad igual a 0, parámetro de escala igual a 1, cota inferior de truncamiento igual a 0 y cota superior igual ∞. La distribución conjunta de U y Z es:

$$f_{U,Z}(u, z) = f_{U|Z}(u|z, \rho) f_Z(z).$$

Es decir:

$$\begin{aligned} f_{U,Z}(u, z|\rho) &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{1-\rho^2}} exp\left\{-\frac{1}{2(1-\rho^2)}(u-\rho z)^2\right\} \\ &\times \frac{1}{\sqrt{2\pi}} \left\{-\frac{1}{2}z^2\right\} I_{(0,\infty)}(z), \quad u \in \mathbb{R}. \end{aligned} \tag{4.2}$$

Observe que la distribución de U puede ser obtenida integrando $f(u, z|\rho)$ con respecto a z, esto es,

$$f_U(u|\rho) = \int_{-\infty}^{\infty} f(u, z|\rho) dz$$

4.2.4. Regresión con errores aleatorios normal asimétricos

El análisis de regresión es una técnica estadística para investigar y modelar la relación entre variables. Azzalini y Capitanio (1999); Rusell y González (2002) propusieron un modelo de regresión lineal donde los errores, e_i, son independientes e idénticamente distribuidos $SN_D(0, \omega, \lambda)$. El modelo propuesto es:

$$y_i = \mu + \beta_1 x_i + e_i.$$

Usando las propiedades de la distribución normal asimétrica, se sigue que $Y_i \sim SN_D(\mu + \beta_1 x_i, \omega, \lambda)$. El modelo se puede extender para incluir más covariables denotadas por x_{ij}, es decir:

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + e_i$$

Azzalini y Capitanio (1999) así como Rusell y González (2002) usaron el método de máxima verosimilitud a fin de obtener la estimación de los parámetros en el modelo, bajo el supuesto de que $n > p$. Ya se ha mencionado que en el contexto de SG, $n << p$ y que una forma de resolver el problema de estimación es a través de métodos de regresión penalizada o métodos Bayesianos.

4.2.5. Regresión Bayesiana con Errores Normal-Asimétricos

En el contexto Bayesiano, se supone una distribución a priori para los parámetros de regresión, sin embargo, la elección de ésta hace que las inferencias posteriores sean sensibles cuando se hacen cambios en la distribución a priori asignada.

Sea $y_i \sim SN_C(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j, \sigma_e^2, \gamma_1), i = 1, 2, \cdots, n$. Entonces la función de verosimilitud está dada por:

$$p(\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \sigma_e^2, \gamma_1) = \prod_{i=1}^{n} SN_C(y_i|\beta_0 + \mathbf{x}_i^t\boldsymbol{\beta}, \sigma_e^2, \gamma_1)$$

Denotando con $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}^t, \sigma_{\boldsymbol{\beta}}^2, \sigma_e^2, \gamma_1)^t$ el vector de parámetros y con $p(\boldsymbol{\theta}|\Omega)$ la distribución a priori de $\boldsymbol{\theta}$ con Ω el conjunto de hiperparámetros, entonces por el teorema de Bayes, la distribución posterior de $p(\boldsymbol{\theta}|datos)$ se obtiene como sigue:

$$\begin{aligned} p(\boldsymbol{\theta}|datos) &\propto p(\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \sigma_e^2, \gamma_1)p(\boldsymbol{\theta}|\Omega) \\ &= \prod_{i}^{n} SN_C(y_i; \beta_0 + \mathbf{x}_i^t\boldsymbol{\beta}, \sigma_e^2, \gamma_1)p(\boldsymbol{\theta}|\Omega) \end{aligned}$$

Como se puede observar, ni la distribución posterior ni la distribución condicional de los parámetros de interés tienen forma cerrada, por lo tanto, para realizar la estimación se propone la técnica de truncamiento oculto dentro de un Muestreador de Gibbs (Geman y Geman, 1984), Caminata Aleatoria y Muestreador Metropolis (Gilks *et al.*, 1995).

4.2.6. Distribuciones a *priori*, a *posteriori* y condicionales completas, propuestas para el modelo normal sesgado

En la sección anterior se mencionó que $f_{U,Z}(u,z|\rho) = f_{U|Z}(u,z)f_Z(z)$ donde, de resultados de la distribución Normal Bivariada $f_{U|Z}(u,z)f_Z(z) = N(u;\rho z, 1-\rho^2)NT(z;0,1,0,\infty)$, por tanto la distribución conjunta de U y Z es la dada en 4.2. Usando el esquema de la parametrización centrada, $Y = \mu+\sigma_e\left(\frac{U-E(U)}{\sqrt{Var(U)}}\right)$ con $\mu \in \mathbb{R}$ y $\sigma_e^2 \in \mathbb{R}^+$, entonces se puede mostrar que $Y \sim SN_C(\mu, \sigma_e^2, \gamma_1)$. La distribución conjunta de las variables aleatorias Y y Z puede ser obtenida de (4.2) usando Jacobianos (Casella y Berger, 2002) y está dada por:

$$\begin{aligned} f_{Y,Z}(y,z|\rho) &= \frac{2}{2\pi\sqrt{1-\rho^2}}\frac{S_u}{\sigma_e} exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{y-\mu}{\sigma_e}\right)S_u + E_u - \rho z\right]^2\right\} \\ &\times \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}z^2\right\}. \end{aligned} \tag{4.3}$$

Con $y \in \mathbb{R}, z \geq 0$ y $S_u = \sqrt{Var(U)} = \sqrt{1-\frac{2}{\pi}\rho}$ y $E_u = E(U) = \sqrt{\frac{2}{\pi}}\rho$.

Observe que la función de densidad de Y se obtiene de integrar $f_{Y,Z}(y,z;\mu,\sigma_e^2,\rho)$ con respecto a z. Es decir

$$f_Y(y|\mu,\sigma_e^2,\gamma_1) = \int_0^\infty f_{Y,Z}(y,z|\mu,\sigma_e^2,\rho)dz,$$

y la expresión asociada al resultado de esta integral es (Azevedo *et al.*, 2011):

$$f_Y(y|\mu,\sigma_e^2,\gamma_1) = \frac{1}{\sqrt{\sigma_e^{2*}}}\phi\left(\frac{y-\mu^*}{\sqrt{\sigma_e^{2*}}}\right)\Phi\left[\lambda^*\left(\frac{y-\mu^*}{\sqrt{\sigma_e^{2*}}}\right)\right],$$

donde $\mu^* = \mu - s\gamma_1^{1/3}$, $\sigma_e^{2*} = \sigma_e^2(1+s^2\gamma_1^{2/3})$, $\lambda^* = \frac{s\gamma_1^{1/3}}{\sqrt{r^2+s^2\gamma_1^{2/3}(r^2-1)}}$ con $r = \sqrt{\frac{2}{\pi}}$, $s = \left(\frac{2}{4-\pi}\right)^{1/3}$. Por lo tanto, $Y \sim SN_C(\mu^*, \sigma_e^{2*}, \gamma_1^*)$.

En el contexto de regresión μ, representa el valor esperado de la variable respuesta, esto es, $\mu_i = \beta_0 + \mathbf{x}_i^t\boldsymbol{\beta}$, por lo tanto la función de verosimilitud aumentada es:

$$L(\boldsymbol{\theta}, \mathbf{z}|\boldsymbol{y}, \boldsymbol{X}) = \prod_{i=1}^n f(y_i, z_i|\mu_i, \sigma_e^2, \rho) \tag{4.4}$$

donde $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}^t, \sigma_e^2, \sigma_{\boldsymbol{\beta}}^2, \rho)^t$, $\mu = \beta_0 + \mathbf{x}^t\boldsymbol{\beta}$. Combinando (4.3) y (4.4), se puede escribir:

$$L(\boldsymbol{\theta}, \mathbf{z}|\boldsymbol{y}, \boldsymbol{X}) = \prod_{i=1}^{n} \frac{2}{2\pi\sqrt{1-\rho^2}} \frac{S_u}{\sigma_e} exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{y_i - \mu_i}{\sigma_e} \right) S_u + E_u - \rho z_i \right]^2 \right\} \times \frac{1}{\sqrt{2\pi}} exp\left\{ -\frac{1}{2} z_i^2 \right\} I_{(0,\infty)}(z_i) \tag{4.5}$$

A fin de tener completamente especificado el modelo, las distribuciones a *priori* propuestas son: en el primer nivel de la jerarquía, para β_0 asignamos una a priori no informativa, $\beta_0|\sigma_{\beta_0}^2 \sim N(0, \sigma_{\beta_0}^2)$; para $\boldsymbol{\beta}$, proponemos una distribución Normal $\boldsymbol{\beta}|\sigma_\beta^2 \sim N(\mathbf{0}, \sigma_\beta^2\mathbf{I})$; en el segundo nivel, para σ_β^2, se propone una distribución Chi-cuadrada invertida-escalada, $\sigma_\beta^2 \sim \chi^{-2}(df_\beta, S_\beta)$; para ρ proponemos usar una distribución a *priori* tipo Beta, esto es, $R = 1 - 2B$, donde $B \sim Beta(a_0, b_0)$ y la denotaremos como $p(\rho|a_0, b_0)$; y al parámetro de escala una distribución Chi-cuadrada invertida-escalada, esto es, $\sigma_e^2|df_e, S_e \sim \chi^{-2}(df_e, S_e)$. Entonces, la distribución de los parámetros de interés se puede escribir como:

$$p(\boldsymbol{\theta}|\Omega) = p(\beta_0|\sigma_{\beta_0}^2)p(\boldsymbol{\beta}|\sigma_\beta^2\mathbf{I})p(\sigma_\beta^2|df_\beta, S_\beta)p(\sigma_e^2|df_e, S_e)p(\rho|a_0, b_0)$$

Combinando (4.5) y aplicando el teorema de Bayes, la distribución posterior queda expresada como:

$$\begin{aligned} p(\boldsymbol{\theta}|resto) \propto & \prod_{i=1}^{n} \left[\frac{1}{2\pi\sqrt{1-\rho^2}} \frac{S_u}{\sigma_e} exp\left\{ -\frac{1}{2(1-\rho^2)} S_u^2 \left(y_i - \mu_i - \frac{\sigma_e}{S_u}\rho z_i + \frac{\sigma_e}{S_u} E_u \right)^2 \right\} \right. \\ & \left. \times \frac{1}{\sqrt{2\pi}} exp\left\{ -\frac{1}{2} z_i^2 \right\} I_{(0,\infty)}(z_i) \right] \\ & \times \frac{1}{\sqrt{2\pi\sigma_{\beta_0}^2}} exp\left\{ -\frac{1}{\sigma_{\beta_0}^2} \beta_0^2 \right\} N(\boldsymbol{\beta}|\mathbf{0}, \sigma_\beta^2\mathbf{I}) \chi^{-2}(\sigma_e^2|df_e, S_e) \\ & \times \chi^{-2}(\sigma_\beta^2|df_\beta, S_\beta) p(\rho|a_0, b_0) \end{aligned} \tag{4.6}$$

Las distribuciones condicionales completas, que se utilizaron para la implementación del Muestreador Gibbs se obtuvieron de (4.6) y un resumen de ellas se encuentra en el anexo B. Como se puede observar de (4.6) para algunos parámetros $(\beta_0, \boldsymbol{\beta}, \sigma_\beta^2)$, es posible reconocer el núcleo de su distribución condicional, por ejemplo:

$$\beta_0|resto \sim N\left(\beta_0; \frac{\sum_{i=1} y_i^*}{n+\frac{(1-\rho^2)\sigma_e^2}{S_u^2\sigma_{\beta_0}^2}}, \frac{(1-\rho^2)\sigma_e^2}{S_u^2\left(n+\frac{(1-\rho^2)\sigma_e^2}{S_u^2\sigma_{\beta_0}^2}\right)}\right),$$

$$\beta_j|resto \sim N\left(\frac{\sum_{i=1}^n x_{ij}y_i^*}{\sum_{i=1}^n x_{ij}^2+\frac{(1-\rho^2)\sigma_e^2}{S_u^2\sigma_\beta^2}}, \frac{(1-\rho^2)\sigma_e^2}{S_u^2\left[\sum_{i=1}^n x_{ij}^2+\frac{(1-\rho^2)\sigma_e^2}{S_u^2\sigma_\beta^2}\right]}\right),$$

$$\sigma_\beta^2 \sim \chi^{-2}(df_\beta + p, \boldsymbol{\beta}^t\boldsymbol{\beta} + S_\beta).$$

Para construir la verosimilitud aumentada, se utilizó una variable latente, cuya función de distribución condicional está dada por:

$$p(z_i|resto) \propto exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{y_i-\beta_0-\mathbf{x}_i^t\boldsymbol{\beta}}{\sigma_e}\right)S_u + E_u - \rho z_i\right]^2\right\} exp\left\{-\frac{1}{2}z_i^2\right\},$$

donde $y_i^* = \left(\frac{y_i-\beta_0-\mathbf{x}_i^t\boldsymbol{\beta}}{\sigma_e}\right)S_u + E_u$. Entonces

$$p(z_i|resto) \quad \propto \quad exp\left\{-\frac{1}{2(1-\rho^2)}\left(z_i-\rho y_i^*\right)^2\right\} I_{(0,\infty)}(z_i), i = 1, \ldots, n.$$

Se reconoce el núcleo de una distribución Normal-Truncada con parámetro de localidad ρy_i^* y parámetro de escala $1-\rho^2$, lo cual se expresa de la siguiente manera:

$$z_i \sim NT(\mu = \rho y_i^*, \sigma_e^2 = 1-\rho^2, a = 0, b = \infty).$$

En el caso de (σ_e^2, ρ) no fue posible identificar el núcleo de alguna distribución univariada conocida, pues su distribución condicional no tiene forma cerrada,

$$\begin{aligned} p(\sigma_e^2|resto) \quad &\propto \quad (\sigma_e^2)^{\left(-\frac{df_e+n}{2}+1\right)} exp\left\{-\frac{1}{2(1-\rho^2)\sigma_e^2}S_u^2\left[\sum_{i=1}^n\left(y_i-\beta_0-\mathbf{x}_i^t\boldsymbol{\beta}-\frac{\sigma_e}{S_u}\rho z_i+\frac{\sigma_e}{S_u}E_u\right)^2\right]\right\} \\ &\times \quad exp\left\{\frac{(1-\rho^2)}{S_u^2}S_e\right\}, \end{aligned}$$

$$p(\rho|resto) \propto \left(\frac{1-\rho}{2}\right)^{a_0-1}\left(1-\frac{1-\rho}{2}\right)^{b_0-1}\left(\frac{S_u}{\sqrt{1-\rho^2}}\right)^n I_{(-1,1)}(\rho).$$

Se hizo uso de varias técnicas computacionales para realizar las simulaciones, Muestreador de Gibbs y en el caso de σ_e^2 se utilizó el algoritmo Metropolis con Caminata Aleatoria para lograr obtener muestras de σ_e^2 (Kim, 2005), para obtener muestras de ρ se usó la transformación de Fisher (1915) y el algoritmo Metropolis con Caminata Aleatoria, detalles de las transformaciones y del algoritmo implementado se encuentran en el anexo B. En el mismo anexo, se proporcionan las reglas para fijar los

hiper-parámetros, $\sigma^2_{\beta_0}, S_e, df_e, S_\beta, df_\beta, a_0, b_0$. Las reglas para $S_e, df_e, S_\beta, df_\beta$ se basan en las reportadas por de los Campos *et al.* (2013) y de los Campos y Pérez-Rodríguez (2015).

4.3. Simulación

En esta sección se presentan los resultados de simulación de los estimadores obtenidos de los parámetros de interés.

Aplicación

Se simularon datos usando marcadores genotípicos de un conjunto de datos de trigo el cual se hizo disponible públicamente por Crossa *et al.* (2010). El conjunto de datos incluye información genotípica de 599 líneas de trigo cuyo rendimiento fue evaluado en cuatro ambientes. Las líneas fueron genotipadas para 1279 marcadores DArT codificados como 0 y 1. Se simularon los fenotipos (variable respuesta) utilizando el siguiente modelo genético aditivo:

$$y_i = \beta_0 + \sum_{j=1}^{1279} x_{ij}\beta_j + e_i \;\; i = 1, \ldots, n, \tag{4.7}$$

donde

$$e_i \sim SN_C(0, 1.5^2, \gamma_1),$$

con

$$\gamma_1 = \sqrt{\frac{2}{\pi}}\rho^3 \left(\frac{4}{\pi} - 1\right)\left(1 - \frac{2}{\pi}\rho^2\right)^{-3/2},$$

y $\rho \in \{0, 0.5, 0.75, 0.90, 0.95, 0.99\}$ lo que permitió obtener diferentes valores para el parámetro asociado al sesgo. β_0 fue fijado en 3 y el efecto de 10 marcadores se muestrearon de una distribución normal con media 0 y varianza $\sigma^2_\beta = \frac{0.5}{10}$ (Pérez-Rodríguez *et al.*, 2015), el resto de lo marcadores se fijo en 0, es decir:

$$\beta_j = \begin{cases} N(0, \frac{0.5}{10}) & j \in \{156, 207, 327, 472, 771, 879, 965, 976, 1050, 1252\} \\ 0 & \text{de otra forma.} \end{cases}$$

El objetivo es verificar, utilizando la simulación, si la metodología propuesta funciona en

forma satisfactoria. A través de la simulación se obtienen estimadores puntuales para β_0, $\boldsymbol{\beta}$, σ_e^2 y ρ.

Con la misma finalidad, se ajustó el modelo de regresión Ridge Bayesiana y se compararon las estimaciones de los coeficientes de regresión, las predicciones y las estimaciones de los valores genéticos en ambos modelos. La correlación de Pearson entre los valores observados ($\boldsymbol{y}$) y los predichos ($\beta_0\mathbf{1} + \mathbf{X}\hat{\boldsymbol{\beta}}$) es una medida de bondad de ajuste que cuantifica qué tan bien el modelo estima los valores "verdaderos", la correlación de Pearson entre los valores genéticos "verdaderos"($\mathbf{X}\boldsymbol{\beta}$) y los predichos ($\mathbf{X}\hat{\boldsymbol{\beta}}$) es una medida que cuantifica la estimación de los valores genéticos y por último la correlación de Pearson entre los efectos de los marcadores "verdaderos"($\boldsymbol{\beta}$) y los estimados ($\hat{\boldsymbol{\beta}}$) evalúa qué tan bueno es el modelo para estimar el efecto de los marcadores (de los Campos *et al.*, 2009b).

El algoritmo que se utilizó para realizar la simulación se describe a continuación:

1. Fijar β_0, $\boldsymbol{\beta}$, σ_e^2 y ρ.
2. Simular los fenotipos usando la ecuación (4.7).
3. Ajustar el modelo de regresión con errores aleatorios normal asimétricos y obtener estimadores puntuales para β_0, $\boldsymbol{\beta}$, σ_e^2 y ρ, esto es $\hat{\beta}_{0SN}$, $\hat{\boldsymbol{\beta}}_{SN}$, $\hat{\sigma}^2_{eSN}$ y $\hat{\rho}_{SN}$.
4. Ajustar el modelo de regresión Ridge Bayesiana y obtener estimadores puntuales para β_0, $\boldsymbol{\beta}$ y σ_e^2, es decir $\hat{\beta}_{0RR}$, $\hat{\boldsymbol{\beta}}_{RR}$, $\hat{\sigma}^2_{eRR}$.
5. Calcular la correlación entre los fenotipos observados y predichos, los valores genéticos "verdaderos" y predichos y los coeficientes de regresión "verdaderos" y estimados con ambos modelos de regresión.
6. Repertir los pasos 1 a 5, 100 veces y obtener el promedio de las correlaciones, intercepto β_0, σ_e^2 y ρ.

Para obtener los estimadores de los parámetros, los modelos se ajustaron usando los datos completos y las inferencias se hicieron con base en 10,000 muestras (obtenidas después de descartar 5000 muestras como calentamiento) de la densidad posterior y 100 réplicas Monte Carlo para cada uno de los seis valores de ρ. Los valores estimados se obtuvieron con respecto a la media de las 100 réplicas y también se obtuvo la desviación estándar de los estimadores entre las réplicas.

En la Tabla 4.1 se reportan los resultados obtenidos, por cada valor de ρ, así como la media posterior (obtenida de las 100 réplicas Monte Carlo) del parámetro de localidad, β_0 y de la varianza residual σ_e^2.

Los resultados en la Tabla 4.1 muestran que los estimadores puntuales de los parámetros de localidad ($\hat{\beta}_0$) y de escala ($\hat{\sigma}_e^2$) son estables, en promedio, a cambios en el valor de ρ.

Del mismo modo, se puede observar que la varianza del efecto de los marcadores ($\boldsymbol{\beta}$) es igual o menor en el modelo propuesto SNB (Skew-Normal Bayesiana) que con el modelo BRR (Regresión Ridge Bayesiana), esto último también se puede observar en el cálculo del cociente de varianzas $\frac{\sigma_e^2}{\sigma_{\boldsymbol{\beta}}^2}$. La correlación de Pearson entre valores observados (y) y valores ajustados ($\hat{y}$), en promedio de las 100 muestras Monte Carlo, fue casi igual en ambos modelos. En esta tabla se presenta una medida de bondad de ajuste estándar, el CME (Cuadrado Medio del Error), con la finalidad de comparar el ajuste de ambos modelos, se observa que en promedio, el modelo propuesto realiza estimaciones semejantes que el modelo de Regresión Ridge Bayesiana, en otras palabras, las estimaciones del efecto de los marcadores son similares en ambos modelos.

Tabla 4.1: Estimación de los parámetros cuando se ajustaron los modelos, Regresión con errores asimétricos Bayesiana (SNB) y Regresión Ridge Bayesiana (RRB), de localidad β_0, de escala σ_e^2, de la varianza del efecto de los marcadores $\sigma_{\boldsymbol{\beta}}^2$, del cociente de varianzas λ_1, de la correlación entre valores verdaderos y ajustados $Corr(y, \hat{y})$, y del Cuadrado Medio del Error. Entre paréntesis se encuentra el valor de la desviación estándar entre réplicas.

$\rho = 0, \ \ \hat{\rho} = 0.016(0.207)$						
Modelo	$\hat{\beta}_0$	$\hat{\sigma}_e^2$	$\hat{\sigma}_{\boldsymbol{\beta}}^2$	λ_1	$Corr(y, \hat{y})$	CME
SNB	3.075(0.854)	2.257(0.052)	0.003(0.001)	833.72	0.479	2.441
RRB	3.113(0.975)	2.207(0.155)	0.003(0.001)	627.33	0.531	3.036
$\rho = 0.5, \ \ \hat{\rho} = 0.075(0.270)$						
SNB	3.009(0.771)	2.218(0.047)	0.003(0.001)	803.35	0.648	3.274
RRB	2.991(0.905)	2.167(0.133)	0.003(0.001)	602.89	0.667	2.714
$\rho = 0.75, \ \ \hat{\rho} = 0.329(0.261)$						
SNB	2.972(0.714)	2.210(0.048)	0.003(0.001)	816.71	0.442	2.219
RRB	2.945(0.828)	2.168(0.139)	0.003(0.001)	614.19	0.506	2.154
$\rho = 0.90, \ \ \hat{\rho} = 0.841(0.115)$						
SNB	3.094(0.821)	2.219(0.054)	0.003(0.001)	833.48	0.648	3.112
RRB	3.120(0.858)	2.175(0.155)	0.003(0.001)	621.59	0.676	2.639
$\rho = 0.95, \ \ \hat{\rho} = 0.943(0.023)$						
SNB	3.055(0.942)	2.27(0.061)	0.003(0.001)	872.98	0.662	1.676
RRB	3.067(0.900)	2.196(0.169)	0.003(0.001)	631.79	0.696	1.642
$\rho = 0.99, \ \ \hat{\rho} = 0.987(0.008)$						
SNB	2.83(1.28)	2.167(0.056)	0.003(0.001)	668.16	0.578	2.593
RRB	2.89(0.878)	2.165(0.153))	0.004(0.001)	606.84	0.631	2.893

$\lambda_1 = \frac{\hat{\sigma_e}^2}{\hat{\sigma_\beta}^2}$.

En la Tabla 4.2 se presenta el cálculo de dos estadísticos que permiten evaluar la bondad del ajuste de los modelos. El pD, número efectivo de parámetros que se define como el promedio de la devianza menos la devianza evaluada en el promedio de los parámetros, es decir:

$$pD = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}),$$

siendo $\boldsymbol{\theta}$ el vector de parámetros y el DIC (Deviance Information Criterion) de Spiegelhalter *et al.* (2002), el cual es una versión generalizada de los criterios AIC y BIC. El DIC se define como:

$$DIC = \overline{D(\boldsymbol{\theta})} + pD.$$

Por otra parte, el DIC mide la complejidad del modelo en términos del número efectivo de parámetros en él. El DIC consiste en incorporar información a priori de los parámetros a través de $-2\log\left[p(\boldsymbol{y}|\boldsymbol{\theta}_k)\right]$, la cual tiene en cuenta la función completa de versomilitud. La medida de bondad de ajuste es la devianza promedio:

$$\overline{D(\boldsymbol{\theta})} = -2\int \log p(\boldsymbol{y}|\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

De igual modo se obtiene el cálculo del coeficiente de correlación de Pearson, entre el "verdadero"valor del efecto de los marcadores y su valor estimado, $Corr(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}})$, así como también la correlación de Pearson entre el "verdadero"valor del efecto genético y su estimación, $Corr(\mathbf{X}\boldsymbol{\beta}, \mathbf{X}\hat{\boldsymbol{\beta}})$.

Los resultados demuestran que en general los valores del pD y el DIC, son menores en el modelo propuesto, SNB, que cuando se ajustó el modelo BRR. En el mismo sentido se observan ligeras ganancias en cuanto a los coeficientes de correlación de Pearson, para valores de $\rho = 0, 0.5, 0.75$, sin embargo cuando el valor de ρ se incrementa ($\rho = 0.90, 0.95, 0.99$), el valor de la correlación aumenta significativamente a favor del modelo propuesto.

Tabla 4.2: Estimación del pD y el DIC, de la correlación entre el efecto de los marcadores verdaderos y estimados $(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}})$ y de la correlación entre valores verdaderos y ajustados $Corr(\mathbf{X}\boldsymbol{\beta}, \mathbf{X}\hat{\boldsymbol{\beta}})$ para comparar el ajuste de los modelos. Entre paréntesis se encuentra el valor de la desviación estándar entre réplicas.

$\rho = 0, \ \hat{\rho} = 0.016(0.207)$				
Modelo	pD	DIC	$Corr(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}})$	$Corr(\mathbf{X}\beta, \mathbf{X}\hat{\beta})$
SNB	66.767(16.632)	2,252.66(41.181)	0.192(0.046)	0.697(0.116)
RRB	80.867(14.346)	2,253.035(41.216)	0.193(0.049)	0.689(0.115)
$\rho = 0.5, \ \hat{\rho} = 0.074(0.270)$				
SNB	68.176(16.385)	2,243.132(40.814)	0.207(0.049)	0.718(0.119)
RRB	82.580(14.619)	2,243.742(39.972)	0.207(0.050)	0.714(0.117)
$\rho = 0.75, \ \hat{\rho} = 0.330(0.261)$				
SNB	67.630(16.632)	2,240.843(39.755)	0.194(0.051)	0.717(0.104)
RRB	81.971(13.803)	2,243.381(38.985)	0.195(0.052)	0.708(0.104)
$\rho = 0.90, \ \hat{\rho} = 0.841(0.115)$				
SNB	69.448(15.506)	2,225.026(41.124)	0.203(0.049)	0.718(0.114)
RRB	81.479(13.814)	2,244.544(40.879)	0.198(0.052)	0.706(0.115)
$\rho = 0.95, \ \hat{\rho} = 0.943(0.023)$				
SNB	71.151(17.001)	2,201.591(46.887)	0.203(0.046)	0.734(0.109)
RRB	80.443(14.292)	2,249.003(45.466)	0.191(0.047)	0.707(0.116)
$\rho = 0.99, \ \hat{\rho} = 0.987(0.008)$				
SNB	80.682(14.495)	2,118.477(52.505)	0.216(0.055)	0.747(0.098)
RRB	82.584(13.475)	2,242.839(42.418)	0.196(0.052)	0.703(0.109)

4.4. Aplicación con datos reales

Una aplicación del modelo propuesto consistió en ajustar el modelo de regresión normal asimétrico a un conjunto de datos de maíz, el cual se describe a continuación.

Conjunto de Datos de Maíz

Los datos provienen del proyecto de Maíz de Tolerancia a la Sequía (DTMA, por sus siglas en inglés) del Programa Global de Maíz de CIMMyT (http://www.cimmyt.org). Para ilustrar la aplicación de la metodología propuesta, se consideró un conjunto de datos que provienen de un estudio a gran escala que tiene como objetivo detectar las regiones cromosómicas que afectan la tolerancia a la sequía. Los datos genotípicos consisten en información de 300 líneas endogámicas tropicales que fueron genotipadas utilizando 1,152 SNPs (Single Nucleotide Polymorphisms).

El rasgo analizado es la mancha gris de plomo (GLS) causada por el hongo *Cercospora zeae-maydis* que se evaluó en Kakamega, Kenia; San Pedro Lagunillas, México y Pereira, Colombia. El rasgo se midió usando una escala ordinal de 1 a 5 con incrementos de 0.5, donde 1 representa ninguna infección y 5 para la infección completa.

Un subconjunto de estos datos fue analizado por Crossa *et al.* (2011), quienes transformaron los datos originales usando la transformación de Box-Cox, (Box y Cox, 1964) y luego los estudiaron. La Figura 4.1 muestra gráficos de densidad para la clasificación GLS en las tres localidades, el núcleo de la densidad se estimó utilizando un núcleo Gaussiano y el ancho de banda para el núcleo se estimó de acuerdo a Venables y Ripley (2002). La Figura 4.1 también muestra el índice de asimetría, $\gamma_1 = \frac{m_3}{s^3}$, donde $m_3 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^3$, $\bar{y}$ es la media muestral y s es la desviación estándar muestral (ver Joanes y Gill, 1998). En los tres casos, la distribución es asimétrica a la derecha, por lo que la masa de la distribución se concentra alrededor de valores pequeños de la variable respuesta.

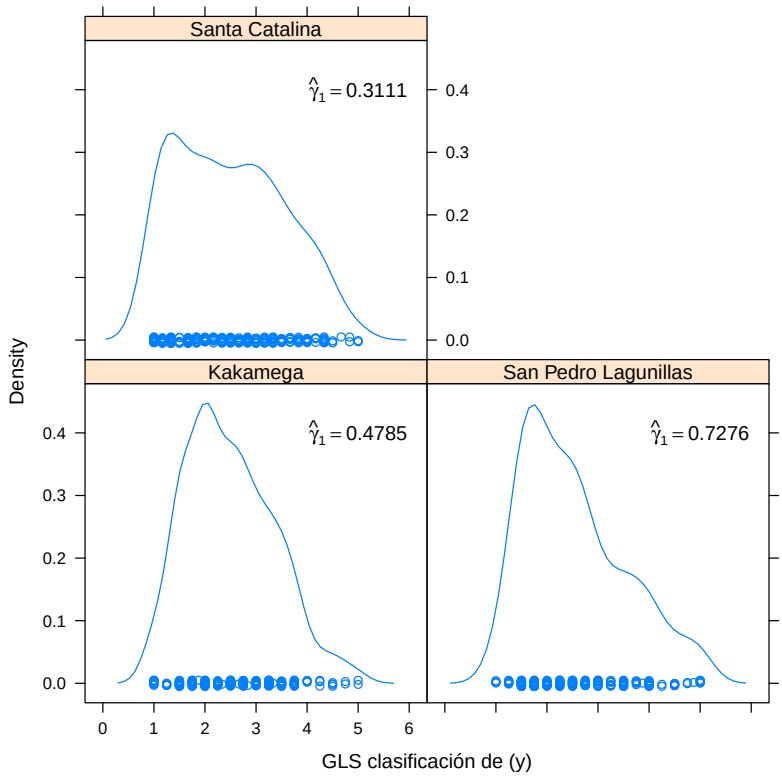

Figura 4.1: Gráficos de la densidad para la clasificación de la mancha gris de plomo (GLS, por sus siglas en inglés) para Santa Catalina (Colombia), Kakamega (Kenia) y San Pedro Lagunillas (México).

Se ajustaron dos modelos: 1) La regresión Ridge Bayesiana, donde los errores son *IID* independientes e idénticamente distribuidos, $e_i \sim N(0, \sigma_e^2), i = 1, \ldots, n$ y 2) El modelo propuesto, un modelo de regresión con errores aleatorios *IID* Normal Asimétricos, $e_i \sim SN_C(0, \sigma_e^2, \lambda_1), i = 1, \ldots, n$. La regresión Ridge Bayesiana se ajustó usando el paquete BGLR (de los Campos y Pérez-Rodríguez, 2015), mientras que el modelo propuesto se ajustó usando el algoritmo descrito en el Anexo B. Ambos modelos se ajustaron usando los datos completos y posteriormente se generaron 100 particiones al azar con 80 % de las observaciones en el conjunto de entrenamiento y 20 % en el conjunto de prueba. Para cada partición al azar, se ajustaron los dos modelos y posteriormente se obtuvieron predicciones para los fenotipos en el conjunto de prueba.

La capacidad predictiva de cada modelo se evaluó utilizando la correlación de Pearson entre los valores observados y los predichos. Los estimadores de los parámetros de interés, resultaron de 100,000 muestras MCMC y después de haber descartado 50,000 que se utilizaron como calentamiento. La convergencia se verificó inspeccionando los gráficos de las trazas.

La Tabla 4.3 reporta la media posterior de $\sigma_e^2, \sigma_\beta^2$ y ρ, así como los valores obtenidos para los dos criterios de información el pD y el DIC.

Tabla 4.3: Estimación de la media posterior de los parámetros $\sigma_e^2, \sigma_\beta^2$ y ρ, cuando se ajustaron los modelos, Regresión con errores asimétricos (SNB) y Regresión Ridge (RRB), utilizando los datos completos. Entre paréntesis se encuentra el valor de la desviación estándar entre réplicas.

Parámetro							
Localidad	Modelo	$\hat{\sigma}_e^2$	$\hat{\sigma}_{\boldsymbol{\beta}}^2$	$\lambda_1 = \frac{\hat{\sigma}_e^2}{\hat{\sigma}_{\boldsymbol{\beta}}^2}$	pD	DIC	$\hat{\rho}$
Kakamega	SNB	0.505(0.045)	0.0003(0.0001)	1569.389	62.572	594.65	0.978(0.033)
	RRB	0.423(0.071)	0.0005(0.0001)	798.127	98.28	628.91	
San Pedro Lagunillas	SNB	0.527(0.062)	0.0004(0.0001)	1161.611	54.09	562.566	0.938(0.198)
	RRB	0.404(0.071)	0.0007(0.0001)	533.364	112.408	595.180	
Santa Catalina	SNB	0.351(0.064)	0.0009(0.0002)	359.945	132.383	599.844	0.307(0.554)
	RRB	0.330(0.069)	0.001(0.0002)	317.425	144.066	597.680	

Como se puede observar las varianzas correspondientes al efecto de los marcadores, ($\boldsymbol{\beta}$), son menores para el modelo propuesto, SNB, que con las obtenidas cuando se ajustó el BRR. En el mismo sentido, los valores obtenidos para los dos criterios de información, pD y DIC son menores para el modelo propuesto, SNB, que para el modelo de regresión Ridge Bayesiana, BRR.

Hay que recordar que estos indicadores cuantifican la complejidad y la capacidad predictiva de un modelo. Cuanto más complejo es un modelo, la capacidad predictiva se reduce. Un valor menor del DIC de un modelo comparado con otro, se puede considerar como una buena medida para encontrar un modelo parsimonioso y que ya está penalizado por su grado de complejidad.

4.5. Discusión de resultados

A partir de la Tabla 4.3 se puede observar que la estimación del efecto de los marcadores es más precisa para el modelo propuesto (SNB) que para el modelo (RRB). La estimación

del parámetro ρ apoya la suposición de errores normal-asimétricos porque el estimador puntual no está alrededor de 0, excepto quizás para la localidad de San Pedro Lagunillas.

La Figura 4.2 muestra diagramas de dispersión para la estimación de GLS usando ambos modelos, SNB y RRB. Como se esperaba, la correlación de Pearson entre las predicciones es muy alta, mayor a 0.95. Esto implica que si los datos son asimétricos y se ajusta un modelo de RRB para obtener candidatos para la selección, se espera obtener aproximadamente a los mismos individuos.

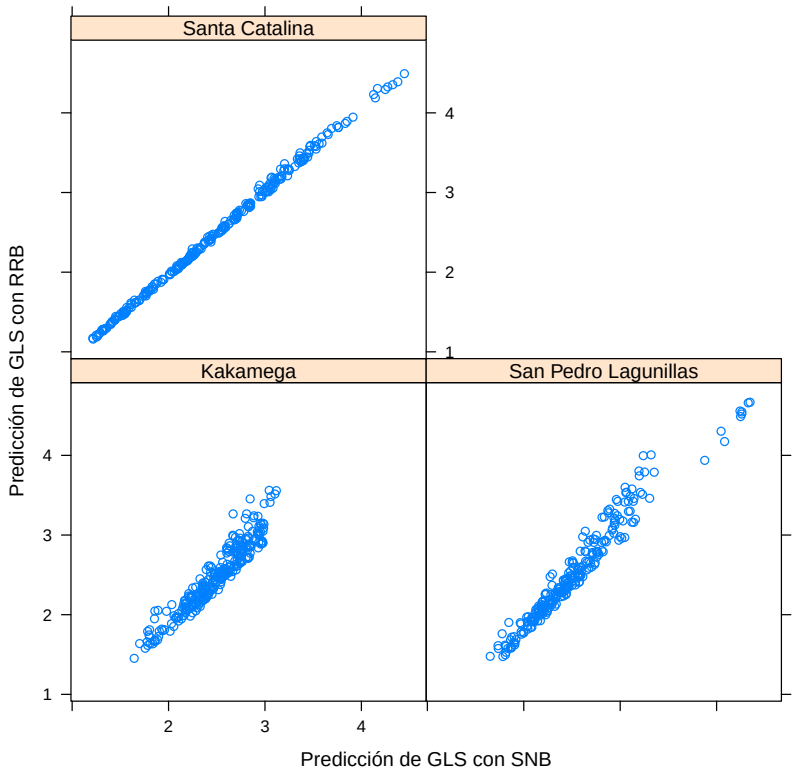

Figura 4.2: Gráficos de dispersión del rasgo GLS estimado, de los modelos SNB y RRB. En los tres casos considerados, la correlación de Pearson fue mayor a 0.95.

Validación Cruzada

En la Figura 4.3 se muestran diagramas de dispersión para la correlación de Pearson entre los valores observados y predichos con los individuos del conjunto de prueba. La correlación se obtuvo después de ajustar los dos modelos SNB y RRB para las tres localidades.

Cuando las correlaciones fueron mayores en favor de SNB, éstas se representaron con un círculo lleno y un círculo vacío en caso contrario. En la misma Figura se muestra el número de veces que la correlación de Pearson fue mayor con respecto al modelo de regresión Normal-Asimétrico versus el modelo de regresión Ridge Bayesiana. A través de este gráfico se aprecia que el modelo SNB realiza mejores predicciones que el modelo RRB.

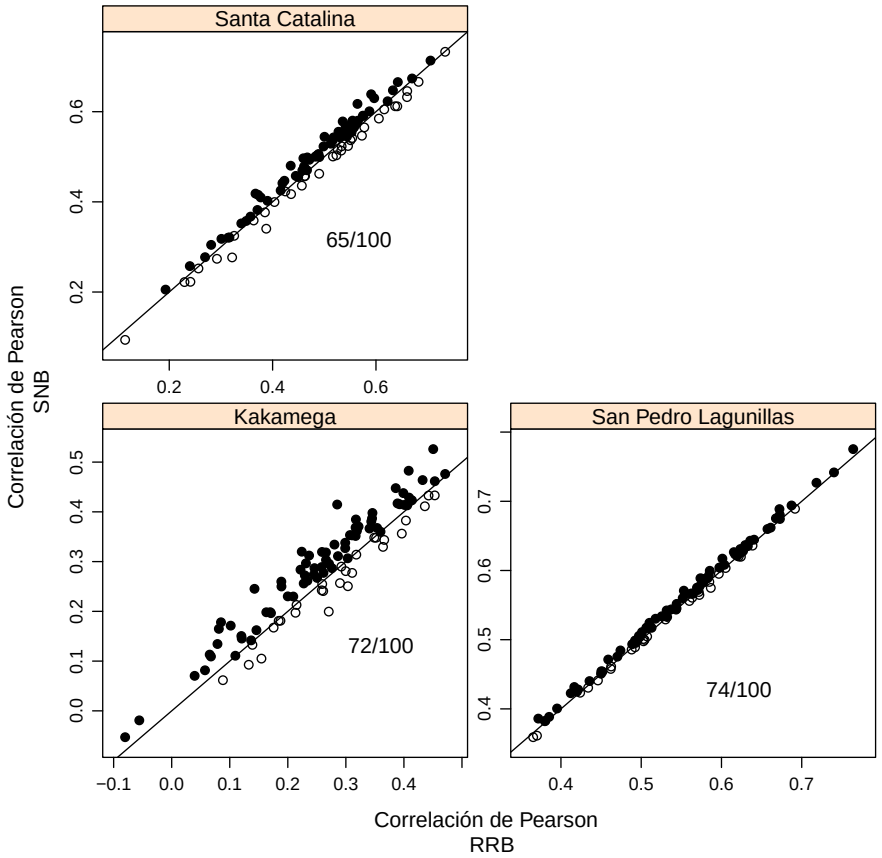

Figura 4.3: Gráficos de la correlación predictiva para cada una de las 100 validaciones cruzadas en las tres localidades. El círculo sólido representa cuando el modelo SNB fue mejor y cuando el modelo RRB fue mejor, está representado por el círculo vacío. El número de veces que la correlación de Pearson en el modelo SNB fue mejor que la correlación de Pearson en el modelo RRB, se muestra en cada gráfico.

La Figura 4.4 muestra diagramas de dispersión para los cuadrados medios del error (CME) obtenidos en el conjunto de prueba en las tres localidades. Cuando el CME fue menor para SNB que el CME para RRB se representa por un círculo vacío y en caso contrario por un círculo sólido. De igual forma, se reporta el número de veces que el CME para RRB es mayor que el CME para SNB. Se puede observar en la Figura 4.4 que en general el CME para el modelo RRB es mayor que el CME para el modelo SNB.

La Tabla 4.4 reporta las correlaciones promedio y el cuadrado medio del error entre valores observados y predichos en el conjunto de prueba cuando ambos modelos se ajustaron. El promedio y la desviación estándar son muy similares para ambos modelos y las diferencias entre los modelos no son significativas, pero las figuras sugieren que el modelo SNB predice mejor que el modelo RRB.

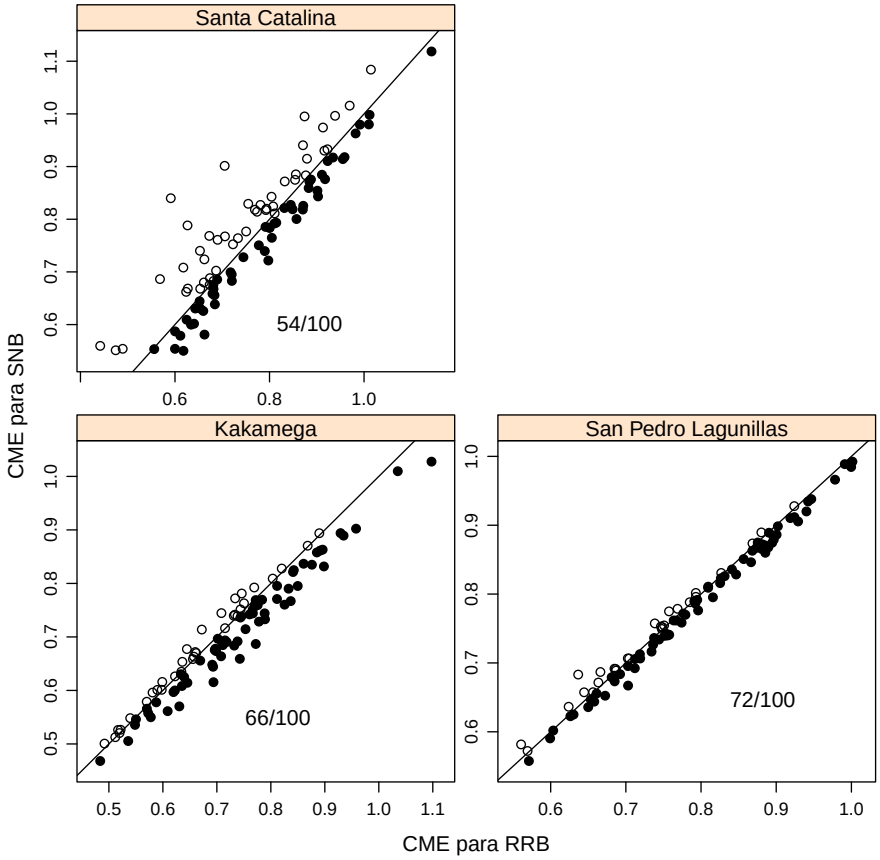

Figura 4.4: Gráficos del Cuadrado Medio del Error en el conjunto de prueba para cada una de las 100 validaciones cruzadas en las tres localidades. Cuando el CME en el modelo SNB fue menor que el CME en el modelo RRB este está representado por un círculo vacío y cuando el CME en el modelo RRB es mayor que el CME en el modelo SNB este se representa por un círculo sólido. El número de veces que el CME en el modelo RRB es mayor que el CME en el modelo SNB se encuentra reportado en cada gráfica.

Tabla 4.4: Correlaciones promedio y error cuadrado medio (CME) entre valores observados y estimados en el conjunto de prueba. Las predicciones fueron obtenidas después de ajustar los modelos SNB y RRB. El promedio es entre las 100 particiones aleatorias con 80 % de las observaciones en el conjunto de entrenamiento y 20 % en el conjunto de prueba. Las desviaciones estándar están dadas entre paréntesis.

Parámetro			
Localidad	Modelo	Coeficiente de Correlación de Pearson	CME
Kakamega	SNB	0.284(0.116)	0.702(0.113)
	RRB	0.261(0.089)	0.719(0.121)
San Pedro Lagunillas	SNB	0.548(0.089)	0.775(0.103)
	RRB	0.545(0.088)	0.780(0.106)
Santa Catalina	SNB	0.487(0.124)	0.780(0.106)
	RRB	0.480(0.122)	0.768(0.134)

4.6. Resumen

En este capítulo se propuso un modelo de regresión Bayesiana para predicción cuando la variable respuesta es asimétrica, con aplicaciones en Selección Genómica y en el caso donde el número de marcadores excede por mucho al número de individuos ($p >> n$), sin embargo esta condición no resta su aplicación en el caso de ($p < n$).

La representación estocástica de la variable aleatoria normal asimétrica utilizada en el ajuste del modelo propuesto, se hizo con la finalidad de hacer menos costoso el cálculo computacional, lo cual permitió utilizar técnicas de Cadenas de Markov (MCMC) para ajustar el modelo propuesto.

Se realizaron pruebas de simulación y se ajustó el modelo propuesto a datos reales, los resultados obtenidos sugieren que el modelo propuesto ajusta mejor a los datos y predice mejor que el modelo de regresión Ridge Bayesiano, el cual es un caso particular del modelo propuesto cuando $\rho = 0$.

De los resultados obtenidos, se observa que el modelo de regresión Ridge Bayesiano es un modelo muy robusto, dado que en las simulaciones se sabía que era un modelo incorrecto para ajustar y obtener predicciones, aún así, con el modelo BRR, se obtuvieron buenos resultados.

La representación estocástica, utilizada en este capítulo, se puede extender a modelos RKHS, (Reproducing Kernel Hilbert Spaces, de los Campos *et al.*, 2010, por sus siglas en inglés), modelos que pueden conducir a predicciones más precisas que la regresión Ridge, LASSO bayesiano, entre otros (por ejemplo Pérez-Rodríguez *et al.*, 2012).

Capítulo 5

Conclusiones y Recomendaciones

De la teoría desarrollada y el estudio de simulación realizado en el presente trabajo y con los resultados obtenidos del ajuste del modelo para predecir rendimiento de híbridos de maíz, empleado en el capítulo 3, se puede concluir:

- La inclusión del término de interacción G×A en el modelo incrementa su poder predictivo. Esto es debido a que las condiciones ambientales modulan el rendimiento de híbridos de maíz.
- La función de covarianza empleada, la propuesta por Jarquín *et al.* (2014), permitió mostrar que un gran porcentaje de la varianza fenotípica es explicada por el efecto de los ambientes, como se observa de los resultados reportados en las Tablas 3.1 y 3.2, y en las Tablas A.1-A.4, para los diferentes caracteres analizados, que se hallan en el Anexo A.
- El error residual disminuyó consistentemente al ajustar el modelo $GBLUP + Amb + Híbrido \times Amb + Padres \times Amb$, este resultado es similar al reportado por Jarquín *et al.* (2014) quienes afirman que los términos de interacción reducen significativamente la varianza del error e incrementan la capacidad predictiva del modelo.
- La capacidad predictiva de los dos modelos ajustados $GBLUP + Amb$ y $GBLUP + Amb + Híbrido \times Amb + Padres \times Amb$, fue medida a partir de la correlación de Pearson promedio entre fenotipos observados y fenotipos predichos, mostrando con esto, que el segundo modelo tiene una mayor capacidad de predicción.

  Estos resultados también se observan en los valores de la Tabla 3.3, que reporta el cambio en el porcentaje de la predicción predictiva del modelo (3.1) versus el modelo (3.2), así como en las Tablas A.5 y A.6, para los demás caracteres analizados.

  En promedio, el modelo que incluye los términos de interacción genotipo×ambiente, aptitud combinatoria general (padres×ambiente) y aptitud combinatoria específica

(híbrido×ambiente) es decir el modelo 3.2, tiene mayor capacidad predictiva que el modelo 3.1 que no incluye ningún término de interacción, el porcentaje de cambio osciló entre el 12 % y 22 % dependiendo del caracter.

- Por último, la incorporación del término de interacción, G×A, en los modelos empleados para predicción genómica de rendimiento de híbridos de maíz presentados en este estudio, son posibles de usar con cualquier cultivo y en la mayoría de los modelos GBLUP.

Con respecto a los resultados obtenidos con el ajuste del modelo de regresión con errores cuya distribución es Normal-Asimétrica empleado en el capítulo 4 concluimos que:

- De los resultados de la simulación reportados en la Tabla 4.1 se observa que, los valores estimados del parámetro de localidad $\hat{\beta}_0$ y de escala $\hat{\sigma}_e^2$ son estables, a cambios en el valor de ρ.

- La varianza estimada del efecto de los marcadores σ_β^2, es igual o menor en el modelo propuesto SNB (Normal-asimétrico Bayesiano) que la que se obtuvo con el modelo RRB (regresión Ridge Bayesiana), lo cual se puede observar con el valor del parámetro λ, valores grandes de este parámetro están asociados con estimaciones más precisas de los coeficientes de regresión para el modelo propuesto, SNB, por sobre el modelo, RRB.

  Las correlación entre valores observados y valores predichos, así como el Cuadrado Medio Error, es bastante similar en ambos modelos y se observa una ligera ganancia en el modelo propuesto.

- Con respecto a los valores reportados en la Tabla 4.2 se concluye que, el número efectivo de parámetros, pD y el Criterio de Información de la Devianza, DIC, permiten afirmar que el modelo propuesto, SNB es mejor que el modelo RRB (valores pequeños de estos criterios, favorecen al modelo propuesto).

  La correlación entre el efecto "verdadero"de los marcadores y el estimado, es ligeramente mayor para el modelo SNB que cuando se ajustó el modelo RRB. Este mismo patrón se observó en la correlación entre los valores genómicos observados ($\mathbf{X}\boldsymbol{\beta}$) y los estimados ($\mathbf{X}\hat{\boldsymbol{\beta}}$).

En relación a los resultados obtenidos del ajuste del modelo SNB, al conjunto de datos reales se concluye:

- El análisis gráfico de la densidad de la variable respuesta, resistencia a la enfermedad, en los tres ambientes, Santa Catalina, Kakamega y San Pedro Lagunillas, reportan asimetría hacia la derecha en cada caso y la distribución se concentra alrededor de valores pequeños. El grado de asimetría se reporta en los tres ambientes y se puede visualizar en la Figura 4.1.

- La estimación de las medias posteriores de los parámetros $\sigma_e^2, \sigma_\beta^2$ y ρ, así como el número efectivo de parámetros, pD y el Criterio de Información de la Devianza, DIC, indican que el modelo propuesto SNB es mejor. Tabla 4.3.

  La estimación del parámetro ρ, soporta el supuesto de asimetría en los errores, excepto quizás en el caso de Santa Catalina.

- Se puede observar de la Figura 4.2 que los resultados de la correlación de Pearson entre valores predicho con ambos modelos es muy alta (mayor a 0.95), lo que implica que, aunque los datos son asimétricos, si se ajusta el modelo RRB para selección de candidatos se esperaría obtener los mismos individuos obtenidos con el ajuste del modelo SNB.

- La predicción obtenida con el modelo SNB es ligeramente mayor que la obtenida cuando se ajustó el modelo RRB, para individuos en el conjunto de prueba. Esto es posible visualizarlo en la Figura 4.3 donde además se reporta el número de veces que el coeficiente de Pearson es más grande para el modelo SNB que el obtenido usando el modelo RRB.

  En el mismo sentido, el Cuadrado Medio del Error (CME) fue menor cuando se ajustó el modelo SNB que cuando se utilizó el modelo RRB, en el conjunto de individuos de prueba en los tres ambientes, favoreciendo ligeramente al modelo propuesto por sobre el modelo RRB, esta última afirmación se justifica con los valores reportados en la Tabla 4.4 y con la Figura 4.4.

- El modelo propuesto, SNB, se analizó en el contexto de Selección Genómica y en el caso $p >> n$, pero se puede emplear en otros contextos y por supuesto cuando $p < n$.

Finalmente:

La predicción de caracteres en Selección Genómica es posible realizarla a través de diferentes metodologías, en el presente trabajo se propusieron dos nuevos modelos: $GBLUP + Amb + Híbrido \times Amb + Padres \times Amb$ para la predicción de rendimiento de híbridos de maíz utilizando información genotípica de padres y un modelo de regresión con errores Normal-Asimétrico (SNB) para predecir resistencia a una enfermedad del maíz, GLS.

Los modelos propuestos presentaron buena capacidad predictiva al ser comparados con GBLUP y RRB (regresión Ridge Bayesiana) respectivamente.

En el caso del modelo SNB, se utilizó la representación estocástica de una variable aleatoria Normal-Asimétrica a fin de facilitar los cálculos. Esta representación se puede emplear en modelos RKHS (de los Campos *et al.*, 2010) que han permitido hacer predicciones más precisas que los modelos de regresión Ridge Bayesiano y LASSO Bayesiano entre otros (por ejemplo Pérez-Rodríguez *et al.*, 2012), por lo tanto es un

tema abierto a futuros estudios, pues se requiere de más investigación para realizar comparaciones de otros modelos con el SNB.

Referencias

Acosta-Pech, R., Crossa, J., de los Campos, G., Teyssèdre, S., Claustres, B., Pérez-Elizalde, S. y Pérez-Rodríguez, P. (2017). Genomic models with genotype × environment interaction for predicting hybrid performance: an application in maize hybrids. *Theoretical and Applied Genetics*, 130, 7, 1431–1440.

Arellano-Valle, R., Bolfarine, H. y Lachos, V. (2005). Skew-normal linear mixed models. *Journal of Data Science*, 3, 4, 415–438.

Arnold, B. C. y Beaver, R. J. (2000). The skew-Cauchy distribution. *Statistics & probability letters*, 49, 3, 285–290.

Azevedo, C. L., Bolfarine, H. y Andrade, D. F. (2011). Bayesian inference for a skew-normal IRT model under the centred parameterization. *Computational Statistics & Data Analysis*, 55, 1, 353–365.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 171–178.

Azzalini, A. y Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61, 3, 579–602.

Azzalini, A. y Genton, M. G. (2008). Robust Likelihood Methods Based on the Skew-t and Related Distributions. *International Statistical Review*, 76, 1, 106–129.

Balzarini, M., Bruno, C. y Arroyo, A. (2005). Análisis de ensayos agrícolas multiambientales. *Brujas. Córdoba, Argentina*.

Bellman, R. y Kalaba, R. (1961). Reduction of dimensionality, dynamic programming, and control processes. *J. Basic Eng*, 88, 82.

Bernardo, R. (1994). Prediction of maize single-cross performance using RFLPs and information from related hybrids. *Crop Science*, 34, 1, 20–25.

Bernardo, R. (1996a). Best linear unbiased prediction of maize single-cross performance. *Crop Science*, 36, 1, 50–56.

Bernardo, R. (1996b). Best linear unbiased prediction of the performance of crosses between untested maize inbreds. *Crop science*, 36, 4, 872–876.

Referencias

Bernardo, R. (1999). Marker-assisted best linear unbiased prediction of single-cross performance. *Crop Science*, 21, 4, 1277–1282.

Bernardo, R. (2002). *Breeding for quantitative traits in plants*. 576.5 B523. Stemma Press.

Bernardo, R. y Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. *Crop Science*, 47, 3, 1082–1090.

Box, G. E. y Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 211–252.

Browning, B. L. y Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics*, 84, 2, 210–223.

Burgueño, J., de los Campos, G., Weigel, K. y Crossa, J. (2012). Genomic Prediction of Breeding Values when Modeling Genotype × Environment Interaction Using Pedigree and Dense Molecular Markers. *Crop Science*, 52, 707–719.

Carlin, B. P. y Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis*. Chapman & Hall/CRC,.

Casella, G. y Berger, R. L. (2002). *Statistical inference*, tomo 2. Duxbury Pacific Grove, CA.

Chiogna, M. (1998). Some results on the scalar skew-normal distribution. *Journal of the Italian Statistical Society*, 7, 1, 1–13.

Cohen, A. C. (2016). *Truncated and censored samples: theory and applications*. CRC press.

Covarrubias-Pazaran, G. (2016). Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer. *PloS one*, 11, 6, e0156744.

Crossa, J. (2012). From Genotype × Environment interaction to Gene × Enviornment interaction. *Current Genomics*, 13, 3, 225–244.

Crossa, J., Cornelius, P. L. y Vargas, M. (2000). *Modelos Estadísticos Multiplicativos para el Análisis de la Interacción Genotipo × Ambiente*.

Crossa, J., de los Campos, G., Pérez-Rodríguez, P., Gianola, D., Burgueño, J., Araus, J. L., Makumbi, D., Singh, R. P., Dreisigacker, S., Yan, J., Arief, V., Banziger, M. y Braun, H.-J. (2010). Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers. *Genetics*, 186, 2, 713–724.

Crossa, J., Pérez-Rodríguez, P., de los Campos, G., Mahuku, G., Dreisigacker, S. y Magorokosho, C. (2011). Genomic Selection and Prediction in Plant Breeding. *Journal of Crop Improvement*, 25, 3, 239–261.

Crossa, J., Vargas, M., Van Eeuwijk, F., Jiang, C., Edmeades, G. y Hoisington, D. (1999). Interpreting Genotype × Environment interaction in tropical maize using linked molecular markers and environmental covariables. *Theoretical and Applied Genetics*, 99, 3-4, 611–625.

Dawson, J. C., Endelman, J. B., Heslot, N., Crossa, J., Poland, J., Dreisigacker, S., Manés, Y., Sorrells, M. E. y Jannink, J.-L. (2013). The use of unbalanced historical data for genomic selection in an international wheat breeding program. *Field Crops Research*, 154, 12 – 22.

de los Campos, G., Gianola, D. y Rosa, G. (2009a). Reproducing Kernel Hilbert Spaces Regression: a general framework for genetic evaluation. *Journal of Animal Science*, 87, 6, 1883–1887.

de los Campos, G., Gianola, D., Rosa, G. J., Weigel, K. A. y Crossa, J. (2010). Semi-parametric genomic-enabled prediction of genetic values using Reproducing Kernel Hilbert Spaces methods. *Genetics Research*, 92, 04, 295–308.

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D. y Calus, M. P. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, 193, 2, 327–345.

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K. y Cotes, J. M. (2009b). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, 182, 1, 375–385.

de los Campos, G. y Pérez-Rodríguez, P. (2015). BGLR: Bayesian Generalized Linear Regression. R package version 1.0.5.

Duvick, D. (2005). Genetic progress in yield of United States maize (Zea mays L.). *Maydica*, 50, 3/4, 193.

Eberhart, S. T. y Russell, W. (1966). Stability parameters for comparing varieties. *Crop science*, 6, 1, 36–40.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. *et al.* (2004). Least Angle Regression. *The Annals of Statistics*, 32, 2, 407–499.

FAO, C. y. G. (2013). www.fao.org/in-action/inpho/crop-compendium/cereals-grains.

Federer, W. T. y Raghavarao, D. (1975). On augmented designs. *Biometrics*, 29–35.

Fernandes, E., Pacheco, A. y Penha-Gonçalves, C. (2007). Mapping of quantitative trait loci using the skew-normal distribution. *Journal of Zhejiang University-Science B*, 8, 11, 792–801.

Finlay, K. y Wilkinson, G. (1963). The analysis of adaptation in a plant-breeding programme. *Crop and Pasture Science*, 14, 6, 742–754.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10, 4, 507–521.

Geman, S. y Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, 6, 6, 721–741.

Gianola, D. (2013). Priors in whole-genome regression: the Bayesian alphabet returns. *Genetics*, 194, 3, 573–596.

Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E. y Fernando, R. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics*, 183, 1, 347–363.

Gianola, D., Fernando, R. L. y Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics*, 173, 3, 1761–1776.

Gianola, D. y van Kaam, J. B. (2008). Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics*, 178, 4, 2289–2303.

Gilks, W. R., Best, N. y Tan, K. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statistics*, 455–472.

Goddard, M. E. y Hayes, B. (2007). Genomic selection. *Journal of Animal breeding and Genetics*, 124, 6, 323–330.

González-Recio, O., Gianola, D., Long, N., Wiegel, K., Rosa, G. y Avendaño, S. (2008). Non parametric methods for incorporating genomic information into genetic evaluation: an application to mortality in broilers. *Genetics*, 178, 2305–2313.

Hastie, T. J., Tibshirani, R. J. y Friedman, J. H. (2011). *The elements of statistical learning: data mining, inference, and prediction*. Springer.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 1, 97–109.

Hayes, B., Goddard, M. *et al.* (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157, 4, 1819–1829.

Heffner, E. L., Sorrells, M. E. y Jannink, J.-L. (2009). Genomic selection for crop improvement. *Crop Science*, 49, 1, 1–12.

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics*, 9, 2, 226–252.

Henderson, C. R. (1963). Selection index and expected genetic advance. *Statistical genetics and plant breeding*, 982, 141–163.

Henderson, C. R. (1984). Applications of linear models in animal breeding.

Heslot, N., Akdemir, D., Sorrells, M. E. y Jannink, J.-L. (2014). Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theoretical and Applied Genetics*, 127, 2, 463–480.

Heslot, N., Yang, H.-P., Sorrells, M. E. y Jannink, J.-L. (2012). Genomic selection in plant breeding: a comparison of models. *Crop Science*, 52, 1, 146–160.

Hoerl, A. E. y Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 1, 55–67.

Jarquín, D., Crossa, J., Lacaze, X., Cheyron, P., Daucourt, J., Lorgeou, J., Piraux, F., Guerreiro, L., Pérez-Rodríguez, P., Calus, M., Burgueño, J. y Campos, G. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics*, 127, 3, 595–607.

Joanes, D. y Gill, C. (1998). Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47, 1, 183–189.

Kadam, D. C., Potts, S. M., Bohn, M. O., Lipka, A. E. y Lorenz, A. J. (2016). Genomic Prediction of Single Crosses in the Early Stages of a Maize Hybrid Breeding Pipeline. *G3: Genes— Genomes— Genetics*, 6, 11, 3443–3453.

Kandus, M., Almorza, D., Boggio Ronceros, R., Salerno, J. *et al.* (2010). Statistical models for evaluating the genotype-environment interaction in maize (Zea mays L.). *Phyton-Revista Internacional de Botanica Experimental*, 79, 39–46.

Kim, H.-J. (2005). Bayesian Estimation for Skew Normal Distributions Using Data Augmentation. *Communications for Statistical Applications and Methods*, 12, 2, 323–333.

Landfors, M., Philip, P., Rydén, P. y Stenberg, P. (2011). Normalization of high dimensional genomics data where the distribution of the altered variables is skewed. *PloS one*, 6, 11, e27942.

Li, Z., Möttönen, J. y Sillanpää, M. (2015). A robust multiple-locus method for quantitative trait locus analysis of non-normally distributed multiple traits. *Heredity*, 115, 6, 556–564.

Liseo, B. y Loperfido, N. (2006). A note on reference priors for the scalar skew-normal distribution. *Journal of Statistical Planning and Inference*, 136, 2, 373–389.

Liseo, B. y Parisi, A. (2013). Bayesian inference for the multivariate skew-normal model: A population Monte Carlo approach. *Computational Statistics & Data Analysis*, 63, 125–138.

López-Cruz, M., Crossa, J., Bonnett, D., Dreisigacker, S., Poland, J., Jannink, J.-L., Singh, R. P., Autrique, E. y de los Campos, G. (2015). Increased prediction accuracy in wheat breeding trials using a marker× environment interaction genomic selection model. *G3: Genes— Genomes— Genetics*, 569–582.

Lorenzana, R. E. y Bernardo, R. (2009). Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theoretical and Applied Genetics*, 120, 1, 151–161.

Lynch, M., Walsh, B. *et al.* (1998). *Genetics and analysis of quantitative traits*, tomo 1. Sinauer Sunderland, MA.

Malosetti, M., Ribaut, J.-M. y van Eeuwijk, F. A. (2014). The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. *Drought phenotyping in crops: From theory to practice*, 4, 44, 53.

Massman, J. M., Gordillo, A., Lorenzana, R. E. y Bernardo, R. (2013). Genomewide predictions from maize single-cross data. *Theoretical and Applied Genetics*, 126, 1, 13–22.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. y Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21, 6, 1087–1092.

Meuwissen, T. H. E., Hayes, B. J. y Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics*, 157, 4, 1819–1829.

Park, T. y Casella, G. (2008). The Bayesian LASSO. *Journal of the American Statistical Association*, 103, 482, 681–686.

Pérez-Rodríguez, P., Crossa, J., Bondalapati, K., De Meyer, G., Pita, F. y de los Campos, G. (2015). A Pedigree-Based Reaction Norm Model for Prediction of Cotton Yield in Multienvironment Trials. *Crop Science*, 55, 1143–1151.

Pérez-Rodríguez, P., de los Campos, G., Crossa, J. y Gianola, D. (2010). Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian Linear Regression package in R. *The Plant Genome*, 3, 2, 106–116.

Pérez-Rodríguez, P., Gianola, D., González-Camacho, J. M., Crossa, J., Manès, Y. y Dreisigacker, S. (2012). Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3: Genes— Genomes— Genetics*, 2, 12, 1595–1605.

Pérez-Rodríguez, P., Villaseñor, J. A., Pérez, S. y Suárez, J. (2017). Bayesian Estimation for the Centered Parameterization of the Skew-Normal Distribution. *Revista Colombiana de Estadística*, 40, 1, 123–140.

Pewsey, A. (2000). Problems of inference for Azzalini's skew- normal distribution. *Journal of Applied Statistics*, 27, 7, 859–870.

Piepho, H.-P. (2009). Ridge regression and extensions for genomewide selection in maize. *Crop Science*, 49, 4, 1165–1176.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ruppert, D., Wand, M. P. y Carroll, R. J. (2003). *Semiparametric regression*. 12. Cambridge University Press.

Rusell, N. G. y González, G., Farías (2002). Análisis de Regresión Lineal con Errores Distribuidos Normal Sesgados. *CIMAT*, I-02-30/10-12-2002, 1–44.

Sánchez, F. (1974). *El problema de la interacción genético-ambiental en genotecnia vegetal*. PATENA.

Schrag, T. A., Möhring, J., Melchinger, A. E., Kusterer, B., Dhillon, B. S., Piepho, H.-P. y Frisch, M. (2010). Prediction of hybrid performance in maize using molecular markers and joint analyses of hybrids and parental inbreds. *Theoretical and applied genetics*, 120, 2, 451–461.

Smith, A. y Roberts, G. (1993). Bayesian Computation Via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society*, 55, 1, 3–23.

Sorensen, D. y Gianola, D. (2007). *Likelihood, Bayesian, and MCMC methods in quantitative genetics*. Springer Science & Business Media.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. y Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 4, 583–639.

Technow, F. y Melchinger, A. E. (2013). Genomic prediction of dichotomous traits with Bayesian logistic models. *Theoretical and Applied Genetics*, 126, 4, 1133–1143.

Technow, F., Riedelsheimer, C., Schrag, T. A. y Melchinger, A. E. (2012). Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. *Theoretical and Applied Genetics*, 125, 6, 1181–1194. ISSN 1432-2242.

Technow, F., Schrag, T. A., Schipprack, W., Bauer, E., Simianer, H. y Melchinger, A. E. (2014). Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. *Genetics*, 197, 4, 1343–1355.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions. *Ann. Statist.*, 22, 4, 1701–1728.

Tikhonov, A. y Arsenin, V. (1977). *Solution of ill-posed problems*. Wiley.

Tikhonov, A. N. (1963). Regularization of incorrectly posed problems. En *Soviet Math. Dokl*, tomo 4, 1624–1627.

Trautmann, H., Steuer, D., Mersmann, O. y Bornkamp, B. (2012). truncnorm: Truncated normal distribution.

van Eeuwijk, F. A., Boer, M., Totir, L. R., Bink, M., Wright, D., Winkler, C. R., Podlich, D., Boldman, K., Baumgarten, A., Smalley, M. *et al.* (2010). Mixed model approaches for the identification of QTLs within a maize hybrid breeding program. *Theoretical and Applied Genetics*, 120, 2, 429–440.

VanRaden, P. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91, 11, 4414–4423.

Vélez, T. M. (2015). *Rendimiento de cruzas simples de líneas de maíz de ACG contrastante y su predicción mediante SNPs*. Tesis Doctoral, Colegio de Postgraduados.

Venables, W. N. y Ripley, B. D. (2002). Random and mixed effects. En *Modern applied statistics with S*, 271–300. Springer.

Weigel, K., de los Campos, G., González-Recio, O., Naya, H., Wu, X., Long, N., Rosa, G. y Gianola, D. (2009). Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. *Journal of dairy science*, 92, 10, 5248–5257.

Witkovský, V. (2012). Estimation, testing, and prediction regions of the fixed and random effects by solving the Henderson's mixed model equations. *Measurement Science Review*, 12, 6, 234–248.

Xu, S., Zhu, D. y Zhang, Q. (2014). Predicting hybrid performance in rice using genomic best linear unbiased prediction. *Proceedings of the National Academy of Sciences*, 111, 34, 12456–12461.

Yates, F. y Cochran, W. (1938). The analysis of groups of experiments. *The Journal of Agricultural Science*, 28, 04, 556–580.

Zhao, Y., Zeng, J., Fernando, R. y Reif, J. C. (2013). Genomic prediction of hybrid wheat performance. *Crop Science*, 53, 3, 802–810.

Zou, H. y Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 2, 301–320.

Anexos

Anexo A: Ajuste de modelos con el paquete `BGLR` de `R`

Este anexo contiene el código de `R` utilizado para ajustar los modelos presentados en el capítulo 3.

Los modelos se ajustaron usando el paquete estadístico `BGLR` (de los Campos y Pérez-Rodríguez, 2015). Para hacer más eficiente el cómputo, en `BGLR`, se utilizó la descomposición en valores propios de las matrices de varianzas y covarianzas de los modelos 1 y 2. Si $\mathbf{Zu}, \mathbf{u} \sim N(\mathbf{0}, \sigma_a^2\mathbf{A})$, con $\mathbf{A}$ matriz de varianzas-covarianzas, y sea $\boldsymbol{\Gamma\Lambda\Gamma^t} = \boldsymbol{A}$, con $\boldsymbol{\Gamma}$ representando los eigen-vectores de $\mathbf{A}$ y $\boldsymbol{\Lambda}$ sus correspondientes eigen-valores; entonces $\mathbf{Zu} \overset{d}{=} \mathbf{Z\Gamma\Lambda}^{\frac{1}{2}}\mathbf{u}^*$, donde $\mathbf{u}^* \sim N(\mathbf{0}, \sigma_a^2\mathbf{I})$ y $\overset{d}{=}$ representa igualdad en distribución. Usando este resultado, el modelo $\boldsymbol{y} = \mathbf{Zu} + \boldsymbol{e}$ se puede escribir como $\boldsymbol{y} = \mathbf{Z\Lambda\Gamma}^{\frac{1}{2}}\mathbf{u}^* + \boldsymbol{e}$, el cual es conocido como regresión Ridge Bayesiana, un modelo que se puede ajustar muy eficientemente en el paquete `BGLR`. Los modelos 1-2 se re-escribieron usando la representación descrita arriba. Abajo se describen los datos y los códigos de `R` necesarios para ajustar los modelos.

Descripción de los Datos

```

#X_D: Una matriz con información genotípica para las líneas Dent. Las entradas son 0 para
#homocigoto recesivo y 2 para homocigoto dominante. El nombre de las filas de
#X_D corresponden a los ids de las líneas.
#X_F: Una matriz con información genotípica para las líneas Flint. Las entradas son 0 para
#homocigoto recesivo y 2 para homocigoto dominante. El nombre de las filas de
#X_F corresponden a los ids de las líneas.
#Pheno: Una tabla de datos con 4 columnas:
#y: Variable respuesta.
#P1:Ids para líneas Dent.
#P2: Ids para líneas Flint.
#Loc: Ids para localidades.

#Matrices de Incidencias y de Relaciones Genómicas
#El siguiente fragmento de código de R genera todas las matrices requeridas
```

```
#para ajustar los modelos 1 y 2 .

#G_D y G_F

X_D = scale(X_D, center=TRUE, scale=TRUE)
X_F = scale(X_F, center=TRUE, scale=TRUE)

G_D=tcrossprod(X_D)/ncol(X_D)
G_F=tcrossprod(X_F)/ncol(X_F)

#Matriz asociada a los híbridos: H
P1 = rownames(G_D)
P2 = rownames(G_F)}

#Híbridos únicos.
hybrid=paste(Pheno\$P1, Pheno\$P2, sep="@")
hybrid=unique(hybrid)

#Construyendo la matriz de relaciones Genómicas para los híbridos.
nh=length(hybrid)
H=matrix(NA, nrow=nh, ncol=nh)

  for(k in 1:nh)
{
    for(l in 1:nh)
  {
    tmp1=strsplit(hybrid[k],"@")[[1]]
    tmp2=strsplit(hybrid[l],"@")[[1]]
    i=which(P1==tmp1[1])
    istar=which(P2==tmp2[1])
    j=which(P2==tmp1[2])
    jstar=which(P2==tmp2[2])
    H[k,l]= G_D[i,istar]*G_F[j,jstar]
  }
}
rownames(H)=colnames(H)=hybrid

#Ordenando filas y columnas de G_D,G_F y H
#Haciendo factor al híbrido y ordenando las filas y columnas de H.

hybrid=as.factor(hybrid)
index=order(hybrid)
H=H[index,index]

#Haciendo factor a P1 y ordenando las filas y columnas de G_D.
P1=as.factor(P1)
index=order(P1)
G_D=G_D[index,index]

#Haciendo factor a P2 y ordenando las filas y columnas de G\_F.
P2=as.factor(P2)
index=order(P2)
G_F=G_F[index,index]

```

```
#Matrices de incidencias Z_E,Z_D,Z_F,Z_H y variable respuesta.

hybrid=as.factor(paste(Pheno\$P1, Pheno\$P2, sep="@"))
Loc=as.factor(Pheno$Loc)
index=order(Loc, hybrid)
Pheno=Pheno[index,]
hybrid=hybrid[index]
Loc=Loc[index]
P1=as.factor(Pheno$P1)
P2=as.factor(Pheno$P2)

#Variable respuesta

y=Pheno$y

#Matrices de Incidencia

Z_D=model.matrix(~ P1-1)
Z_F=model.matrix(~P2-1)
Z_h=model.matrix(~hybrid-1)
Z_E=model.matrix(~ Loc-1)}
Z_EZ_Et=tcrossprod(Z_E)}


#Ajustando los modelos
#Ajustando el modelo 1: GBLUP+Amb
#Cargando el Software

library (BGLR)

#Constuyendo las matrices de incidencia.

#Descomposición en eigen-valores para hacer más rápidos los cálculos.

EVD_G_D=eigen(G_D)
index=EVD_G_D$values>1e-10
EVD_G_D$vectors=EVD_G_D$vectors[,index]
EVD_G_D$values=EVD_G_D$values[index]}

EVD_G_F=eigen(G_F)
index=EVD_G_F$values>1e-10
EVD_G_F$vectors=EVD_G_F$vectors[,index]
EVD_G_F$values=EVD_G_F$values[index]

EVD_H=eigen(H)
index=EVD_H$values>1e-10
EVD_H$vectors=EVD_H$vectors[,index]
EVD_H$values=EVD_H$values[index]

Z_Dstar=Z_D\ %*\ %EVD_G_D$vectors\ %*\ %sqrt(diag(EVD_G_D$values))
Z_Fstar=Z_F\ %*\ %EVD_G_F$vectors\ %*\ %sqrt(diag(EVD_G_F$values))}
Z_hstar=Z_h\ %*\ %EVD_H$vectors\ %*\ %sqrt(diag(EVD_H$values))}

#Predictor Lineal
```

```
124
125 ETA1=list(list(X=Z_E, model="BRR"),
126 list(X=Z_Dstar, model="BRR"),
127 list(X=Z_Fstar, model="BRR"),
128 list(X=Z_hstar, model="BRR"))
129
130 #Ajustando el modelo 1
131
132 m1=BGLR(y=y, ETA=ETA1, nIter=30000, burnIn=5000)
133
134 #Ajustando el modelo 2: GBLUP+Amb+Híbrido x Amb+ Padres x Amb}
135 #Matrices de incidencia adicionales
136
137 V1=tcrossprod(Z_Dstar)*Z_EZ_Et
138 EVD_V1=eigen(V1)
139 index=EVD_V1$values>1e-10
140 EVD_V1$vectors=EVD_V1$vectors[,index]
141 EVD_V1$values=EVD_V1$values[index]
142 Z_DxEstar=EVD_V1$vectors\%*\%sqrt(diag(EVD_V1\$values))}
143
144 V2=tcrossprod(Z_Fstar)*Z_EZ_Et
145 EVD_V2=eigen(V2)
146 index=EVD_V2$values>1e-10
147 EVD_V2$vectors=EVD_V2$vectors[,index]
148 EVD_V2$values=EVD_V2$values[index]
149 Z_FxEstar=EVD_V2$vectors\%*\%sqrt(diag(EVD_V2$values))}
150
151 #Predictor Lineal
152
153 ETA2=list(list(X=Z_E, model="BRR"),
154 list(X=Z_Dstar, model="BRR"),
155 list(X=Z_Fstar, model="BRR"),
156 list(X=Z_hstar, model="BRR")),
157 list(X=Z_hxEstar, model="BRR")),
158 list(X=Z_DxEstar, model="BRR")),
159 list(X=Z_FxEstar, model="BRR"))
160
161 #Ajustando el modelo 2
162
163 m2=BGLR(y=y, ETA=ETA2, nIter=30000, burnIn=5000)
```

Generando los conjuntos para la validación cruzada

El siguiente fragmento de código, genera 5 particiones del esquema 2 de validación cruzada (vea Burgueño *et al.*, 2012, para más detalles). El paquete `BGLR` acepta valores faltantes (NAs) en el vector de respuesta, las entradas con valores perdidos se predicen durante el ajuste y se pueden extraer una vez que el modelo se ha ajustado.

```
1 nFolds=5
2 set.seed(123)
```

```
IDs=as.character(unique(hybrid))
IDy=as.character(hybrid)
sets=rep(NA, length(IDy))
  for(i in 1:IDs)
{
  tmp=which(IDy==i)
  ni=length(tmp)
  tmpFold=sample(1:nFolds, size=ni, replace=ni>nFolds)
  sets[tmp]=tmpFold
}

#Ajustando el modelo 1 con valores perdidos para el conjunto 1

yNA=y
yNA[sets==1]=NA
m1=BGLR(y=yNA, ETA=ETA1, nIter=30000, burnIn=5000)
#Predicciones para el conjunto 1
m1$yHat[sets==1]
```

Estimación de los parámetros de varianza para %SC y %DMC

Tabla A.1: Parámetros de varianza estimados (A=ambientes; D=Dent; F=Flint; H=Híbridos, Res=Residual), y desviación estándar (en paréntesis) y el porcentaje de varianza dentro de los ambientes explicada por cada efecto aleatorio para %SC estimado, ajustando el Modelo 3.1.

Modelo 1: GBLUP+Amb					
Año	A	D	F	H	Res
2004	3(1.6) --	2(0.7) 12.1	2.9(1.8) 17	1.8(0.7) 11.3	9.6(0.8) 59.7
2005	15.8(6.4) --	3.7(1.2) 14.7	6.2(2.9) 24.1	2.8(0.8) 11.4	12.3(0.7) 49.8
2006	10.9(5) --	3.8(1.3) 13.6	4.8(2.4) 16.7	3.1(1) 11.2	16.1(1) 58.5
2007	8.4(3.5) --	2.6(0.8) 12.7	3.6(1.6) 17.3	2.4(0.7) 11.8	11.8(0.6) 58.2
2008	5.6(2.1) --	3.3(1.1) 14	3.7(1.7) 15.6	2.6(0.7) 11.1	13.7(0.6) 59.3
2009	9.2(3.5) --	2.4(0.7) 14.3	5(2.5) 28.1	2(0.5) 12.1	7.6(0.3) 45.4
2010	7.4(2.7) --	3.2(0.7) 17.4	4.5(1.7) 24.1	1.5(0.3) 8.1	9.1(0.3) 50.4
2011	6.5(2.4) --	2.8(0.7) 15.7	3.5(1.3) 19.3	1.3(0.3) 7.6	10(0.4) 57.3
2012	9.5(3.6) --	3.3(0.7) 16.9	5.1(1.8) 25.6	2.5(0.4) 12.7	8.6(0.3) 44.7
2013	29.3(10) --	4.1(0.9) 11.8	8(2.9) 22.3	2.5(0.5) 7.3	20.3(0.6) 58.6
2014	7.9(2.8) --	2.9(0.6) 10.2	7.8(2.8) 26.5	1.5(0.3) 5.3	16.5(0.5) 58
2015	14.9(5.9) --	3.6(0.9) 14.5	9.3(3.3) 36.2	2.2(0.5) 8.9	9.9(0.4) 40.4

Tabla A.2: Parámetros de varianza estimados (E=ambientes; D=Dent; F=Flint; H=Híbridos; H×E=Híbridos×Env; D×E=Dent×Env; F×E=Flint×Env;Res=Residual), y desviación estándar (en paréntesis) y el porcentaje de varianza dentro de los ambientes explicada por cada efecto aleatorio para %SC estimado, ajustando el Modelo 3.2.

Modelo 2: GBLUP+Amb+Híbrido×Amb+Padres×Amb								
Año	A	D	F	H	H ×A	D × A	F × A	Res
2004	2.3(1.3)	1.4(0.6)	2.1(1.5)	1.4(0.6)	1.2(0.5)	1.2(0.5)	1.4(0.6)	7.7(0.8)
	—	8.6	12.2	8.4	7.6	7.4	8.5	47.4
2005	12.4(5.4)	2.6(0.9)	3.3(1.8)	1.9(0.6)	1.9(0.6)	2.4(0.8)	3.2(1.1)	9(0.6)
	—	10.8	13.3	7.9	7.9	9.8	13	37.4
2006	8.1(3.9)	2.6(1.1)	3(1.7)	2.4(0.9)	1.8(0.6)	2(0.7)	2.4(0.9)	13.6(1)
	—	9.4	10.6	8.6	6.4	7.1	8.5	49.4
2007	6.9(2.9)	1.8(0.7)	2.5(1.3)	1.6(0.5)	1.8(0.6)	1.4(0.4)	2.4(0.8)	9.3(0.5)
	—	8.8	11.7	7.8	8.8	6.6	11.4	44.9
2008	5.2(2.2)	2.8(1)	2.7(1.4)	1.5(0.5)	1.8(0.5)	2.3(0.6)	3.4(0.9)	8.9(0.5)
	—	11.8	11.4	6.3	7.7	10	14.5	38.1
2009	8.8(3.5)	2(0.6)	3.5(1.8)	1.5(0.4)	0.9(0.2)	0.9(0.3)	1.9(0.5)	6(0.3)
	—	12.3	20	9.1	5.7	5.5	11.2	36.3
2010	6.3(2.4)	3.3(0.7)	4(1.7)	1(0.2)	1(0.2)	0.9(0.2)	2.8(0.5)	6.4(0.3)
	—	17	20.4	5	5.3	4.5	14.6	33.1
2011	5.3(2.1)	2.7(0.7)	2.8(1.1)	1(0.3)	0.9(0.2)	1(0.2)	1.4(0.4)	8.4(0.3)
	—	15	15.4	5.4	4.7	5.3	7.7	46.4
2012	7.7(2.9)	3.1(0.6)	3.7(1.4)	1.6(0.3)	1.1(0.3)	0.9(0.2)	2.2(0.4)	6.4(0.3)
	—	16.4	18.9	8.7	6	5	11.4	33.7
2013	22.6(8)	3.4(0.8)	6(2.3)	2(0.5)	1.6(0.4)	1.5(0.3)	3.2(0.7)	16.6(0.5)
	—	9.9	17.2	5.8	4.6	4.3	9.4	48.8
2014	7.8(2.7)	2.4(0.6)	6.9(2.6)	1.2(0.3)	1.3(0.3)	1.5(0.4)	1.9(0.4)	13.8(0.5)
	—	8.2	23.3	4.1	4.4	5.3	6.6	48.1
2015	13.3(5.3)	2.9(0.8)	7.7(3)	1.6(0.5)	1(0.3)	1.2(0.3)	1.4(0.4)	8.7(0.4)
	—	11.9	30.5	6.7	4.2	5.2	5.6	35.9

Tabla A.3: Parámetros de varianza estimados (E=ambientes; D=Dent; F=Flint; H=Híbridos, Res=Residual), y desviación estándar (en paréntesis) y el porcentaje de varianza dentro de los ambientes explicada por cada efecto aleatorio para %DMC estimado, ajustando el Modelo 3.1.

Modelo 1: GBLUP+Amb					
Año	E	D	F	H	Res
2004	5.7(2) --	1.5(0.4) 22.3	1.7(0.9) 24.5	0.7(0.2) 9.9	2.8(0.1) 43.3
2005	9.3(3.2) --	2.2(0.6) 20.7	3.3(1.5) 30.1	0.9(0.2) 8.9	4.2(0.2)) 40.3
2006	7.5(2.2) --	1(0.3) 14.7	1.5(0.6) 20.6	0.8(0.2) 11.9	3.6(0.1) 52.7
2007	5.8(2.0) --	2.3(0.6) 19.5	2.5(1.1) 20.4	1.2(0.3) 10.7	5.7(0.2) 49.4
2008	10.5(3.1) --	2(0.5) 16.8	2.2(1) 18.4	1.4(0.3) 11.6	6.3(0.2) 53.2
2009	11.1(3.4) --	1.4(0.4) 11.8	2.8(1.2) 23.2	0.9(0.2) 7.7	6.7(0.2) 57.3
2010	4.3(1.3) --	1.8(0.4) 16.2	3.1(1.2) 26.7	1(0.2) 8.8	5.4(0.2) 48.3
2011	8.6(2.7) --	1.4(0.3) 13.8	2.3(0.8) 21.7	1(0.2) 10	5.6(0.2) 54.6
2012	11.5(3.2) --	1.7(0.3) 14	4.1(1.3) 33.4	1.5(0.2) 12.9	4.7(0.1) 39.7
2013	10.9(3.1) --	2.5(0.5) 15	5(1.7) 28.9	0.9(0.2) 5.5	8.4(0.2) 50.6
2014	12.5(3.3) --	1.3(0.2) 12.7	3.5(1.2) 32.2	0.8(0.1) 7.9	4.9(0.1) 47.2
2015	12(3.1) --	1.8(0.3) 15.6	4.2(1.5) 34.6	0.9(0.1) 7.3	5(0.1) 42.5

Tabla A.4: Parámetros de varianza estimados (E=ambientes; D=Dent; F=Flint; H=Híbridos; H×E=Híbridos×Env; D×E=Dent×Env; F×E=Flint×Env;Res=Residual), y desviación estándar (en paréntesis) y el porcentaje de varianza dentro de los ambientes explicada por cada efecto aleatorio para %DMC estimado, ajustando el Modelo 3.2.

Modelo2: GBLUP+Amb+Híbrido×Amb+Padres×Amb								
Año	A	D	F	H	H ×A	D × A	F × A	Res
2004	5.3(2)	1(0.3)	1.4(0.8)	0.5(0.1)	0.5(0.1)	0.5(0.1)	1(0.3)	1.8(0.1)
	—	15.1	20.1	7.6	7.7	6.9	15.2	27.4
2005	8.7(3.1)	1.8(0.5)	2.8(1.3)	0.6(0.2)	0.6(0.1)	0.6(0.1)	1(0.3)	3.2(0.2)
	—	16.9	25.4	6	5.6	5.7	9.5	31
2006	7.1(2.1)	0.9(0.3)	0.9(0.5)	0.6(0.2)	0.5(0.1)	0.6(0.1)	0.5(0.1)	2.7(0.1)
	—	13	13.6	9.1	7.5	8.4	8.2	40.1
2007	6(2)	2(0.5)	1.9(0.9)	0.8(0.2)	0.7(0.2)	1.2(0.2)	1.3(0.3)	3.3(0.1)
	—	18	16.7	7	6.6	10.6	11.4	29.8
2008	10.5(3.1)	1.8(0.5)	1.7(0.8)	1(0.3)	1(0.2)	0.9(0.2)	1(0.2)	4.1(0.2)
	—	15.5	14.8	8.5	8.4	7.9	8.9	36
2009	11.4(3.4)	1.3(0.4)	2.3(1.1)	0.7(0.2)	0.9(0.2)	1(0.2)	1.3(0.3)	4.6(0.2)
	—	10.7	18.4	5.5	7.9	8.1	10.9	38.6
2010	4(1.2)	2(0.4)	3(1.2)	0.7(0.1)	0.6(0.1)	0.7(0.1)	1.5(0.2)	3.4(0.1)
	—	16.7	24.5	5.8	5.1	6	12.6	29.3
2011	7.7(2.5)	1.3(0.3)	1.7(0.7)	0.9(0.2)	0.6(0.1)	0.7(0.1)	1.7(0.3)	3.9(0.1)
	—	12.2	15.6	8	5.7	6.4	16	36.1
2012	10.4(3.1)	1.6(0.3)	3.4(1.2)	0.9(0.2)	0.5(0.1)	0.7(0.1)	1.9(0.3)	2.9(0.1)
	—	13.3	28	7.6	4.3	6.2	15.8	24.9
2013	10(3)	2.5(0.5)	4.2(1.5)	0.8(0.2)	0.6(0.1)	0.8(0.1)	2.1(0.3)	5.8(0.2)
	—	15.2	24.6	4.7	3.4	4.8	12.4	34.9
2014	11.8(3.3)	1.2(0.2)	3.2(1.2)	0.7(0.1)	0.4(0.1)	0.7(0.1)	1.5(0.2)	3.6(0.1)
	—	15	13.6	10.3	5.2	4.1	10.4	41.6
2015	11.9(3.3)	1.7(0.3)	3.9(1.4)	0.7(0.1)	0.9(0.1)	0.8(0.1)	1.3(0.2)	3.1(0.1)
	—	13.7	30.7	5.9	7.5	6.4	10.8	25

Tabla A.5: [a] %Cambio M1 vs M2 $= (r_{M2} - r_{M1})/r_{M1} \times 100$, donde r_{M1} y r_{M2} son los coeficientes de correlación de Pearson para los modelos M1-M2 respectivamente. Variable respuesta %SC.

Año	M1	M2	$\%Cambio\ M1\text{vs}M2^a$
2004	0.2091	0.2327	11.28
2005	0.3680	0.4634	25.93
2006	0.3638	0.3946	8.47
2007	0.3742	0.41816	11.75
2008	0.3407	0.4606	35.19
2009	0.4604	0.5293	14.97
2010	0.5717	0.6445	12.72
2011	0.4584	0.5060	10.39
2012	0.5798	0.6202	6.96
2013	0.4544	0.4912	8.10
2014	0.4981	0.5094	2.26
2015	0.5410	0.5387	-0.43
Promedio	0.4350	0.4840	16.73

Tabla A.6: [a] %Cambio M1 vs M2 $= (r_{M2} - r_{M1})/r_{M1} \times 100$, donde r_{M1} y r_{M2} son los coeficientes de correlación de Pearson para los modelos M1-M2 respectivamente. Variable respuesta %DMC.

Año	M1	M2	$\%Cambio\ M1\text{vs}M2^a$
2004	0.4178	0.5580	33.56
2005	0.4450	0.5563	25.02
2006	0.4329	0.5340	23.37
2007	0.5668	0.6986	23.27
2008	0.4130	0.5552	34.43
2009	0.4000	0.5456	36.40
2010	0.5770	0.6572	13.90
2011	0.4969	0.5960	19.95
2012	0.6709	0.7404	10.35
2013	0.5247	0.6155	17.30
2014	0.5558	0.6164	10.92
2015	0.5692	0.6397	12.37
Promedio	0.5058	0.6094	21.74

Anexo B: Distribuciones condicionales de los parámetros.

En este anexo, se presenta el desarrollo para obtener las distribuciones condicionales completas, utilizadas para estimar los parámetros de la Regresión Normal Asimétrica del capítulo 4.

Parametrización Centrada

El propósito principal al utilizar la parametrización centrada de una variable Normal-Asimétrica, es que permite resolver el problema de singularidad de la matriz de información y la forma de la verosimilitud es generalmente más simple (Azzalini y Capitanio, 1999). La reparametrización se realiza de la siguiente manera. Primero se parte del esquema de la parametrización directa, cuando la variable latente se agrega al modelo y se obtiene la siguiente distribución conjunta de U y Z:

$$f_{U,Z}(u,z|\rho) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{1-\rho^2}} exp\left\{-\frac{1}{2(1-\rho^2)}(u-\rho z)^2\right\}\frac{1}{\sqrt{2\pi}}\left\{-\frac{1}{2}z^2\right\} I_{(0,\infty)}(z), \quad u \in \mathbb{R}.$$

Ahora, considere la siguiente transformación:

$$Y = \mu + \sigma_e\left(\frac{U-E(U)}{\sqrt{Var(U)}}\right) \quad y \quad W = Z,$$

entonces:

$$\frac{Y-\mu}{\sigma_e} = \frac{U-E(U)}{\sqrt{Var(U)}} = \frac{U-E(U)}{S_u},$$

por lo tanto:

$$U = \left(\frac{Y-\mu}{\sigma_e}\right) S_u + E_u = g_1^{-1}(Y,W),$$

y

$$g_2^{-1}(Y,W) = W = Z.$$

Utilizando el teorema de cambio de variable (Casella y Berger, 2002), se obtiene la distribución conjunta de $f(y,w) = f_{(U,Z)}(g_1^{-1}(y,w), g_2^{-1}(y,w))|J^{-1}|$. Siendo esta igual a:

$$f(y,z) = \frac{2}{\sqrt{1-\rho^2}}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{2\pi}}\frac{S_u}{\sigma_e} exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{y-\mu}{\sigma_e}\right)S_u + E_u - \rho z\right]^2\right\} exp\left\{-\frac{1}{2}z^2\right\},$$

$y \in \mathbb{R}, z \geq 0.$

Verosimilitud Aumentada

En este apartado se presenta la función de verosimilitud asociada al vector de parámetros en el contexto de regresión y se incorpora la expresión asociada a las variables latentes, generando la verosimilitud aumentada, observe que esto permite que la versión final de la función de verosimilitud tenga un aspecto más tratable en la identificación del núcleo de las distribuciones condicionales posteriores de cada uno de los parámetros de interés. Si $L(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{y}, \boldsymbol{X})$ denota la función de verosimilitud entonces:

$$L(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{y}, \boldsymbol{X}) =$$
$$\prod_{i=1}^{n} \frac{2}{\sqrt{1-\rho^2}} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2\pi}} \frac{S_u}{\sigma_e} exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{y_i-\mu_i}{\sigma_e} \right) S_u + E_u - \rho z_i \right]^2 \right\} \times exp\left\{ -\frac{1}{2} z_i^2 \right\} I_{(0,\infty)}(z_i)$$
$$L \propto \prod_{i=1}^{n} \frac{1}{\sqrt{1-\rho^2}} \frac{S_u}{\sigma_e} exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{y_i-\mu_i}{\sigma_e} \right) S_u + E_u - \rho z_i \right]^2 \right\} \times exp\left\{ -\frac{1}{2} z_i^2 \right\} I_{(0,\infty)}(z_i),$$

donde $\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = \beta_0 + \mathbf{x}_i^t \boldsymbol{\beta}$.

Distribuciones a *priori*

Las distribuciones a priori asignadas a cada uno de los parámetros de interés fueron las siguientes:

$$\begin{aligned}
\beta_0|\sigma^2_{\beta_0} &\sim N(0, \sigma^2_{\beta_0}), \\
\boldsymbol{\beta}|\sigma^2_{\beta} &\sim N(\mathbf{0}, \sigma^2_{\beta}\mathbf{I}), \\
p(\sigma^2_{\beta}|S_{\beta}, df_{\beta}) &= \chi^{-2}(S_{\beta}, df_{\beta}), \\
p(\sigma^2_{e}|S_{e}, df_{e}) &= \chi^{-2}(S_{e}, df_{e}), \\
p(\rho|a_0, b_0) &\propto \left(\frac{1-\rho}{2} \right)^{(a_0-1)} \left(1 - \frac{1-\rho}{2} \right)^{(b_0-1)}.
\end{aligned}$$

La parametrización de la distribución χ^{-2} es la utilizada en Sorensen y Gianola (2007): Si $T \sim \chi^{-2}(\nu, S)$ entonces $f_T(t|\nu, S) \propto t^{-\left(\frac{\nu}{2}+1\right)} exp\left\{ -\frac{S}{2t} \right\}$

A Priori Conjunta

La expresión final de la distribución a priori conjunta, se obtuvo de la siguiente manera:

$$\begin{aligned} p(\beta_0, \boldsymbol{\beta}, \sigma^2_\beta, \sigma^2, \rho|\Omega) &= p(\beta_0|\sigma^2_{\beta_0}) p(\boldsymbol{\beta}|\sigma^2_\beta) \\ &\times p(\sigma^2_\beta|df_\beta, S_\beta) \\ &\times p(\sigma^2_e|df_e, S_e) \\ &\times p(\rho|a_0, b_0). \end{aligned}$$

Teorema de Bayes

El resultado principal en el cual se basa la inferencia Bayesiana es el teorema de Bayes

$$p(\boldsymbol{\theta}|datos) \propto p(datos|\boldsymbol{\theta}) \times p(\boldsymbol{\theta}),$$

donde $\boldsymbol{\theta}$ representa al vector de parámetros de interés. Aplicando este resultado a nuestros datos, obtenemos:

$$p(\boldsymbol{\theta}|datos) \propto L \times Priori's,$$

donde L representa a la función de verosimilitud aumentada. Por lo tanto obtenemos la siguiente expresión final:

$$\begin{aligned} p(\boldsymbol{\theta}|datos) &= \left[\prod_{i=1}^{n} \frac{2}{\sqrt{1-\rho^2}} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2\pi}} \frac{S_u}{\sigma_e} exp\left\{ -\frac{1}{2(1-\rho^2)\sigma^2_e} S^2_u \left(y_i - \mu_i - \frac{\sigma_e}{S_u}\rho z_i + \frac{\sigma_e}{S_u} E_u \right)^2 \right\} \right. \\ &\left. \times\ exp\left\{ -\frac{1}{2} z_i^2 \right\} I_{(0,\infty)}(z_i) \right] \\ &\times p(\beta_0, \boldsymbol{\beta}, \sigma^2_\beta, \sigma^2, \rho|\Omega). \end{aligned}$$

Distribución Condicional de β_0

De la expresión anterior y con un poco de álgebra obtenemos la expresión para la distribución condicional del parámetro de localidad, dada por:

$$\begin{aligned} p(\beta_0|resto) &= L(\beta_0, \boldsymbol{\beta}, \sigma^2_\beta, \sigma^2_e, \rho, \boldsymbol{z}|datos) \times p(\beta_0) \\ &= \prod_{i=1}^{n} \frac{2}{\sqrt{1-\rho^2}} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2\pi}} \frac{S_u}{\sigma_e} exp\left\{ -\frac{1}{2(1-\rho^2)\sigma^2_e} S^2_u \left(y_i - \mu_i - \frac{\sigma_e}{S_u}\rho z_i + \frac{\sigma_e}{S_u} E_u \right)^2 \right\} \\ &\times\ exp\left\{ -\frac{1}{2} z_i^2 \right\} I_{(0,\infty)}(z_i) \frac{1}{\sqrt{2\pi\sigma^2_{\beta_0}}} exp\left\{ -\frac{1}{2\sigma^2_{\beta_0}} \beta_0^2 \right\} \\ &\propto \prod_{i=1}^{n} exp\left\{ -\frac{1}{2(1-\rho^2)\sigma^2_e} S^2_u \left(y_i - \beta_0 - \mathbf{x}_i^t \boldsymbol{\beta} - \frac{\sigma_e}{S_u}\rho z_i + \frac{\sigma_e}{S_u} E_u \right)^2 \right\} exp\left\{ -\frac{1}{2\sigma^2_{\beta_0}} \beta_0^2 \right\}. \end{aligned}$$

Sea $y_i^* = y_i - \mathbf{x}_i^t\boldsymbol{\beta} - \frac{\sigma_e}{S_u}\rho z_i + \frac{\sigma_e}{S_u}E_u$, entonces la condicional de β_0 dado el resto, se puede escribir de la siguiente manera:

$$\begin{aligned}
p(\beta_0|y_i^*,\beta,\sigma_\beta^2,\sigma_e^2,\rho,z_i) &\propto exp\left\{-\tfrac{1}{2(1-\rho^2)\sigma_e^2}S_u^2\textstyle\sum_{i=1}^n(y_i^*-\beta_0)^2\right\}exp\left\{-\tfrac{1}{\sigma_{\beta_0}^2}\beta_0^2\right\}\\
&= exp\left\{-\tfrac{1}{2(1-\rho^2)\sigma_e^2}S_u^2\left[\textstyle\sum_{i=1}^n(y_i^{*2}-2y_i^*\beta_0+\beta_0^2)+\tfrac{(1-\rho^2)\sigma_e^2}{S_u^2\sigma_{\beta_0}^2}\beta_0^2\right]\right\}\\
&\propto exp\left\{-\tfrac{1}{2(1-\rho^2)\sigma_e^2}S_u^2\left(-2\beta_0\textstyle\sum_{i=1}^n y_i^*+n\beta_0^2+\tfrac{(1-\rho^2)\sigma_e^2}{S_u^2\sigma_{\beta_0}^2}\beta_0^2\right)\right\}\\
&= exp\left\{-\tfrac{1}{2(1-\rho^2)\sigma_e^2}S_u^2\left(n\beta_0^2-2\beta_0\textstyle\sum_{i=1}^n y_i^*+\tfrac{(1-\rho^2)\sigma_e^2}{S_u^2\sigma_{\beta_0}^2}\beta_0^2\right)\right\}\\
&= exp\left\{-\tfrac{1}{2(1-\rho^2)\sigma_e^2}S_u^2\left[\beta_0^2\left(n+\tfrac{(1-\rho^2)\sigma_e^2}{S_u^2\sigma_{\beta_0}^2}\right)-2\beta_0\textstyle\sum_{i=1}^n y_i^*\right]\right\}\\
&\propto exp\left\{-\tfrac{1}{2(1-\rho^2)\sigma_e^2}S_u^2c_1\left[\beta_0^2-2\tfrac{\beta_0}{c_1}\textstyle\sum_{i=1}^n y_i^*\right]\right\}\\
&= exp\left\{-\tfrac{1}{2(1-\rho^2)\sigma_e^2}S_u^2c_1\left[\beta_0-\tfrac{\sum_{i=1}^n y_i^*}{c_1}\right]^2\right\},
\end{aligned}$$

en la última expresión $c_1 = n + \frac{(1-\rho^2)\sigma_e^2}{S_u^2\sigma_{\beta_0}^2}$, y se identifica el núcleo de una distribución normal para la distribución condicional de β_0 :

$$N\left(\beta_0;\frac{\sum_{i=1}y_i^*}{n+\frac{(1-\rho^2)\sigma_e^2}{S_u^2\sigma_{\beta_0}^2}},\frac{(1-\rho^2)\sigma_e^2}{S_u^2\left(n+\frac{(1-\rho^2)\sigma_e^2}{S_u^2\sigma_{\beta_0}^2}\right)}\right).$$

Distribución Condicional de $\beta_j, j = 1,\ldots,p$

La distribución condicional asociada al efecto de cada marcador está dada de la siguiente manera:

$$p(\beta_j|resto) \propto exp\left\{-\tfrac{1}{2(1-\rho^2)\sigma_e^2}S_u^2\left[\textstyle\sum_{i=1}^n(y_i^*-x_{ij}\beta_j)\right]^2\right\}exp\left\{-\tfrac{1}{2\sigma_{\boldsymbol{\beta}}^2}\beta_j^2\right\}$$

Observe que

$$\mathbf{x}_i^t\boldsymbol{\beta} = x_{ij}\beta_j + \mathbf{x}_{i,-j}^t\boldsymbol{\beta}_{-j},$$

y haciendo

$$y_i^* = y_i - \beta_0 - \mathbf{x}_{i,-j}^t\boldsymbol{\beta}_{-j} - \frac{\sigma_e}{S_u}\rho z_i + \frac{\sigma_e}{S_u}E_u,$$

se pueden sustituir las expresiones anteriores, y se obtiene:

$$\begin{aligned} p(\beta_j|\text{resto}) &\propto exp\left\{-\tfrac{1}{2(1-\rho^2)\sigma_e^2}S_u^2\sum_{i=1}^n(y_i^*-x_{ij}\beta_j)^2\right\}exp\left\{-\tfrac{1}{2\sigma_\beta^2}\beta_j^2\right\} \\ &= exp\left\{-\tfrac{1}{2(1-\rho^2)\sigma_e^2}S_u^2\left[-2\beta_j\sum_{i=1}^n x_{ij}y_i^*+\beta_j^2\left(\sum_{i=1}^n x_{ij}^2+\tfrac{(1-\rho^2)\sigma_e^2}{S_u^2\sigma_\beta^2}\right)\right]\right\} \\ &\propto exp\left\{-\tfrac{1}{2(1-\rho^2)\sigma_e^2}S_u^2\left(\sum_{i=1}^n x_{ij}^2+\tfrac{(1-\rho^2)\sigma_e^2}{S_u^2\sigma_\beta^2}\right)\left[\left(\beta_j-\tfrac{\sum_{i=1}^n x_{ij}y_i^*}{\sum_{i=1}^n x_{ij}^2+\frac{(1-\rho^2)\sigma_e^2}{S_u^2\sigma_\beta^2}}\right)^2\right]\right\}. \end{aligned}$$

por tanto la condicional de β_j es:

$$\beta_j|resto \sim N\left(\frac{\sum_{i=1}^n x_{ij}y_i^*}{\sum_{i=1}^n x_{ij}^2+\frac{(1-\rho^2)\sigma_e^2}{S_u^2\sigma_\beta^2}}, \frac{(1-\rho^2)\sigma_e^2}{S_u^2\left[\sum_{i=1}^n x_{ij}^2+\frac{(1-\rho^2)\sigma_e^2}{S_u^2\sigma_\beta^2}\right]}\right).$$

Distribución Condicional de las Variables Latentes

Para construir la verosimilitud aumentada, se utilizó una variable latente, cuya función de distribución condicional posterior está dada por:

Distribución condicional de la variable latente $z_i, i = 1, \ldots, n$.

$$p(z_i|resto) \propto exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{y_i-\beta_0-\mathbf{x}_i^t\boldsymbol{\beta}}{\sigma_e}\right)S_u+E_u-\rho z_i\right]^2\right\}exp\left\{-\frac{1}{2}z_i^2\right\},$$

donde $y_i^* = \left(\frac{y_i-\beta_0-\mathbf{x}_i^t\boldsymbol{\beta}}{\sigma_e}\right)S_u + E_u$.

$$\begin{aligned} p(z_i|resto) &\propto exp\left\{-\tfrac{1}{2(1-\rho^2)}(y_i^*-\rho z_i)^2\right\}exp\left\{-\tfrac{1}{2}z_i^2\right\} \\ &= exp\left\{-\tfrac{1}{2(1-\rho^2)}(-2y_i^*\rho z_i+z_i^2)\right\} \\ &= exp\left\{-\tfrac{1}{2(1-\rho^2)}\left(z_i-\rho y_i^*\right)^2\right\}I_{(0,\infty)}(z_i), i = 1, \ldots, n. \end{aligned}$$

En la última expresión se puede reconocer el núcleo de una distribución Normal-Truncada con parámetro de localidad ρy_i^* y parámetro de escala $1-\rho^2$, lo cual se expresa de la siguiente manera:

$$z_i \sim NT(\mu = \rho y_i^*, \sigma_e^2 = 1-\rho^2, a = 0, b = \infty).$$

Se utilizó la librería *truncnorm* (Trautmann *et al.*, 2012) de `R` para generar muestras de esta distribución.

Distribución Condicional de σ^2_β

Para la varianza asociada al efecto de los marcadores, la distribución condicional obtenida fue:

$$\begin{aligned} p(\sigma^2_\beta|resto) &\propto exp\left\{-\tfrac{1}{2\sigma^2_\beta}\boldsymbol{\beta}^t\boldsymbol{\beta}\right\}\left(\tfrac{1}{\sqrt{2\pi\sigma^2_\beta}}\right)^p(\sigma^2_\beta)^{-(df_\beta/2+1)}exp\left\{-\tfrac{S_\beta}{2\sigma^2_\beta}\right\} \\ &= exp\left\{-\tfrac{1}{2\sigma^2_\beta}[\boldsymbol{\beta}^t\boldsymbol{\beta}+S_\beta]\right\}(\sigma^2_\beta)^{-(\frac{df_\beta+p}{2}+1)} \\ \sigma^2_\beta|resto &\sim \chi^{-2}(df_\beta+p,\boldsymbol{\beta}^t\boldsymbol{\beta}+S_\beta). \end{aligned}$$

Distribución Condicional de σ^2_e

En el caso de la varianza asociada a los errores aleatorios, la distribución condicional que se obtuvo tiene la siguiente representación:

$$\begin{aligned} p(\sigma^2_e|resto) &\propto \textstyle\prod_{i=1}^n \frac{1}{\sigma_e}exp\left\{-\frac{1}{2(1-\rho^2)\sigma^2_e}S^2_u\left(y_i-\mu_i-\frac{\sigma_e}{S_u}\rho z_i+\frac{\sigma_e}{S_u}E_u\right)^2\right\}(\sigma^2_e)^{-(df_e/2+1)}exp\left\{-\tfrac{S_e}{2\sigma^2_e}\right\} \\ &= \left(\tfrac{1}{\sigma^{2/2}_e}\right)^n exp\left\{-\tfrac{1}{2(1-\rho^2)\sigma^2_e}S^2_u\textstyle\sum_{i=1}^n\left(y_i-\beta_0-\mathbf{x}^t_i\boldsymbol{\beta}-\frac{\sigma_e}{S_u}\rho z_i+\frac{\sigma_e}{S_u}E_u\right)^2\right\} \\ &\times (\sigma^2_e)^{-(df_e/2+1)}exp\left\{-\tfrac{S_e}{2\sigma^2_e}\right\} \\ &\propto (\sigma^2_e)^{(-\frac{df_e+n}{2}+1)} \\ &\times exp\left\{-\tfrac{1}{2(1-\rho^2)\sigma^2_e}S^2_u\left[\textstyle\sum_{i=1}^n\left(y_i-\beta_0-\mathbf{x}^t_i\boldsymbol{\beta}-\frac{\sigma_e}{S_u}\rho z_i+\frac{\sigma_e}{S_u}E_u\right)^2+\frac{(1-\rho^2)}{S^2_u}S_e\right]\right\}. \end{aligned}$$

De la última expresión, se observa que la distribución condicional posterior de σ^2_e, es muy compleja y su núcleo no corresponde a ninguna función de densidad univariada conocida.

Se obtuvieron muestras de la distribución condicional utilizando el algoritmo Metropolis y otras técnicas MCMC. Siguiendo la propuesta de (Kim, 2005) se consideró el algoritmo Metropolis con Caminata Aleatoria.

Dado que $\sigma^2_e > 0$, sea $\xi = log(\sigma^2_e)$, ξ tiene soporte en $\mathbb{R}$. Observe que la función de densidad de ξ puede ser obtenida usando el método de transformación (Casella y Berger, 2002) y está dada por:

$$p(\xi|resto) \propto p(\sigma^2_e|resto)exp(\xi).$$

En el algoritmo Metropolis con Caminata Aleatoria, se generó ξ eligiendo un núcleo de transición propuesto para añadir ruido al estado actual. Suponiendo que el valor real de ξ es ξ_k se desea actualizar su valor en la siguiente iteración a ξ_{k+1}, entonces se pueden realizar los pasos a-c

mostrados abajo para este fin, es decir:

a. Muestrear ξ, $\xi = \xi_k + \psi_1$ donde $\psi_1 \sim N(0, \eta^2)$

b. Muestrear u, $U \sim U(0,1)$

c. Si $u < \dfrac{p(\xi|resto)}{p(\xi_k|resto)}$) entonces $\xi_{k+1} = \xi$, de otro modo $\xi_{k+1} = \xi_k$.

Una vez que se ha obtenido ξ_{k+1}, calcular $\sigma^2_{e,k+1} = exp(\xi_{k+1})$. El parámetro η^2 se puede modificar para obtener una tasa de aceptación óptima.

Distribución Condicional de ρ

Finalmente, para el parámetro asociado a la asimetría de la distribución de los fenotipos y, la distribución condicional asociada es de tipo Beta y tiene la siguiente expresión:

Sea $R = 1 - 2B$ y $B \sim Beta(a_0, b_0)$; entonces

$$f_R(r) = \left(\frac{1-r}{2}\right)^{a_0-1}\left(1 - \frac{1-r}{2}\right)^{b_0-1}$$

$$\begin{aligned} p(\rho|resto) &\propto \prod_{i=1}^n \frac{1}{\sqrt{1-\rho^2}} S_u exp\left\{-\frac{1}{2(1-\rho^2)\sigma_e^2} S_u^2 \left[\left(y_i - \beta_0 - \mathbf{x}_i^t\boldsymbol{\beta} - \frac{\sigma_e}{S_u}\rho z_i + \frac{\sigma_e}{S_u}E_u\right)^2\right]\right\} \\ &\times \left(\frac{1-\rho}{2}\right)^{a_0-1}\left(1 - \frac{1-\rho}{2}\right)^{b_0-1} \\ &= exp\left\{-\frac{1}{2(1-\rho^2)\sigma_e^2} S_u^2 \left[\left(y_i - \beta_0 - \mathbf{x}_i^t\boldsymbol{\beta} - \frac{\sigma_e}{S_u}\rho z_i + \frac{\sigma_e}{S_u}E_u\right)^2\right]\right\} \\ &\times \left(\frac{1-\rho}{2}\right)^{a_0-1}\left(1 - \frac{1-\rho}{2}\right)^{b_0-1}\left(\frac{S_u}{\sqrt{1-\rho^2}}\right)^n I_{(-1,1)}(\rho) \end{aligned}$$

Observe que $p(\rho|resto)$ es una función compleja de ρ y su núcleo no se identifica con ninguna función de densidad univariada conocida, por lo que, para obtener muestras de la función se debe utilizar el algoritmo Metropolis u otra técnica MCMC.

Aquí se propone utilizar la transformación de Fisher (1915) de ρ definida como $\vartheta = \dfrac{1}{2} log\left(\dfrac{1+\rho}{1-\rho}\right) = tanh^{-1}(\rho)$, el soporte de ϑ está en $\mathbb{R}$.

La función de densidad de ϑ se puede obtener usando el método de transformación (Casella y Berger, 2002) y está dada por:

$$p(\vartheta|resto) \propto p(\rho|resto) \times tanh(\vartheta).$$

En el algoritmo Metropolis con Caminata Aleatoria, se generó ϑ eligiendo un núcleo de transición propuesto para añadir ruido al estado actual. Suponiendo que el valor real de ϑ es ϑ_k se desea actualizar su valor en la siguiente iteración a ϑ_{k+1}, entonces se pueden realizar los pasos a-c mostrados abajo para este fin, es decir:

a. Muestrear ϑ, $\vartheta = \vartheta_k + \psi_2$ donde $\psi_2 \sim N(0, \nu^2)$

b. Muestrear u, $U \sim U(0,1)$

c. Si $u < \dfrac{p(\vartheta|resto)}{p(\vartheta_k|resto)}$) entonces $\vartheta_{k+1} = \vartheta$, de otro modo $\vartheta_{k+1} = \vartheta_k$.

Una vez que se ha obtenido ϑ_{k+1}, calcular $\rho = tanh(\vartheta_{k+1})$. El parámetro ν^2 se puede modificar para obtener una tasa de aceptación óptima.

Las muestras para ambos parámetros, σ_e^2 y ρ, de la distribución posterior se obtuvieron utilizando el Muestreador de Gibbs (Geman y Geman, 1984) y Metropolis con Caminata Aleatoria. En el algoritmo, se tomaron muestras de cada una de las distribuciones condicionales completas hasta obtener muestras del tamaño deseado. Se implementó el algoritmo en el paquete estadístico `R` (R Core Team, 2017). Para acelerar los cálculos, las rutinas donde se tomaron las muestras de $p(\beta_j|resto), j = 1, \ldots, p$ se implementaron en lenguaje de programación C y se generó una biblioteca compartida, posteriormente las rutinas compiladas fueron utilizadas en `R`.

Fijando los hiper-parámetros para las distribuciones a priori

Se presenta una descripción de la forma en la cual los valores de los hyper-parámetros son ajustados, cuando éstos no son proporcionados por el usuario, las reglas son similares a las utilizadas en el software BGLR (de los Campos y Pérez-Rodríguez, 2015). Con estas reglas se asignan distribuciones a priori adecuadas pero débilmente informativas, de tal manera que se particiona la varianza total de los fenotipos en dos componentes i) el error y ii) el predictor lineal, es decir:

$$\begin{aligned} Var(y_i) &= Var\left(\sum_{j=1}^{p} x_{ij}\beta_j + e_i\right) \\ &= Var\left(\sum_{j=1}^{p} x_{ij}\beta_j\right) + Var(e_i) \\ &= \sum_{j=1}^{p} x_{ij}^2 Var(\beta_j) + Var(e_i) \\ &= \sigma_\beta^2 \sum_{j=1}^{p} x_{ij}^2 + \sigma_e^2 \\ &= Var(g_i) + Var(e_i), \end{aligned} \tag{B.1}$$

donde $g_i = \sum_{j=1}^{p} x_{ij}\beta_j$. A priori, la varianza total genética es $\sum_{i=1}^{n} Var(g_i) = \sigma_\beta^2 \sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}^2$ y a priori la varianza genética promedio es:

$$\begin{aligned} V_g &= \frac{\sigma_\beta^2}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}^2 \\ &= \sigma_\beta^2 MS_x, \end{aligned} \tag{B.2}$$

donde $MS_x = \dfrac{1}{n}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{p} x_{ij}^2$. De (B.1) y (B.2) la partición de la varianza fenotípica está dada por:

$$\begin{aligned} V_y &= V_g + V_e \\ &= \sigma_\beta^2 MS_x + \sigma_e^2, \end{aligned}$$

donde $V_e = Var(e_i) = \sigma_e^2$.

Fijando los hiper-parámetros para $\chi^{-2}(\sigma_\beta^2|S_\beta, df_\beta)$.

La distribución a priori asignada al parámetro de localización $\boldsymbol{\beta}$, es una distribución Normal con media y varianza σ_β^2. En el segundo nivel de la jerarquía del modelo, a σ_β^2, se le asigna una densidad Chi-cuadrada invertida-escalada con parámetros df_β y S_β, $\chi^{-2}(\sigma_\beta^2|S_\beta, df_\beta)$. La parametrización de esta densidad, es tal que el valor esperado a priori y la moda son:

$$\mathbb{E}(\sigma_\beta^2|S_\beta, df_\beta) = \frac{S_\beta}{df_\beta - 2} \quad \text{y} \quad \text{moda}(\sigma_\beta^2|S_\beta, df_\beta) = \frac{S_\beta}{df_\beta + 2}.$$

Definimos $df_\beta = 5$ para que la distribución a prior tenga una media finita. De la ecuación (B.2,)

$$\sigma_\beta^2 = \frac{V_g}{MS_x}, \tag{B.3}$$

reemplazando el lado derecho de (B.3) con la moda de $\sigma_\beta^2|S_\beta, df_\beta$ entonces:

$$\frac{S_\beta}{df_\beta + 2} = \frac{V_g}{MS_x}. \tag{B.4}$$

De (B.4) se resuelve para $S_\beta = \frac{V_g \times (df_\beta + 2)}{MS_x}$. De la definición de heredabilidad

$$h^2 = \frac{V_g}{V_y} \Longrightarrow V_g = h^2 V_y,$$

entonces:

$$S_\beta = \frac{h^2 V_y \times (df_\beta + 2)}{MS_x} = \frac{R^2 V_y \times (df_\beta + 2)}{MS_x}$$

Una vez que se establece df_β se puede fijar S_β y sólo se necesita calcular la varianza fenotípica $(V_y), MS_x$ y fijar R^2 como la proporción de la varianza a priori que es explicada por los

marcadores, por defecto fijamos $R^2 = 0.5$.

Fijando los hiper-parámetros para $\chi^{-2}(\sigma_e^2|S_e, df_e)$.

En el caso de la varianza residual, σ_e^2,

$$\mathbb{E}(\sigma_e^2|S_e, df_e) = \frac{S_e}{df_e - 2} \;\; y \;\; \text{moda}(\sigma_e^2|S_e, df_e) = \frac{S_e}{df_e + 2},$$

fijamos $df_e = 5$ y $S_e = (1 - R^2)V_y \times (df_e + 2)$.

Fijando los hiper-parámetros para $N(\beta_0|0, \sigma_{\beta_0}^2)$.

Fijamos $\sigma_{\beta_0}^2 = 1 \times 10^6$ de tal modo que la a priori asignada al intercepto es efectivamente plana.

Fijando los hiper-parámetros para $p(\rho|a_0, b_0)$.

La función de densidad de la a priori asignada a ρ es:

$$p(\rho|a_0, b_0) = \frac{1}{2Beta(a_0, b_0)} \left(\frac{1-\rho}{2}\right)^{a_0-1} \left(1 - \frac{1-\rho}{2}\right)^{b_0-1} I_{(-1,1)}(\rho), a_0 > 0, b_0 > 0$$

La Figura B.1 muestra la función de densidad de ρ para diferentes valores de los parámetros a_0, b_0. Fijando diferentes valores para esos hiper-parámetros, se pueden obtener diversas formas, como en el caso de la distribución Beta. Si fijamos $a_0 = b_0 = 1$ entonces obtenemos la distribución uniforme, que es la a priori que se utilizó en este trabajo.

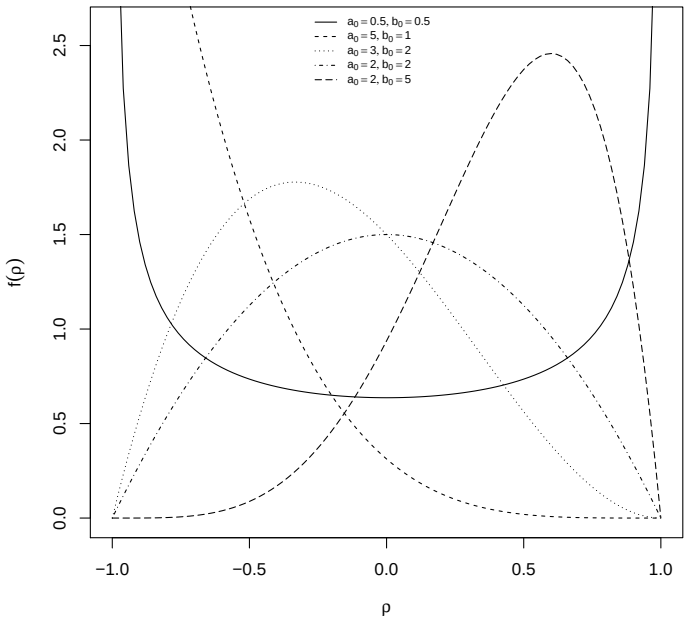

Figura B.1: Función de densidad de una variable tipo Beta con soporte en el intervalo $(-1, 1)$ para diferentes valores de los parámetros a_0, b_0.

Código en C

El siguiente código de programación es utilizado para hacer el ajuste de un modelo de regresión cuyos errores se distribuyen de acuerdo a una Distribución Normal Asimétrica, $y_i = \beta_0 + \mathbf{x}^t\boldsymbol{\beta} + e_i$, $e \sim SN_C(\beta_0, \sigma_e^2, \gamma_1)$. Se presenta el código en `C` que se utilizó para muestrear los parámetros $\beta_j|resto$ del modelo ajustado en el capítulo 4.

```
1
2  /*
3
4  File:  util_sample_sn.c
5
6  Compilation in Windows:
7
8  R CMD SHLIB util_sample_sn.c -lRlapack -lRblas
9
10 Compilation in UNIX (macOS, Linux):
11
12 R CMD SHLIB util_sample_sn.c
13
14 */
15
16
17 #include <R.h>
```

```
18 #include <Rmath.h>
19 #include <Rinternals.h>
20 #include <Rdefines.h>
21 #include <Rconfig.h>
22 #include <R_ext/Lapack.h>
23
24
25 SEXP sample_beta_c(SEXP n, SEXP p, SEXP X, SEXP x2, SEXP sx, SEXP beta, SEXP e, SEXP
       sigma, SEXP z, SEXP rho, SEXP sigma2b, SEXP EU, SEXP SU)
26 {
27     double *xj;
28     double *pX, *px2, *psx, *pb, *pe, *pz;
29     double b, s, r, s2b, eu, su;
30     double m, v, c1, c2;
31     int inc=1;
32     int j, rows, cols;
33
34     SEXP list;
35
36     GetRNGstate();
37
38     rows=INTEGER_VALUE(n);
39     cols=INTEGER_VALUE(p);
40     s=NUMERIC_VALUE(sigma);
41     r=NUMERIC_VALUE(rho);
42     s2b=NUMERIC_VALUE(sigma2b);
43     eu=NUMERIC_VALUE(EU);
44     su=NUMERIC_VALUE(SU);
45
46     PROTECT(X=AS_NUMERIC(X));
47     pX=NUMERIC_POINTER(X);
48
49     PROTECT(x2=AS_NUMERIC(x2));
50     px2=NUMERIC_POINTER(x2);
51
52     PROTECT(sx=AS_NUMERIC(sx));
53     psx=NUMERIC_POINTER(sx);
54
55     PROTECT(beta=AS_NUMERIC(beta));
56     pb=NUMERIC_POINTER(beta);
57
58     PROTECT(e=AS_NUMERIC(e));
59     pe=NUMERIC_POINTER(e);
60
61     PROTECT(z=AS_NUMERIC(z));
62     pz=NUMERIC_POINTER(z);
63
64     for(j=0; j<cols;j++)
65     {
66         xj=pX+j*rows;
67         b=pb[j];
68
69         c2=F77_NAME(ddot)(&rows,xj,&inc,pe,&inc);
70         c2+=px2[j]*b;
```

```
71          c2-=s*r/su*F77_NAME(ddot)(&rows,xj,&inc,pz,&inc);
72          c2+=s*eu/su*psx[j];
73
74          c1=px2[j]+(1-r*r)*s*s/(su*su*s2b);
75
76          m=c2/c1;
77          v=(1-r*r)*s*s/(su*su*c1);
78
79        pb[j]= m + sqrt(v)*norm_rand();
80
81          b-=pb[j];
82
83          F77_NAME(daxpy)(&rows, &b,xj,&inc, pe,&inc);
84      }
85
86      // Creating a list with 2 vector elements:
87      PROTECT(list = allocVector(VECSXP, 2));
88      // attaching b vector to list:
89      SET_VECTOR_ELT(list, 0, beta);
90      // attaching e vector to list:
91      SET_VECTOR_ELT(list, 1, e);
92
93      PutRNGstate();
94
95      UNPROTECT(7);
96
97      return(list);
98  }
```

Código en R

Para poder usar el código C en R que se creó para realizar las estimaciones de los parámetros β_j, se debe crear una biblioteca compartida (`.dll` en Windows o `.so` en Unix), para más detalles ver el manual *Writing R Extensions* incluido en los manuales de `R`. Esto se puede hacer usando la rutina `R CMD SHLIB` incluida en el paquete `R` desde la línea de comandos del sistema operativo. Esto dará como resultado un archivo llamado ‘‘`util_sample_sn.dll`’’ o ‘‘`util_sample_sn.so`’’. Una vez creada la biblioteca, las funciones se pueden usar combinando las rutinas `"dyn.load"` y `.Call`

El siguiente código en `R` fue usado para realizar las estimaciones de todos los parámetros a estimar en el modelo Normal-Asimétrico, del capítulo (4).

```
1
2  ##################################################
3  #Load requiered libraries
4  ##################################################
5  library(truncnorm)
6  library(sn)
7  library(MASS)
8  library(BGLR)
```

```
9
10
11 ##################################################################
12 ###Auxiliary functions
13 ##################################################################
14
15 log_L=function(y,mu,sigma,rho)
16 {
17   EU=sqrt(2/pi)*rho
18   VU=1-2/pi*rho^2
19   SU=sqrt(VU)
20   ystar=y-mu+sigma*EU/SU
21   c=log(2)-1/2*log(2*pi)+log(SU)-log(sigma)
22   n=length(y)
23   tmp1=0.5*(log(pi-2*rho^2)-log(pi*(1-rho^2)))
24   tmp2=exp(tmp1)
25   n*c-1/2*VU/sigma^2*sum(ystar^2)+sum(pnorm(q=rho*tmp2*ystar/sigma,mean = 0, sd = 1,
       lower.tail = TRUE, log.p = TRUE))
26 }
27
28 rsn=function(n,rho)
29 {
30   m=c(0,0)
31   V=matrix(c(1,rho,rho,1),nrow=2,ncol=2)
32   d=mvrnorm(n=n,mu=m,Sigma=V)
33   ifelse(d[,1]>0,d[,2],-d[,2])
34 }
35
36 rho_to_lambda=function(rho)
37 {
38   rho/sqrt(1-rho^2)
39 }
40
41 lambda_to_rho=function(lambda)
42 {
43   lambda/sqrt(1+lambda^2)
44 }
45
46 moda=function(z)
47 {
48   dens=density(z)
49   dens$x[dens$y == max(dens$y)] #da la moda
50 }
51
52 #eta1=log(sigma^2)
53 log_eta1=function(eta1,n,df,s,z,error,SU,EU,rho)
54 {
55   sigma2=exp(eta1)
56   sigma=sqrt(sigma2)
57   delta=-sigma*rho*z/SU+sigma*EU/SU
58   tmp1=-(df/2+n/2+1)*log(sigma2)
59   tmp2=-SU^2/(2*(1-rho^2)*sigma2)*sum((error+delta)^2)-s/(2*sigma2)
60   tmp3=log(sigma2)
61   ans=tmp1+tmp2+tmp3
```

```
  return(-ans)
}

log_eta2=function(eta2,n,sigma,error,z)
{
  sigma2=sigma^2
  rho=(exp(eta2)-1)/(1+exp(eta2))
  EU=sqrt(2/pi)*rho
  VU=1-2/pi*rho^2
  SU=sqrt(VU)
  delta=-sigma*rho*z/SU+sigma*EU/SU
  tmp1=-VU/(2*(1-rho^2)*sigma2)*sum((error+delta)^2)
  tmp2=-n/2*log(1-rho^2)
  tmp3=n*log(SU)
  tmp4=eta2-2*log(1+exp(eta2))
  ans=tmp1+tmp2+tmp3+tmp4
  return(-ans)
}

########################################################
#### End of auxiliary functions
########################################################


#NAs not allowed
fit_sn_centered_no_miss=function(y,X,B=5000,burnIn=2500, thin=10,
                                 R2=0.5,dfb=5,dfsigma2=5,
                                 v1=0.04,v2=0.04)
{

  #Fit the model y=beta0+e, where e~ SN(0,sigma,gamma_1)
  fit_sn=selm(y~1)
  cp=as.vector(fit_sn@param$cp)
  dp=as.vector(fit_sn@param$dp)

  beta0=cp[1]
  sigma=cp[2]
  lambda=dp[3]
  rho=lambda_to_rho(lambda)

  EU=sqrt(2/pi)*rho
  VU=1-2/pi*rho^2
  SU=sqrt(VU)

  p=ncol(X)
  n=length(y)

  x2=colSums(X*X)
  sx=colSums(X)
  sumMeanXSq = sum((apply(X,2L,mean))^2)
  MSx=sum(x2)/n-sumMeanXSq
  Vy=sigma^2
  sb=(Vy*R2)/(MSx)*(dfb-2)
  cat("sb=",sb,"\n")
```

```
116
117  ssigma2 = Vy*(1 - R2)*(dfsigma2+2)
118  cat("ssigma2=",ssigma2,"\n")
119
120  beta=rnorm(n=ncol(X),0,sd=0.1)
121  z=abs(rnorm(n))
122
123  sigma2beta=var(y)/4/sum(x2/n)
124
125  error=as.vector(y-X%*%beta)
126  error=error-beta0
127
128  #Objects to store the samples
129  Betas=matrix(NA,nrow=B,ncol=ncol(X))
130  Sigma=rep(NA,B)
131  Rho=rep(NA,B)
132  Beta0=rep(NA,B)
133  Sigma2beta=rep(NA,B)
134
135
136  accept_rate_sigma2=0
137  accept_rate_rho=0
138
139  Xv=as.vector(X)
140
141  nSums=0
142  post_logLik=0
143
144
145
146 ##########################################################
147 ###Gibbs and Metropolis
148 ##########################################################
149
150 #y: response variable
151 #X: incidence matrix
152 #nIter: number of iterations
153 #burnIn: number of samples discarded
154 #dfb: degrees of freedom to the prior assigned to sigma2beta
155 #sb: scale parameter to the prior assigned to sigma2beta
156 #dfsigma2: degrees of freedom for the prior assigned to Vy
157 #ssigma2: scale parameter to the prior assigned to Vy
158 #a,b: parameters to prior assigned to rho
159 #R2: R-squared of the model
160
161 #B=1000
162 #burnIn=B/2
163
164 #dfb=5
165 #R2=0.5
166 #dfsigma2=5
167
168
169
```

```
for(iter in 1:B)
{
 cat("iter=",iter,"beta0=",round(beta0,4),"sigma=",round(sigma,4),"rho=",round(rho,4),"\n")

 ##########################################
 #Sample from beta0
 ##########################################

 delta=-sigma*rho*z/SU+sigma*EU/SU

 yc=error+beta0+delta
 c1=n+(1-rho^2)*sigma^2/(SU^2*1000)
 c2=sum(yc)
 m=c2/c1
 v=(1-rho^2)*sigma^2/(SU^2*c1)

 beta0_old=beta0
 beta0_new=rnorm(1,m,sqrt(v))
 beta0=beta0_new

 #Update the error
 error=error+(beta0_old-beta0_new)
 Beta0[iter]=beta0

 ##########################################
 #Sample from beta1,...,betap
 ##########################################

 out=.Call("sample_beta_c",n, p, Xv, x2, sx, beta, error, sigma, z,
           rho, sigma2beta, EU, SU)
 beta=out[[1]]
 error=out[[2]]

 Betas[iter,]=beta

 #for(j in 1:p)
 #{
 # yc=error+X[,j]*beta[j]+delta
 # c1=x2[j]+(1-rho^2)*sigma^2/(SU^2*sigma2beta)
 # c2=sum(yc*X[,j])
 # m=c2/c1
 # v=(1-rho^2)*sigma^2/(SU^2*c1)
 #
 # bold=beta[j]
 # bnew=rnorm(1,m,sqrt(v))
 # beta[j]=bnew
 #
 # #Update the error
 # error=error+(bold-bnew)*X[,j]
 #}
 #Betas[iter,]=beta

 ##########################################
 #Sample from z
```

```
224  ########################################
225
226  z=rtruncnorm(n=n,a=0, b=Inf, mean = rho*((error/sigma)*SU+EU), sd = sqrt(1-rho^2))
227
228  ########################################
229  #Sample from sigma^2_\beta
230  ########################################
231
232  escala=as.numeric(crossprod(beta)) + sb
233  sigma2beta=escala/rchisq(1,df=dfb+p)
234  Sigma2beta[iter]=sigma2beta
235
236
237  ########################################
238  #Sample from sigma
239  ########################################
240
241
242  eta1=log(sigma^2)
243  eta1p=eta1+rnorm(n=1,mean=0,sd=sqrt(v1))
244
245  U_eta1=log_eta1(eta1,n,dfsigma2,ssigma2,z,error,SU,EU,rho)
246  U_eta1p=log_eta1(eta1p,n,dfsigma2,ssigma2,z,error,SU,EU,rho)
247
248  accept = (U_eta1-U_eta1p) > log(runif(1))
249
250  if(accept)
251  {
252    sigma2=exp(eta1p)
253    sigma=sqrt(sigma2)
254    accept_rate_sigma2=accept_rate_sigma2+1
255  }
256
257  Sigma[iter]=sigma
258
259  ########################################
260   #Sample from rho
261   ########################################
262
263  eta2=log((1+rho)/(1-rho))
264  eta2p=eta2+rnorm(n=1,mean=0,sd=sqrt(v2))
265
266  U_eta2=log_eta2(eta2,n,sigma,error,z)
267  U_eta2p=log_eta2(eta2p,n,sigma,error,z)
268
269  accept = (U_eta2-U_eta2p) > log(runif(1))
270
271  if (accept) {
272    rho=(exp(eta2p)-1)/(1+exp(eta2p))
273    accept_rate_rho=accept_rate_rho+1;
274  }
275
276  EU=sqrt(2/pi)*rho
277  VU=1-2/pi*rho^2
```

```
278     SU=sqrt(VU)
279     Rho[iter]=rho
280
281
282     if(iter>burnIn & iter %%thin == 0)
283     {
284       #mu=y-error
285       nSums=nSums+1
286       k=(nSums-1)/nSums
287       logLik=log_L(y,y-error,sigma,rho)
288       post_logLik = post_logLik * k + logLik/nSums
289     }
290
291   }
292
293   #Burn-in and thin
294
295   Beta0=Beta0[(burnIn+1):B]
296   index=seq(1,length(Beta0),by=thin)
297   Beta0=Beta0[index]
298   Betas=Betas[(burnIn+1):B,]
299   Betas=Betas[index,]
300   Sigma=Sigma[(burnIn+1):B]
301   Sigma=Sigma[index]
302   Rho=Rho[(burnIn+1):B]
303   Rho=Rho[index]
304   Sigma2beta=Sigma2beta[(burnIn+1):B]
305   Sigma2beta=Sigma2beta[index]
306   rm(index)
307
308   #LogLikAtPostMean
309
310   logLikAtPostMean=log_L(y,mean(Beta0)+as.vector(X %* %colMeans(Betas)),mean(Sigma),
        mean(Rho))
311
312   pD = -2 * (post_logLik - logLikAtPostMean)
313   DIC = pD - 2 * post_logLik
314
315   fit=list()
316   fit$logLikAtPostMean=logLikAtPostMean
317   fit$postMeanLogLik=post_logLik
318   fit$pD=pD
319   fit$DIC=DIC
320
321   return(list(beta0=Beta0,betas=Betas,sigma=Sigma,rho=Rho,
322               sigma2beta=Sigma2beta,
323               accept_rate_rho=accept_rate_rho/B, accept_rate_sigma=accept_rate_sigma2/B,
324               fit=fit))
325 }
326
327
328
329
330 ##########################################
```

```
331 ##########################################
332
333 ##########################################
334 #Function for missing data
335 ##########################################
336
337 #NAs allowed
338 #loss can be ''mse'' or ''0-1''
339
340 fit_sn_centered=function(y,X,B=5000,burnIn=2500, thin=10,
341                          R2=0.5,dfb=5,dfsigma2=5,
342                          v1=0.04,v2=0.04,loss="mse")
343 {
344   isNA=any(is.na(y))
345
346   if(isNA)
347   {
348     cat("Missing values FOUND\n");
349
350     index=is.na(y)
351
352     #Training
353     ytrn=y[!index]
354     Xtrn=X[!index,]
355
356     fit=fit_sn_centered_no_miss(y=ytrn,X=Xtrn,B=B,burnIn=burnIn, thin=thin,
357                                 R2=R2,dfb=dfb,dfsigma2=dfsigma2,
358                                 v1=v1,v2=v2)
359
360   }else{
361     cat("Missing values NOT FOUND\n")
362     fit=fit_sn_centered_no_miss(y=y,X=X,B=B,burnIn=burnIn, thin=thin,
363                                 R2=R2,dfb=dfb,dfsigma2=dfsigma2,
364                                 v1=v1,v2=v2)
365   }
366
367   if(loss=="mse")
368   {
369     beta0=mean(fit$beta0)
370     betas=colMeans(fit$betas)
371
372   }else{
373     beta0=moda(fit$beta0)
374     betas=apply(fit$betas,2,moda)
375   }
376
377   yHat=beta0+as.vector(X %*% betas)
378   fit$yHat=yHat
379   return(fit)
380 }
381
382 ##########################################################
383 ##########################################################
384
```

```
385
386 Example 1, wheat dataset
387
388 if(FALSE)  ##### Cambiar (FALSE) a (TRUE) para poder correr el ejemplo  #####
389 {
390   set.seed(123)
391
392  dyn.load("util_sample_sn.dll")
393
394   library(BGLR)
395   data(wheat)
396   X=wheat.X
397   y=wheat.Y[,1]
398   sets=wheat.sets
399   yNa=y
400   whichNa=(sets==2)
401   yNa[whichNa]=NA
402   fm=fit_sn_centered(yNa,X,B=10000,burnIn=5000,thin=1)
403
404   plot(y,fm$yHat,xlab="Phenotype",
405        ylab="Pred. Gen. Value" ,cex=.8,bty="L")
406   points(x=y[whichNa],y=fm$yHat[whichNa],col=2,cex=.8,pch=19)
407   legend("topleft", legend=c("training","testing"),bty="n",
408          pch=c(1,19),col=c("black","red"))
409
410     #log_L(y,fm$yHat,mean(fm$sigma),mean(fm$rho))
411
412     fm2=BGLR(y=yNa,ETA=list(list(X=X,model="BRR")),nIter=10000,burnIn=5000,thin=10)
413   unlink("*.dat")
414
415   dyn.unload("util_sample_sn.dll")
416 }
417
418 #Example 2, GLS data
419 if(FALSE)  ##### Cambiar (FALSE) a (TRUE) para poder correr el ejemplo #####
420 {
421   dyn.load("util_sample_sn.dll")
422
423   load("gls_4_1k.RData")
424   testing=sample(1:length(y),size=25,replace=FALSE)
425   training=setdiff(1:length(y),testing)
426
427   yNa=y
428   yNa[testing]=NA
429
430   fm=fit_sn_centered(yNa,X,B=10000,burnIn=5000,thin=10)
431
432   par(mfrow=c(1,2))
433   plot(fm$beta0,xlab="Iteration",ylab=expression(beta[0]),type="l")
434   hist(fm$beta0,xlab=expression(beta0),main="")
435
436   par(mfrow=c(1,2))
437   plot(fm$sigma,xlab="Iteration",ylab=expression(sigma),type="l")
438   hist(fm$sigma,xlab=expression(sigma),main="")
```

```
439
440   par(mfrow=c(1,2))
441   plot(fm$rho,xlab="Iteration",ylab=expression(rho),type="l")
442   hist(fm$rho,xlab=expression(rho),main="")
443
444   plot(fm$yHat,y)
445
446   fm2=BGLR(y=yNa,ETA=list(list(X=X,model="BRR")),nIter=10000,burnIn=5000,thin=10)
447   unlink("*.dat")
448
449   dyn.unload("util_sample_sn.dll")
450 }
451
452 #Example 3, simulated data using wheat data
453 if(FALSE)   ##### Cambiar (FALSE) a (TRUE) para poder correr el ejemplo #####
454 {
455   dyn.load("util_sample_sn.dll")
456   library(BGLR)
457   data(wheat)
458   X=wheat.X
459
460   X=scale(X,center=TRUE,scale=TRUE)
461   n=nrow(X)
462   p=ncol(X)
463   nQTL=10
464   h2=0.5
465   whichQTL=seq(from=floor(p/nQTL/2),by=floor(p/nQTL),length=nQTL)
466   b0=rep(0,p)
467   b0[whichQTL]=rnorm(n=nQTL,sd=sqrt(h2/nQTL))
468   signal=as.vector(X %* %b0)
469
470   rho=0.99
471   s=1.5
472   loc=3
473
474   EU=sqrt(2/pi)*rho
475   VU=1-2/pi*rho^2
476   SU=sqrt(VU)
477
478   e=rsn(n,rho=rho)
479   e=(e-EU)/SU
480
481   y=loc+signal+s*e
482
483   fm=fit_sn_centered(y,X,B=10000,burnIn=5000,thin=10)
484
485   fm2=BGLR(y=y,ETA=list(list(X=X,model="BRR")),nIter=10000,burnIn=5000,thin=10)
486   unlink("*.dat")
487
488 dyn.unload("util_sample_sn.dll")
489 }
```