



COLEGIO DE POSTGRADUADOS

INSTITUCIÓN DE ENSEÑANZA E INVESTIGACIÓN
EN CIENCIAS AGRÍCOLAS

CAMPUS MONTECILLO

SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA
ESTADÍSTICA

**Selección genómica basada en la teoría de la
decisión**

Bartolo de Jesús Villar Hernández

T E S I S

PRESENTADA COMO REQUISITO PARCIAL PARA
OBTENER EL GRADO DE:

DOCTOR EN CIENCIAS

MONTECILLO, TEXCOCO, EDO. DE MÉXICO
2018

CARTA DE CONSENTIMIENTO DE USO DE LOS DERECHOS DE AUTOR Y DE LAS REGALÍAS COMERCIALES DE PRODUCTOS DE INVESTIGACIÓN

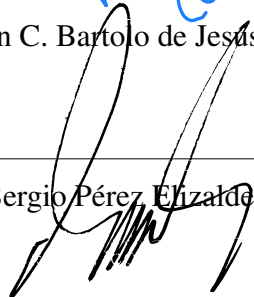
En adición al beneficio ético, moral y académico que he obtenido durante mis estudios en el Colegio de Postgraduados, el que suscribe, **Bartolo de Jesús Villar Hernández**, Alumno de esta Institución, estoy de acuerdo en ser partícipe de las regalías económicas y/o académicas, de procedencia nacional e internacional, que se deriven del trabajo de investigación que realicé en esta institución, bajo la dirección del Profesor **Sergio Pérez Elizalde**, por lo que otorgo los derechos de autor de mi tesis **Selección genómica basada en la teoría de la decisión**, y de los productos de dicha investigación al Colegio de Postgraduados. Las patentes y secretos industriales que se puedan derivar serán registrados a nombre del 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.

ALUMNO



M. en C. Bartolo de Jesús Villar Hernández

VoBo. CONSEJERO



Dr. Sergio Pérez Elizalde

Montecillo, Texcoco, México, Julio de 2018.

La presente tesis titulada: **Selección genómica basada en la teoría de la decisión**, realizada por el alumno: **Bartolo de Jesús Villar Hernández**, bajo la dirección del Consejo Particular indicado ha sido aprobada por el mismo y aceptada como requisito parcial para obtener el grado de:

DOCTOR EN CIENCIAS

**SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA
ESTADÍSTICA**

CONSEJO PARTICULAR

CONSEJERO



Dr. Sergio Pérez Elizalde

DIRECTOR DE TESIS



Dr. José Luis Francisco Crossa Hiriart

ASESOR



Dr. Paulino Pérez Rodríguez

ASESOR



Dr. José Andrés Christen Gracia

ASESOR



Dra. Martha Elva Ramírez Guzmán

Montecillo, Texcoco, México, Julio de 2018

Selección genómica basada en la teoría de la decisión.

Bartolo de Jesús Villar Hernández, Dr.

Colegio de Postgraduados, 2018.

RESUMEN

Los mejoradores de plantas y animales se interesan en seleccionar a los mejores individuos de un conjunto de candidatos para los siguientes ciclos de mejora. En esta investigación se propone abordar la selección genómica (SG) como un problema de decisión desde la perspectiva Bayesiana de la estadística. Se proponen tres funciones de pérdida univariadas (Kullback-Leibler, KL; Continuous Ranked Probability Score, CRPS; y Lineal-Lineal, LinLin), así como sus correspondientes generalizaciones multivariadas (Kullback-Leibler multivariada, KL; Energy Score, EnergyS; y la Función de Pérdida Asimétrica Multivariada, MALF). Todas las funciones de pérdida se expresan en términos de la heredabilidad y se midieron sus desempeños en un conjunto de datos reales para un ciclo de selección y en un estudio de simulación de un programa de mejora. La ganancia obtenida con cada función de pérdida se comparó con la de la forma estándar de selección que no emplea funciones de pérdida. El contraste se realizó en términos de la respuesta a la selección así como de la reducción de la varianza genética en cada ciclo de selección. Los resultados obtenidos en la selección de un solo rasgo sugieren que es posible obtener mayor progreso genético a medida que el programa de mejora transcurre en el tiempo para una presión de selección del 30 % de individuos, pero no cuando dicha presión fue del 10 %. En el contexto de la selección multirasgo, los resultados mostraron ganancias en las medias poblacionales en todos los rasgos bajo selección, aún en presencia de rasgos negativamente correlacionados. Así mismo, las varianzas poblacionales al fin del programa de selección no fueron menores a las que se obtuvieron por medio del método estándar (índices de selección). El uso de funciones de pérdida puede ser un criterio útil cuando se seleccionan los mejores individuos para el siguiente ciclo de mejora.

Palabras clave y frases : Teoría de la decisión, Selección Genómica, Función de Pérdida.

Genomic selection based on decision theory

Bartolo de Jesús Villar Hernández, Dr.

Colegio de Postgraduados, 2018.

ABSTRACT

Plant and animal breeders are interested in selecting the best individuals from a candidate set for the next breeding cycle. In this research, we propose a formal method under the Bayesian decision theory framework to tackle the selection problem based on genomic selection (GS) in single- and multi-trait settings. We proposed and tested three univariate loss functions (Kullback-Leibler, KL; Continuous Ranked Probability Score, CRPS; Linear-Linear loss, LinLin) and their corresponding multivariate generalizations (Kullback-Leibler, KL; Energy Score, EnergyS; and the Multivariate Asymmetric Loss Function, MALF). We derived and expressed all the loss functions in terms of heritability and tested them on a real wheat dataset for one cycle of selection and in a simulated selection program. The performance of each univariate loss function was compared with the standard method of selection (Std) that does not use loss functions. We compared the performance in terms of the selection response and decrease of the population's genetic variance during recurrent breeding cycles. Results suggest that it is possible to obtain better performance in a long-term breeding program in the single-trait scheme by selecting 30% of the best individuals in each cycle but not when selecting 10% of the best individuals. For the multi-trait approach, results show that the population mean for all traits under consideration had positive gains, even though two of the traits were negatively correlated. The corresponding population variances were not statistically different from the different loss function during the 10th selection cycle. The use of loss function should be a useful criterion when selecting the candidates to selection for the next breeding cycle.

Keywords and phrases : Bayesian Decision Theory, Genomic Selection, Loss Function.

AGRADECIMIENTOS

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico brindado para realizar mis estudios de doctorado. También a mi amado país México por darme oportunidades a lo largo de mi formación profesional.

Al Colegio de Postgraduados, por haberme brindado la oportunidad de seguir mi formación académica y profesional en sus aulas.

A los integrantes de mi Consejo Particular:

Dr. Sergio Pérez Elizalde, por su excelente dirección y apoyo para la realización del presente trabajo. Este trabajo no hubiera sido posible sin su invaluable apoyo y su amistad.

Dr. Paulino Pérez Rodríguez, por su disposición, apoyo y acertadas sugerencias a esta investigación. Su contribución en mi formación académica fue vital para llevar a buen fin este proyecto.

Dr. José Luis Francisco Crossa Hiriart, por motivarme en todo momento a dar lo mejor de mí. La pasión con la que trabaja fue contagiosa para poder escribir y re-escribir numerosas ocasiones el artículo derivado de esta investigación.

Dra. Martha Elva Ramírez Guzmán, por sus atinados comentarios hechos tanto en las reuniones de consejo como en la revisión de la presente tesis.

Dr. José Andrés Christen Gracia, por su colaboración y revisión crítica a la presente tesis.

Y también quiero agradecer a todos aquellos amigos y amigas que de una u otra forma me ayudaron a concluir ésta etapa de mi vida, en especial a Magín, Jazmín, Nalleli y Angela; así como a las/los administrativos de estadística, sobre todo a Isabel y Griselda.

DEDICATORIA

A mi amiga-confidente-esposa: Claudia. Las grandes batallas las hemos dado juntos.

Para mis amados hijos: Aleshka y Caín. Los sueños cobran sentido gracias a ustedes

A mis amados padres: Evodio y Carmen. Porque toda historia tiene un comienzo, y ustedes empezaron a escribir la mía.

A mis queridos hermanos: César, Bertin Paúl y Roberto.

A las y los profesores que fueron partícipes de mi formación profesional y humana a lo largo de mi vida.

CONTENIDO

RESUMEN	iv
ABSTRACT	v
AGRADECIMIENTOS	vi
DEDICATORIA	vii
1. INTRODUCCIÓN	1
2. REVISIÓN DE LITERATURA	4
2.1. Inferencia Bayesiana	4
2.1.1. ¿Qué es la verosimilitud?	4
2.1.2. Transitando de la distribución <i>a priori</i> y la verosimilitud a la distribución <i>a posteriori</i>	7
2.1.3. Muestreador de Gibbs	8
2.1.4. Distribución predictiva <i>a posteriori</i>	9
2.2. Selección genómica asistida por marcadores moleculares	10

CONTENIDO

2.3.	Breve revisión sobre modelos de regresión y predicción	12
2.3.1.	Métodos de regresión penalizada	13
2.3.2.	Modelo de regresión para múltiples rasgos	16
2.4.	Utilidad y pérdida, conceptos de teoría de la decisión	18
2.4.1.	Función <i>score</i>	18
2.4.2.	Medidas de divergencia	20
2.4.3.	Relación con el concepto de función de pérdida	21
2.4.4.	Algunas funciones de pérdida univariadas y multivariadas	21
2.5.	Índices de selección	27
3.	MATERIALES Y MÉTODOS	30
3.1.	La selección como un problema de decisión	30
3.1.1.	Funciones de pérdida univariadas	33
3.1.2.	Funciones de pérdida multivariadas	39
3.1.3.	Pérdida esperada <i>a posteriori</i>	44
3.2.	Estudio de aplicación de las funciones de pérdida	46
3.2.1.	Aplicación en datos de trigo	46
3.2.2.	Aplicación en un estudio de simulación	47
4.	RESULTADOS	52
4.1.	Resultados en datos de trigo utilizando funciones de pérdida univariadas	52
4.2.	Resultados en datos de trigo utilizando funciones de pérdida multivariadas	54

CONTENIDO

4.3. Resultados de estudio simulación	56
4.3.1. Resultados de funciones de pérdida univariadas	56
4.3.2. Resultados de funciones de pérdida multivariadas	62
5. DISCUSIÓN Y CONCLUSIONES	70
5.1. Discusión	70
5.2. Conclusiones	74
LITERATURA CITADA	74
ANEXOS	80
A1. Desarrollo de la función de pérdida Kullback-Leibler para fines de selección genómica	80
A1.1. Función de pérdida Kullback-Leibler univariada	80
A1.2. Función de pérdida Kullback-Leibler Multivariada	81
A2. Derivación de la función CRPS cuando la distribución predictiva es normal .	82
A3. Código R de las funciones de pérdida propuestas	83
A3.1. Funciones univariadas	84
A3.2. Funciones multivariadas	86

LISTA DE CUADROS

4.1. Prueba T^2 de Hotelling para resultados en datos de trigo.	55
4.2. Prueba de t para contrastar diferencias en 1) la media poblacional estandarizada y 2) la media de la varianza poblacional escalada, al 10mo ciclo de selección, empleando las funciones de pérdida KL, CRPS y LinLin, así como el método estándar	62
4.3. Promedio de las diferencias de las medias poblacionales en el 10mo ciclo de selección menos las medias poblacionales en el primer ciclo de selección para el rasgo 1 (T1), rasgo 2 (T2) y rasgo 3 (T3). Las heritabilidades para todos los rasgos se fijó en 0.3 y 0.6 (Errores estándar respectivos se presentan en paréntesis).	67
4.4. Promedio de las diferencias correspondientes a las varianzas poblacionales en el 10mo ciclo de selección menos las varianzas poblacionales en el primer ciclo para el rasgo 1 (T1), rasgo 2 (T2) y rasgo 3 (T3). Las heritabilidades para todos los rasgos se fijó en 0.3 y 0.6 (Errores estándar respectivos se presentan en paréntesis).	67

LISTA DE FIGURAS

2.1. Esquema simplificado de selección genómica asistida por marcadores moleculares.	11
2.2. Penalizaciones para la regresión Ridge y la regresión LASSO	14
2.3. Ilustración de la función CRPS	24
2.4. Función de pérdida LinLin para diferentes niveles de α	26
2.5. Gráficos de contorno de la Función de Pérdida Asimétrica Multivariada . . .	28
3.1. Distribución base, proporción de seleccionados y distribución en la segunda generación	31
3.2. Selección por truncamiento.	34
3.3. Selección usando funciones de pérdida.	35
3.4. Selección por truncamiento para dos rasgos	41
3.5. Representación bivariada de las funciones de pérdida KL, EnergyS y MALF	44
3.6. Esquema de simulación empleado.	49
3.7. Árbol de carpetas utilizadas en el esquema de simulación	50

LISTA DE FIGURAS

4.1. Boxplots de los valores de cría estimados para los datos de trigo usando las funciones de pérdida univariadas	53
4.2. Boxplots de los valores de cría estimados para los datos de trigo usando las funciones de pérdida multivariadas	55
4.3. Respuesta a la selección en la simulación univariada del top10 % de seleccionados	57
4.4. Respuesta a la selección en la simulación univariada del top30 % de seleccionados	58
4.5. Varianza poblacional en la simulación univariada del top10 % de seleccionados	59
4.6. Varianza poblacional en la simulación univariada del top30 % de seleccionados	59
4.7. Resultados de la selección univariada para la media y la varianza poblacional al 10mo ciclo de selección del top10 % de seleccionados	60
4.8. Resultados de la selección univariada para la media y la varianza poblacional al 10mo ciclo de selección del top30 % de seleccionados	61
4.9. Resultado de estudio de simulación multivariado - medias poblacionales . . .	65
4.10. Resultado de estudio de simulación multivariada - varianzas poblacionales .	66

El proceso de mejoramiento consiste en seleccionar individuos con determinados rasgos de interés para cruzarlos entre sí, ya que de este modo se identifican aquellos individuos con las mejores características de ambos padres. La selección de plantas empleando el enfoque convencional (selección fenotípica y con base en el pedigrí) se realiza por medio de selección por truncamiento que consiste en escoger aquellas líneas (en mejoramiento de plantas) o individuos (en mejoramiento de animales) de una población base que exhiben un mayor rendimiento. Las líneas son elegidas con base el valor de cría (VC) de uno o más rasgos de interés auxiliándose de los valores fenotípicos observados. Mientras tanto en selección genómica (SG), se emplean marcadores moleculares para predecir los VC de las líneas candidatas a la selección en una población que ha sido genotipada¹, pero de las cuales se desconocen los valores fenotípicos (Meuwissen *et al.*, 2001). La SG parte de seleccionar como padres aquellas líneas o individuos cuyos VC predichos sean los más altos para uno o varios rasgos.

Cuando se selecciona un solo rasgo, el diferencial de selección (S) corresponde a la diferencia entre la media de los individuos seleccionados (μ_s) y la media de la población base (μ_1), mientras que la respuesta a la selección (R) es igual a la diferencia de la media de los descendientes (μ_2) de las líneas seleccionadas y μ_1 ; esto es, R es la fracción de S que se ganaría sobre la media de la población base cuando la selección se ha efectuado. Bos y Caligari (2008) señalan que $R = h^2 S$, donde h^2 es la heredabilidad en sentido estricto del rasgo de interés. Es evidente que cuando $h \rightarrow 1$, la media de la descendencia se aproxima a la media de los padres seleccionados, y por lo tanto, $R \rightarrow S$, mientras que cuando $h^2 \rightarrow 0$, la media de la descendencia tiende a la media de la población base, y en consecuencia $R \ll S$. Por lo tanto, el diferencial de selección permite a los mejoradores estimar el

¹Por genotipado, o genotipificación o caracterización genética, se entiende el proceso de determinación del genotipo o contenido genómico, en forma de ADN, específico de un organismo biológico, mediante un procedimiento de laboratorio.

1. INTRODUCCIÓN

progreso esperado antes de llevar a cabo la selección.

Note que cuando la selección se emplea para mejorar el valor económico y el mérito genético de un cultivo, los programas de mejoramiento se aplican simultáneamente a varios rasgos (Falconer y Mackay, 1996). Esto es un factor que dificulta la selección, sobre todo cuando se presentan rasgos antagónicos. La estrategia más común para salvar esta dificultad es la utilización de índices de selección (IS) y de Índices de selección moleculares (ISM). Los primeros involucran pesos subjetivos para cada rasgo que el mejorador determina con base en su experiencia, mientras que los segundos incorporan información genómica a través del uso de marcadores moleculares.

La mayoría de los trabajos realizados para seleccionar a las/los mejores líneas o individuos a cruzar se basan en la selección por truncamiento, donde la distribución base se restringe a un subconjunto dado un punto de truncamiento. Por lo tanto, el subconjunto de individuos seleccionados que serán padres de una siguiente generación tiene una distribución truncada. Otro enfoque para seleccionar a los mejores individuos plantea la selección mediante un enfoque de optimización y emplea la información del pedigrí y los valores genéticos aditivos (Shepherd, R.K. y Kinghorn, B.P., 1998). También se han desarrollado algoritmos e índices de selección de parejas para encontrar el mejor conjunto de padres, maximizando el mérito genético de la progenie y minimizando la endogamia y la co-ancestría (Brisbane y Gibson, 1995, Wray y Goddard, 1994).

Ambos enfoques (selección por truncamiento y selección como un problema de optimización) pueden plantearse formalmente como un problema de decisión bajo incertidumbre. Este es el planteamiento central del presente trabajo de investigación, donde el criterio de decisión utilizado para seleccionar un subconjunto de líneas se basa en elegir aquellas cuya pérdida esperada sea la mínima, dada una función de pérdida (FP) que represente la preferencia del mejorador. Cuando la selección se realiza para un solo rasgo, la función de pérdida debe considerar factores tales como la variabilidad del rasgo así como su media; mientras que en la selección multirasgo se debe considerar la estructura de dependencia y variabilidad (covarianzas) entre los diferentes rasgos, así como sus respectivas medias, siempre tratando de maximizar la respuesta a la selección.

En la literatura existen algunas medidas de divergencia que pueden adecuarse como funciones de pérdida en el contexto de la selección. Las líneas o individuos candidatos a la selección cuyas distribuciones divergan poco con respecto a la distribución teórica parental, tendrán las menores pérdidas esperadas (R alta) y el mejorador debe preferir estas líneas

1. INTRODUCCIÓN

dado que alcanzan las medias deseadas y mantienen la variabilidad genética (h^2 alta).

Por lo tanto, el objetivo principal de esta investigación es mostrar como la teoría de la decisión Bayesiana puede emplearse para seleccionar las mejores líneas, minimizando la divergencia entre las distribuciones de las líneas o individuos candidatos a la selección y la distribución teórica (maximizando R). En primera instancia se desarrolla el planteamiento para la selección de un solo rasgo y posteriormente se generaliza para cuando el interés de la selección está en más de un rasgo.

Los objetivos específicos en esta investigación son: (1) presentar una metodología formal con sustento en la teoría de la decisión bayesiana, adaptado al problema de la selección de plantas y animales; (2) describir las funciones de pérdida univariadas (selección en solo rasgo) KL (Kullback Leibler), CRPS (Continuous Ranked Probability Score) y LinLin (Lineal Lineal), así como sus generalizaciones al contexto multivariado (selección multirasgo) - KL, EnergyS (Energy Score) y MALF (Función de Pérdida Asimétrica Multivariada), respectivamente - en el contexto de la selección genómica para seleccionar a las mejores líneas para los ciclos futuros de selección en un programa de mejoramiento; y (3) aplicar la metodología propuesta en un conjunto de datos reales, y mediante un estudio de simulación, para así ilustrar y comparar los resultados obtenidos con los métodos estándar de selección.

En este capítulo discutiremos aspectos muy generales de la selección genómica asistida por marcadores moleculares. Se dará un bosquejo superficial del papel que juegan los modelos estadísticos y la inferencia bayesiana utilizados en SG. La revisión de literatura también abarca los aspectos generales de la teoría de la decisión y su relación con los conceptos de función de pérdida y función *score*. También se presentan la formulación teórica de las funciones de pérdida utilizadas en esta investigación para fines de selección.

2.1. Inferencia Bayesiana

2.1.1. ¿Qué es la verosimilitud?

La verosimilitud de una hipótesis (H) dados algunos datos (D), es proporcional a la probabilidad de obtener D dado que H es verdadera, multiplicada por una constante (Cte) arbitraria. En otras palabras, $L(H|D) = Cte \times P(D|H)$. Por lo tanto, la verosimilitud no es una probabilidad y tampoco obedece varias reglas de la probabilidad, por ejemplo, la verosimilitud no necesita sumar 1. Una diferencia fundamental entre la probabilidad y la verosimilitud es la interpretación de lo que es fijo y de lo que varía. En el caso de la probabilidad condicional, $P(D|H)$, la hipótesis es fija y los datos pueden variar libremente. La verosimilitud sin embargo, es lo opuesto. La verosimilitud de una hipótesis, $L(H|D)$, condiciona los datos como si fueran fijos, al tiempo que permite que las hipótesis varíen.

Un problema recurrente en estadística es la estimación de un parámetro (o grupo de parámetros) dado un conjunto de observaciones. A partir de los datos se construyen aseveraciones basándose en un modelo estadístico que intenta describir algún aspecto del “estado de las

2.1. Inferencia Bayesiana

cosas”. Desde el punto de vista paramétrico, estos modelos están gobernados por parámetros (y que a su vez pueden ser función de otros parámetros). Los valores que toman estos parámetros son desconocidos y se infieren a partir de los datos observados. Un ejemplo de datos observados en genética, son los registros fenotípicos para algún rasgo en particular, ya sean cuantitativos o cualitativos.

Formalizando el concepto de verosimilitud desde un punto de vista matemático, suponga que Y_1, Y_2, \dots, Y_n son n variables aleatorias resultado de un proceso aleatorio que puede ser caracterizado en términos de una función de densidad de probabilidad (fdp) $f(y_i|\theta)$ dependiente del parámetro θ (escalar, vector o matriz) que toma valores en Ω . La fdp conjunta de n observaciones independientes $\mathbf{y} = (y_1, y_2, \dots, y_n)'$ está dada por,

$$p(\mathbf{y}|\theta) = \prod_{i=1}^n p(y_i|\theta) = L(\theta|\mathbf{y}). \quad (2.1)$$

El producto en la ecuación 2.1 se debe al supuesto de independencia entre las y_i 's, y se le conoce como función de verosimilitud. Es común que en lugar de trabajar la función de verosimilitud se trabaje con la función de log-verosimilitud, matemáticamente más tratable y más fácil de maximizar

$$\log L(\theta|\mathbf{y}) = l(\theta|\mathbf{y}) = \sum_{i=1}^n \log p(\theta|y_i). \quad (2.2)$$

En el caso de que \mathbf{Y} sea un vector aleatorio discreto, entonces $L(\theta|\mathbf{y}) = P_{\theta}(\mathbf{Y} = \mathbf{y})$. Si comparamos la función de verosimilitud en dos puntos y encontramos que $P_{\theta_1}(\mathbf{Y} = \mathbf{y}) = L(\theta_1|\mathbf{y}) > L(\theta_2|\mathbf{y}) = P_{\theta_2}(\mathbf{Y} = \mathbf{y})$ entonces la muestra que hemos observado es más probable que haya ocurrido si $\theta = \theta_1$ que si $\theta = \theta_2$, lo que se interpreta diciendo que θ_1 es un valor más plausible para el valor verdadero de θ que θ_2 .

Los estimadores de máxima verosimilitud (EMV) se definen formalmente como aquellos valores $\hat{\theta}$ tal que

$$l(\hat{\theta}|\mathbf{y}) \geq l(\theta|\mathbf{y}) \quad \forall \theta \in \Omega.$$

2.1. Inferencia Bayesiana

El vector score. La primera derivada de la función de log-verosimilitud es conocida como función score y está dada por:

$$u(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta}}.$$

Note que el *score* corresponde al vector de primeras derivadas parciales, una por cada elemento de $\boldsymbol{\theta}$. Los EMV se obtienen igualando a cero el *score*, es decir, resolviendo el sistema de ecuaciones $u(\hat{\boldsymbol{\theta}}) = \mathbf{0}$.

Matriz de información. El vector score es un vector aleatorio con algunas propiedades interesantes desde el punto de vista estadístico. En particular, el score evaluado en el valor real del parámetro $\boldsymbol{\theta}$ tiene media cero, es decir, $E(u(\boldsymbol{\theta})) = \mathbf{0}$, y la matriz de varianzas y covarianzas dada por la matriz de información de Fisher

$$\text{Var}(u(\boldsymbol{\theta})) = E(u(\boldsymbol{\theta})u(\boldsymbol{\theta})') = I(\boldsymbol{\theta}).$$

Bajo ciertas condiciones de regularidad, la matriz de información también puede obtenerse como el negativo del valor esperado de las segundas derivadas parciales de la función de log-verosimilitud

$$I(\boldsymbol{\theta}) = -E\left(\frac{\partial^2 l(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right),$$

esta matriz resultante se denomina en ocasiones como matriz de información observada.

Newton-Raphson y Score de Fisher. El cálculo del EMV frecuentemente requiere la utilización de procedimientos iterativos. Considere expandir la función *score* evaluada en el EMV $\hat{\boldsymbol{\theta}}$ en un valor propuesto $\hat{\boldsymbol{\theta}}_0$ usando una serie de Taylor de primer orden, de modo que

$$u(\hat{\boldsymbol{\theta}}) \approx u(\hat{\boldsymbol{\theta}}_0) + \frac{\partial u(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}}(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_0).$$

Sea H la matriz Hesiana de segundas derivadas parciales de la función de log-verosimilitud

2.1. Inferencia Bayesiana

$$H(\boldsymbol{\theta}) = \frac{\partial^2 l(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \frac{\partial u(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \quad (2.3)$$

Igualando a cero la ecuación 2.3 y resolviendo para $\boldsymbol{\theta}$ nos da la siguiente aproximación:

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 - H^{-1}(\boldsymbol{\theta}_0)u(\boldsymbol{\theta}_0).$$

Este resultado proporciona la base para el esquema iterativo de cálculo de EMV conocido como Newton-Raphson. Dado un valor inicial, empleando la ecuación anterior se obtiene una mejor estimación para $\boldsymbol{\theta}$, y el proceso se repite hasta que la diferencia entre las estimaciones sucesivas sea cercana a cero. Este procedimiento converge muy rápido siempre que la log-verosimilitud sea aproximadamente cuadrática en la vecindad del máximo y siempre que los valores iniciales sean cercanos al EMV. Un procedimiento alternativo consiste en reemplazar el negativo del Hesiano por su valor esperado, la matriz de información

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + I^{-1}(\boldsymbol{\theta}_0)u(\boldsymbol{\theta}_0). \quad (2.4)$$

Estas y otras consideraciones en relación a los EMV pueden consultarse en [Rodríguez \(2007\)](#).

2.1.2. Transitando de la distribución *a priori* y la verosimilitud a la distribución *a posteriori*

En estadística bayesiana, las conclusiones acerca de algún parámetro $\boldsymbol{\theta} \in \Omega$, o de algún dato no observado \tilde{y} , se dan en términos de probabilidades condicionadas en los datos observados \mathbf{y} . $p(\boldsymbol{\theta}|\mathbf{y})$ y $p(\tilde{y}|\mathbf{y})$ representan las fdp de $\boldsymbol{\theta}$ y de \tilde{y} condicionadas por \mathbf{y} . Partiendo de

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

donde $p(\boldsymbol{\theta})$ corresponde a la distribución *a priori* para $\boldsymbol{\theta}$ y $p(\mathbf{y}|\boldsymbol{\theta})$ representa la distribución

2.1. Inferencia Bayesiana

para los datos, y usando el Teorema de Bayes se tiene que

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\boldsymbol{\theta}, \mathbf{y})}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}. \quad (2.5)$$

La expresión 2.5, corresponde a la distribución *a posteriori* de $\boldsymbol{\theta}$, donde $p(\mathbf{y})$ es el factor de normalización $p(\mathbf{y}) = \int_{\boldsymbol{\theta} \in \Omega} p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}$ en el supuesto de que $\boldsymbol{\theta}$ sea continua y que no depende de $\boldsymbol{\theta}$; por tanto, una forma equivalente de esta expresión resulta en la distribución *a posteriori* “no normalizada”

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \quad (2.6)$$

Como podemos apreciar en el primer término de la ecuación 2.6, $p(\mathbf{y}|\boldsymbol{\theta})$ es una función que depende de $\boldsymbol{\theta}$, pero no depende de \mathbf{y} . Estas simples fórmulas, capturan la esencia de la inferencia bayesiana: la primera tarea en cualquier aplicación es desarrollar un modelo de la forma $p(\boldsymbol{\theta}, \mathbf{y})$ para después realizar los cálculos necesarios y resumir $p(\boldsymbol{\theta}|\mathbf{y})$ en formas apropiadas al interés particular (Gelman *et al.*, 2014).

Si $\mathbf{y} = (y_1, y_2, \dots, y_n)'$ es una muestra aleatoria, entonces $p(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^n p(y_i|\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\mathbf{y})$ corresponde a la función de verosimilitud. Note que al elegir una distribución para los datos, implica que los datos \mathbf{y} se reflejan en la inferencia *a posteriori* únicamente a través de $p(\mathbf{y}|\boldsymbol{\theta})$. Por tanto, la inferencia bayesiana obedece al *principio de verosimilitud*, el cual sostiene que dado una muestra, dos modelos de probabilidad $p(\mathbf{y}|\boldsymbol{\theta})$ que tienen la misma función de verosimilitud proporcionarán la misma inferencia para $\boldsymbol{\theta}$. De 2.5 y 2.6 se observa que la distribución *a posteriori* de $\boldsymbol{\theta}$ combina dos fuentes de información para $\boldsymbol{\theta}$; por un lado el conocimiento *a priori* (subjetivo) acerca de $\boldsymbol{\theta}$ mediante una distribución, y por otra lado la información contenida en los datos mediante la función de verosimilitud.

2.1.3. Muestreador de Gibbs

En modelos multiparamétricos (como los que se emplean en selección genómica), la distribución *a posteriori* conjunta no tiene forma analítica y es difícil muestrear de la misma directamente. Suponga que $p(\boldsymbol{\theta}|\mathbf{y})$ es la distribución de interés, donde $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$. Cada componente de $\boldsymbol{\theta}$ puede ser un escalar, vector o matriz. Considérese también que las

2.1. Inferencia Bayesiana

distribuciones condicionales $p_i(\theta_i|\theta_{-i}, \mathbf{y})$, $i = 1, \dots, p$ son conocidas, lo que implica que podemos tomar muestras de ellas. En estos casos, la distribución *a posteriori* $p(\boldsymbol{\theta}|\mathbf{y})$ puede aproximarse mediante el muestreador de Gibbs (Casella y George, 1992) que se presenta a continuación.

1. Inicializar el contador $j = 1$ de iteraciones de la cadena y proponer valores iniciales $\boldsymbol{\theta}^0 = (\theta_1^{(0)}, \dots, \theta_p^{(0)})'$;
2. Obtener un nuevo valor $\boldsymbol{\theta}^{(j)} = (\theta_1^{(j)}, \dots, \theta_p^{(j)})'$ de $\boldsymbol{\theta}^{(j-1)}$ a través de la generación sucesiva de valores

$$\begin{aligned}\theta_1^{(j)} &\sim p(\theta_1|\theta_2^{(j-1)}, \dots, \theta_p^{(j-1)}, \mathbf{y}), \\ \theta_2^{(j)} &\sim p(\theta_2|\theta_1^{(j)}, \theta_3^{(j-1)}, \dots, \theta_p^{(j-1)}, \mathbf{y}), \\ &\vdots \\ \theta_p^{(j)} &\sim p(\theta_p|\theta_1^{(j)}, \dots, \theta_{p-1}^{(j)}, \mathbf{y});\end{aligned}$$

3. Incrementar el contador j en $j + 1$ y repetir el paso 2 hasta que el algoritmo alcance la convergencia, es decir, hasta que las cadenas simuladas pertenezcan a sus respectivas distribuciones estacionarias.

El algoritmo anterior no toma muestras directamente de $p(\boldsymbol{\theta})$, en lugar de ello, simula las muestras a través de todas las condicionales *a posteriori*, parámetro por parámetro. Dado que se necesita asumir valores iniciales para comenzar el proceso iterativo, las muestras obtenidas en las primeras iteraciones generalmente no son representativas de la distribución *a posteriori*, sin embargo, la teoría de Cadenas de Markov Monte Carlo (MCMC) garantiza que después de cierto número de iteraciones (llamado periodo de calentamiento), la distribución estacionaria de las muestras generadas corresponde a la distribución *a posteriori* conjunta de interés (Gilks *et al.*, 1995).

2.1.4. Distribución predictiva *a posteriori*

Una vez que se ha observado una muestra $\mathbf{y} = (y_1, y_2, \dots, y_n)'$ de una población, quizá el interés este en predecir una nueva observación \tilde{y} (o nuevas observaciones). La distribución predictiva de interés para nuevas observaciones es

2.2. Selección genómica asistida por marcadores moleculares

$$\begin{aligned} p(\tilde{y}|\mathbf{y}) &= \int_{\mathbb{R}^p} p(\tilde{y}, \boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \\ &= \int_{\mathbb{R}^p} p(\tilde{y}|\boldsymbol{\theta}, \mathbf{y})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \\ &= \int_{\mathbb{R}^p} p(\tilde{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}. \end{aligned}$$

Esta distribución es llamada “distribución predictiva *a posteriori*”, y la razón de llamarla predictiva se debe a que es una predicción para una cantidad observable \tilde{y} pero desconocida, y *a posteriori* dado que está condicionada en los datos observados. En muchas ocasiones será fácil muestrear de $p(\boldsymbol{\theta}|\mathbf{y})$ y de $p(\mathbf{y}|\boldsymbol{\theta})$, sin embargo, muestrear de $p(\tilde{y}|\mathbf{y})$ puede ser complicado. En éstos casos toman muestras de la distribución predictiva *a posteriori* indirectamente, a través de la técnica Monte Carlo expuesta en la sección anterior, donde un conjunto de muestras de \tilde{Y} se obtendría del siguiente modo,

$$\begin{aligned} \text{muestrear } \boldsymbol{\theta}^{(1)} &\sim p(\boldsymbol{\theta}|\mathbf{y}) \quad , \quad \text{muestrear } \tilde{y}^{(1)} \sim p(\tilde{y}|\boldsymbol{\theta}^{(1)}) \\ \text{muestrear } \boldsymbol{\theta}^{(2)} &\sim p(\boldsymbol{\theta}|\mathbf{y}) \quad , \quad \text{muestrear } \tilde{y}^{(2)} \sim p(\tilde{y}|\boldsymbol{\theta}^{(2)}) \\ &\vdots \\ \text{muestrear } \boldsymbol{\theta}^{(S)} &\sim p(\boldsymbol{\theta}|\mathbf{y}) \quad , \quad \text{muestrear } \tilde{y}^{(S)} \sim p(\tilde{y}|\boldsymbol{\theta}^{(S)}) \end{aligned}$$

La secuencia $\{(\boldsymbol{\theta}, \tilde{y})^{(1)}, \dots, (\boldsymbol{\theta}, \tilde{y})^{(S)}\}$ forman S muestras de la distribución *a posteriori* conjunta de $(\boldsymbol{\theta}, \tilde{Y})$, y la secuencia $\{\tilde{y}^{(1)}, \dots, \tilde{y}^{(S)}\}$ constituyen S muestras de la distribución marginal *a posteriori* de \tilde{Y} , la cual es la distribución predictiva *a posteriori*.

2.2. Selección genómica asistida por marcadores moleculares

Entre los diferentes métodos que utilizan marcadores moleculares para fines de selección y mejoramiento, la SG ha recibido considerable atención en la última década (Rincent *et al.*, 2012). El objetivo de este enfoque es predecir los VC de individuos candidatos con base en su información genotípica empleando marcadores moleculares. Un modelo de regresión con fines de predicción se construye con base en los genotipos y fenotipos de individuos de

2.2. Selección genómica asistida por marcadores moleculares

una población base formando así un conjunto de calibración [o entrenamiento] (Meuwissen *et al.*, 2001). La SG potencialmente incluye todos los efectos de los marcadores. Si la densidad de los marcadores es suficientemente grande de modo que cubra todo el genoma, estos permiten al modelo capturar una parte importante de la varianza genética (Yang *et al.*, 2010).

La SG se utilizó primero en programas de mejoramiento animal, específicamente en vacas lecheras, y su utilización claramente mejoró la eficiencia de selección (Rincent *et al.*, 2012). Hoy en día es ampliamente utilizada en programas de mejoramiento de plantas, con resultados prometedores (Crossa *et al.*, 2010, Jannink *et al.*, 2010).

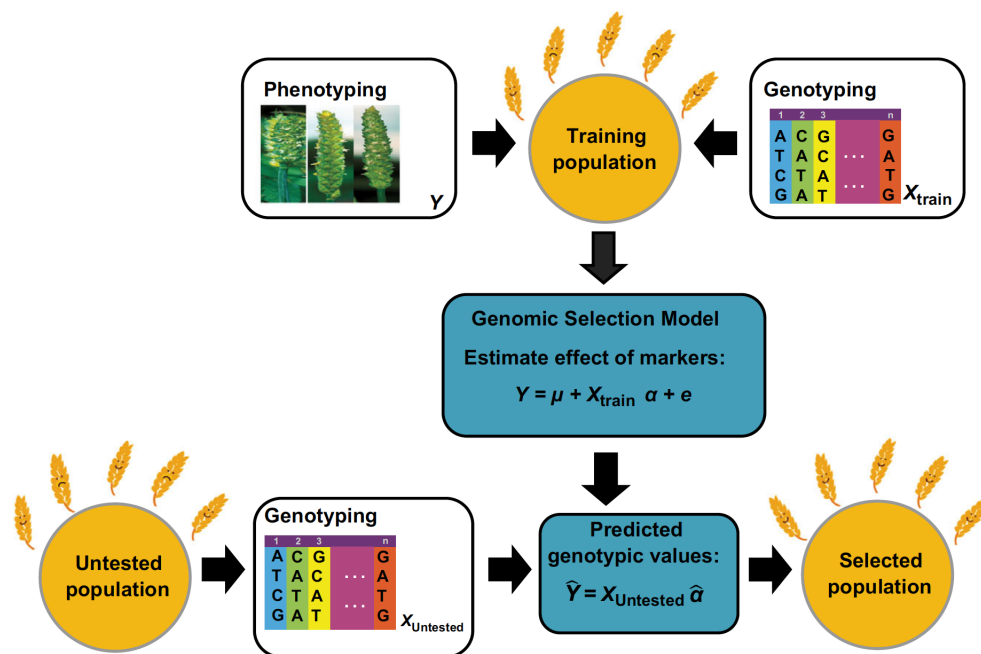


Figura 2.1: Esquema simplificado de selección genómica asistida por marcadores moleculares.

En la SG, se estiman los efectos de los marcadores en el conjunto de datos de entrenamiento (ver figura 2.1), para después predecir los valores de cría en el conjunto de prueba, multiplicando sus genotipos por los efectos de los marcadores. Este enfoque es utilizado por ejemplo, en el modelo de regresión RR-BLUP (regresión ridge-mejor predictor lineal insesgado¹).

La implementación de la SG se ha facilitado gracias a los recientes avances en secuenciación. Ahora se puede acceder a arreglos genotípicos de alta calidad y bajo costo (Rincent

¹en inglés *best linear unbiased predictions*

2.3. Breve revisión sobre modelos de regresión y predicción

et al., 2012). En conjunto, la biología molecular y la genómica, tienen el potencial para iniciar una nueva “revolución verde”, que a su vez, es de vital importancia para el desarrollo acelerado del germoplasma de cultivos (Salgotra *et al.*, 2014). La vinculación cada vez más precisa de marcadores y genes a rasgos, está resultando en la reproducción de plantas más eficientes. Las ventajas principales de la SG sobre la selección convencional son la reducción del costo por ciclo de selección, así como el menor tiempo entre ciclos de mejora (Crossa *et al.*, 2017). La SG selecciona como padres aquellas líneas o individuos cuyos VC estimados son los más mayores para uno o varios rasgos.

En contraste, aunque ha habido importantes avances en la automatización del proceso de fenotipado, es aún muy caro obtener fenotipos reelevantes, con alta heredabilidad para un conjunto grande de individuos (Rincent *et al.*, 2012); de aquí surge la necesidad de emplear modelos de predicción siguiendo el esquema de la figura 2.1.

2.3. Breve revisión sobre modelos de regresión y predicción

Los modelos de regresión y predicción usados en SG incorporan información de marcadores moleculares como variables explicativas del valor fenotípico de algún rasgo de interés, esto lo hacen a través de una función $f(\cdot)$ que puede ser paramétrica o no. El modelo estándar de regresión se expresa como $y_i = \mu + \sum_{j=1}^p x_{ij}\beta_j + \varepsilon_i$, donde μ es el intercepto, x_{ij} es el genotipo correspondiente al i -ésimo individuo o línea, en el j -ésimo marcador ($j = 1, 2, \dots, p$), y β_j corresponde a los efectos del marcador. En su forma matricial queda representado como

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \quad (2.7)$$

donde \mathbf{y} representa el vector de registros fenotípicos de dimensión $n \times 1$, μ corresponde a una media general, $\mathbf{1}$ es un vector de 1's de orden $n \times 1$, \mathbf{X} es la matriz de incidencias o matriz de marcadores moleculares de orden $n \times p$, $\boldsymbol{\beta}$ es el vector de efectos de marcadores (no conocido) de dimensión $p \times 1$, y $\boldsymbol{\varepsilon}$ es el vector de errores de dimensión $n \times 1$. Se asume que los errores se distribuyen normales y que son homocedásticos, es decir, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_{\varepsilon}^2 \mathbf{I})$.

2.3. Breve revisión sobre modelos de regresión y predicción

2.3.1. Métodos de regresión penalizada

Con el desarrollo y crecimiento de las tecnologías de secuenciación, el número de marcadores moleculares y sus correspondientes parámetros a estimar, exceden por mucho el número de observaciones (líneas o individuos), $p \gg n$ (un problema conocido como “la maldición de la dimensión”)(Bellman, 1961). En este contexto, los modelos de regresión convencionales (mínimos cuadrados y máxima verosimilitud) no consiguen soluciones estables para los parámetros del modelo 2.7. Esto ha obligado y popularizado el uso de modelos de regresión que penalizan los efectos de cada marcador. A estos métodos también se les conoce como técnicas de regularización. Las técnicas de regularización imponen restricciones en el conjunto de soluciones admisibles, de modo que se soluciona

$$\hat{\beta}_{PLS} = \arg \min_{\beta^*} [(\mathbf{y} - \mathbf{X}^* \beta^*)'(\mathbf{y} - \mathbf{X}^* \beta^*) + \lambda \cdot pen(\beta)] \quad (2.8)$$

donde $pen(\beta)$ es una función que contempla el tipo de penalización, mientras que $\lambda \geq 0$ corresponde al parámetro de suavizamiento que gobierna el impacto de la penalización. Note que por facilidad en la exposición, hemos supuesto que el modelo en 2.7 está reescrito como $\beta^* = (\mu \mathbf{1}, \beta)'$ y con el correspondiente reajuste de la matriz diseño \mathbf{X}^* , ahora de orden $n \times 1 + p$, ya que generalmente no se desea penalizar el intercepto.

Regresión Ridge

En la regresión Ridge la función $pen(\beta) = \sum_{j=1}^p \beta_j^2 = \beta' \beta$, donde $pen(\beta)$ corresponde a la norma- ℓ_2 de β . Derivando y resolviendo 2.8 con respecto a β se obtiene

$$\hat{\beta}_{PLS} = (\mathbf{X}^{*'} \mathbf{X}^* + \lambda \mathbf{P})^{-1} \mathbf{X}^{*'} \mathbf{y}$$

con $\mathbf{P} = \text{diag}(0, 1, \dots, 1)$ y donde $\hat{\beta}_{PLS}$ difiere de la bien conocida $\hat{\beta} = (\mathbf{X}^{*'} \mathbf{X}^*)^{-1} \mathbf{X}^{*'} \mathbf{y}$ (MCO²) por el término $\lambda \mathbf{P}$ que proviene de la penalización y que cuando $\lambda \rightarrow 0$, $\hat{\beta}_{PLS} = \hat{\beta}$.

²Mínimos Cuadrados Ordinarios

2.3. Breve revisión sobre modelos de regresión y predicción

Regresión LASSO

La regresión LASSO³ (de los Campos *et al.*, 2009, Park y Casella, 2008) es una técnica de regularización que combina la estimación de un modelo y la selección de variables al mismo tiempo. En la regresión LASSO, a diferencia de la regresión Ridge que contrae el efecto de algunas covariables hacia el cero, los efectos de algunas covariables son exactamente cero, lo que sirve como herramienta de selección de modelos. La función $pen(\beta)$ en la regresión LASSO se expresa como $pen(\beta) = \sum_{j=1}^p |\beta_j|$ y

$$\hat{\beta}_{PLS} = \arg \min_{\beta} \left[(\mathbf{y} - \mathbf{X}^* \beta)' (\mathbf{y} - \mathbf{X}^* \beta) + \lambda \sum_{j=1}^p |\beta_j| \right] \quad (2.9)$$

donde el intercepto no se penaliza. Al observar el tipo de penalización entre la regresión Ridge y la regresión LASSO, observamos que la regresión Ridge impone una penalización cuadrática, lo que implica que estimaciones grandes sufren mayor penalización en contraste con las pequeñas. Por el contrario, en la regresión LASSO, la penalización valor absoluto se incrementa en pequeñas proporciones para coeficientes grandes, pero se acerca rápido al cero para valores pequeños. A diferencia de la regresión Ridge, no existe una solución cerrada para las estimaciones de las β 's en la regresión LASSO. La figura 2.2 muestra la diferencia entre las penalizaciones para la regresión Ridge y LASSO.

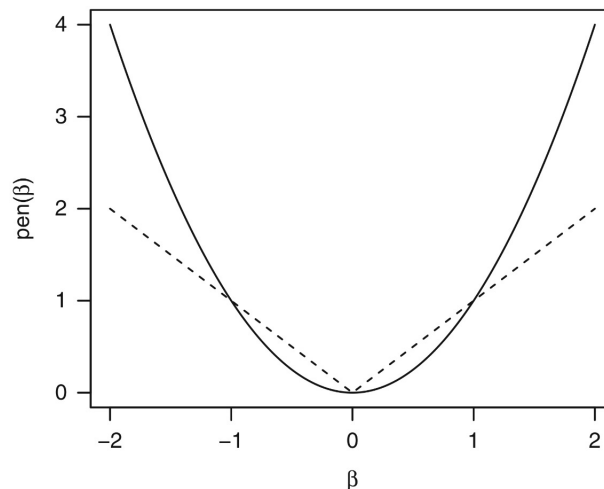


Figura 2.2: Penalizaciones para la regresión Ridge (línea discontinua) y la regresión LASSO (línea sólida).

³Siglas en inglés de *Least Absolute Shrinkage and Selection Operator*.

2.3. Breve revisión sobre modelos de regresión y predicción

Parametrización Bayesiana

La solución del modelo 2.7 en el contexto de inferencia bayesiana se realiza asignando distribuciones *a priori* que tengan efectos similares a las penalizaciones. Por ejemplo, la librería BGLR (de los Campos y Pérez Rodríguez, 2015) de R (R Core Team, 2016) emplea distribuciones *a prioris* gaussianas independientes para cada elemento de β (excluyendo al intercepto), es decir, $\beta|\tau^2 \sim N(\mathbf{0}, \tau^2\mathbf{I})$, y de este modo consigue el encojimiento similar a la regresión Ridge, y que BGLR llama modelo BRR⁴. Al hiperparámetro τ se le asigna una distribución $\chi_{Inv}^{-2}(\nu, S)$ con ν grados de libertad y parámetro de escala S , mientras que $\varepsilon|\sigma_\varepsilon^2 \sim N(\mathbf{0}, \sigma_\varepsilon^2\mathbf{I})$. De este modo, maximizar la distribución *a posteriori* con respecto a β equivale a minimizar el criterio de mínimos cuadrados penalizados para $\lambda = \sigma^2/\tau^2$ fijo (Fahrmeir *et al.*, 2013).

El modelo presentado en 2.7 puede escribirse como

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{u} + \varepsilon, \quad (2.10)$$

donde $\mathbf{u} = \mathbf{X}\beta \sim N(\mathbf{0}, \sigma_u^2\mathbf{G})$ y $\mathbf{G} = \mathbf{X}\mathbf{X}'/p$ es la matriz de relaciones genómicas. Esta parametrización del modelo se estima a través de la teoría BLUP⁵ (Mejor predictor líneal insesgado) y que es computacionalmente más eficiente que el modelo en 2.7.

Similar a la regresión Ridge bayesiana, para la regresión LASSO se define una distribución *a priori* para los parámetros, de modo que la moda *a posteriori* se obtiene al minimizar 2.9. Nuevamente se asume independencia (condicional) de modo que

$$\beta|\tau_1^2, \dots, \tau_p^2 \sim N(\mathbf{0}, \text{diag}(\tau_1^2, \dots, \tau_p^2)),$$

donde ahora, cada coeficiente de regresión β_j tiene su propia varianza τ_j^2 . El parámetro de varianza τ_j^2 se asume mutuamente independiente con distribución *a priori* $\tau_j^2|\lambda \stackrel{iid}{\sim} \text{Expo}(0.5\lambda^2)$ y con hiperparámetro $\lambda \sim G(a_\lambda, b_\lambda)$. Tanto la regresión Ridge como la regresión LASSO en su versión bayesiana puede ajustarse empleando la librería BGLR de R.

⁴siglas en inglés de *Bayesian Ridge Regression*

⁵Siglas en inglés de *Best Linear Unbiased Prediction*.

2.3. Breve revisión sobre modelos de regresión y predicción

2.3.2. Modelo de regresión para múltiples rasgos

Como se ha advertido previamente, la selección se emplea para mejorar el valor económico y el mérito genético de un cultivo, por lo tanto, los programas de mejoramiento se aplican simultáneamente a varios rasgos (Falconer y Mackay, 1996). En estos casos, además de modelar y predecir los VC en cada rasgo, es necesario modelar la estructura de dependencia entre los rasgos involucrados. Enseguida se presenta un Modelo Multirasgo cuyos detalles pueden encontrarse en de los Campos, G. y Grüneberg, A. (2016).

El modelo multirasgo se escribe como enseguida

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_T \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{n_1}\mu_1 \\ \mathbf{1}_{n_2}\mu_2 \\ \vdots \\ \mathbf{1}_{n_T}\mu_T \end{pmatrix} + \begin{pmatrix} \mathbf{X}_1 & 0 & \dots & 0 \\ 0 & \mathbf{X}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{X}_T \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_T \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_T \end{pmatrix}, \quad (2.11)$$

donde \mathbf{y}_t ($t = 1, 2, \dots, T$) es un vector n_t -dimensional de valores fenotípicos del t -ésimo rasgo, \mathbf{X}_t es la matriz de orden $n_t \times p$ que contiene información genotípica para el t -ésimo rasgo, β_t es el vector p -dimensional de efectos de marcadores específico para el t -ésimo rasgo, μ_t es una media general para cada rasgo, y ε es un vector n_t -dimensional de residuales, uno por cada rasgo. Se asume que los residuales no están correlacionados y que se distribuyen normales, i.e., $\varepsilon_t \sim N_{n_t \times n_t}(\mathbf{0}, \mathbf{I}\sigma_{\varepsilon_t}^2)$. Si se supone distribuciones *a priori* $\sigma_{\varepsilon_t}^2 \sim \chi_{\nu, S}^{-2}$ para cada uno de los componentes de ε_t , entonces, la distribución *a priori* conjunta para el vector completo de residuales $\varepsilon = (\varepsilon_1, \dots, \varepsilon_T)'$ en 2.11 es

$$p(\varepsilon, \sigma_1^2, \dots, \sigma_T^2) = \prod_{t=1}^T \prod_{i=1}^{n_T} N(\varepsilon_{ti} | 0, \sigma_{\varepsilon_t}^2) \chi^{-2}(\sigma_{\varepsilon_t}^2 | \nu, S).$$

Sea $\beta = (\beta_1', \dots, \beta_T)'$ el vector completo de efectos de marcadores, y se asume que $\beta \sim MVN_{T \cdot p \times T \cdot p}(\mathbf{0}, \mathbf{B} \otimes \mathbf{I}_{p \times p})$, donde \otimes denota al producto Kronecker y \mathbf{B} corresponde a la matriz de varianzas y covarianzas de efectos entre marcadores,

2.3. Breve revisión sobre modelos de regresión y predicción

$$\mathbf{B} = \begin{pmatrix} \sigma_{\beta_1}^2 & \sigma_{\beta_{12}} & \cdots & \sigma_{\beta_{1T}} \\ \sigma_{\beta_{21}} & \sigma_{\beta_2}^2 & \cdots & \sigma_{\beta_{2T}} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\beta_{T1}} & \sigma_{\beta_{T2}} & \cdots & \sigma_{\beta_T}^2 \end{pmatrix}$$

donde $\sigma_{\beta_{jt}}$ ($j = 1, \dots, T; t = 1, \dots, T; j \neq t$) es la covarianza entre efectos de marcadores de los rasgos j y t . [Lehermeier et al. \(2015\)](#) mostraron que el modelo en (2.11) puede parametrizarse en términos de un modelo GBLUP. Sea $\mathbf{g}^* = (\mathbf{g}_1^*, \dots, \mathbf{g}_T^*)'$ el vector $(n \times T)$ -dimensional, que contiene los valores genotípicos para cada línea y cada rasgo, donde $\mathbf{g}_t^* = \mathbf{X}\boldsymbol{\beta}_t$ y $\mathbf{X} = (\mathbf{X}'_1, \dots, \mathbf{X}'_T)'$ es la matriz que contiene todos los genotipos. Usando el producto Kronecker $\mathbf{g}^* = (\mathbf{I}_{T \times T} \otimes \mathbf{X})\boldsymbol{\beta}$. Dado que \mathbf{g}^* es una combinación lineal del vector de efectos de marcadores que se asumió MVN, por lo tanto \mathbf{g}^* también se distribuye MVN. Puede mostrarse que $\mathbf{g}^* | \boldsymbol{\Sigma}_g \sim MVN_{(n \cdot T) \times (n \cdot T)}(\mathbf{0}, \boldsymbol{\Sigma}_g \otimes \mathbf{G})$, donde $\boldsymbol{\Sigma}_g = \mathbf{B}p$ es la matriz de covarianzas entre rasgos,

$$\boldsymbol{\Sigma}_g = \begin{pmatrix} \sigma_{g_1}^2 & \sigma_{g_{12}} & \cdots & \sigma_{g_{1T}} \\ \sigma_{g_{21}} & \sigma_{g_2}^2 & \cdots & \sigma_{g_{2T}} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{g_{T1}} & \sigma_{g_{T2}} & \cdots & \sigma_{g_T}^2 \end{pmatrix}.$$

Por lo tanto, el modelo planteado en 2.11 es equivalente al siguiente modelo

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_T \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{n_1}\mu_1 \\ \mathbf{1}_{n_2}\mu_2 \\ \vdots \\ \mathbf{1}_{n_T}\mu_T \end{pmatrix} + \begin{pmatrix} \mathbf{Z}_1\mathbf{g}_1^* \\ \mathbf{Z}_2\mathbf{g}_2^* \\ \vdots \\ \mathbf{Z}_T\mathbf{g}_T^* \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_T \end{pmatrix}, \quad (2.12)$$

dado que solo algunas entradas de \mathbf{g}^* están ligadas a los fenotipos. \mathbf{Z}_t son la matrices que están relacionadas a los fenotipos de \mathbf{g}_t^* ; cada una de estas matrices tiene n_t filas y n columnas. Para finalizar es necesario asignar una distribución *a priori* a $\boldsymbol{\Sigma}_g$ para formar un modelo bayesiano completamente especificado, en este sentido $\boldsymbol{\Sigma}_g \sim W^{-1}(\boldsymbol{\Psi}, \nu)$ donde $\boldsymbol{\Psi}$ es la matriz de escala y ν son los grados de libertad de la distribución Wishart invertida. Este modelo puede ajustarse empleando la librería MTM de R, disponible en github (<https://>

[//github.com/QuantGen/MTM](https://github.com/QuantGen/MTM)).

2.4. Utilidad y pérdida, conceptos de teoría de la decisión

En términos generales, el objetivo de la teoría de la decisión es ayudar a elegir entre un conjunto de acciones cuyas consecuencias no pueden ser anticipadas completamente, generalmente porque dependen del futuro o de algún estado desconocido de la naturaleza. De acuerdo a la teoría de la decisión, la *utilidad esperada* representa una utilidad cuantitativa asociada a cada consecuencia, una probabilidad para el estado de la naturaleza, y luego seleccionar una acción que maximice el valor esperado de dicha utilidad (Giovanni Parmigiani, 2009). Como definiremos más adelante, dicha utilidad puede interpretarse también como una pérdida asociada a cada decisión. Una decisión es mejor en la medida que produce más satisfacción a quien toma decisiones.

Un problema de decisión se define en términos de tres componentes: un espacio de resultados o consecuencias, un espacio de acciones y una función de pérdida (FP). A partir de esto se pueden construir funciones *score* (FS) [también denominadas funciones de utilidad], funciones de entropía y medidas de divergencia. La función de pérdida define una dualidad entre el espacio de acción y el de resultados (Dawid, 2007). Dado que cualquier problema de decisión genera una FS [y consecuentemente una función de pérdida], existe una amplia variedad de estas funciones (Dawid, 1998, Parry *et al.*, 2012).

En este capítulo se presenta una revisión general de algunas FS. De aquí en adelante los términos FS y FP se tratarán como conceptos equivalentes (utilidades negativamente interpretadas); además se aborda la relación que guarda con las medidas de divergencia, y al final se presentan algunas FPs útiles en el presente trabajo de investigación.

2.4.1. Función *score*

El concepto de función de pérdida (FP) está estrictamente relacionado con otro concepto denominado función *score* (FS). La idea fundamental detrás de una FS es encontrar formas de incentivar al tomador de decisiones de modo que sea acertado en la elección de las mismas. La penalización o pérdida a la cual estará sujeto se hará a la luz de los resultados. En este sentido, una FS es una FP que mide la calidad de cierta distribución Q para una

2.4. Utilidad y pérdida, conceptos de teoría de la decisión

variable aleatoria X , cuando una realización de X es x .

Imagine un juego entre la *Naturaleza* y *Usted* (un tomador de decisiones). Sea X una variable aleatoria con valores en \mathcal{X} , y sea \mathcal{P} una familia de distribuciones que gobierna a \mathcal{X} . Después de cierto tiempo, la *Naturaleza* revelará x , de X . La tarea de *Usted* es escoger una distribución $Q \in \mathcal{P}$ que refleje la incertidumbre acerca de qué X podría resultar. Una vez que la naturaleza haya revelado x , entonces usted estará sujeto a una penalización $S(x, Q)$, que depende tanto de X , como del valor x . La función S es una FS. Al apostar por Q , la penalización esperada resulta en $S(P, Q) = E_{X \sim P} S(x, Q) = \int S(x, Q) dP(x)$, y representa el “score esperado” con respecto a P cuando la predicción probabilística es Q y siendo P la distribución verdadera.

Visto como un problema de decisión usted debería apostar por aquella Q que minimice el score esperado $S(P, Q)$. El score esperado puede interpretarse como una medida de la información del pronosticador.

Definición formal y sus propiedades

Considérese predicciones probabilísticas en un espacio muestral Ω . Sea \mathcal{A} una σ -álgebra de subconjuntos de Ω , y sea \mathcal{P} una clase convexa de medidas de probabilidad en (Ω, \mathcal{A}) . Tal como sostiene [Gneiting y Raftery \(2007\)](#), una función definida en Ω y que toma valores en la línea real extendida $\overline{\mathbb{R}} = [-\infty, \infty]$, es P -cuasi-integrable si esta es medible con respecto a \mathcal{A} y es cuasi-integrable con respecto a todo $Q \in \mathcal{P}$. Una “predicción probabilística” es cualquier medida de probabilidad $Q \in \mathcal{P}$.

Definición 2.1 Una función score es cualquier función valuada real ampliada $S : \mathcal{P} \times \Omega \rightarrow \overline{\mathbb{R}}$, de modo que $S(\cdot, Q)$ es \mathcal{P} -cuasi-integrable para todo $Q \in \mathcal{P}$.

Una propiedad deseable en toda FS es que esta sea *propia*, es decir, que la única forma de minimizar la pérdida esperada sea reportando $Q = P$. Más aún, se desea que toda función score sea *estrictamente propia*, que en palabras vagas significa que la mínima pérdida esperada sea única. Formalmente:

Definición 2.2 La función score es *propia* con respecto a \mathcal{P} si, para toda $P, Q \in \mathcal{P}$, el

2.4. Utilidad y pérdida, conceptos de teoría de la decisión

score esperado $S(P, Q)$ se minimiza en $Q = P$. Además, S es estrictamente propia si este es el único mínimo: $S(P, Q) > S(P, P)$ para $Q \neq P$.

El score esperado lo entenderemos aquí como una cantidad negativamente orientada, es decir, preferiremos aquellas Q con cuyos scores sean los menores.

Otra característica de las funciones *score propias*, es que cualquier transformación válida conducirá a una nueva FS con las mismas propiedades, es decir, será *propia* o más aún, *estrictamente propia*.

Definición 2.3 S_1 y S_2 son equivalentes si $S_2(Q, x) = cS_1(Q, x) + h(x)$, donde $c > 0$ es una constante y h es un función \mathcal{P} -integrable.

Se dice que una FS es *local* si $S(x, Q)$ depende de la distribución predictiva Q , solo a través de su comportamiento en un entorno infinitesimal a la observación x .

La teoría matemática de las FSs sigue en continuo desarrollo, y el rango de sus aplicaciones es muy amplio. Sus incipientes aplicaciones sucedieron en la meteorología (Brier, 1950), actualmente se dan también en la selección de modelos en inferencia bayesiana (Dawid *et al.*, 2015).

2.4.2. Medidas de divergencia

La esperanza de una FS propia está íntimamente ligada a una medida de distancia entre distribuciones de probabilidad conocida como *divergencia de Brègman*, la cual es una generalización de la divergencia de Kullback-Leibler (Richmond *et al.*, 2008). De acuerdo a Parry *et al.* (2012), asociada a cualquier FS propia, existe una función de entropía (generalizada) $H(P) = S(P, P)$ y una función de divergencia $d(P, Q) = S(P, Q) - H(P)$. Los elementos de esta última expresión son intercambiables y su interpretación es en el contexto de utilidades o recompensas, opuesto al concepto de funciones de pérdida. Si dos FS difieren únicamente por una función de x , entonces conducirán a funciones de divergencia idénticas, y se dice que son equivalentes.

2.4. Utilidad y pérdida, conceptos de teoría de la decisión

2.4.3. Relación con el concepto de función de pérdida

Cualquier función *score* puede verse como una función de pérdida. Las siguientes líneas tomadas de [Parry et al. \(2012\)](#) explican la relación entre estos dos conceptos. Ejemplificando nuevamente con un juego entre la *Naturaleza* y *Usted*, considere al juego como un problema de decisión con espacios de resultados \mathcal{X} , y un espacio de acción arbitrario \mathcal{A} . Nuevamente la tarea de *Usted* es tomar una decisión, esta vez de la forma $a \in \mathcal{A}$, después la naturaleza revelará x de X y entonces usted estará sujeto a un pérdida $L(x, a)$.

Sea \mathcal{P} una familia de distribuciones gobernando \mathcal{X} tal que para cada $P \in \mathcal{P}$, exista un *acto de Bayes*:

$$a_P := \arg \min_{a \in \mathcal{A}} L(P, a),$$

donde $L(P, a) := E_{X \sim P} L(X, a)$. Al comparar dos acciones a_Q y a_P , después de que los datos sean observados, la acción preferida será aquella para la cual la pérdida esperada *a posteriori* sea la más pequeña. Aquella acción a^* que minimiza la pérdida esperada se le denomina acto de Bayes. Dicho en otras palabras, si el acto de Bayes a_P no es único, arbitrariamente nominaremos otro acto, a_Q . Ahora definimos una FS S

$$S(x, Q) = L(x, a_Q) \quad (x \in \mathcal{X}, Q \in \mathcal{P}).$$

Entonces, $S(P, Q) = L(P, a_Q) \geq L(P, a_P) = S(P, P)$. En consecuencia, S es una función estrictamente propia con respecto a \mathcal{P} . Salvada esta aclaración, de aquí en adelante nos referiremos a las funciones *score* como funciones de pérdida.

2.4.4. Algunas funciones de pérdida univariadas y multivariadas

En la literatura se encuentran un sinnúmero de funciones de pérdida derivado de que cualquier problema de decisión genera una función *score* propia ([Dawid, 1998](#), [Parry et al., 2012](#)). Algunas funciones son expresiones simples y de fácil interpretación, otras en cambio no tienen forma analítica cerrada. Las FPs que se discuten enseguida están directamente relacionadas con el presente trabajo de investigación.

2.4. Utilidad y pérdida, conceptos de teoría de la decisión

Divergencia de Kullback-Leibler univariada y multivariada

El primer ejemplo de FP es la pérdida logarítmica definida como $S(x, Q) = LS(x, Q) = -\log q(x)$, y su valor esperado es $LS(p, q) = \int S(x, Q)p(x)dx$, donde q corresponde a la distribución predictiva y p representa la distribución verdadera.

Al evaluar el valor esperado con respecto a x se obtiene $H(P) = LS(p, p) = -\int p(x)S(x, P)dx$, que corresponde a la entropía de Shannon. La pérdida logarítmica fue propuesta originalmente por [Good \(1952\)](#) y es una FS estrictamente propia. Ha sido ampliamente utilizada aunque no está exenta de críticas por las penalizaciones infinitas que da en ciertos casos ([Selten, 1998](#)).

La pérdida logarítmica conduce a la divergencia de Kullback-Leibler ([Kullback y Leibler, 1951](#)), que mide la divergencia en distribución de Q a P :

$$KL(Q||P) = LS(q, p) - LS(p, p) = \int \log \frac{p(x)}{q(x)} p(x) dx.$$

La divergencia de Kullback-Leibler tiene algunas propiedades interesantes, por ejemplo, la $KL(Q||P) \geq 0$, y es 0, si y solo si $q(x) = p(x)$. Esto significa que es una función de pérdida estrictamente propia; la divergencia de KL es una función asimétrica ya que $KL(Q||P) \neq KL(P||Q)$. Además, si X y Y son dos variables aleatorias independientes, y P y Q son funciones de distribución conjuntas, es decir, $p(x, y)$ y $q(x, y)$, entonces $KL_{XY}(P||Q) = KL_X(P||Q) + KL_Y(P||Q)$, lo cual significa que la entropía relativa de P y Q corresponde a la suma de las entropías marginales. Finalmente la divergencia de KL es invariante bajo transformaciones de las variables aleatorias. En [Cover y Thomas \(2006\)](#) pueden consultarse los detalles de estas y otras propiedades.

Para introducir la generalización de la divergencia de Kullback-Leibler al contexto multivariado, sea $\mathbf{Z} \in \mathbb{R}^m$ un vector aleatorio con distribución $f_{\mathbf{Z}}(\mathbf{z})$, la entropía de Shannon es $H(\mathbf{Z}) = -E[\log f_{\mathbf{Z}}(\mathbf{z})] = -\int_{\mathbb{R}^m} f_{\mathbf{Z}}(\mathbf{z}) \log f_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z}$. Ahora suponga que $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^m$ son dos vectores aleatorios con distribuciones $f_{\mathbf{X}}(\mathbf{x})$ y $f_{\mathbf{Y}}(\mathbf{y})$ respectivamente, asumiendo que tienen el mismo soporte. La entropía cruzada que compara la medida de información en \mathbf{Y} con respecto a \mathbf{X} está definida como $C(\mathbf{X}, \mathbf{Y}) = -E[\log f_{\mathbf{Y}}(\mathbf{y})] = -\int_{\mathbb{R}^m} f_{\mathbf{X}}(\mathbf{x}) \log f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{x}$. Note que $C(\mathbf{X}, \mathbf{X}) = H(\mathbf{X})$, sin embargo $C(\mathbf{X}, \mathbf{Y}) \neq C(\mathbf{Y}, \mathbf{X})$ a menos que $\mathbf{X} \stackrel{d}{=} \mathbf{Y}$.

2.4. Utilidad y pérdida, conceptos de teoría de la decisión

Tal como en el caso univariado, la divergencia de Kullback-Leibler entre la distribución de \mathbf{X} y \mathbf{Y} se plantea como

$$KL(p(\mathbf{x})||q(\mathbf{y})) = E[\log(f_{\mathbf{X}}(\mathbf{x})/f_{\mathbf{Y}}(\mathbf{y}))] = \int_{\mathbb{R}^m} \log \frac{f_{\mathbf{X}}(\mathbf{x})}{f_{\mathbf{Y}}(\mathbf{y})} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}.$$

Note que $KL(p(\mathbf{x})||q(\mathbf{y})) = C(\mathbf{X}, \mathbf{Y}) - H(\mathbf{X})$ y que $KL(p(\mathbf{x})||q(\mathbf{x})) = 0$. Por otra parte $KL(\mathbf{X}, \mathbf{Y}) \neq KL(\mathbf{Y}, \mathbf{X})$ a menos que $\mathbf{X} \stackrel{d}{=} \mathbf{Y}$, y tal como en el caso univariado la divergencia de Kullback-Leibler no es simétrica (Contreras-Reyes y B. Arellano-Valle, 2012).

Continuous Ranked Probability Score y Energy Score

Otra función *score* es la denominada *Continuous Ranked Probability Score* (CRPS) que de aquí en adelante nos referiremos a esta simplemente como FP CRPS. No traduciremos el término a español para evitar una traducción poco estética.

La función CRPS es una FP definida en términos de distribuciones predictivas acumuladas, y está definida como

$$CRPS(F, x) = \int_{-\infty}^{\infty} [F(y) - \mathbf{1}(y \geq x)]^2 dy, \quad (2.13)$$

donde X es una variable aleatoria y F representa la función de distribución acumulada (fda) de X , tal que $F(y) = Pr(X \leq y)$, x es una observación y F representa la fda asociada con una predicción probabilística empírica. El término $\mathbf{1}(y \geq x)$ denota una variable indicadora que toma el valor 1 si $y \geq x$, y 0 de otro modo.

La FP CRPS mide la diferencia entre la fda predicha y la ocurrida. El valor más pequeño que toma es cero cuando la predicción es perfecta, por tanto es una FS propia. La función CRPS en 2.13 se puede escribir de forma equivalente como:

$$CRPS(F, x) = E_F|X - x| - \frac{1}{2}E_F|X - X'|, \quad (2.14)$$

2.4. Utilidad y pérdida, conceptos de teoría de la decisión

donde X y X' son copias independientes de una variable aleatoria con fda F y primer momento finito (Baringhaus y Franz, 2004, Gneiting y Raftery, 2004). La FP CRPS es una generalización de la FP error absoluto y por lo tanto proporciona una forma directa de comparar un amplio rango de predicciones probabilísticas usando una medida sencilla (Grimit *et al.*, 2006). De acuerdo a Gneiting y Raftery (2007), las aplicaciones de esta función han sido limitadas ya que no siempre se tienen soluciones analíticas de la expresión en 2.13.

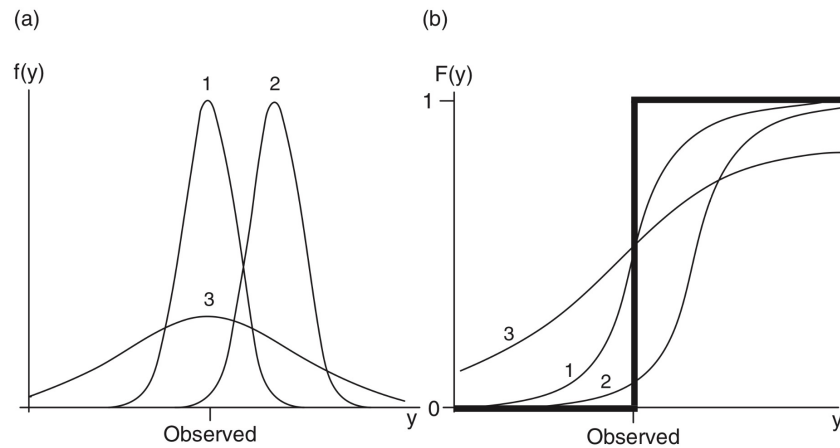


Figura 2.3: Ilustración de la función CRPS. En (a) se ilustran tres funciones predictivas para un resultado observado y , mientras que en (b) se presentan las respectivas funciones de distribución acumulada junto con la función de salto.

La figura 2.3 ejemplifica la FP CRPS⁶. En 2.3-(a) se presentan tres funciones predictivas $f(y)$ para un solo evento cuyo valor observado es y . Las fdp 1 y 3 están centradas en la observación pero la distribución 1 concentra toda su masa de probabilidad alrededor de y , en contraste en la distribución 3 la incertidumbre es mucho mayor. Otra fdp para y es la número 2 que aunque tiene poca incertidumbre, la distribución está centrada lejos de y . Por lo tanto la distribución 1 tendrá la menor pérdida. En 2.3-(b) se presentan las tres fda respectivas así como la función de salto. Dado que la FP CRPS corresponde a la integral del cuadrado de las diferencias entre las respectivas fda's y la función de salto, entonces la fda 1 es la que más se aproxima a la función salto y por lo tanto tendrá la menor pérdida. Es importante resaltar que la FP CRPS es estrictamente propia (Gneiting y Raftery, 2007).

La generalización al contexto multivariado de la FP CRPS es el *energy score* (*EnergyS*). El *EnergyS* mide la distancia entre distribuciones de vectores aleatorios. La distancia es cero si y solo si las distribuciones son idénticas (Székely y Rizzo, 2013). Sea $\mathcal{P}_\beta \in (0, 2)$ la

⁶tomada de Wilks (2011)

2.4. Utilidad y pérdida, conceptos de teoría de la decisión

clase de medidas de probabilidad de Borel P en \mathbb{R}^m tal que $E_F \|\mathbf{X}\|^\beta$ sea finita, donde $\|\cdot\|$ corresponde a la norma ℓ_2 . [Gneiting y Raftery \(2007\)](#) definen el *EnergyS* como:

$$ES(F, \mathbf{x}) = E_F \|\mathbf{X} - \mathbf{x}\|_\beta - \frac{1}{2} E_F \|\mathbf{X} - \mathbf{X}'\|_\beta, \quad (2.15)$$

donde \mathbf{X} y \mathbf{X}' son vectores aleatorios independientes con distribución $F \in \mathcal{P}_\beta$. Cuando $\beta = 1$ y $m = 1$, el *EnergyS* se reduce a la FP CRPS. Igual al caso univariado, es difícil obtener expresiones analíticas del *EnergyS*, sin embargo la expresión 2.15 puede aproximarse mediante simulación Monte Carlo.

Funciones de pérdida asimétricas: LinLin y MALF

En determinadas circunstancias interesa penalizar de distinta forma los errores por sobreestimación y subestimación. Esto se logra empleando alguna función de pérdida asimétrica.

Considere una variable aleatoria cuantitativa $X \in \mathcal{X}$, donde X tiene función de distribución $F(X)$ tal que $F(X) = Pr(X \leq x)$. Sea $\alpha \in (0, 1)$ el α -ésimo cuantil de X de modo que $F^{-1}(\alpha) = \inf[x : \alpha \leq F(x)]$. El escenario más común es cuando $\alpha = 0.5$ y representa la mediana. En otras palabras, la mediana es el valor de x bajo el cual X cae con una probabilidad de 0.5.

Al definir $e = X - x$, siendo x una realización de X , entonces e representa la desviación calculada para un cuantil α , y esto representado mediante una función de pérdida se escribe como

$$L(X, x) = (\alpha - \mathbf{1}(e < 0))e, \quad (2.16)$$

donde $\mathbf{1}$ denota una variable indicadora que vale 1 cuando se cumple que $(e < 0)$ y 0 de otro modo. Esta función es conocida como FP lineal-lineal (LinLin) en estadística, y en otros contextos la denominan *check loss* o *tick loss* ([Giacomini y Komunjer, 2005](#)). Esta FP se ha empleado en finanzas ([Lee, 2008](#)) y también en hidrología ([Zellner, 1986](#)).

La FP LinLin es asimétrica por construcción. La asimetría de las penalizaciones se aprecia en la figura 2.4 para diferentes valores de α . El primer registro de aplicación documentado

2.4. Utilidad y pérdida, conceptos de teoría de la decisión

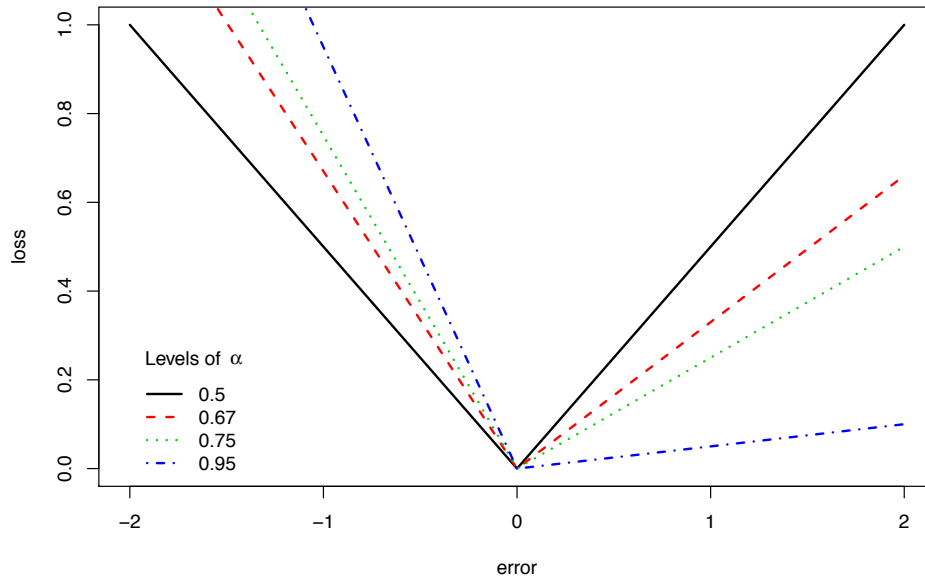


Figura 2.4: Función de pérdida LinLin para diferentes niveles de α

se da en (Granger, 1969).

En hidrología se ha utilizado al momento de dimensionar la altura de la cortina de una presa de almacenamiento, donde la subestimación de la altura de la cortina puede resultar en un problema mucho más serio que la sobrestimación. La sobrestimación incrementa los costos de construcción, sin embargo la subestimación puede causar desbordamiento y colapso de la presa, ocasionando incuantificables pérdidas humanas y materiales.

Su generalización al contexto multivariado es conocida como Función de pérdida asimétrica multivariada (MALF, por sus siglas en inglés ⁷) (Komunjer y Owyang, 2011). La FP MALF está definida como

$$L_{\beta}(\boldsymbol{\tau}, \mathbf{e}) = (\|\mathbf{e}\|_{\beta} + \boldsymbol{\tau}'\mathbf{e})\|\mathbf{e}\|_{\beta}^{\beta-1}, \quad (2.17)$$

donde $\mathbf{e} = (\mathbf{X} - \mathbf{x})'$ es el vector de desvíos entre \mathbf{x} y \mathbf{X} , $\|\mathbf{e}\|_{\beta} = (|e_1|^{\beta} + |e_2|^{\beta} + \dots + |e_n|^{\beta})^{1/\beta}$ y cuando $\beta = 2$, $\|\mathbf{e}\|_2$ corresponde a la norma Euclídeana (ℓ_2). Note que $\mathbf{e} \in \mathbb{R}^m$, y el vector n -dimensional ($1 < \boldsymbol{\tau} < 1$) controla el grado de la asimetría, cuando $\boldsymbol{\tau} = \mathbf{0}$ la

⁷MALF: *Multivariate Asymmetric Loss Function*

2.5. Índices de selección

FP es simétrica. Si $\beta = 1$ la función se reduce a:

$$L_1(\boldsymbol{\tau}, \mathbf{e}) = |\mathbf{e}| + \boldsymbol{\tau}'\mathbf{e}, \quad (2.18)$$

y en el caso univariado, haciendo $\tau = 2\alpha - 1$, tiene como caso particular a la FP LinLin.

Asimetría y dependencia. Consideremos el caso cuando nos interesa obtener los errores al predecir dos variables X_1 y X_2 . Suponiendo $\beta = 2$ se tiene

$$L_2(\boldsymbol{\tau}, \mathbf{e}) = e_1^2 + e_2^2 + (\tau_1 e_1 + \tau_2 e_2)(e_1^2 + e_2^2)^{1/2}.$$

La figura 2.5-(izquierda) presenta las isocurvas de esta función bivariada, mientras que 2.5-(derecha) se grafica la suma de $L_2(\tau_1, e_1)$ y $L_2(\tau_2, e_2)$ para distintos valores de $\boldsymbol{\tau}$. Esto ilustra que la función de pérdida MALF no es separable, es decir,

$$L_2(\boldsymbol{\tau}, \mathbf{e}) \neq L_2(\tau_1, e_1) + L_2(\tau_2, e_2),$$

a menos que $\tau_1 = \tau_2 = 0$. En otras palabras, la función de pérdida bivariada difiere de la simple suma de las pérdidas individuales. Por lo tanto, al minimizar la función de pérdida *n*-variada $L_\beta(\boldsymbol{\tau}, \mathbf{e})$ producirá un vector de errores \mathbf{e}^* cuyas coordenadas e_i^* no necesariamente se minimizan utilizando $L_\beta(\tau_i, e_i)$; así $L_\beta(\boldsymbol{\tau}, \mathbf{e})$ captura tanto la asimetría como la dependencia entre las coordenadas de \mathbf{e} (Komunjer y Owyang, 2011).

2.5. Índices de selección

El método estándar de selección multirasgo emplea índices de selección (IS). Un IS es una combinación lineal de los valores fenotípicos predichos. El objetivo de un índice de selección es predecir el mérito genético $H = \mathbf{w}'\mathbf{a}$, donde $\mathbf{a}' = (a_1, a_2, \dots, a_t)$ (t es el número de rasgos) es el vector de los VC de individuos candidatos a la selección, y $\mathbf{w}' = (w_1, w_2, \dots, w_t)$ es el vector de ponderación de los rasgos (Ceron-Rojas *et al.*, 2015). Un ejemplo de IS es el índice de Smith (Smith, 1936). Sea $\mathbf{p}' = (p_1, p_2, \dots, p_t)$ un vector que

2.5. Índices de selección

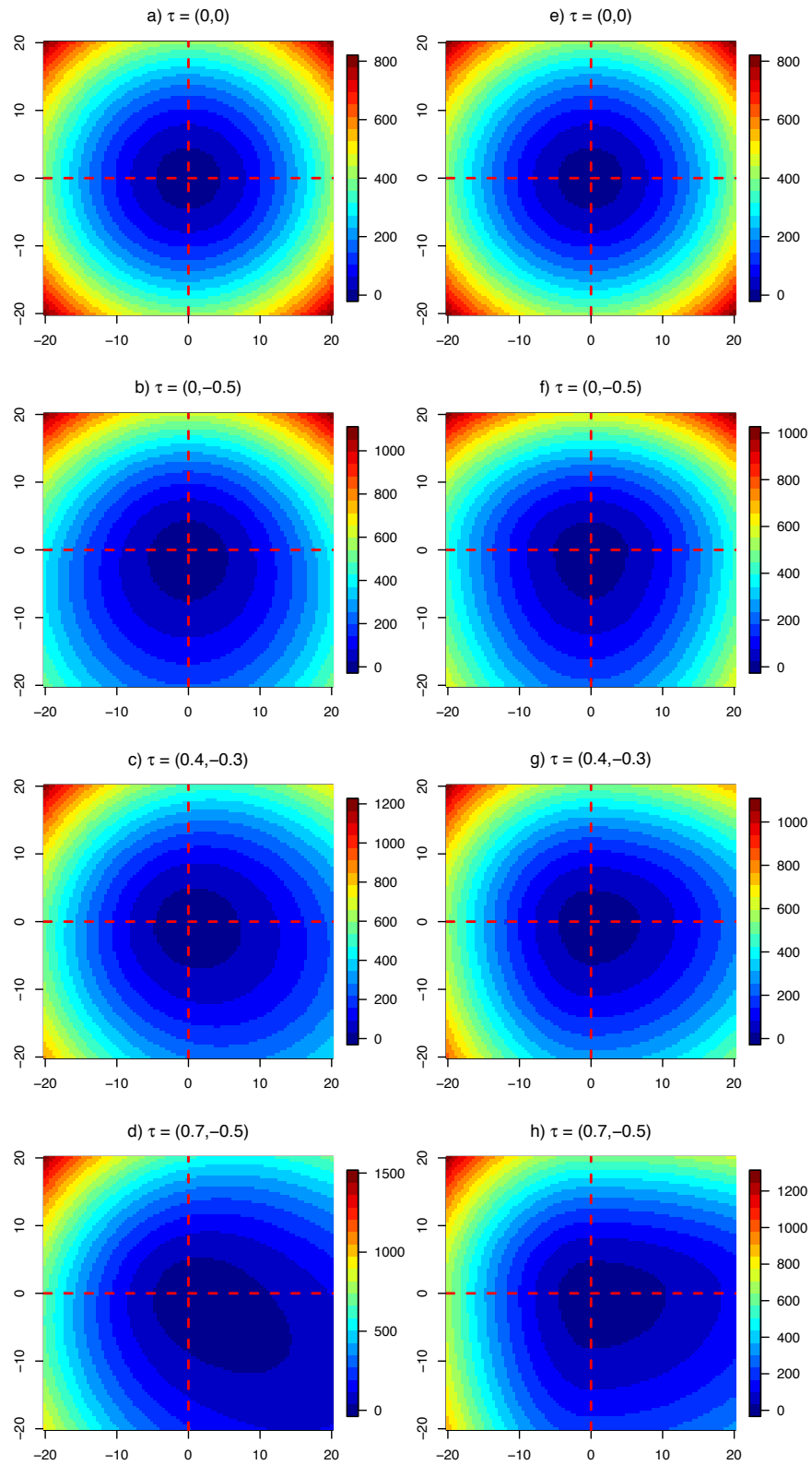


Figura 2.5: Gráficos de contorno de la pérdida biviada $L_2(\tau, \cdot)$ con $\tau(\tau_1, \tau_2)'$ (de a hasta d), y la suma de las pérdidas univariadas $L_2(\tau_1, \cdot) + L_2(\tau_2, \cdot)$ (de e hasta h).

2.5. Índices de selección

contiene los valores fenotípicos para cada rasgo de interés, el índice de Smith se expresa como

$$SI = \mathbf{b}'\mathbf{p}, \quad (2.19)$$

donde \mathbf{b} es un vector de ponderación. Para maximizar la correlación entre H y el IS, se hace que $\mathbf{G}\mathbf{w} = \mathbf{K}\mathbf{b}$, donde \mathbf{K} es la matriz de varianzas-covarianzas fenotípica y \mathbf{G} es la matriz de varianzas-covarianzas genotípica entre los rasgos. Esta última expresión se resuelve como $\hat{\mathbf{b}} = \hat{\mathbf{K}}^{-1}\hat{\mathbf{G}}\mathbf{w}$, en el entendido de que se utilizan predicciones para \mathbf{G} y \mathbf{K} . Finalmente, $\hat{\mathbf{b}}$ se sustituye en 2.19 y se seleccionan aquellos individuos cuyo valor del IS sean los mayores. Como el lector ya lo habrá advertido, no es fácil calibrar las ponderaciones para cada rasgo, ya que se necesita tiempo y un análisis detallado para determinarlos apropiadamente.

Otro índice de selección frecuentemente empleado es el ESIM⁸. Este índice resulta de la solución de $(\mathbf{K} - \lambda\mathbf{I})\mathbf{b} = \mathbf{0}$, donde \mathbf{b} corresponde al primer eigen-vector y λ es el eigenvalor de \mathbf{K} , respectivamente. En el ESIM los pesos para cada rasgo corresponden a los componentes del eigenvector \mathbf{b} y ya no los determina el mejorador. Los detalles de este IS se pueden consultar en [Ceron-Rojas et al. \(2008\)](#).

⁸Del inglés *Eigenanalysis Selection Index Method*.

En este capítulo abordaremos el problema de la selección desde un enfoque formal de la teoría de la decisión. Como ya se ha advertido, un problema de decisión está definido en términos de un espacio de resultados, un espacio de acciones, y una función de pérdida (o recompensa) (Dawid, 2007). En el contexto de la SG, el espacio de acciones está conformado por las líneas candidatas a la selección, y las acciones son las elecciones de los individuos que serán padres del siguiente ciclo de mejora. Las funciones de pérdida son el método de selección que penaliza cada acción posible en función de la preferencia del mejorador.

3.1. La selección como un problema de decisión

El proceso de mejoramiento consiste en seleccionar a los mejores individuos con determinados rasgos de interés para cruzarlos entre sí. El cruzamiento permite que los individuos seleccionados intercambien alelos y la diversidad se observa en las generaciones futuras. De este modo se identifican aquellos individuos con las mejores características de ambos padres. Como ya se expuso anteriormente, la respuesta a la selección (R) depende de la heredabilidad (h^2) del rasgo y del diferencial de selección (S), ya que $R = h^2 S$. Cuando la h^2 se aproxima a uno, la respuesta a la selección se aproxima a S . Por lo tanto el diferencial de selección permite al mejorador estimar el progreso esperado antes de llevar a cabo la selección. Esta información ayuda en los programas de mejora a predecir el éxito del método de selección.

La selección por truncamiento que se discutió en la introducción de esta tesis se ilustra en la figura 3.1a, donde una proporción pequeña de individuos (cuyos valores fenotípicos son los más altos) son seleccionados como padres de la siguiente generación. Una vez realizada la

3.1. La selección como un problema de decisión

cuando se cruza entre los individuos seleccionados, el éxito de la selección se medirá en la distribución de los valores fenotípicos de los descendientes, figura 3.1b.

Como se observa, la distribución de los descendientes posee menor variabilidad que la población base, esto es porque se han seleccionado a los individuos superiores (una fracción de la población total). El problema fundamental de este planteamiento es que en ciclos futuros de selección será más difícil encontrar mejores individuos que el promedio, disminuyendo la respuesta a la selección a medida que transcurre el programa de mejora (Oldenbroek y van der Waaij, 2015).

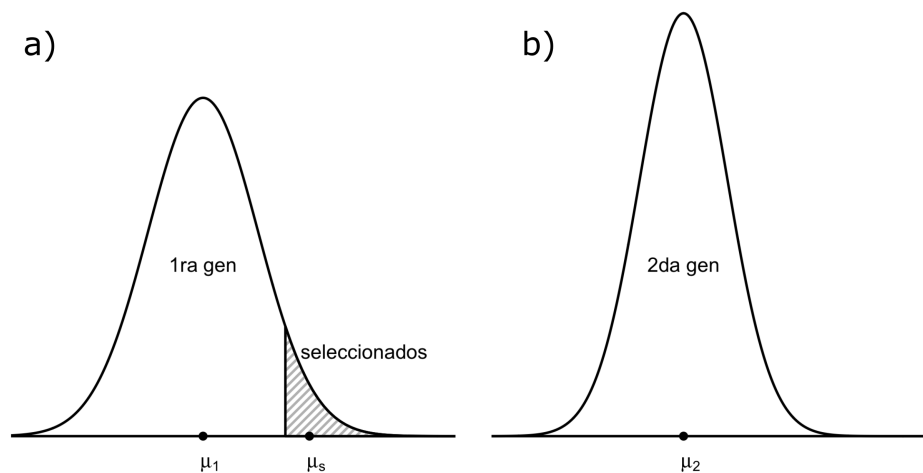


Figura 3.1: a) Distribución parental mostrando la fracción de los seleccionados. b) Distribución de los descendientes resultado de cruzar a los seleccionados en la primera generación

Otro método de selección consiste en optimizar un subconjunto de candidatos a la selección con base en el pedigrí y los valores genéticos aditivos (Shepherd, R.K. y Kinghorn, B.P., 1998). También se han desarrollado algoritmos para la selección de parejas y encontrar el mejor conjunto de padres que maximicen el mérito genético de la progenie. Un enfoque planteado como un problema de optimización cuadrática para minimizar la endogamia y la co-ancestría fue abordado por Wray y Goddard (1994) y por Brisbane y Gibson (1995). Recientemente, Akdemir y Sánchez (2016) propusieron un algoritmo que optimiza las cruas genéticas entre padres empleando SG. Los autores usan un índice de riesgo (de un plan de cruzamiento) y minimizan una función que además contempla la consanguineidad. En la propuesta de Akdemir y Sánchez (2016), se incluyen los efectos de los marcadores, la

3.1. La selección como un problema de decisión

relación genética entre padres, y la matriz de varianzas-covarianzas esperada de la proge-
nie.

Ambos enfoques discutidos - la selección por truncamiento y la selección como un proble-
ma de optimización - pueden plantearse formalmente como un problema de decisión bajo
incertidumbre. Este es el planteamiento en el presente trabajo de investigación, donde el
criterio de decisión utilizado para la selección de los mejores individuos se basa en elegir
aquellos cuya pérdida esperada *a posteriori* sea la mínima, dada una función de pérdida
(FP) que represente las preferencias del mejorador. La FP debe tomar en consideración
factores tales como la media y la variabilidad del rasgo cuando la selección se efectúa en
un solo rasgo; mientras que en la selección multirasgo debe considerar la estructura de de-
pendencia y variabilidad (covarianzas) entre los diferentes rasgos, así como sus respectivas
medias, siempre tratando de maximizar la respuesta a la selección.

Desde la perspectiva Bayesiana de la estadística, el problema de selección de los mejores
padres es con base a minimizar la pérdida esperada *a posteriori*. Supongamos que el valor
fenotípico de un rasgo observable es descrito en términos de un modelo estadístico que
involucra un conjunto de parámetros, para los cuales se dan distribuciones *a priori*. Una
vez que los datos se han observado, la incertidumbre de los parámetros está reflejada en
sus distribuciones *a posteriori*, y la incertidumbre de los valores fenotípicos de los indi-
viduos candidatos está implícita en sus respectivas distribuciones predictivas *a posteriori*.
Cualquier inferencia o decisión en relación a cada candidato debe contemplar por lo tanto
las distribuciones predictivas *a posteriori*. Sin embargo, en SG y fenotípica, las decisiones
se toman con base al *ranking* de los VC predichos (puntuales) e ignorando la incertidum-
bre asociada. Aquí es donde la selección vista como un problema de decisión debe jugar
un papel central. La selección empleando funciones de pérdida puede ser útil, ya que de
este modo el mejorador puede incorporar los criterios antes mencionados, es decir, la va-
riabilidad de los rasgos, la estructura de dependencia entre los rasgos, las medias, y las
distribuciones predictivas *a posteriori* de cada individuo o línea candidata a la selección.

En la revisión de literatura se han presentado algunas medidas de divergencia que pueden
ser útiles adecuándolas al problema de la selección. En términos simples, las medidas de
divergencia cuantifican la divergencia entre las distribuciones de variables aleatorias. En
selección, las variables aleatorias son tanto los valores fenotípicos de los rasgos de interés,
así como los parámetros de los modelos estadísticos empleados. Dadas estas considera-
ciones, es este capítulo se introducen los elementos de incertidumbre que permitan tomar
mejores decisiones en el contexto del mejoramiento de plantas y animales. Las FP son la

3.1. La selección como un problema de decisión

vía propuesta para transitar desde un espacio de acciones (líneas candidatas a la selección), al espacio de resultados (líneas seleccionadas), de modo que las decisiones sean óptimas. En la siguiente sección nos centraremos en el planteamiento teórico de las funciones de pérdida en el contexto univariado (selección en un solo rasgo) y posteriormente, se generalizaran a la selección multirasgo. En ambos casos centraremos la discusión en rasgos cuyos VC deseamos incrementar a medida que transcurre el programa de selección.

3.1.1. Funciones de pérdida univariadas

Sea $Y \in \mathbb{R}$ una variable aleatoria que representa el valor fenotípico de algún rasgo de interés en una población base. Supondremos que $Y \sim N(\mu_1, \sigma^2)$, un supuesto ampliamente utilizado en genética y en genómica para muchos rasgos estudiados. Retomando la idea de la selección por truncamiento o censura, diremos que aquellas realizaciones de Y que sean superiores a un punto de censura y_c son elegidas como padres de la siguiente generación. Por tanto, $Y_s \in \mathbb{R}$ es una variable aleatoria truncada que resulta de censurar por la izquierda en y_c . Por propiedades de la distribución normal, $Y_s \sim N_T(\mu_1, \sigma^2, a = y_c, b = \infty)$, donde N_T denota una distribución normal truncada con parámetros μ_1, σ^2, a y b . Por simplicidad en la notación, de aquí en adelante representaremos como $Y_s \sim N(\mu, \sigma^2, y_c)$. La función de densidad de probabilidad (fdp) de Y_s es igual a $\pi(y|\mu_1, \sigma^2, y_c) = \frac{1}{z\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2\sigma^2}(y - \mu_1)^2\}$, siempre que $y_c \leq y \leq \infty$, donde $z = 1 - \Phi\left(\frac{y_c - \mu_1}{\sigma}\right)$ y $\Phi(\cdot)$ representa la función de distribución acumulada (fda) de una normal estándar.

Por otra parte, considere a $Y_o \in \mathbb{R}$ como la variable aleatoria que representa los valores fenotípicos de una línea candidata a la selección. Dado que Y es normal, es válido suponer que $Y_o \sim N(\mu_2, \sigma^2)$. Note que se asume que tanto Y como Y_o tienen diferentes medias, pero la misma varianza, un supuesto válido, en el sentido de que se desea que se mantenga la varianza genética del rasgo bajo consideración.

La figura 3.2a ilustra el planteamiento expuesto y se esquematiza la distribución base y la distribución teórica de los seleccionados. En la figura 3.2b se presenta la distribución de una línea candidata a la selección. μ_s corresponde a la media de la distribución teórica de los seleccionados, mientras que $S = \mu_s - \mu_1$ representa el diferencial de selección y $R = \mu_2 - \mu_1$ se conoce como la respuesta a la selección. Desde un enfoque formal de decisión, cualquier función de pérdida debe tomar en cuenta nuestra preferencia por aquellas líneas cuya distribución esté centrada en, o lo más próxima a μ_s , de modo que a

3.1. La selección como un problema de decisión

su vez también se garantice la mayor respuesta a la selección posible ($R = h^2 S$).

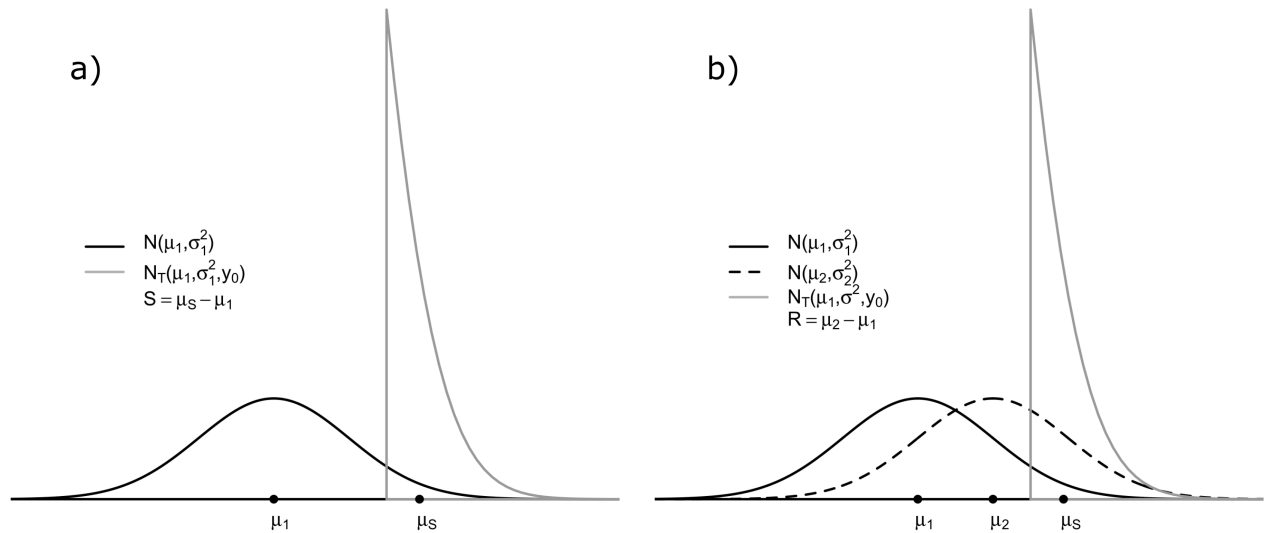


Figura 3.2: Selección por truncamiento. **a)** corresponde a la distribución base (línea negra sólida) y la distribución truncada (línea gris). En **b)** se presenta además la distribución de una línea candidata a la selección (línea negra discontinua). μ_s es la media de la distribución truncada, μ_1 es la media de población base, y μ_2 es la media de la línea candidata.

Para ilustrar el concepto de selección mediante funciones de pérdida, considere la figura 3.3a-b. Como puede apreciarse en a), se presenta la distribución de la población base, la distribución truncada en y_c , y tres fdp de tres líneas candidatas a la selección (líneas discontinuas). La línea discontinua en rojo sería la mejor de las tres, dado que se acerca más a μ_s y además su varianza es parecida a la de la distribución base, por lo tanto una función de pérdida asignaría menor pérdida esperada a dicha línea y esta sería la línea seleccionada. En la figura 3.3-b, se presentan la forma en que penalizan las tres funciones de pérdida univariadas que se proponen en este trabajo y que más adelante discutiremos a detalle.

Hasta este punto pareciera que hemos vuelto más complejo el problema de la selección, sin embargo, la justificación y las ventajas de verlo desde este enfoque se verá más adelante, en la versión multivariada del problema. Procedemos ahora a la propuesta y discusión de tres funciones de pérdida: Kullback-Leibler, LinLin y CRPS.

3.1. La selección como un problema de decisión

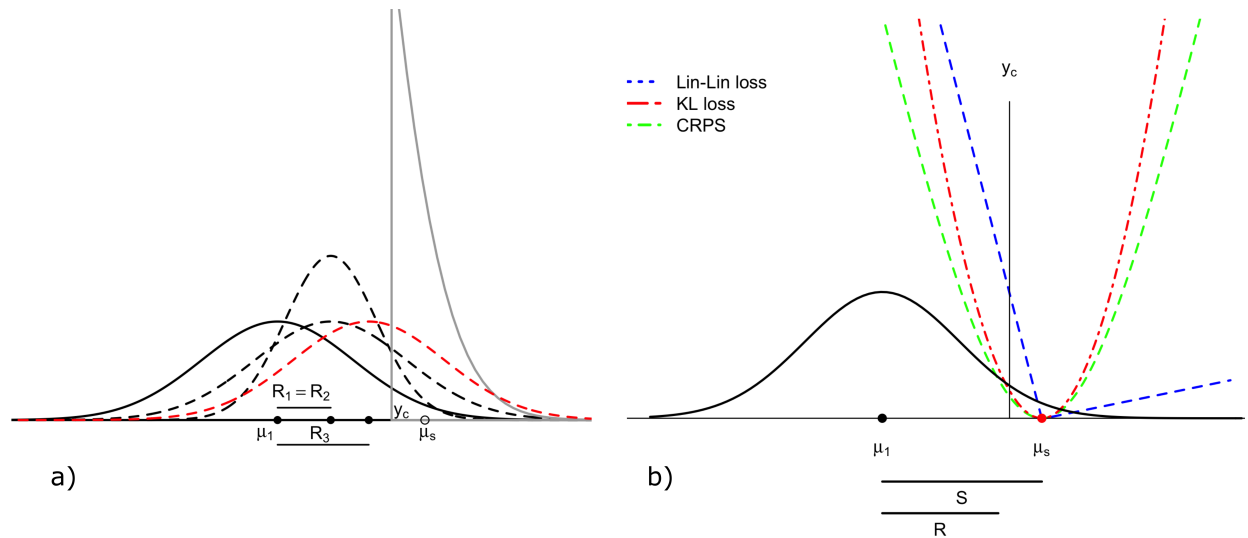


Figura 3.3: Selección usando funciones de pérdida. **a)** La línea sólida en negro representa la población base, la línea gris corresponde a la distribución truncada después de censurar en y_c (representando la preferencia del mejorador), las líneas discontinuas representan tres distribuciones de tres posibles líneas candidatas a la selección. Aquella línea cuya distribución está centrada lo más próxima a la media teórica de los seleccionados μ_s (y que tiene mayor respuesta a la selección R) y varianza similar a la población parental tiene la menor pérdida esperada *a posteriori*. **b)** Representación de las tres funciones de pérdida propuestas. Las funciones de pérdida se minimizan en μ_s para favorecer así a las líneas candidatas con mayor respuesta a la selección. Las pérdidas fueron escaladas restando el valor mínimo para mejor representación.

3.1. La selección como un problema de decisión

Función de pérdida basada en la divergencia de Kullback-Leibler

Dadas la consideraciones anteriores, es posible construir una función de pérdida con base en la divergencia de Kullback-Leibler. Como se vió en el capítulo de revisión de literatura, la divergencia de KL se deriva del *score* logarítmico. La divergencia de KL refleja que tanto divergen dos distribuciones. Si ambas distribuciones son parecidas, la pérdida de información de una con respecto a la otra será mínima, y la pérdida será cero si y solo si las dos distribuciones son idénticas. Por lo tanto, podemos medir que tan cercana son las distribuciones de los VC de las líneas candidatas a la selección con respecto a la distribución teórica que contempla las preferencias del mejorador, y que formalmente hemos definido como una distribución truncada en y_c .

La divergencia de KL para propósitos de SG entre la distribución de Y_s y la distribución de Y_o , y que deseamos minimizar se plantea como

$$D_{KL}(F_{Y_o}, F_{Y_s}) = \int_{y_c}^{\infty} \log \frac{N_T(\mu_1, \sigma^2, y_c)}{N(\mu_2, \sigma^2)} N_T(\mu_1, \sigma^2, y_c) dy. \quad (3.1)$$

Después de evaluar la integral en 3.1 y simplificar algebraicamente (ver Anexo A1.1), la función de pérdida Kullback-Leibler se expresa como

$$D_{KL}(F_{Y_o}, F_{Y_s}) = -\log(z) + \frac{1}{2\sigma^2} [(\mu_s - \mu_2)^2 - (\mu_s - \mu_1)^2], \quad (3.2)$$

donde $\mu_s = E(y|y \geq y_c) = \mu_1 + \sigma \left[\frac{\phi\left(\frac{y_c - \mu_1}{\sigma}\right)}{1 - \Phi\left(\frac{y_c - \mu_1}{\sigma}\right)} \right]$ corresponde a la media de la distribución normal truncada, ϕ y Φ denotan la fdp y la fda de una normal estándar, respectivamente. Note que la esta función de pérdida se expresa en términos de diferencias de medias al cuadrado escalada por la varianza, más el término $-\log z = \log \frac{1}{Pr(y > y_c)}$, y dado que $\log \frac{1}{Pr(y > y_c)}$ es una transformación monótona de $\frac{1}{Pr(y > y_c)}$, éste término induce penalizaciones más pequeñas a medida que la probabilidad de que $Y_o > y_c$ crezca. Reordenando algunos términos en función de S y R , la ecuación 3.2 se expresa como

3.1. La selección como un problema de decisión

$$D_{KL}(F_{Y_o}, F_{Y_s}) = \log \frac{1}{Pr(y > y_c)} + \frac{1}{2} \left(\frac{(S - R)^2}{\sigma^2} - i^2 \right) \quad (3.3)$$

$$= \log \frac{1}{Pr(y > y_c)} + \frac{1}{2} \{i^2 [h^2(h^2 - 2)]\} \quad (3.4)$$

donde $S = \mu_s - \mu_1$ es del diferencial de selección, $R = \mu_2 - \mu_1$ es la respuesta a la selección, y $i = S/\sigma$ denota la intensidad de selección. El segundo término del lado derecho de la ecuación 3.3 implica que cuando R se aproxima a S y la intensidad de selección se incrementa, la divergencia entre la población teórica de los seleccionados y la distribución de las líneas candidatas se minimiza, y como puede verse en 3.4, la $D_{KL}(F_{Y_o}, F_{Y_s})$ depende de la intensidad de selección y es una función decreciente de la heredabilidad.

La figura 3.3b (línea punteada en rojo) corresponde a la gráfica de la FP Kullback-Leibler expresada en la ecuación 3.2. Note que la función de pérdida se minimiza en la media de la distribución teórica de los seleccionados μ_s , de modo que estará favoreciendo a aquellos individuos cuya distribución esté lo más cercana a la distribución teórica de los seleccionados. La función de pérdida es del tipo simétrica ya que penaliza igual a ambos lados del valor objetivo y estará seleccionando aquellos individuos cuya variabilidad sea en la medida de lo posible, igual que la variabilidad presente en la población base.

Función de pérdida con base en el CRPS

La FP CRPS (*Continuous Ranked Probability Score*) se discutió a detalle en el capítulo de revisión de literatura. El CRPS es una métrica que refleja la divergencia entre dos variables aleatorias en términos de sus respectivas distribuciones acumuladas. Para fines de SG, la FP CRPS permitó asociar una pérdida a cada línea candidata a la selección, en función de sus respectivas distribuciones acumuladas de los VC. A medida que las distribuciones estén lo más cerca posible a la distribución teórica de los seleccionados (que refleja la preferencia del mejorador) tendrán menor pérdida y serán seleccionadas.

La función CRPS para fines de selección genómica se expresa como $CRPS(F_{Y_o}, \mu_s) = \int_{-\infty}^{\infty} (F_{Y_o}(y_o) - \mathbf{1}(y_o \geq \mu_s))^2 dy_o$ o de forma alternativa mediante la expresión propuesta por [Baringhaus y Franz \(2004\)](#),

3.1. La selección como un problema de decisión

$$CRPS(F_{Y_o}, \mu_s) = E_F |Y_o - \mu_s| - \frac{1}{2} E |Y_o - Y'_o| \quad (3.5)$$

donde Y_o y Y'_o son variables aleatorias independientes con distribución $F_{Y_o}(y_o)$. Si mantenemos el supuesto de normalidad para Y_o y Y'_o , con media μ_2 y varianzas σ^2 , i.e., $F_{Y_o}(y_o) = N(\mu_2, \sigma^2)$, y resolvemos las esperanzas en la expresión 3.5 (ver desarrollo en Anexo A2), la FP CRPS queda expresada de la siguiente forma,

$$CRPS(F_{Y_o}, \mu_s) = -\sigma \left[\frac{1}{\sqrt{\pi}} - 2\phi\left(\frac{\mu_s - \mu_2}{\sigma}\right) - \left(\frac{\mu_s - \mu_2}{\sigma}\right) \left(2\Phi\left(\frac{\mu_s - \mu_2}{\sigma}\right) - 1\right) \right] \quad (3.6)$$

$$= -\sigma \left[\frac{1}{\sqrt{\pi}} - 2\phi\left(\frac{S - R}{\sigma}\right) - \left(\frac{S - R}{\sigma}\right) \left(2\Phi\left(\frac{S - R}{\sigma}\right) - 1\right) \right] \quad (3.7)$$

$$= -\sigma \left[\frac{1}{\sqrt{\pi}} - 2\phi(i(1 - h^2)) - (i(1 - h^2)) (2\Phi(i(1 - h^2)) - 1) \right] \quad (3.8)$$

donde ϕ y Φ corresponden a la fdp y fda de una normal estándar. Note que la pérdida que asocia la FP CRPS incrementa a medida que la intensidad de selección lo hace, y es una función decreciente de la heredabilidad del rasgo. La representación de la FP CRPS se presenta en la figura 3.3b (línea discontinua en verde).

Función de pérdida LinLin

Hasta aquí, las funciones de pérdida adaptadas al problema de selección han sido FPs simétricas. Por lo tanto, las penalizaciones a distancias iguales a ambos lados del valor objetivo μ_s son las mismas. Sin embargo, es natural pensar que en la selección, (convencional y en SG) aquellas realizaciones de Y_o mayores a μ_s deben ser penalizadas de manera distinta a aquellas que son menores o iguales; siempre que el objetivo de la selección sea incrementar los VC. Aquí es donde las funciones de pérdida asimétricas deben jugar un rol preponderante, ya que asignan pequeñas penalizaciones a aquellas líneas candidatas a la selección cuyas distribuciones de sus VC estén a la derecha de la distribución teórica de los seleccionados.

3.1. La selección como un problema de decisión

Una función de pérdida asimétrica sencilla es la conocida como pérdida Lineal-Lineal (Lin-Lin), que como su nombre lo indica, induce penalizaciones lineales. Sin embargo, consta de un parámetro de asimetría $\alpha \in (0, 1)$ responsable de inducir diferentes penalizaciones a ambos lados del valor objetivo (Berk, 2011). En selección, la función LinLin queda definida como

$$L(F_{Y_o}, \alpha) = (\alpha - 1(e < 0)) \times e \quad (3.9)$$

donde $e = \mu_S - \mu_2 = S - R = \sigma i(1 - h^2)$. La FP LinLin es también una función decreciente en términos de la heredabilidad del rasgo. Note que cuando $\alpha = 0.5$, la FP LinLin se reduce a una función de pérdida lineal simétrica. La función de pérdida LinLin se ilustra en la figura 3.3b (línea discontinua en azul).

3.1.2. Funciones de pérdida multivariadas

Cuando la selección opera para incrementar los valores económicos y el mérito genético de un cultivo, los programas de mejoramiento son aplicados a varios rasgos de manera simultánea, y no solo a uno (Falconer y Mackay, 1996). Sin embargo, surge un problema cuando algunos rasgos están correlacionados negativamente con otros, algo que conoce como antagonismo, y que a medida que en un rasgo existe ganancia genética es a expensas de sacrificar el progreso en uno o más rasgos.

Para hacer frente a esta situación, los mejoradores emplean con frecuencia IS que asignan pesos subjetivos a cada rasgo. Por ejemplo, el índice de selección de Smith (Smith, 1936) previamente discutido en la revisión de literatura; o propuestas más elaboradas que tratan de evitar asignar pesos económicos subjetivos a cada rasgo, reemplazando dichos pesos por los componentes del eigen-vector resultantes de la descomposición eigen-valor eigen-vector de la matriz de varianzas y covarianzas genéticas (Ceron-Rojas *et al.*, 2008).

Sin embargo, es posible extender el concepto de selección mediante un enfoque formal de divergencias entre distribuciones multivariadas. Para ello, generalizamos cada función de pérdida univariada al contexto multirasgo, reduciendo la subjetividad en el proceso de selección.

3.1. La selección como un problema de decisión

Sea $\mathbf{Y} = (Y_1, Y_2, \dots, Y_t)' \in \mathbb{R}^t$ el vector aleatorio que representa los valores fenotípicos de 1, 2, \dots , t rasgos de interés en la población base. Si suponemos normalidad como en el caso univariado, entonces $\mathbf{Y} \sim MVN(\boldsymbol{\mu}_1, \mathbf{K})$, donde $\boldsymbol{\mu}_1 = (\mu_1, \mu_2, \dots, \mu_t)'$ corresponde al vector de medias poblacionales, y $\mathbf{K}_{t \times t} = \sigma_{ij}$ positiva definida, es la matriz de varianzas y covarianzas que captura la asociación entre los rasgos.

Como en el caso univariado, supondremos que la selección por truncamiento puede extenderse al caso multivariado, es decir, el truncamiento de la distribución base multivariada ocurre en un vector de censura. Sea entonces $\mathbf{Y}_s = (Y_{s1}, Y_{s2}, \dots, Y_{st})' \in \mathbb{R}^t$ el vector aleatorio después del truncamiento en $\mathbf{y}_c = (y_{c1}, y_{c2}, \dots, y_{ct})' \in \mathbb{R}^t$ ocurre. Por lo tanto, $\mathbf{Y}_s \sim TMVN(\boldsymbol{\mu}_1, \mathbf{K}, \mathbf{a} = \mathbf{y}_c, \mathbf{b} = \infty)$, donde $TMVN$ denota a la distribución Normal Multivariada Truncada. Por simplicidad en la notación, de aquí en adelante escribiremos $\mathbf{Y}_s \sim TMVN(\boldsymbol{\mu}_1, \mathbf{K}, \mathbf{y}_c)$. La fdp de \mathbf{Y}_s es

$$\pi(\mathbf{y} | \boldsymbol{\mu}_1, \mathbf{K}, \mathbf{y}_c) = \left(\frac{1}{z} \right) (2\pi)^{-t/2} |\mathbf{K}|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu}_1)' \mathbf{K}^{-1} (\mathbf{y} - \boldsymbol{\mu}_1) \right\}; \quad \mathbf{y} \in \mathbb{R}_{\geq \mathbf{y}_c}^t$$

donde $z = Pr(\mathbf{y} \geq \mathbf{y}_c) = (2\pi)^{-t/2} |\mathbf{K}|^{-1/2} \int_{\mathbf{y}_c}^{\infty} \exp \left\{ -\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu}_1)' \mathbf{K}^{-1} (\mathbf{y} - \boldsymbol{\mu}_1) \right\} d\mathbf{y}$, y $\int_{\mathbf{y}_c}^{\infty}$ es una integral de Riemann t dimensional de \mathbf{y}_c a ∞ , y $\mathbb{R}_{\geq \mathbf{y}_c}^t = \{\mathbf{y} \in \mathbb{R}^t : \mathbf{y} \geq \mathbf{y}_c\}$. La figura 3.4a-b ilustra el escenario cuando la selección opera en dos rasgos simultáneamente. Una vez ocurrida la censura, la distribución resultante refleja nuestra preferencia por aquellas líneas superiores, y corresponde a una distribución bivariada truncada. Esta idea puede generalizarse a cualquier número de rasgos.

Sea ahora $\mathbf{Y}_o = (Y_{o1}, Y_{o2}, \dots, Y_{ot})' \in \mathbb{R}^t$ el vector aleatorio que representa los valores fenotípicos de una línea candidata a la selección, y asumimos que $\mathbf{Y}_o \sim MVN(\boldsymbol{\mu}_2, \mathbf{K})$. También podemos definir $\mathbf{S} = (\boldsymbol{\mu}_s - \boldsymbol{\mu}_1)'_{t \times 1}$ como el vector diferencial de selección y $\mathbf{R} = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'_{t \times 1} = \mathbf{SGP}^{-1}$ como el vector respuesta a la selección, donde \mathbf{G} corresponde a la matriz de varianzas-covarianzas genotípica.

Dadas estas consideraciones, nos enfocaremos de aquí en adelante al planteamiento de las funciones de pérdida multivariadas, la apuesta de selección para multiples rasgos.

Divergencia de Kullback-Leibler multivariada

Similar al caso univariado, una función de pérdida basada en la divergencia de Kullback-Leibler se plantea como

3.1. La selección como un problema de decisión

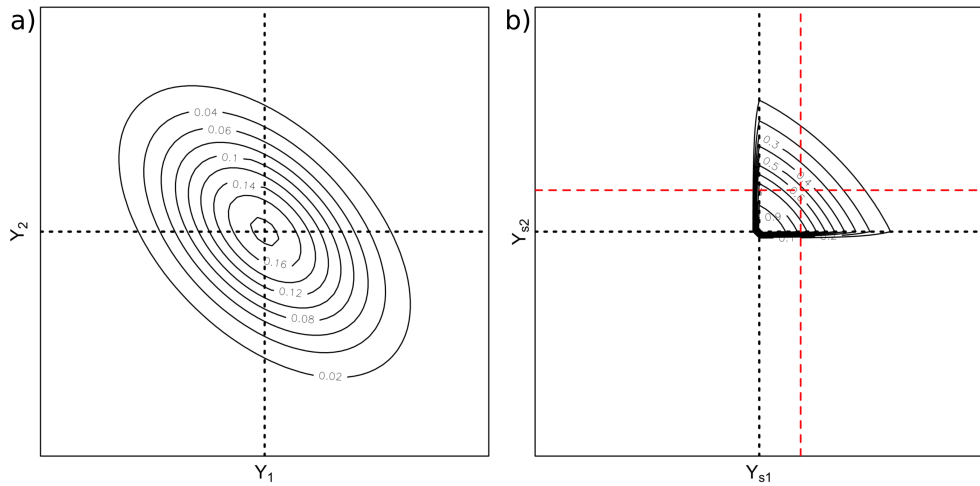


Figura 3.4: Selección por truncamiento para dos rasgos con distribución normal bivariada. a) Distribución base bivariada. b) Distribución teórica de los seleccionados. Las líneas discontinuas en negro representan los valores de censura, mientras que las líneas discontinuas en rojo indican los valores para la media μ_s .

$$D_{KL}(F_{Y_o}, F_{Y_s}) = \int_{\mathbf{y}_c}^{\infty} \log \frac{TMVN(\boldsymbol{\mu}_1, \mathbf{K}; \mathbf{y}_c)}{MVN(\boldsymbol{\mu}_2, \mathbf{K})} TMVN(\boldsymbol{\mu}_1, \mathbf{K}; \mathbf{y}_c) d\mathbf{y}. \quad (3.10)$$

La integral en 3.10 se desarrolla en el Anexo A1.2, y después de manipularla algebraicamente, se resume en

$$D_{KL}(F_{Y_o}, F_{Y_s}) = -\log(z) + \frac{1}{2} [(\boldsymbol{\mu}_s - \boldsymbol{\mu}_2)' \mathbf{K}^{-1} (\boldsymbol{\mu}_s - \boldsymbol{\mu}_2) - (\boldsymbol{\mu}_s - \boldsymbol{\mu}_1)' \mathbf{K}^{-1} (\boldsymbol{\mu}_s - \boldsymbol{\mu}_1)] \quad (3.11)$$

$$= -\log(z) + \frac{1}{2} \mathbf{S}' [(\mathbf{I} - \mathbf{G}\mathbf{K}^{-1}) \mathbf{K}^{-1} (\mathbf{I} - \mathbf{G}\mathbf{K}^{-1}) - \mathbf{K}^{-1}] \mathbf{S}, \quad (3.12)$$

donde z es el factor de normalización de la distribución $TMVN$ definida previamente, \mathbf{S} es vector diferencial de selección y \mathbf{G} es la matriz de varianzas-covarianzas genotípica. Por lo tanto, en la medida en que las matrices de varianzas-covarianzas fenotípica y genotípica tiendan a explicar la misma cantidad de variación y asociación entre los rasgos, $\mathbf{I} - \mathbf{G}\mathbf{K}^{-1} = \mathbf{0}$, la divergencia entre la distribución de los padres y la distribución de la línea candidata tenderá a disminuir. Note que $\mathbf{G}\mathbf{K}^{-1}$ es equivalente a la heredabilidad de los rasgos con la matriz de varianzas-covarianzas genotípica en el numerador, y en el de-

3.1. La selección como un problema de decisión

nominador la matriz de varianzas-covarianzas fenotípicas; por lo tanto cuando $GK^{-1} = I$ la heredabilidad de cada rasgo es 1 y $R = S$ y $\mu_2 = \mu_s$.

Similar al caso univariado, la función de pérdida KL multivariada contiene el término $-\log(z)$ que es una transformación monótona de la probabilidad conjunta para aquellas realizaciones de Y_o superiores a y_c , las cuales tendrán menor penalización. La segunda parte de la expresión involucra distancias cuadráticas en términos de μ_s, μ_2 y μ_1 , escaladas por la inversa de la matriz de varianzas y covarianzas K . La diagonal principal de K actúa como pesos para cada rasgo, mientras que los elementos fuera de la diagonal toman en cuenta la dependencia entre cada rasgo. Dado que K está expresada en términos de su inversa, entonces aquellos rasgos con mayor varianza inducen menos penalización, contribuyendo así a mantener las varianzas genéticas respectivas.

A través de la FP KL multivariada se tiene una forma de selección apropiada que además de capturar la asociación entre los rasgos bajo selección, también asigna los pesos para cada rasgo involucrado en la selección de forma inmediata como función de las varianzas genéticas respectivas, reduciendo así, la subjetividad en la selección. La ilustración de la FP multivariada con base en la divergencia de KL se presenta en la figura 3.5a.

Función *Energy Score*

Cuando la selección se efectúa en más de dos rasgos simultáneamente, se puede utilizar la FP Energy Score (EnergyS), la cual es la generalización de la FP SCRPS (Gneiting y Raftery, 2007, Székely y Rizzo, 2013). De acuerdo al objetivo de la selección la FP EnergyS toma la siguiente forma

$$ES(F_{Y_o}, \mu_s) = E_F \|Y_o - \mu_s\| - \frac{1}{2} E_F \|Y_o - Y'_o\| \quad (3.13)$$

donde $\|\cdot\|$ denota la norma Euclidiana, Y_o y μ_s fueron definidas previamente, y Y'_o corresponde a un vector aleatorio independiente con la misma distribución que Y_o , es decir, F_{Y_o} ; por lo tanto, la FP EnergyS penaliza en función de la distancia euclidiana entre vectores aleatorios. A diferencia de la FP KL, con la FP EnergyS podemos pasar por alto el supuesto de normalidad multivariada, manteniendo únicamente el supuesto de que los valores fenotípicos de los rasgos están en una escala de intervalo. Esta es una gran ventaja ya

3.1. La selección como un problema de decisión

que muchos datos en SG violan el supuesto de normalidad. En la figura 3.5b se ilustra la forma de penalización de la función EnergyS en un contexto bivariado.

Función de Pérdida Asimétrica Multivariada

En la selección multirasgo, si definimos $e = (\boldsymbol{\mu}_s - \boldsymbol{\mu}_2)'_{t \times 1}$ como el vector aleatorio de desvíos de $\boldsymbol{\mu}_2$ con respecto a $\boldsymbol{\mu}_s$, entonces la FP LinLin multivariada puede generalizarse en lo que se conoce como Función de Pérdida Asimétrica Multivariada (MALF¹) (Komunjer y Owyang, 2011). Con fines de SG, la FP MALF se expresa como

$$L_2(F_{Y_o}, \boldsymbol{\mu}_s, \boldsymbol{\tau}) = (\|e\|_2 + \boldsymbol{\tau}'e) \|e\|_2^1, \quad (3.14)$$

donde L_2 implica que usamos la norma euclidiana, por tanto $\|e\|_2 = (|e_1|^2 + |e_2|^2 + \dots + |e_t|^2)^{\frac{1}{2}}$, donde $\boldsymbol{\tau}$ es el vector-parámetro que controla la asimetría en las penalizaciones, y que en la FP LinLin equivale a α . En su versión más simple ocurre cuando se utiliza la norma- L^1 para los desvíos y está expresada como

$$L_1(F_{Y_o}, \boldsymbol{\mu}_s, \boldsymbol{\tau}) = |e| + \boldsymbol{\tau}'e. \quad (3.15)$$

Note que el vector de desvíos e , puede expresarse como $e = S - R = S(I - GK^{-1})$, y es $\mathbf{0}$ cuando las matrices de varianzas-covarianzas genotípicas y fenotípicas tiendan a ser iguales, explicando la misma cantidad de variabilidad y entonces, $GK^{-1} = I$.

En el caso univariado, haciendo $\tau = 2\alpha - 1$, la expresión en 3.15 se reduce a dos veces la función de pérdida LinLin, $L_1(e, \tau) = 2L(e, \alpha)$. En éste trabajo de investigación se utilizó la expresión 3.15. En la Figura 3.5c se presenta la forma en que penaliza la función MALF, y se muestra la asimetría en las penalizaciones.

¹Siglas en inglés de *Multivariate asymmetric loss function*.

3.1. La selección como un problema de decisión

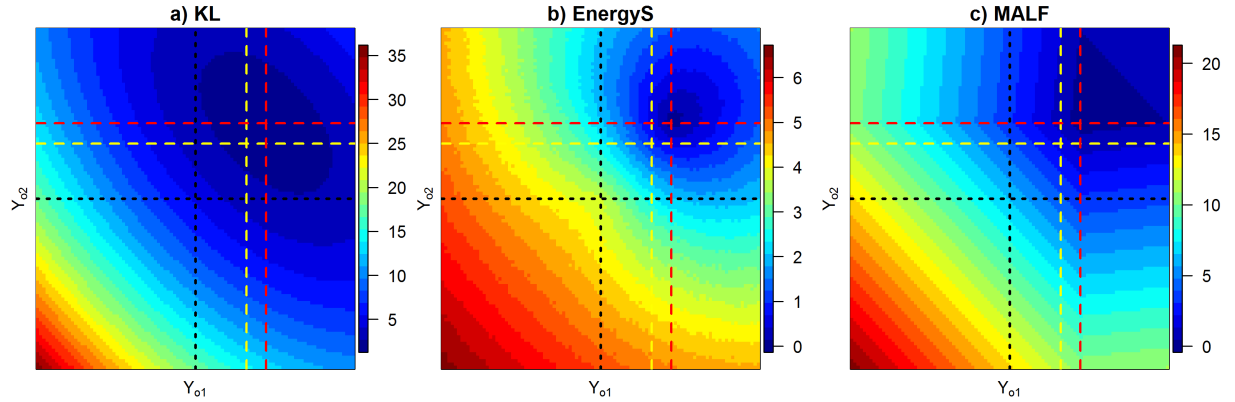


Figura 3.5: Representación bivariada de las funciones de pérdida multivariadas. **a)** Kullback-Leibler (KL), **b)** Energy score (EnergyS) y **c)** Función de pérdida asimétrica multivariada (MALT). La distribución base es $N_2(\boldsymbol{\mu}_1, \mathbf{K})$, con $\boldsymbol{\mu}_1 = (0, 0)'$ (líneas en negro) y $\mathbf{K} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$. Después de truncar en $\mathbf{y}_c = (1.3, 1.3)'$ (líneas en amarillo), $\boldsymbol{\mu}_s = (1.79, 1.79)'$ (líneas en rojo). Las líneas candidatas cerca de la región donde cada función de pérdida se minimiza tendrán las menores pérdidas esperadas *a posteriori*, y serán las seleccionadas en un programa de mejoramiento.

3.1.3. Pérdida esperada *a posteriori*

Cada función de pérdida discutida en esta investigación se puede utilizar ya sea en selección convencional y/o en SG, para seleccionar las mejores líneas con base en las preferencias del mejorador. Enseguida se expone de forma sencilla, concisa y genérica, el enfoque formal para evaluar las pérdidas esperadas *a posteriori* de cada una de las FPs. Nos centraremos en el contexto univariado, sin embargo, en el contexto multirasgo es equivalente.

Dado un modelo de predicción (por ejemplo, la Regresión Ridge Bayesiana) se tiene la distribución *a posteriori* conjunta

$$p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) \propto Lik(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X})p(\boldsymbol{\theta}), \quad (3.16)$$

donde $Lik(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X})$ es la función de verosimilitud, y $p(\boldsymbol{\theta})$ representa la distribución *a priori* para $\boldsymbol{\theta}$, siendo $\boldsymbol{\theta} = (\mu_1, \boldsymbol{\beta}, \sigma^2)$, donde μ_1 corresponde a la media de la población base, $\boldsymbol{\beta}$ (vector) representa los efectos de los marcadores o cualesquiera otros efectos (fijos o aleatorios) y σ^2 la varianza fenotípica. El vector de registros fenotípicos de las líneas en la población base está denotado por \mathbf{y} , mientras que \mathbf{X} corresponde a la matriz de incidencias

3.1. La selección como un problema de decisión

que contiene la información genotípica de cada línea de la población base.

Sea Y_o con distribución $F_{Y_o}(y_o; \boldsymbol{\theta})$, la variable aleatoria representando el valor fenotípico (no observado) de cualquier línea candidata a la selección, cuyo vector de covariables (información genotípica observada) es \mathbf{x}_o . La función predictiva *a posteriori* para y_o está dada por

$$p(y_o | \mathbf{x}_o, \mathbf{y}, \mathbf{X}) = \int_{\boldsymbol{\theta} \in \Theta} f(y_o | \boldsymbol{\theta}, \mathbf{x}_o) p(\boldsymbol{\theta} | \mathbf{y}, \mathbf{X}) d\boldsymbol{\theta}. \quad (3.17)$$

Por lo tanto, dada la información fenotípica y genómica observada $(\mathbf{y}, \mathbf{X}, \mathbf{x}_o)$, todo lo que se desee predecir sobre cualquier línea candidata está dada por 3.17, y cualquier decisión de selección debe tomar en cuenta la incertidumbre contenida en $p(y_o | \mathbf{x}_o, \mathbf{y}, \mathbf{X})$.

De acuerdo a la teoría de la decisión Bayesiana, la decisión óptima es aquella que minimiza la pérdida esperada *a posteriori*. En este caso, una vez observado los valores fenotípicos \mathbf{y} y la matriz de covariables \mathbf{X} se tiene que promediar la FP con base en todo lo desconocido en el modelo, esto es, $\boldsymbol{\theta}$ y lo que es observable pero aún desconocido que corresponde al valor fenotípico de la línea candidata o . Dado que μ_s está en función de μ, β y σ^2 , de aquí en adelante denotaremos a las FPs previamente descritas como $L(F_{Y_o}, \boldsymbol{\theta})$. Por lo tanto, la pérdida esperada *a posteriori* bajo cada FP para el candidato o está dada por

$$\bar{L}_o = \int_{y_o \in \mathcal{Y}} \int_{\boldsymbol{\theta} \in \Theta} L(F_{Y_o}, \boldsymbol{\theta}) f(y_o | \boldsymbol{\theta}, \mathbf{x}_o) p(\boldsymbol{\theta} | \mathbf{y}, \mathbf{X}) d\boldsymbol{\theta} dy_o \quad (3.18)$$

Entonces, para cada candidato o , se tiene una pérdida esperada *a posteriori* \bar{L}_o , y seleccionaremos aquellas líneas cuyas pérdidas esperadas a posteriori sean las menores.

Los valores esperados de las FPs dadas en las expresiones 3.2, 3.6 y 3.9, así como sus respectivas generalizaciones al contexto multirasgo, ecuaciones 3.11, 3.13 y 3.15, pueden aproximarse empleando integración vía Cadenas de Markov Monte Carlo (MCMC) tomando S muestras de cada distribución marginal *a posteriori* de $\mu_1, \mu_2, \mu_s, \sigma^2$ y Y_o (en el contexto univariado), o de $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_s, \mathbf{K}$ y \mathbf{Y}_o (contexto multirasgo). En el enfoque multirasgo, $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_s, \mathbf{K})$.

3.2. Estudio de aplicación de las funciones de pérdida

Para ilustrar la metodología de selección a través de FPs, se utilizaron dos enfoques, el primero en un conjunto de datos reales del cultivo de trigo provenientes de un programa de mejoramiento del CIMMyT², y el segundo en un programa de mejoramiento simulado mediante computadora. Las siguientes secciones detallan cada uno de los enfoques utilizados.

3.2.1. Aplicación en datos de trigo

Para ilustrar la metodología de selección propuesta, se utilizó una base de datos del cultivo de trigo de primavera provenientes de un programa de mejoramiento del CIMMyT. La base de datos consta de 320 líneas con registros fenotípicos de cuatro rasgos: a) rendimiento (GY), b) peso de mil granos (TKW), c) concentración de Zn en el grano (GZnC), y d) concentración de Fe en el grano (GFeC). Cada línea con información de 24,497 marcadores tipo DaRT (Velu *et al.*, 2016). Los cuatro rasgos están ligeramente correlacionados; por ejemplo, GY tiene una correlación de 0.21 con TKW pero cero con los otros rasgos, mientras que GZnC y GFeC tienen una correlación de 0.26.

Para el escenario de las FPs univariadas, se empleó el modelo $\mathbf{y} = \mu\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ donde \mathbf{y} representa los registros fenotípicos, μ corresponde a la media general, \mathbf{X} denota la matriz de diseño que contiene la información de marcadores moleculares, $\boldsymbol{\beta}$ el vector de coeficientes asociado a los marcadores y $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ el vector de residuales del modelo. Este modelo fue parametrizado como un modelo de regresión Ridge Bayesiana y ajustado empleando la librería BGLR (de los Campos y Pérez Rodríguez, 2015) de R (R Core Team, 2016). De esta forma se obtuvieron las distribuciones *a posteriori* mediante cadenas MCMC para μ , σ^2 y $\boldsymbol{\beta}$.

Se utilizaron las 320 líneas de trigo como candidatas a la selección y una vez ajustado el modelo de regresión se seleccionó el top10 % (32 líneas) de las líneas cuyas pérdidas esperadas *a posteriori* resultaron las mínimas, mediante las tres FPs univariadas (KL, CRPS, and LinLin). Los resultados empleando las FPs se compararon con los obtenidos mediante el método estándar de selección (Std), que consiste en seleccionar individuos cuyos VC puntuales son los mayores. En el caso de las FPs multivariadas (KL, Energy Score and

²Centro Internacional de Mejoramiento de Maíz y Trigo

3.2. Estudio de aplicación de las funciones de pérdida

MALF) para la selección multirasgo, se empleó el modelo multirasgo MTM (de los Campos, G. y Grüneberg, A., 2016). El objetivo fue seleccionar las “mejores líneas” cuyo rendimiento para todos los rasgos fuera alto.

El parámetro de asimetría se fijó acorde a la proporción de líneas seleccionadas, por lo tanto, para las FPs LinLin $\alpha = 0.9$; y en el caso de la FP MALF $\tau = (0.9, 0.9, 0.9)'$. Es decir, el complemento de la proporción de selección.

3.2.2. Aplicación en un estudio de simulación

Para evaluar el desempeño de las funciones de pérdida univariadas y multivariadas propuestas en esta investigación, se diseñó y se llevó a cabo un estudio de simulación de un programa de selección. El desempeño de las funciones de pérdida univariadas se comparó con el desempeño del método estándar de selección (Std, abreviado así de aquí en adelante). El método Std implica seleccionar aquellas líneas cuyos valores de cría son mayores. En el caso multivariado, el desempeño de las funciones de pérdida se comparó con el desempeño de dos índices de selección canónicos: índice de Smith (Smith, 1936) y el ESIM (Ceron-Rojas *et al.*, 2008). Enseguida, describiremos de manera general el esquema de simulación utilizado.

El componente genético simulado parte de las leyes de segregación mendelianas para especies diploides. El genoma se conformó de 4000 sitios segregando de forma independiente uno del otro. Para representar la evolución histórica e inducir el equilibrio de ligamiento se simularon 200 generaciones de cruzamiento aleatorio en una población de 1000 líneas segregando para todos los *loci*. La frecuencia alélica se fijó en 0.5.

Rasgos y heredabilidades

En la simulación univariada se simuló un solo rasgo cuantitativo con una heredabilidad de 0.5. Los efectos genéticos se simularon de una distribución gamma con parámetro de forma y escala igual a dos. Los valores fenotípicos se construyeron mediante un modelo aditivo que suma los valores genéticos verdaderos más un término representando el efecto medioambiental, este último consistente con la heredabilidad esperada. Esto es $p_i = g_i + e_i = \sum_{j=1}^m x_{ij}b_j + e_i$, donde m corresponde al número de sitios (o genes), x_{ij} es el genotipo para el j -ésimo gen de la i -ésima línea (codificada como -1 para heterocigótico, y 1 para homocigótico), b_j es el efecto verdadero del j -ésimo gen, y e_i el término medioambiental,

3.2. Estudio de aplicación de las funciones de pérdida

como $e_i \sim N(0, \sigma_g^2(1 - h^2)/h^2)$, donde σ_g^2 representa la varianza genética.

En la simulación multivariada se simularon tres rasgos cuantitativos correlacionados, asumiendo un modelo completamente pleitrópico (Zhe Zhang *et al.*, 2015). En este caso, los efectos genéticos se simularon de una distribución normal multivariada con vector de medias cero y una matriz de varianzas-covarianzas positivamente definida, tal que se indujo las siguientes correlaciones genéticas entre los rasgos simulados al inicio del programa de selección: -0.37 entre el rasgo 1 (T1) y el rasgo 2 (T2); 0.34 entre el rasgo 2 (T2) y el rasgo 3 (T3); y -0.02 entre el T1 y el T3. Los efectos de los sitios segregantes se utilizaron para obtener los verdaderos valores de cría (VVC) y los fenotipos individuales se obtuvieron tomando los respectivos VVC y adicionando un término aleatorio con distribución normal, con una varianza consistente con la heredabilidad esperada (h^2). Para emular rasgos simples y complejos, se fijaron dos valores de heredabilidad en sentido estricto, de 0.3 y 0.6, respectivamente. De aquí en lo subsecuente, cuando hablemos de heredabilidad siempre nos referiremos a la heredabilidad en sentido estricto y la denotaremos como h^2 , esto debido a que en el esquema de simulación solo se contemplaron efectos puramente aditivos y no se incluyeron efectos por dominancia.

Ciclos de selección

Se simularon ciclos de selección mediante un esquema de selección recurrente. La figura 3.6 ilustra a grandes rasgos el esquema de simulación. En cada ciclo de selección, una descendencia compuesta por 100 hermanos completos se derivaron de 200 padres tomados al azar de toda la población. De cada descendiente se generaron 10 líneas de dobles haploides aleatoriamente, por lo que en cada ciclo se obtuvieron en total 1000 líneas. Para cada una de las 1000 líneas se simularon los valores fenotípicos tal como se explicó previamente.

Posteriormente, el 70 % de dichas líneas elegidas al azar se usaron como población base y así entrenar los respectivos modelos de regresión (Modelo MTM para el caso multirasgo, o la Regresión Rigde Bayesiana (BRR) en selección de un solo rasgo). El restante 30 % de las líneas se consideraron como el conjunto de líneas candidatas a la selección, y cada línea como la unidad de selección.

Después se seleccionó el 10 % del conjunto de candidatas a la selección cuyas pérdidas esperadas *a posteriori* fueran las mínimas (FPs univariadas o multivariadas) o bien con el método estandar de selección (Std en selección de un solo rasgo) o los índices de selección (selección multirasgo). Las líneas seleccionadas bajo los diferentes criterios se recombinaron.

3.2. Estudio de aplicación de las funciones de pérdida

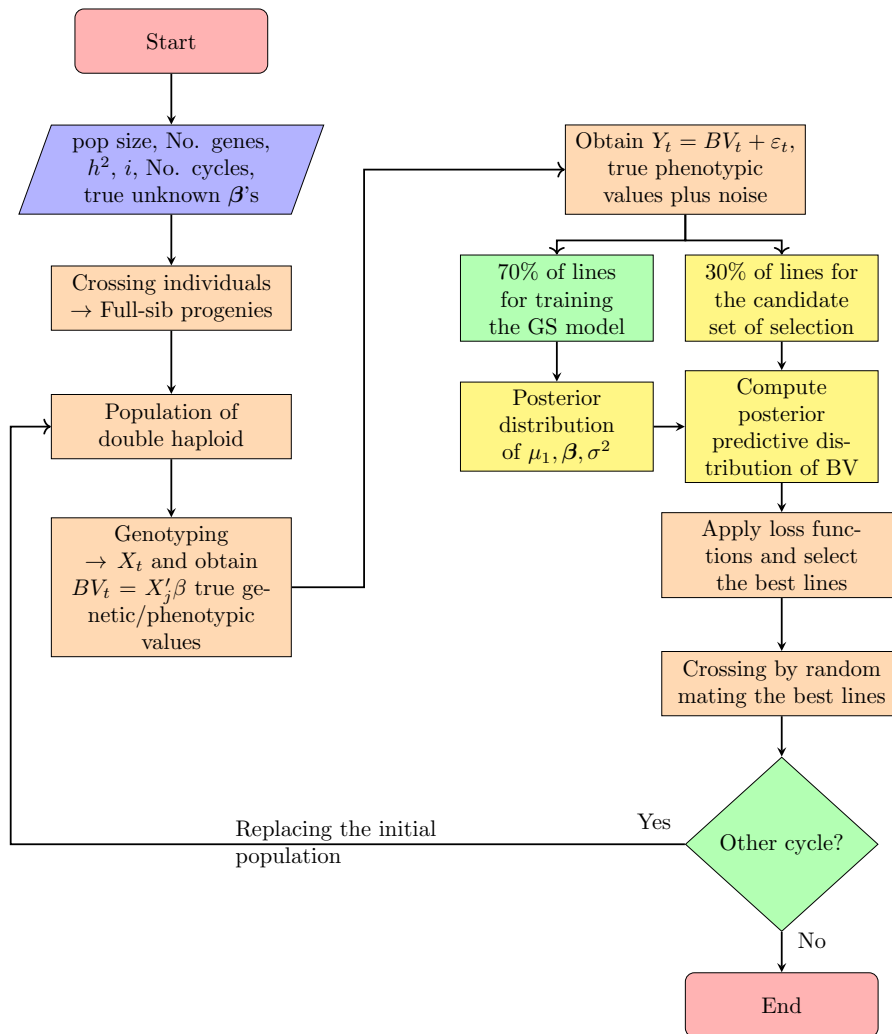


Figura 3.6: Esquema de simulación empleado.

3.2. Estudio de aplicación de las funciones de pérdida

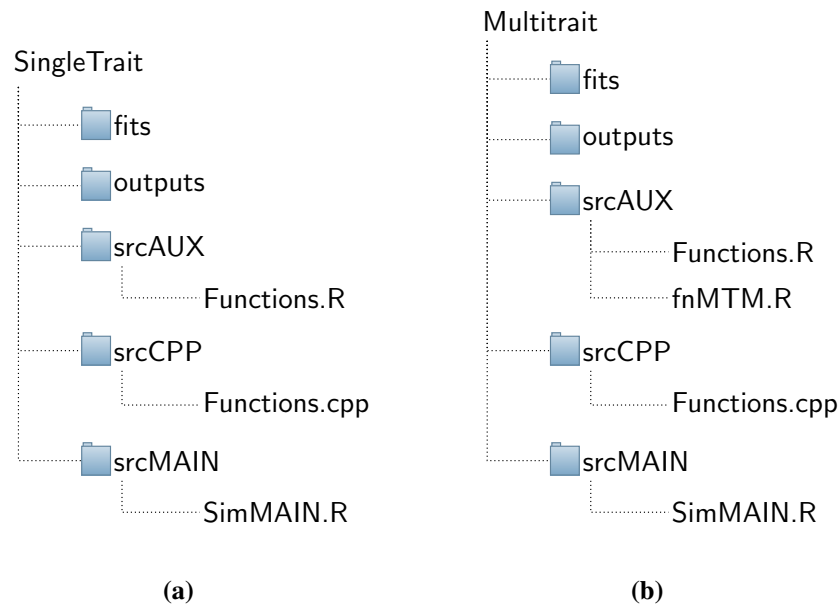


Figura 3.7: Árbol de carpetas utilizadas en el esquema de simulación

ron por cruzamiento aleatorio y así se recobró el tamaño de la población de 1000 líneas para el siguiente ciclo de selección, y así sucesivamente, ciclo a ciclo. La pérdida esperada *a posteriori* se evaluó aproximando la expresión 3.18 utilizando las últimas 10,000 muestras MCMC resultantes de eliminar las primeras 30,000 cadenas y un adelgazamiento a distancia dos. En la simulación univariada se probaron dos proporciones de selección, 10 % y 30 %; mientras que en la simulación multivariada se probó únicamente con el 10 %. Ambos porcentajes con respecto al total de líneas candidatas a la selección.

El parámetro de asimetría en la FP LinLin en la simulación univariada se fijó como $\alpha = 0.9$ que corresponde al complemento de la proporción de selección (también llamada presión de selección) empleada, que fue del 10 %, o bien $\alpha = 0.7$, cuando dicha presión fue del 30 %; mientras que en su versión multivariada MALF, se utilizó un vector de asimetría $\tau = (0.9, 0.9, 0.9)'$ dado que la proporción de selección fue del 10 %. El punto de truncamiento se fijó como el $q_{0.9}$ (cuantil muestral 0.9) de la distribución base, mientras que en el esquema multivariado, el vector de truncamiento correspondió al $q_{0.97}$. Ambos valores corresponden aproximadamente al valor fenotípico del último individuo rankeado en función de su VC (método estándar de selección).

Los resultados se presentan en el siguiente capítulo, y estos corresponden a los resúmenes

3.2. Estudio de aplicación de las funciones de pérdida

después de 20 repeticiones del esquema de simulación previamente descrito. Las simulaciones se implementaron en el lenguaje C++, y fue compilado, vinculado y ejecutado usando el lenguaje de programación estadística R versión 3.3.3 (R Core Team, 2016) aprovechando las facilidades del paquete Rcpp (Eddelbuettel y François, 2011). En la figura 3.7a-b se ilustra la estructura del árbol de carpetas empleadas en el estudio de simulación. El árbol parte de una carpeta principal que contiene cinco carpetas. En la carpeta `fits` se guardaron las cadenas MCMC de las distribuciones *a posteriori*, en el folder `outputs` se alojaron los archivos `*.csv` con las medias y varianzas poblacionales, así como otras cantidades resumen en cada ciclo de selección. La carpeta `srcAUX` alojó algunas funciones auxiliares escritas en lenguaje R. Concretamente, en el archivo `Functions.R` se codificó la programación de las FPs. En los Anexos A3.1 y A3.2 se presentan únicamente el código de las FPs univariadas y multivariadas. El archivo `fnMTM.R` contiene el algoritmo para ajustar el modelo Multirasgo. En el folder `srcCPP` se alojó un archivo con código C++ que realiza la parte intensiva de cruzamiento de las líneas seleccionadas; finalmente, el folder `srcMAIN` alojó el archivo principal (`SimMAIN.R`), que es un script de R que une todo el código de simulación.

4.1. Resultados en datos de trigo utilizando funciones de pérdida univariadas

Aquí se presentan los resultados de la selección correspondientes a las FPs univariadas (KL, CRPS y LinLin) y se comparan con los obtenidos mediante el método estándar de selección (Std). Los resultados se interpretan como si la selección se diera en un rasgo a la vez, es decir, independiente del resto y para un ciclo de selección.

Los boxplots que se presentan en la figura 4.2a-d corresponden al promedio de los valores de cría estimados de las líneas seleccionadas bajo cada criterio de selección (FPs y el método Std) para cada rasgo. Como se aprecia, no se observan diferencias sustanciales en ningún rasgo. Esto obedece a que muy pocas líneas (números en paréntesis) cambiaron empleando las FPs versus el método estándar.

La FP KL fue la que arrojó más líneas diferentes al momento de seleccionar en tres de los cuatros rasgos. Concretamente, en los rasgos GY, TKW y GFeC seleccionó cuatro líneas diferentes en cada rasgo con respecto al método Std, y solo una línea en el rasgo GZnC. En contraste, la FP LinLin fue la que seleccionó menos líneas diferentes en comparación al método Std, ya que en los rasgos GY y GFeC solo seleccionó una línea diferente en cada rasgo, tres líneas diferentes en el rasgo TKW, y ninguna para el rasgo GZnC. Finalmente, la FP CRPS obtuvo un desempeño intermedio a las FPs anteriores en comparación con el método estándar, dado que esta seleccionó 3, 1, 1 y 2 líneas diferentes en los rasgos GY, TKW, GZnC y GFeC, respectivamente.

A pesar de que pocas líneas resultaron diferentes, estas pueden inducir diferencias a largo plazo, es decir, en ciclos subsecuentes de selección, suponiendo que a partir de aquí se

4.1. Resultados en datos de trigo utilizando funciones de pérdida univariadas

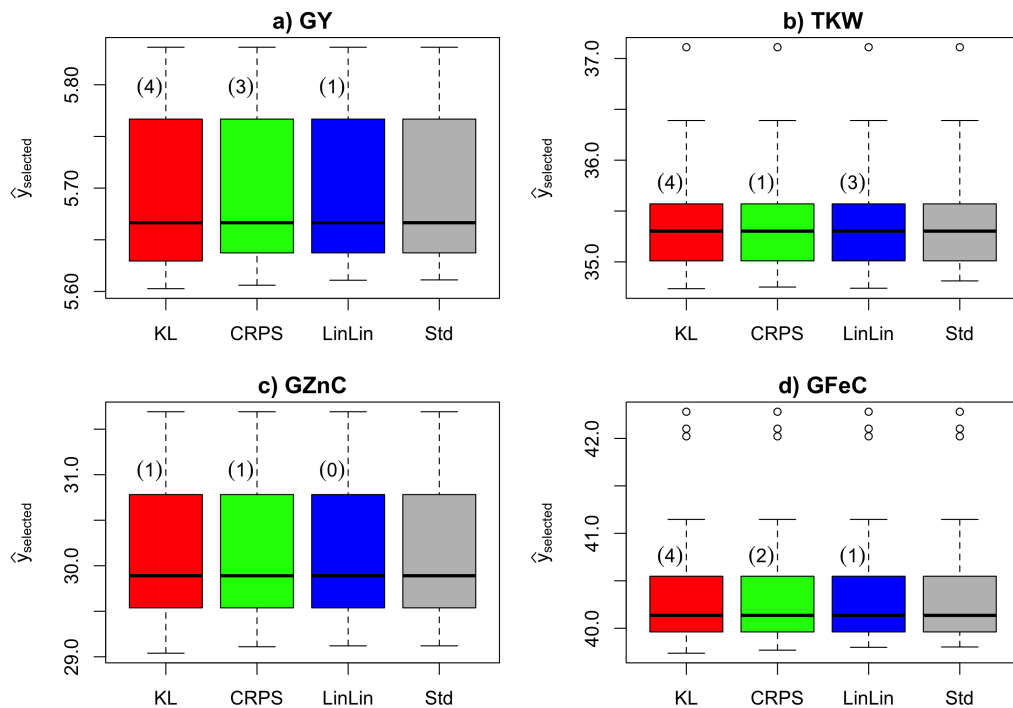


Figura 4.1: Boxplots de los valores de cría estimados para los datos de trigo (con cuatro rasgos) al seleccionar el 10 % de las líneas candidatas a la selección empleando las funciones de pérdida univariadas Kullback-Leibler (KL), Continuous Ranked Probability Score (CRPS) y Lineal-Lineal (LinLin), así como empleando el método estándar de selección (Std) para los rasgos **a)** rendimiento (GY), **b)** peso de mil granos (TKW), **c)** concentración de Zn en el grano y **d)** concentración de Fe en el grano. Los valores en paréntesis corresponden al número de líneas que las funciones de pérdida seleccionaron y el método estándar no.

4.2. Resultados en datos de trigo utilizando funciones de pérdida multivariadas

avance el programa de mejoramiento. Es válido especular que en poblaciones más grandes, habrá más líneas diferentes seleccionando con las FPs contra el método estándar, y si esto se repitiese ciclo a ciclo de selección, puede impactar (para mejor) en la media y/o en la varianza de la población.

4.2. Resultados en datos de trigo utilizando funciones de pérdida multivariadas

Como previamente se comentó, en la selección multirasgo lo que se busca es seleccionar las “mejores líneas”, que se desempeñen bien en todos los rasgos de interés para el mejorador, aún en el caso de rasgos complejos y antagónicos. En esta sección se describen los resultados de comparar las FPs multivariadas (KL, EnergyS y MALF) en los datos de trigo.

Para tal efecto, en la figura 4.2a-d se presentan los boxplot de las medias predichas de las líneas seleccionadas para cada FP. En términos generales, la divergencia de KL fue la FP con mejor desempeño ya que el promedio de las líneas seleccionadas en cada rasgo resultó mayor que la media respectiva de cada rasgo (línea en rojo). Las FPs MALF y EnergyS muestran un desempeño similar en todos los rasgos, con una ligera ventaja de la primera en el rasgo GY.

Puntualmente en el rasgo GY cuya correlación muestral con el rasgo TKW es de 0.21 y prácticamente cero con los otros dos rasgos, la FP KL resultó mejor ya que el promedio de los VC de los seleccionados es superior a la media poblacional. La FP MALF y el EnergyS tuvieron desempeño similar en este rasgo, aunque con una ligera ventaja de la primera sobre la segunda. Para el rasgo TWK, las tres FPs obtuvieron un desempeño similar, lo que se constata también en el promedio de los VC, que resultó superior a la media de la población base. En el rasgo GZnC, la FP MALF y el EnergyS se adjudicaron un desempeño ligeramente superior a la FP KL. Finalmente en el rasgo GFeC el promedio de las tres FPs resultó similar.

Adicionalmente a los resultados gráficos anteriores, se efectuó una prueba de comparación de medias multivariada mediante la prueba de la T^2 de Hotelling, cuyos resultados se resumen en la cuadro 4.1. Note que a un nivel de significancia $\alpha \approx 0.005$ para el error tipo I, existen diferencias significativas entre la media de las líneas seleccionadas entre las FPs

4.2. Resultados en datos de trigo utilizando funciones de pérdida multivariadas

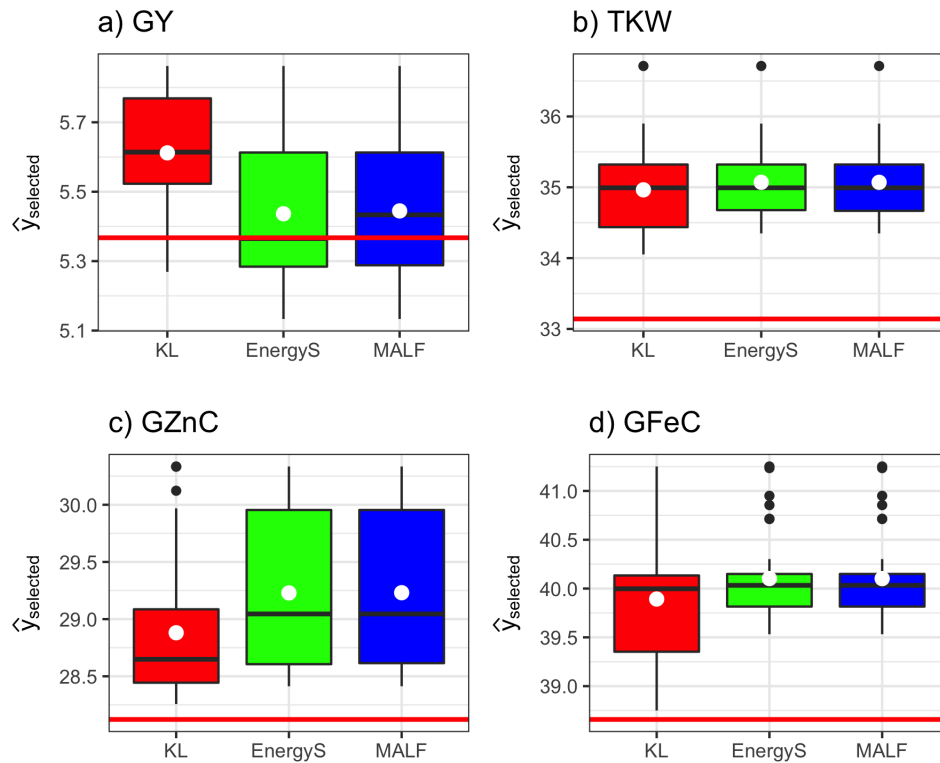


Figura 4.2: Boxplots de los valores de cría estimados para los datos de trigo al seleccionar el 10 % de las líneas candidatas a la selección empleando las funciones de pérdida multivariadas Kullback-Leibler (KL) Energy Score (EnergyS) y Función de Pérdida Asimétrica Multivariada (MALF) para los rasgos **a)** rendimiento (GY), **b)** peso de mil granos (TKW), **c)** concentración de Zn en el grano y **d)** concentración de Fe en el grano. Puntos blancos representan el promedio de las líneas seleccionadas, y las líneas en rojo corresponden a la media de la población base.

Cuadro 4.1: Prueba T^2 de Hotelling para resultados en datos de trigo.

Contraste	No. líneas diferentes	T^2	p-value
KL vs EnergyS	12	4.125	0.005
KL vs MALF	11	3.785	0.008
Energy vs MALF	1	0.021	0.999

4.3. Resultados de estudio simulación

KL y EnergyS. La misma prueba sugiere que existen diferencias al contrastar las FPs KL y MALF pero a un nivel de significancia $\alpha \approx 0.008$. Sin embargo no hay evidencia de que las medias entre las FPs MALF y EnergyS sean diferentes. El cuadro 4.1 también presenta el número de líneas diferentes que se seleccionaron en cada par de FPs. Note que entre la KL y el EnergyS, 12 de 32 líneas (38 %) son diferentes. Entre las FPs KL y MALF, 11 de 32 líneas (34 %) resultaron diferentes. En contraste, entre las FPs EnergyS y MALF solo una línea resultó diferente. La expectativa es que si se cruzan las líneas seleccionadas bajo cada FP, recobrando así las poblaciones, y repitiendo ciclo a ciclo de selección este procedimiento, se espera una ganancia en la media de todos los rasgos a medida que transcurra el programa de mejora, y probablemente con mejor desempeño de las líneas seleccionadas con la FP KL.

4.3. Resultados de estudio simulación

En los siguientes resultados, tanto del estudio de simulación univariada (un solo rasgo) o multivariada (multirasgo), las medias poblacionales en cada ciclo de selección fueron estandarizadas como $(\mu_i - \mu_1)/\sigma_1 = R_i/\sigma_1$. Esta cantidad es adimensional y corresponde a la respuesta a la selección estandarizada, donde μ_i corresponde a la media poblacional en el ciclo i , μ_1 es la media poblacional en el ciclo 1, R_i es la respuesta a la selección en el ciclo i con respecto al primer ciclo de selección, y σ_1 es la desviación estándar poblacional correspondiente al ciclo 1. También las varianzas poblacionales en cada ciclo de selección (σ_i^2) fueron estandarizados con respecto a la varianza poblacional del ciclo 1 (σ_1^2), es decir, σ_i^2/σ_1^2 .

4.3.1. Resultados de funciones de pérdida univariadas

La figura 4.3a-b presenta las tendencias de la media poblacional estandarizada de cada función de pérdida, y del método estándar, separando los ciclos 1-5 y 5-10 para una mejor visualización. Por su parte, la varianza poblacional escalada se ilustra en la figura 4.6a-b, donde se resume los resultados obtenidos cuando la proporción de seleccionados fue del 10 %, de las líneas candidatas a la selección acorde al criterio de mínima pérdida esperada *a posteriori*. Con base en estos resultados, no se encontraron cambios significativos ni en las medias, ni en las varianzas poblacionales al utilizar las FPs como método de selección.

4.3. Resultados de estudio simulación

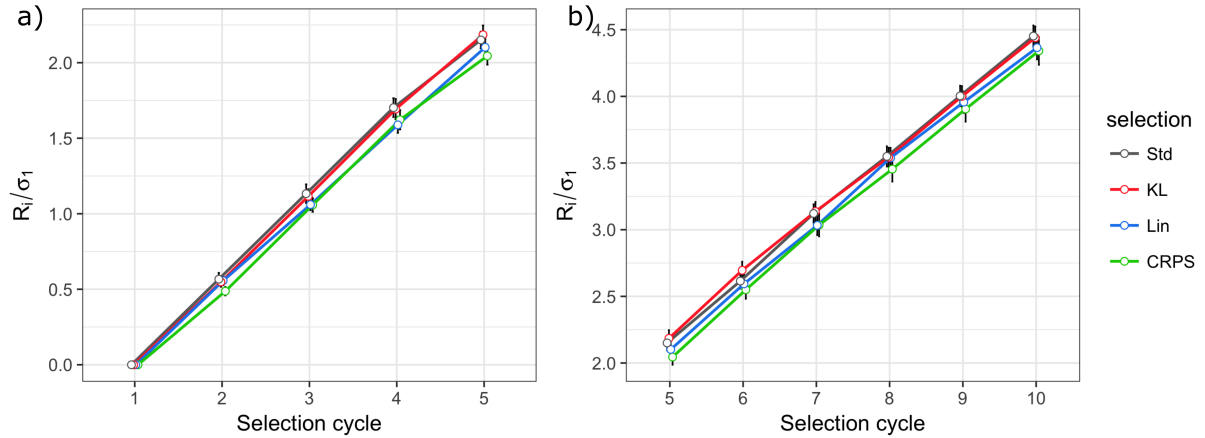


Figura 4.3: Respuesta a la selección estandarizada $R_i/\sigma_1 = (\mu_i - \mu_1)/\sigma_1$. En **a)** se presentan los ciclos de selección 1 al 5, y en **b)** los ciclos 5 al 10. En cada ciclo de selección se seleccionó el 10 % de las líneas candidatas a la selección con base en las mínimas pérdidas esperadas *a posteriori* empleando las funciones de pérdida KL, CRPS, LinLin; y mediante el método estándar de selección (Std). Las líneas seleccionadas se cruzaron entre sí para recobrar el tamaño poblacional en cada ciclo de selección. μ_i y R_i corresponden a la media poblacional y la respuesta a la selección, en el ciclo i , respectivamente; μ_1 y σ_1 denota la media poblacional y la desviación estándar poblacional, en el ciclo 1, respectivamente. Las líneas verticales en negro representan el error estándar de 20 repeticiones del estudio de simulación.

4.3. Resultados de estudio simulación

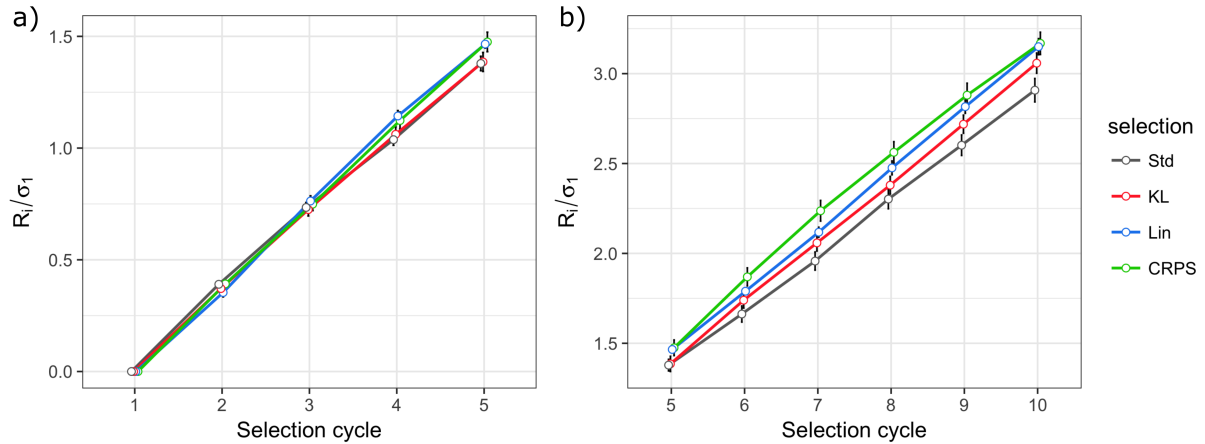


Figura 4.4: Respuesta a la selección estandarizada $R_i/\sigma_1 = (\mu_i - \mu_1)/\sigma_1$. En **a)** se presentan los ciclos de selección 1 al 5, y en **b)** los ciclos 5 al 10. En cada ciclo de selección se seleccionó el 30 % de las líneas candidatas a la selección con base en las mínimas pérdidas esperadas *a posteriori* empleando las funciones de pérdida KL, CRPS, LinLin; y mediante el método estándar de selección (Std). Las líneas seleccionadas se cruzaron entre sí para recobrar el tamaño poblacional en cada ciclo de selección. μ_i y R_i corresponden a la media poblacional y la respuesta a la selección, en el ciclo i , respectivamente; μ_1 y σ_1 denota la media poblacional y la desviación estándar poblacional, en el ciclo 1, respectivamente. Las líneas verticales en negro representan el error estándar de 20 repeticiones del estudio de simulación.

Sin embargo, cuando la presión de selección fue del 30 %, se observa una ligera pero significativa ventaja de los resultados obtenidos con las FPs a medida que los ciclos de selección transcurrieron. Esto último se ilustra en la figura 4.4a-b para la media poblacional estandarizada y en la figura 4.6a-b para la varianza poblacional escalada.

4.3. Resultados de estudio simulación

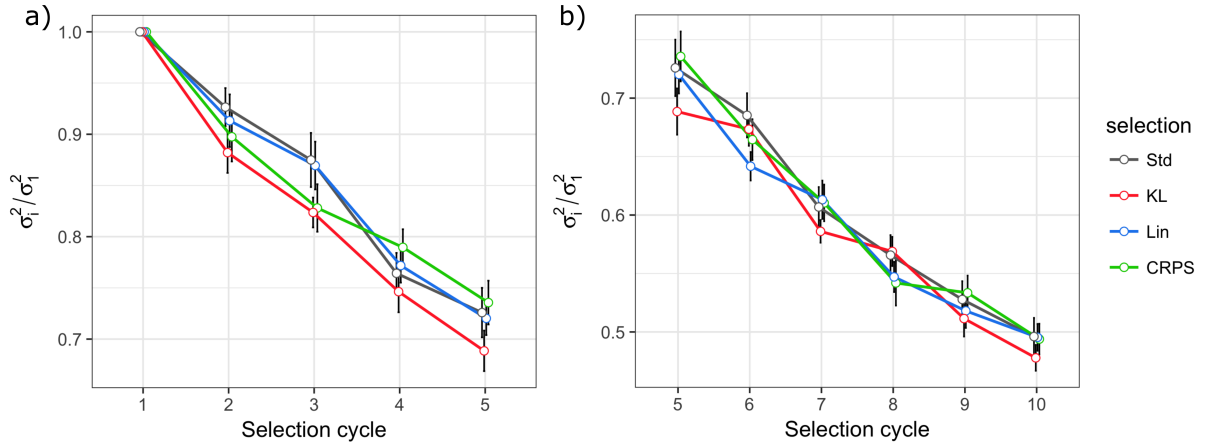


Figura 4.5: Varianza poblacional escalada σ_i^2/σ_1^2 . En **a)** se presentan los ciclos 1 al 5, y en **b)** los ciclos 5 al 10. En cada ciclo de selección se seleccionó el 10 % de las líneas candidatas a la selección con base en las mínimas pérdidas esperadas *a posteriori* empleando las funciones de pérdida KL, CRPS, LinLin; y mediante el método estándar de selección (Std). Las líneas seleccionadas se cruzaron entre si para recobrar el tamaño poblacional en cada ciclo de selección. σ_i^2 y σ_1^2 corresponden a la varianza poblacional en el ciclo i y en el ciclo 1, respectivamente. Las líneas verticales en negro representan el error estándar de 20 repeticiones del estudio de simulación.

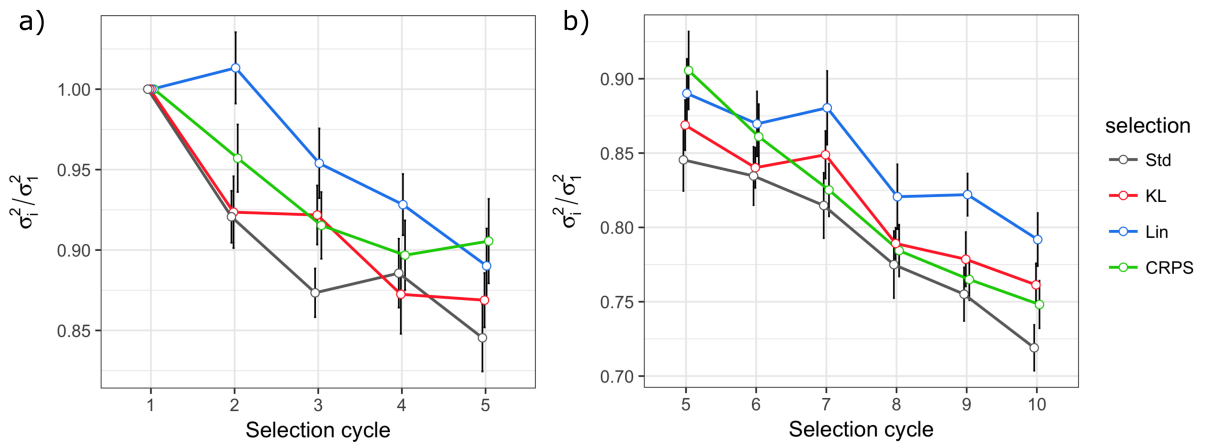


Figura 4.6: Varianza poblacional escalada σ_i^2/σ_1^2 . En **a)** se presentan los ciclos 1 al 5, y en **b)** los ciclos 5 al 10. En cada ciclo de selección se seleccionó el 30 % de las líneas candidatas a la selección con base en las mínimas pérdidas esperadas *a posteriori* empleando las funciones de pérdida KL, CRPS, LinLin; y mediante el método estándar de selección (Std). Las líneas seleccionadas se cruzaron entre si para recobrar el tamaño poblacional en cada ciclo de selección. σ_i^2 y σ_1^2 corresponden a la varianza poblacional en el ciclo i y en el ciclo 1, respectivamente. Las líneas verticales en negro representan el error estándar de 20 repeticiones del estudio de simulación.

4.3. Resultados de estudio simulación

Los gráficos de caja que se presentan en las figuras 4.7-a y 4.8-a corresponden a las medias poblacionales estandarizadas al 10mo ciclo de selección (fin del programa de selección) para las dos proporciones de selección, 10 % y 30 %, respectivamente. Como se aprecia, para el top10 % no se observan diferencias sustantivas entre los resultados de las FPs y el método estándar. En cambio en el top30 % si se aprecia una ligera ventaja en favor de las FPs, dado que las medias de todas las FPs son superiores a la media del método estándar. Las diferencias de mayor a menor son para la FP LinLin, KL y CRPS.

En el cuadro 4.2a-b se presentan los resultados correspondiente a la prueba de t para comparar diferencias en las medias al 10mo ciclo de selección entre las FPs y el método Std, tanto para el 10 % de líneas seleccionadas como para el 30 %. Con base en dicha información se concluye que hay diferencias estadísticamente significativas al seleccionar con las FPs CRPS y LinLin para la presión de selección del 30 %.

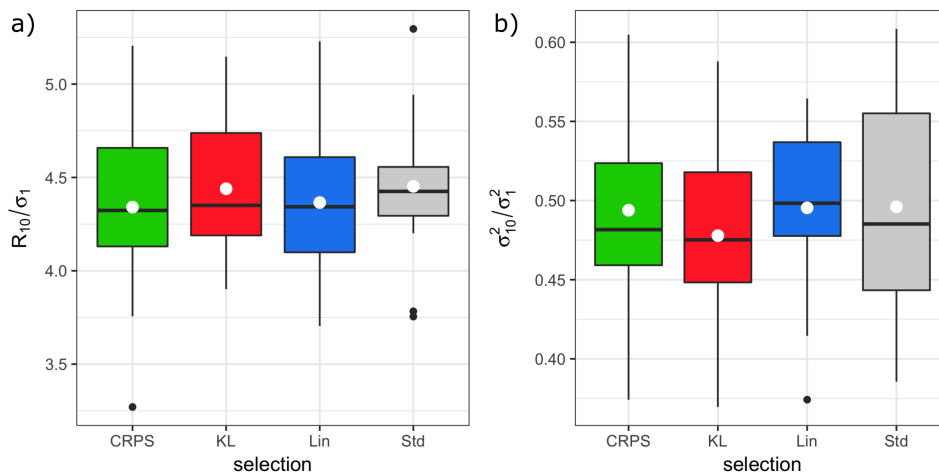


Figura 4.7: Resultados del estudio de simulación univariado al 10mo ciclo de selección del 10 % de proporción de seleccionados. En **a)** los resultados de la media poblacional estandarizada o R_{10}/σ_1 , y en **b)** los resultados de la varianza poblacional escalada o σ_{10}^2/σ_1^2 , empleando las funciones de pérdida KL, CRPS, LinLin, versus el método estándar (Std). Los puntos blancos representan la media y las líneas negras la mediana, de 20 repeticiones de la simulación.

4.3. Resultados de estudio simulación

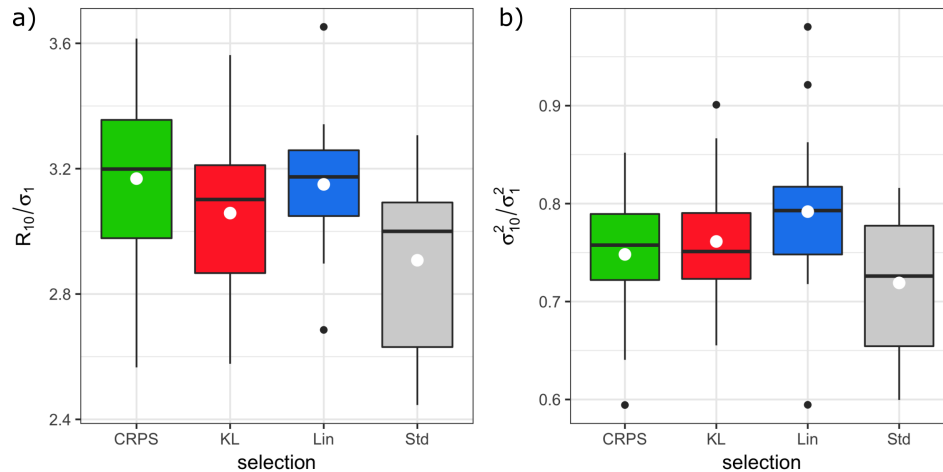


Figura 4.8: Resultados del estudio de simulación univariado al 10mo ciclo de selección del 30 % de proporción de seleccionados. En **a)** los resultados de la media poblacional estandarizada o R_{10}/σ_1 , y en **b)** los resultados de la varianza poblacional escalada o σ_{10}^2/σ_1^2 , empleando las funciones de pérdida KL, CRPS, LinLin, versus el método estándar (Std). Los puntos blancos representan la media y las líneas negras la mediana, de 20 repeticiones de la simulación.

Los resultados correspondientes a la varianza poblacional escalada al fin del programa de selección (10mo ciclo) se ilustran en las figuras 4.7-b y 4.8-b para el 10 % y el 30 % de presión de selección, respectivamente. Note que para el 10 % de seleccionados, los *boxplot* de la figura 4.8-b no muestran ninguna diferencia en las varianzas obtenidas al seleccionar con las FPs versus la varianza obtenida con el método estándar. En contraste, cuando se seleccionó el 30 % de las líneas candidatas, si se observan diferencias sustanciales en las varianzas poblacionales cuando se empleó las FPs. Esto se ilustra en la figura 4.8-b, donde las varianzas poblacionales son mayores que las varianzas obtenida al seleccionar con el método Std.

Para finalizar, el cuadro 4.2c-d presenta la prueba de t para comparación de medias correspondientes a la varianza poblacional al fin del programa de selección, para el 10 % y el 30 % de presión de selección. Los únicos escenarios donde la media de las varianzas poblacionales seleccionado con las FPs fueron superiores a la media de las varianzas poblacionales del procedimiento estándar fue para la FP KL y LinLin, y para el 30 % de presión de selección.

4.3. Resultados de estudio simulación

Cuadro 4.2: Prueba de t para contrastar diferencias en 1) la media poblacional estandarizada y 2) la media de la varianza poblacional escalada, al 10mo ciclo de selección, empleando las funciones de pérdida KL, CRPS y LinLin, así como el método estándar (Std). * se consideran estadísticamente significativas

contraste	a) media del top10 %			b) media del top30 %		
	t	df	p-valor	t	df	p-valor
CRPS vs Std	-0.85	36	0.4	2.9	38	0.006*
KL vs Std	-0.11	38	0.914	1.7	37	0.088
Lin vs Std	-0.73	38	0.469	3.1	34	0.004*
contraste	c) varianza del top10 %			d) varianza del top30 %		
	t	df	p-valor	t	df	p-valor
CRPS vs Std	-0.1	36.6	0.917	1.3	38	0.198
KL vs Std	-0.9	33.8	0.355	2	37.8	0.052*
Lin vs Std	0	34.5	0.973	3.1	37.2	0.003*

4.3.2. Resultados de funciones de pérdida multivariadas

Medias poblacionales

La figura 4.9 muestra la evolución del promedio de la media poblacional estandarizada a través de 10 ciclos de selección después de repetir el esquema de simulación 20 veces, para cada combinación de función de pérdida (o índices de selección) y heredabilidades. Particularmente, la figura 4.9-a muestra la evolución de la media cuando la heredabilidad se fijó en 0.3 para todos los rasgos. En el caso del T1 negativamente correlacionado con el T2, y correlación despreciable con el T3, las funciones de pérdida KL y MALF tuvieron un desempeño similar en todos los ciclos de selección, sus respectivos rendimientos expresados como respuesta a la selección estandarizados, fueron superiores a los índices de selección e inclusive mejores que la función Energy Score.

En el cuadro 4.3, se presentan los promedios de las diferencias entre las medias poblacionales (para cada función de pérdida o índice de selección) en el 10mo ciclo de selección con respecto al primer ciclo cuando la heredabilidad se fijó en 0.3. Note que la función de pérdida KL tuvo una ganancia de 1.583 % contra el 1.211 % que obtuvo la función MALF. Por otra parte, la función de pérdida EnergyS obtuvo una ganancia conservadora de apenas 0.462 %. Por el contrario, los índices de selección tuvieron un rendimiento pobre, ya que no consiguieron el objetivo de incrementar la media poblacional, por el contrario, la media

4.3. Resultados de estudio simulación

disminuyó. Los cambios en la media para el T1 expresados en porcentaje para el ESIM y el índice de Smith fueron de -0.47 % y -1.21 %, respectivamente. A pesar de esto, el ESIM fue el que menos sacrificó la media para este rasgo.

En lo que respecta al segundo rasgo (T2), las tres funciones de pérdida propuestas así como los índices de selección tuvieron un comportamiento satisfactorio. Las medias poblacionales se incrementaron ciclo a ciclo de selección, tal como se aprecia en la figura 4.9-a (en medio). Inspeccionando el cuadro 4.3, en este rasgo, el índice de Smith tuvo el mejor rendimiento, ya que obtuvo una ganancia de 10.141 % al final del programa de selección. En segundo lugar, el ESIM y la función de pérdida Energy Score obtuvieron ganancias similares, de 8.79 % y 8.224 %, respectivamente. Por otra parte las funciones de pérdida MALF y KL tuvieron el menor rendimiento para este rasgo, cuyas ganancias al termino del programa de selección resultaron en 6.493 % y 6.441 %, respectivamente. Hay que recordar que la correlación entre el T1 y el T2 fue de -0.37 aproximadamente, y ambas funciones de pérdida tuvieron el mejor desempeño en el T1; aún así, sus ganancias resultaron positivas para el T2.

Finalmente para el T3, las tres funciones de pérdida así como los índices de selección obtuvieron rendimientos similares ciclo a ciclo de selección (ver figura 4.9-a (abajo)). Las diferencias son muy pequeñas y las tendencias en las medias poblacionales prácticamente se traslapan. En orden decreciente, las ganancias al final de programa de selección fueron de 6.186 % (índice de Smith), 6.14 % (EnergyS), 5.805 % (ESIM), 5.762 % (KL) y 5.499 % (MALF).

Cuando la heredabilidad se fijó en 0.6, los resultados obtenidos fueron similares a los ya descritos previamente. En la figura 4.9-b se presenta la evolución de las medias poblacionales estandarizadas para los tres rasgos. Note que para el T1, las funciones de pérdida KL y MALF tuvieron mejor desempeño que el EnergyS y los dos índices de selección canónicos. Como se aprecia, el EnergyS y el ESIM tuvieron desempeños similares, con ligeras diferencias en favor del segundo. Por otra parte, el índice de Smith tuvo de nuevo un desempeño pobre ya que la media poblacional decreció consistentemente ciclo a ciclo de selección. La incertidumbre medida a través del error estándar, muestra también que el índice de Smith tuvo el comportamiento más inestable, sobre todo al final de los ciclos de selección.

En el cuadro 4.3 se presentan los promedios de las diferencias entre las medias poblacionales en el 10mo ciclo de selección con respecto al primer ciclo. En particular, para el T1,

4.3. Resultados de estudio simulación

estas diferencias fueron de 3.318 % para la KL, 2.347 % para la función MALF, 0.506 % para el EnergyS, 0.829 % para el ESIM, y de -1.501 % para el índice de Smith, confirmando así el pobre rendimiento de este último. En el T2 el mejor desempeño lo obtuvo el índice de Smith (a expensas de sacrificar el T1), seguidos por el EnergyS, el ESIM, la función KL y la MALF. Al fin del ciclo de mejoramiento, las ganancias respectivas fueron de 12.534 % para el índice de Smith, 9.818 % para el EnergyS, 8.832 % para el ESIM, 6.206 % para la KL y finalmente la MALF que obtuvo una ganancia de 5.721 %.

4.3. Resultados de estudio simulación

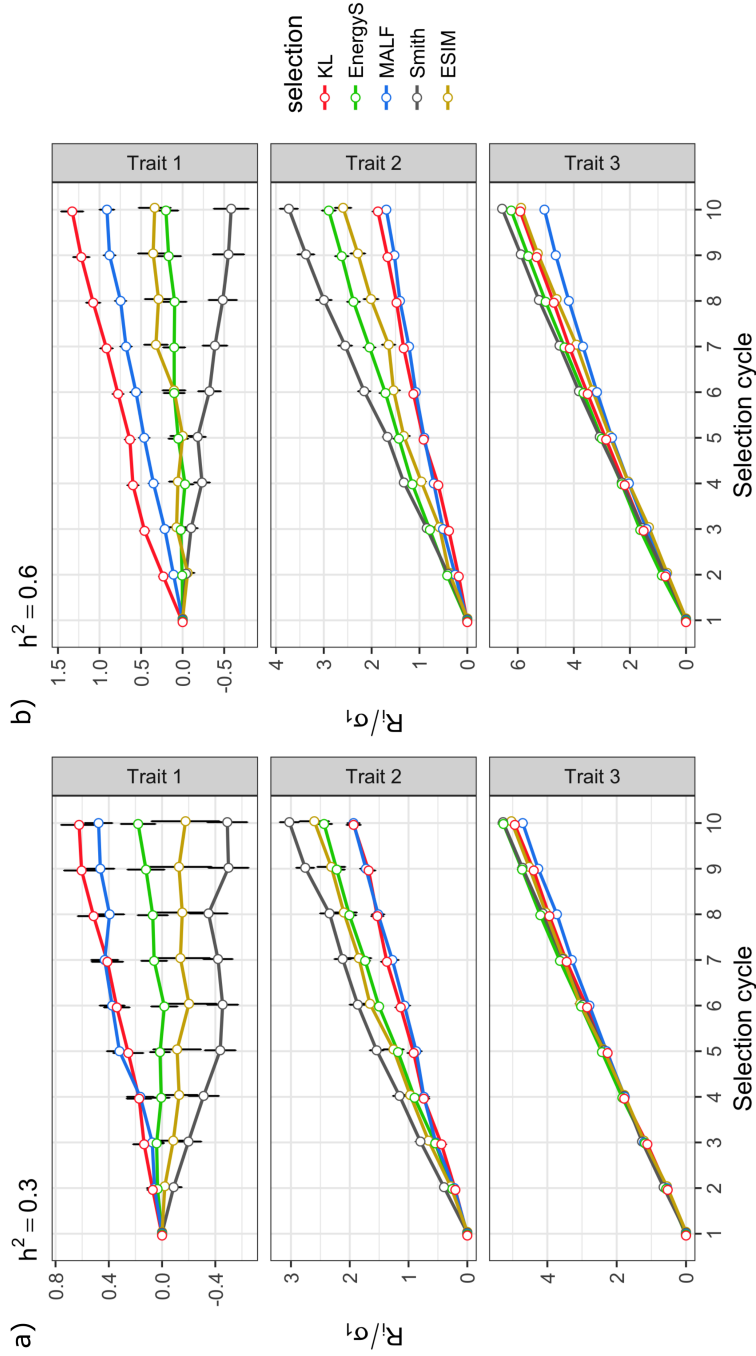


Figura 4.9: Resultados del estudio de simulación multivariado. En **a)** se presenta la media poblacional estandarizada $(\mu_i - \mu_1) / \sigma_1 = R_i / \sigma_1$ para los ciclos de selección cuando la heredabilidad para todos los rasgos fue de 0.3, y en **b)** cuando la heredabilidad de fijó en 0.6. En cada ciclo de mejoramiento se seleccionó el top-10% del total de líneas candidatas a la selección usando la funciones de pérdida multivariadas: Kullback-Leibler (KL), Energy Score (EnergyS), y la Función de pérdida Asimétrica Multivariada (MALF); y mediante los índices de selección: Smith y ESIM. μ_i y R_i corresponde a la media poblacional y la respuesta a la selección en el ciclo i , respectivamente; μ_1 y σ_1 es la media poblacional y la desviación estándar poblacional en el ciclo 1, respectivamente. Las líneas verticales (en negro) corresponde al error estándar de R_i / σ_1 resultado de 20 repeticiones del estudio de simulación.

4.3. Resultados de estudio simulación

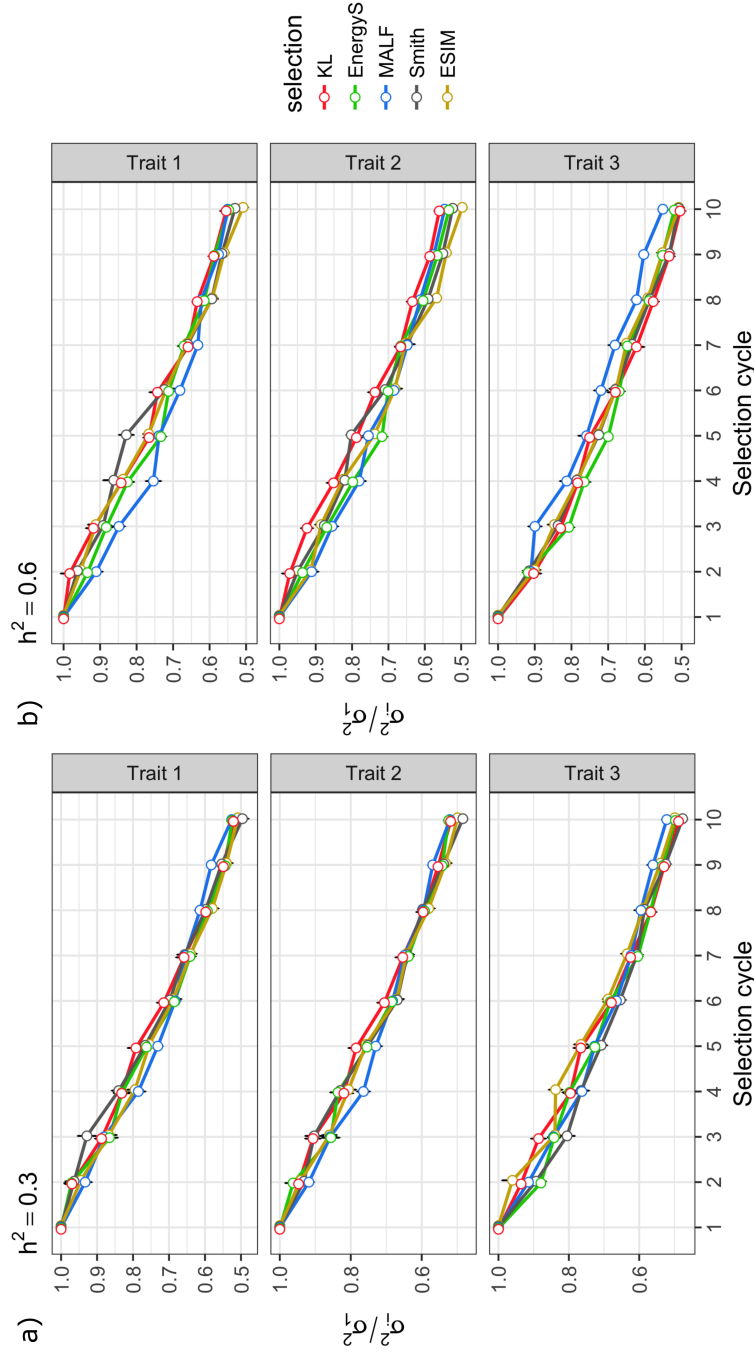


Figura 4.10: Resultados del estudio de simulación multivariado. En **a)** se presenta la varianza poblacional escalada σ_i^2/σ_1^2 para los ciclos de selección, cuando la heredabilidad para todos los rasgos fue de 0.3, y en **b)** cuando la heredabilidad de fijó en 0.6. En cada ciclo de mejoramiento se seleccionó el top-10% del total de líneas candidatas a la selección usando la funciones de pérdida multivariadas: Kullback-Leibler (KL), Energy Score (EnergyS), y la Función de pérdida Asimétrica Multivariada (MALF); y mediante los índices de selección: Smith y ESIM. σ_i^2 y σ_1^2 corresponden a la varianza poblacional en el ciclo i y en el ciclo 1, respectivamente. Las líneas verticales (en negro) indican el error estándar de σ_i^2/σ_1^2 resultado de 20 repeticiones del estudio de simulación.

4.3. Resultados de estudio simulación

Cuadro 4.3: Promedio de las diferencias de las medias poblacionales en el 10mo ciclo de selección menos las medias poblacionales en el primer ciclo de selección para el rasgo 1 (T1), rasgo 2 (T2) y rasgo 3 (T3). Las heritabilidades para todos los rasgos se fijó en 0.3 y 0.6 (Errores estándar respectivos se presentan en paréntesis).

$h^2 = 0.3$			
Loss	T1	T2	T3
KL	1.583 (0.343)	6.441 (0.387)	5.762 (0.12)
MALF	1.211 (0.264)	6.493 (0.344)	5.499 (0.139)
EnergyS	0.462 (0.332)	8.224 (0.477)	6.14 (0.173)
ESIM	-0.47 (0.648)	8.79 (0.849)	5.805 (0.175)
Smith	-1.21 (0.381)	10.141 (0.493)	6.186 (0.153)
$h^2 = 0.6$			
Loss	T1	T2	T3
KL	3.318 (0.327)	6.206 (0.403)	6.935 (0.118)
MALF	2.347 (0.232)	5.721 (0.239)	5.842 (0.121)
EnergyS	0.506 (0.375)	9.819 (0.500)	7.291 (0.139)
ESIM	0.829 (0.491)	8.832 (0.600)	6.833 (0.165)
Smith	-1.501 (0.528)	12.534 (0.633)	7.624 (0.143)

Cuadro 4.4: Promedio de las diferencias correspondientes a las varianzas poblacionales en el 10mo ciclo de selección menos las varianzas poblacionales en el primer ciclo para el rasgo 1 (T1), rasgo 2 (T2) y rasgo 3 (T3). Las heritabilidades para todos los rasgos se fijó en 0.3 y 0.6 (Errores estándar respectivos se presentan en paréntesis).

$h^2 = 0.3$			
Loss	T1	T2	T3
KL	-47.886 (1.678)	-48.128 (1.506)	-51.181 (1.37)
MALF	-47.437 (1.221)	-47.928 (1.496)	-47.82 (1.352)
EnergyS	-47.533 (1.146)	-47.584 (1.372)	-50.287 (1.327)
ESIM	-49.128 (1.637)	-50.108 (1.446)	-50.117 (1.417)
Smith	-50.406 (1.72)	-51.577 (1.27)	-52.335 (0.967)
$h^2 = 0.6$			
Loss	T1	T2	T3
KL	-44.532 (1.672)	-43.975 (1.341)	-49.588 (1.509)
MALF	-44.977 (0.811)	-45.451 (0.827)	-44.91 (1.271)
EnergyS	-45.256 (1.550)	-46.618 (1.456)	-48.025 (1.345)
ESIM	-49.154 (1.416)	-50.296 (1.272)	-49.206 (1.481)
Smith	-46.961 (1.645)	-47.664 (1.63)	-49.205 (1.37)

4.3. Resultados de estudio simulación

Para finalizar, para el T3, todas las funciones de pérdida, así como los dos índices de selección canónicos tuvieron un desempeño similar. Note que el índice de Smith y la función de pérdida EnergyS tuvieron las mayores ganancias, de 7.624 % y 7.291 %, respectivamente. Dichos rendimientos fueron seguidos por los obtenidos con la función de pérdida KL (6.935 %), el índice de selección ESIM (6.833 %), y en último lugar por la función de pérdida MALF (5.842 %).

Varianzas poblacionales

En lo que respecta a los resultados de las varianzas poblacionales, estos se presentan en la figura 4.10a-b, donde se grafican las tendencias del promedio de las varianzas poblacionales. Los resultados corresponden a la varianza poblacional escalada, ciclo a ciclo de selección bajo cada función de pérdida e índice de selección, como ya se comentó previamente.

Cuando el valor de la heredabilidad se fijó en 0.3 (figura 4.10-a), las varianzas poblacional (escalada) disminuye de forma similar en todos los rasgos y en todas las funciones de pérdida e índices de selección. No parece haber una diferencia significativa en el comportamiento de las varianzas entre cada criterio de selección, y esto nos lleva a analizar de manera puntual dichos resultados. El análisis puntual se presenta en el cuadro 4.4, donde se resume el promedio de las diferencias de la varianza poblacional (no escalada) en el 10mo ciclo de selección con respecto al primer ciclo de selección. Si realizamos un ranking de cada función de pérdida e índice de selección de acuerdo a dicho porcentaje y dado que deseamos perder la mínima varianza posible, entonces, en el T1 el ranking queda como enseguida: (1) MALF, (2) EnergyS, (3) KL, (4) ESIM y (5) índice de Smith; en el T2: (1) EnergyS, (2) MALF, (3) KL, (4) ESIM y (5) índice de Smith; y para el T3: (1) MALF, (2) ESIM, (3) Energy, (4) KL, y (5) índice de Smith. Dado lo anterior, las funciones de pérdida en 2 de 3 rasgos ocupan mejor ranking que los índices de selección.

Cuando la heredabilidad se fijó en 0.6, se obtuvieron resultados similares. Esto se observa en la figura 4.10-b donde las tendencias son similares entre si. El ranking con los resultados del 10mo ciclo de selección para el T1 es: (1) EnergyS, (2) KL, (3) MALF, (4) ESIM y (5) índice de Smith; en el T2: (1) KL, (2) MALF, (3) EnergyS, (4) índice de Smith, y (5) ESIM; y finalmente para el T3: (1) MALF, (2) EnergyS, (3) índice de Smith, (4) ESIM, y (5) KL. Nuevamente en 2 de 3 rasgos, las funciones de pérdida obtuvieron mejor ranking que las funciones de pérdida.

4.3. Resultados de estudio simulación

En general, podemos comentar que la disminución de la varianza poblacional en todos los rasgos, y para las dos heredabilidades, tanto las funciones de pérdida como los índices de selección obtuvieron rendimientos similares, con una ligera ventaja de las funciones de pérdida sobre los índices de selección. Además, es de suma importancia resaltar que las funciones de pérdida obtuvieron ganancias en las medias poblacionales de todos los rasgos y los índices de selección no, esto proporciona una clara ventaja a las funciones de pérdida.

5

DISCUSIÓN Y CONCLUSIONES

5.1. Discusión

El objetivo central de este trabajo de investigación fue proponer una metodología formal con sustento en la teoría de la decisión, al problema de seleccionar las/los mejores líneas/individuos en un programa de selección y mejora. Determinar el mejor subconjunto de individuos es crucial en SG, dado que el éxito del programa de selección dependerá completamente de cuáles individuos se seleccionen, ya que estos individuos conformaran la población de padres para el siguiente ciclo de mejora, y así sucesivamente.

El problema planteado corresponde a un problema de decisión. Todo problema de decisión está compuesto fundamentalmente por un espacio de acciones, un espacio de resultados y una función de pérdida que penaliza cada decisión/acción posible. La función de pérdida es por lo tanto, el mecanismo que permite la transición entre un espacio de acciones, al espacio de resultados. En SG, las FPs reflejan la preferencia del mejorador por aquellas/aquellos líneas/individuos con características deseadas, de modo que se garantice el mayor progreso genético. Las pérdidas esperadas, pueden interpretarse entonces como una penalización por alejarnos del progreso genético deseado. Aquellas líneas cuya pérdida esperada *a posteriori* sean menores implica que sus correspondientes distribuciones predictivas *a posteriori* están lo más cerca posible de la distribución (teórica) de los seleccionados. Este enfoque de ver el problema, adquiere notable relevancia cuando la selección opera en muchos rasgos simultáneamente, y donde además, los rasgos están positiva y negativamente correlacionados.

Por tal motivo, en esta investigación se han propuesto tres FPs univariadas (Kullback-Leibler univariada, CRPS y LinLin), así como sus generalizaciones al contexto multirasgo (KL multivariada, Energy Score y MALF), con el propósito de asistir a los mejoradores en

5.1. Discusión

decidir cuáles líneas o individuos serán los mejores padres en cada ciclo de selección. Las FPs univariadas y multivariadas, simétricas y asimétricas, se presentaron de forma simple y clara, y se mostraron como funciones decrecientes de la heredabilidad de los rasgos. Desde el enfoque propuesto, las desviaciones entre las distribuciones de las líneas candidatas a ser padres (líneas candidatas a la selección) y la distribución teórica que refleja las preferencias del mejorador se interpretan como distancias. Así por ejemplo, la FP Kullback-Leibler involucra una pérdida del tipo cuadrática escalada por la varianza fenotípica cuando la selección opera en un solo rasgo, o por la matriz de varianzas-covarianzas en el caso multirasgo. Las FPs CRPS y Energy Score involucran distancias en términos de la norma ℓ_1 y ℓ_2 , respectivamente. Por su parte la FP LinLin y su generalización multivariada MALF (con norma ℓ_1) implican penalizaciones lineales asimétricas, donde la asimetría favorece a aquellas líneas cuyas realizaciones de los VC predichos sean mayores a la media (teórica) de los seleccionados.

En el capítulo anterior se presentaron los resultados obtenidos usando una aplicación de la metodología propuesta en un conjunto de datos reales (trigo de primavera) para un ciclo de selección. También se presentaron los resultados del estudio de simulación univariado (selección en un solo rasgo) y multivariado (selección en varios rasgos). Los resultados obtenidos desde el enfoque univariado se compararon con los obtenidos empleando la forma estándar de selección (que consiste en seleccionar aquellas líneas con más altos VC predichos), mientras que los resultados derivados del enfoque multirasgo se compararon con los obtenidos empleando dos índices de selección muy utilizados en selección: índice de Smith y ESIM.

Los resultados de la simulación univariada fueron satisfactorios, ya que mostraron un mejor desempeño, en términos de la media y la varianza poblacional, de las FPs univariadas sobre el método estándar de selección, siempre que la presión de selección no fue tan restrictiva (30 %); en el caso menos favorable (presión de selección del 10 %) las FPs se desempeñaron similar al método estándar. Estos resultados motivaron la generalización la metodología de selección propuesta, al problema de selección multirasgo. En la selección multivariada, las FPs indujeron ganancias para las respectivas medias poblacionales de todos los rasgos sujetos a la selección, aún en escenarios con rasgos antagónicos (negativamente correlacionados), en contraste con los índices de selección, los cuales funcionaron bien únicamente para los rasgos positivamente correlacionados. Es importante recalcar que las FPs multivariadas y los índices de selección tuvieron un desempeño similar en las varianzas poblacionales. Estas dos consideraciones, dan una notable ventaja a las FPs sobre los índices de selección.

5.1. Discusión

Abordar la selección como un problema de decisión ataca de una forma elegante y formal la selección en uno o más rasgos. En la selección multirasgo, el uso de las FPs reduce la subjetividad en la selección ya que los pesos de los rasgos involucrados en la selección se da de forma automática. No así en algunos índices de selección. Pero esto no es todo, desde la perspectiva de la teoría de la decisión bayesiana, se incorporan los elementos de incertidumbre mediante las respectivas distribuciones predictivas *a posteriori* de las cantidades y parámetros de interés involucradas en las FPs. También es importante resaltar que la metodología propuesta permite a los mejoradores controlar la distribución parental e inducir la selección a favor uno o más rasgos en los que se desee mayor progreso genético, tal como se hace en los índice de selección restringidos, por lo que nuestra metodología es integradora. Por ejemplo, supongamos que para el T1 (en el estudio de simulación multivariada) no se desea progreso genético ciclo a ciclo de selección, basta con no truncar la distribución para dicho rasgo ($y_c = \pm\infty$).

Tal como mencionó previamente, en el estudio de simulación y selección multirasgo, se indujo a que el T1 y el T2 estuviesen negativamente correlacionados, de modo que cuando en el T1 se incrementaran sus valores fenotípicos, en el T2 disminuían, siempre que la selección operaba en favor del T1 y viceversa. Para el caso del T2 y el T3 la correlación inducida fue positiva (y no representan mayores problemas a la selección), y en el caso del T1 y del T3 se simularon independientes (ninguna correlación). Esto se eligió con el propósito de simular rasgos complejos, y así mostrar que la metodología propuesta es aplicable a estos escenarios.

Los resultados obtenidos en términos de la media poblacional, indican que las FPs favorecieron a todos los rasgos a pesar de presencia de rasgos antagónicos; el EnergyS obtuvo mejor desempeño para los rasgos 2 y 3 en comparación con el T1, y este sentido fue el menos prometedor de las tres FPs. Las FPs Kullback y MALF obtuvieron redimientos notables en el T1 en comparación a los IS y el EnergyS.

Por otro lado, en el T2, el EnergyS y los IS se desempeñaron mejor que las otras dos FPs propuestas, lo cual no es de extrañarse ya que sacrificaron progreso genético para el T1. Para el T3, tanto las FPs como los IS obtuvieron desempeños muy similares. La comparativa entre las FPs indica que dichas diferencias, aunque pequeñas, se consideran significativas.

Como ya se mencionó, los índices de selección empleados (Smith y ESIM) tuvieron pobre desempeño para el T1, aunque en el T2 y el T3 obtuvieron ganancias notables. Sin em-

5.1. Discusión

bargo, una desventaja importante de los índices de selección en comparación a las FPs, es que en los primeros, casi siempre se requiere la asignación de los pesos económicos para cada rasgo sujeto a la selección. Esto requiere un análisis previo y cierta experiencia del mejorador con los rasgos involucrados. En cambio, las FPs solucionan este problema de forma automática, lo cual hace evidente su ventaja.

Los comentarios anteriormente expuestos derivan de las figuras 4.9a-b. Un aspecto final a destacar es que aunque las FPs obtuvieron ganancias para las medias poblacionales en los tres rasgos, las varianzas poblacionales reportadas, no se sacrificaron más que las obtenidas con los índices de selección, lo cual, nuevamente puso en ventaja a las FPs. Esto se corrobora en las figuras 4.10a-b así como en el cuadro 4.4.

En resumen, en esta investigación se propuso una metodología formal para la selección de las/los mejores líneas/individuos que serán padres en el siguiente ciclo de reproducción y selección, garantizando el máximo progreso genético posible. El enfoque propuesto, está basado en la teoría de la decisión bayesiana, para la construcción de medidas de divergencia (funciones de pérdida) para rankear las líneas candidatas a la selección. La pérdida espera *a posteriori* es el mecanismo que penaliza cada decisión posible, y esta considera las incertidumbres asociadas tanto a los parámetros como a las predicciones inherentes a los modelos de regresión y predicción. Los resultados obtenidos indican que emplear las FPs en la selección en un solo rasgo, puede mejorar a la respuesta a la selección y al mismo tiempo preservar la varianza genética tanto como sea posible. En la selección multirasgo, las ventajas de las FPs son evidentes. Las medias poblacionales de todos los rasgos considerados obtuvieron ganancias positivas, a pesar de que dos rasgos actuaron como antagonicos. Creemos que la selección con base en las FPs tiene ventajas sustanciales que cuando se utilizan únicamente los valores puntuales de los valores de cría predichos, aunado a que en las FPs, los pesos para cada carácter se da forma automática. Es más fácil fijar un vector de truncamiento, que calibrar los pesos económicos que se utilizan en algunos índices de selección. Estos resultados resultaron válidos tanto para rasgos simples como rasgos complejos.

Finalmente, la metodología propuesta aplicada aquí a la SG, también puede emplearse en el escenario de la selección convencional (que no emplea marcadores moleculares). Como se mostró en el estudio de simulación a largo plazo, los cambios en el rango de unos pocos individuos, puede cambiar la respuesta a la selección final después de varios ciclos de selección.

5.2. Conclusiones

Se propuso una metodología formal con sustento en la teoría de la decisión, al problema de la selección ya sea en un solo rasgo o multirasgo, en selección de plantas y animales empleando selección genómica. Por lo tanto, se presentó la discusión teórica de la metodología y se propusieron tres funciones de pérdida univariadas (Kullback-Leibler, CRPS y LinLin), así como sus generalizaciones multivariadas (Kullback-Leibler multivariada, Energy Score y MALF). Se discutieron cada una de las funciones de pérdida a detalle y se expresaron en términos de la heredabilidad del rasgo o los rasgos.

Se condujo un ejemplo de aplicación en un conjunto de datos reales para un ciclo de selección empleando tanto las funciones de pérdida univariadas y multivariadas, así como en un estudio de simulación de un programa de selección genómica para monitorear tanto de la media poblacional como las varianzas poblacional a lo largo de los ciclos de reproducción. Los resultados obtenidos se compararon con los obtenidos mediante los métodos estándar de selección (basado en predicciones puntuales y mediante el uso de índices de selección). Nuestros resultados de selección en un solo rasgo sugieren que es posible obtener un mejor rendimiento mediante el uso de las funciones de pérdida, cuando la presión de selección no es tan restrictiva (aproximadamente 30 %), en programas de selección y reproducción, a largo plazo.

En la selección multirasgo, las funciones de pérdida garantizaron ganancias positivas en la media poblacional de todos los rasgos involucrados, aún en presencia de rasgos antagónicos, tanto para rasgos simples (alta heredabilidad) como rasgos complejos (baja heredabilidad). A pesar de que las ganancias fueron positivas en las medias, las varianzas poblacionales no se sacrificaron más que las obtenidas con los índices de selección.

LITERATURA CITADA

- Akdemir, D. y Sánchez, J. I. (2016). Efficient Breeding by Genomic Mating. *Frontiers in Genetics*, 7. ISSN 1664-8021.
- Baringhaus, L. y Franz, C. (2004). On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88, 190–206.
- Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton University Press. ISBN 978-0-691-07901-1.
- Berk, R. (2011). Asymmetric Loss Functions for Forecasting in Criminal Justice Settings. *Journal of Quantitative Criminology*, 27, 1, 107–123. ISSN 0748-4518, 1573-7799.
- Bos, I. y Caligari, P. (2008). *Selection Methods in Plant Breeding*. Springer Netherlands, segunda edición. ISBN 978-1-4020-6369-5.
- Brier, A. P. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 1, 78, 1–3.
- Brisbane, J. R. y Gibson, J. P. (1995). Balancing selection response and rate of inbreeding by including genetic relationships in selection decisions. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, 91, 3, 421–431. ISSN 0040-5752.
- Casella, G. y George, E. I. (1992). Explaining the Gibbs Sampler. *The American Statistician*, 46, 3, 167–174. ISSN 00031305.
- Ceron-Rojas, J. J., Crossa, J., Arief, V. N., Basford, K., Rutkoski, J., Jarquín, D., Alvarado, G., Beyene, Y., Semagn, K. y DeLacy, I. (2015). A Genomic Selection Index Applied to Simulated and Real Data. *G3: Genes|Genomes|Genetics*, 5, 10, 2155–2164. ISSN 2160-1836.
- Ceron-Rojas, J. J., Crossa, J., Sahagun Castellanos, J., Castillo Gonzalez, F. y Santacruz Varela, A. (2008). A selection index method based on Eigenanalysis. *Journal of Agricultural, Biological, and Environmental Statistics*, 13, 4, 440–457.

LITERATURA CITADA

- Contreras-Reyes, J. y B. Arellano-Valle, R. (2012). Kullback-Leibler Divergence Measure for Multivariate Skew-Normal Distributions. *Entropy*, 14, 1606.
- Cover, T. M. y Thomas, J. A. (2006). *Elements of information theory 2nd edition*. Wiley-Interscience.
- Crossa, J., de los Campos, G., Pérez, P., Gianola, D., Burgueño, J., Araus, J. L., Makumbi, D., Singh, R. P., Dreisigacker, S., Yan, J. *et al.* (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., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G. d. I., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J. y Varshney, R. K. (2017). Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends in Plant Science*, 22, 11, 961–975. ISSN 1360-1385.
- Dawid, A. P. (1998). Coherent measures of discrepancy, uncertainty and dependence, with applications to bayesian predictive experimental design. Inf. téc., Department of Statistical Science, University College London.
- Dawid, A. P. (2007). The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59, 1, 77–93.
- Dawid, A. P., Musio, M. *et al.* (2015). Bayesian model selection based on proper scoring rules. *Bayesian Analysis*, 10, 2, 479–499.
- de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K. y Cotes, J. M. (2009). Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. *Genetics*, 182, 1, 375–385. ISSN 0016-6731, 1943-2631.
- de los Campos, G. y Pérez Rodríguez, P. (2015). *BGLR: Bayesian Generalized Linear Regression*. R package version 1.0.5.
- de los Campos, G. y Grüneberg, A. (2016). MTM (Multiple-Trait Model) package.
- Eddelbuettel, D. y François, R. (2011). Rcpp: Seamless R and C++ Integration. *Journal of Statistical Software*, 40, 8, 1–18.
- Fahrmeir, L., Kneib, T., Lang, S. y Marx, B. (2013). *Regression: models, methods and applications*. Springer Science & Business Media.
- Falconer, D. S. y Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*. Pearson, Harlow, cuarta edición. ISBN 978-0-582-24302-6.

LITERATURA CITADA

- Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. y Rubin, D. (2014). *Bayesian Data Analysis, Third Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, London, tercera edición. ISBN 1439840954.
- Giacomini, R. y Komunjer, I. (2005). Evaluation and combination of conditional quantile forecasts. *Journal of Business & Economic Statistics*, 23, 4, 416–431.
- Gilks, W., Richardson, S. y Spiegelhalter, D. (1995). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis. ISBN 9780412055515.
- Giovanni Parmigiani, L. I. (2009). *Decision Theory: Principles and Approaches (Wiley Series in Probability and Statistics)*. Wiley, primera edición. ISBN 047149657X,9780471496571.
- Gneiting, T. y Raftery, A. (2004). Strictly proper scoring rules, prediction and estimation. *Technical Report no. 463. Department of Statistics, University of Washington*, 1–30.
- Gneiting, T. y Raftery, A. (2007). Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association*, 102, 477, 359–378.
- Good, I. J. (1952). Rational Decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14, 1, 107–114. ISSN 00359246.
- Granger, C. W. J. (1969). Prediction with a Generalized Cost of Error Function. *OR*, 20, 2, 199–207. ISSN 14732858.
- Grimit, E. P., Gneiting, T., Berrocal, V. J. y Johnson, N. A. (2006). The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society*, 132, 621C, 2925–2942.
- Jannink, J.-L., Lorenz, A. J. y Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *Briefings in functional genomics*, 9, 2, 166–177.
- Komunjer, I. y Owyang, M. T. (2011). Multivariate Forecast Evaluation and Rationality Testing. *The Review of Economics and Statistics*, 94, 4, 1066–1080. ISSN 0034-6535.
- Kullback, S. y Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22, 1, 79–86. ISSN 0003-4851, 2168-8990.
- Lee, T.-H. (2008). Loss functions in time series forecasting. *International encyclopedia of the social sciences*, 9, 495–502.
- Lehermeier, C., Schön, C.-C. y de los Campos, G. (2015). Assessment of genetic heterogeneity in structured plant populations using multivariate whole-genome regression models. *Genetics*, 201, 1, 323–337.

LITERATURA CITADA

- 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.
- Oldenbroek, K. y van der Waaij, L. (2015). *Textbook Animal Breeding and Genetics for BSc students*. Centre for Genetic Resources The Netherlands and Animal Breeding and Genomics Centre.
- Park, T. y Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103, 482, 681–686. ISSN 0162-1459.
- Parry, M., Dawid, P. y Lauritzen, S. (2012). Proper local scoring rules. *The Annals of Statistics*, 40, 1, 561–592.
- R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Richmond, R. J., Nau, R. F. y Winkler, R. L. (2008). Scoring rules, generalized entropy, and utility maximization. *Operations Research*, 56, 5, 1146–1157.
- Rincen, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., Rodríguez, V., Moreno-Gonzalez, J., Melchinger, A., Bauer, E., Schoen, C.-C., Meyer, N., Giauffret, C., Bauland, C., Jamin, P., Laborde, J., Monod, H., Flament, P., Charcosset, A. y Moreau, L. (2012). Maximizing the Reliability of Genomic Selection by Optimizing the Calibration Set of Reference Individuals: Comparison of Methods in Two Diverse Groups of Maize Inbreds (*Zea mays* L.). *Genetics*, 192, 2, 715–728.
- Rodríguez, G. (2007). Lecture notes on generalized linear models. url: <http://data.princeton.edu/wws509/notes/>.
- Salgotra, R., Gupta, B. y Stewart Jr, C. (2014). From genomics to functional markers in the era of next-generation sequencing. *Biotechnology letters*, 36, 3, 417–426.
- Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, 1, 1, 43–62.
- Shepherd, R.K. y Kinghorn, B.P. (1998). A tactical approach to the design of crossbreeding programs. En *6th World Congress on Genetics Applied to Livestock Production*, tomo 25, 431–438. Armidale.
- Smith, H. F. (1936). A discriminant function for plant selection. *Annals of Eugenics*, 7, 3, 240–250. ISSN 2050-1439.
- Székely, G. J. y Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143, 8, 1249–1272. ISSN 0378-3758.

LITERATURA CITADA

- Tsagris, M., Beneki, C. y Hassani, H. (2014). On the folded normal distribution. *Mathematics*, 2, 1, 12–28. ISSN 2227-7390. ArXiv: 1402.3559.
- Velu, G., Crossa, J., Singh, R. P., Hao, Y., Dreisigacker, S., Perez-Rodriguez, P., Joshi, A. K., Chattrath, R., Gupta, V., Balasubramaniam, A., Tiwari, C., Mishra, V. K., Sohu, V. S. y Mavi, G. S. (2016). Genomic prediction for grain zinc and iron concentrations in spring wheat. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, 129, 8, 1595–1605. ISSN 1432-2242.
- Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences*, tomo 100. Academic press.
- Wray, N. y Goddard, M. (1994). Increasing long-term response to selection. *Genetics Selection Evolution*, 26, 431. ISSN 1297-9686.
- Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E. y Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*, 42, 7, 565–569.
- Zellner, A. (1986). Bayesian Estimation and Prediction Using Asymmetric Loss Functions. *Journal of the American Statistical Association*, 81, 394, 446–451. ISSN 01621459.
- Zhe Zhang, Xiujin Li, Xiangdong Ding, Jiaqi Li y Qin Zhang (2015). GPOPSIM: a simulation tool for whole-genome genetic data. *BMC genetics*, 16, 1, 10+.

ANEXOS

A1. Desarrollo de la función de pérdida Kullback-Leibler para fines de selección genómica

A1.1. Función de pérdida Kullback-Leibler univariada

$$\begin{aligned} KL(F_{Y_o}, F_{Y_s}) &= \int_{y_c}^{\infty} \log \frac{N_T(y|\mu_1, \sigma^2, y_c)}{N(y|\mu_2, \sigma^2)} N_T(y|\mu_1, \sigma^2, y_c) dy \\ &= \int_{y_c}^{\infty} \log N_T(y|\mu_1, \sigma^2, y_c) N_T(y|\mu_1, \sigma^2, y_c) dy - \int_{y_c}^{\infty} \log N(y|\mu_2, \sigma^2) N_T(y|\mu_1, \sigma^2, y_c) dy \\ &= -\log(Z) - \frac{1}{2\sigma^2} \int_{y_c}^{\infty} (y - \mu_1)^2 N_T(y|\mu_1, \sigma^2, y_c) dy + \frac{1}{2\sigma^2} \int_{y_c}^{\infty} (y - \mu_2)^2 N_T(y|\mu_1, \sigma^2, y_c) dy \\ &= -\log(Z) - \frac{1}{2\sigma^2} \left\{ \int_{y_c}^{\infty} y^2 N_T(y|\mu_1, \sigma^2, y_c) dy - 2\mu_1 \int_{y_c}^{\infty} y N_T(y|\mu_1, \sigma^2, y_c) dy \right. \\ &\quad \left. + \mu_1^2 \int_{y_c}^{\infty} N_T(y|\mu_1, \sigma^2, y_c) dy \right\} + \frac{1}{2\sigma^2} \left\{ \int_{y_c}^{\infty} y^2 N_T(y|\mu_1, \sigma^2, y_c) dy \right. \\ &\quad \left. - 2\mu_2 \int_{y_c}^{\infty} y N_T(y|\mu_1, \sigma^2, y_c) dy + \mu_2^2 \int_{y_c}^{\infty} N_T(y|\mu_1, \sigma^2, y_c) dy \right\} \\ &= -\log(z) - \frac{1}{2\sigma^2} (V_s + \mu_s^2 - 2\mu_1\mu_s + \mu_1^2) + \frac{1}{2\sigma^2} (V_s + \mu_s^2 - 2\mu_2\mu_s + \mu_2^2) \\ &= -\log(z) + \frac{1}{2\sigma^2} [(\mu_s - \mu_2)^2 - (\mu_s - \mu_1)^2]. \end{aligned}$$

donde $z = 1 - \Phi\left(\frac{y_c - \mu_1}{\sigma}\right)$, μ_s y V_s denotan la media y la varianza de y_s después de que ocurre la censura de y y_c .

A1.2. Función de pérdida Kullback-Leibler Multivariada

$$\begin{aligned}
KL(F_{Y_o}, F_{Y_s}) &= \int_{y_c}^{\infty} \log MVN_T(\mathbf{y}|\boldsymbol{\mu}_1, \mathbf{K}, y_c) MVN_T(\mathbf{y}|\boldsymbol{\mu}_1, \mathbf{K}, y_c) dy - \int_{y_c}^{\infty} \log MVN(\mathbf{y}|\boldsymbol{\mu}_2, \mathbf{K}) MVN_T(\mathbf{y}|\boldsymbol{\mu}_1, \mathbf{K}, y_c) dy \\
&= \int_{y_c}^{\infty} \left[-\log(Z) - \frac{t}{2} \log(2\pi) - \frac{1}{2} \log(\det(\mathbf{K})) - \frac{1}{2} (\mathbf{y} - \boldsymbol{\mu}_1)' \mathbf{K}^{-1} (\mathbf{y} - \boldsymbol{\mu}_1) \right] MVN_T(\mathbf{y}|\boldsymbol{\mu}_1, \mathbf{K}, y_c) dy \\
&\quad - \int_{y_c}^{\infty} \left[-\frac{t}{2} \log(2\pi) - \frac{1}{2} \log(\det(\mathbf{K})) - \frac{1}{2} (\mathbf{y} - \boldsymbol{\mu}_2)' \mathbf{K}^{-1} (\mathbf{y} - \boldsymbol{\mu}_2) \right] MVN_T(\mathbf{y}|\boldsymbol{\mu}_1, \mathbf{K}, y_c) dy \\
&= -\log(Z) - \frac{1}{2} \int_{y_c}^{\infty} \underbrace{(\mathbf{y} - \boldsymbol{\mu}_1)' \mathbf{K}^{-1} (\mathbf{y} - \boldsymbol{\mu}_1)}_{**} MVN_T(\mathbf{y}|\boldsymbol{\mu}_1, \mathbf{K}, y_c) dy + \frac{1}{2} \int_{y_c}^{\infty} \underbrace{(\mathbf{y} - \boldsymbol{\mu}_2)' \mathbf{K}^{-1} (\mathbf{y} - \boldsymbol{\mu}_2)}_{++} MVN_T(\mathbf{y}|\boldsymbol{\mu}_1, \mathbf{K}, y_c) dy.
\end{aligned}$$

Desarrollando **

$$\begin{aligned}
(\mathbf{y} - \boldsymbol{\mu}_1)' \mathbf{K}^{-1} (\mathbf{y} - \boldsymbol{\mu}_1) &= \mathbf{y}' \mathbf{K}^{-1} \mathbf{y} - \mathbf{y}' \mathbf{K}^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_1' \mathbf{K}^{-1} \mathbf{y}_1 + \boldsymbol{\mu}_1' \mathbf{K}^{-1} \boldsymbol{\mu}_1 \\
&= \mathbf{y}' \mathbf{K}^{-1} \mathbf{y} - 2\mathbf{y}' \mathbf{K}^{-1} \boldsymbol{\mu}_1 + \boldsymbol{\mu}_1' \mathbf{K}^{-1} \boldsymbol{\mu}_1
\end{aligned}$$

y ++

$$\begin{aligned}
(\mathbf{y} - \boldsymbol{\mu}_2)' \mathbf{K}^{-1} (\mathbf{y} - \boldsymbol{\mu}_2) &= \mathbf{y}' \mathbf{K}^{-1} \mathbf{y} - \mathbf{y}' \mathbf{K}^{-1} \boldsymbol{\mu}_2 - \boldsymbol{\mu}_2' \mathbf{K}^{-1} \mathbf{y}_1 + \boldsymbol{\mu}_2' \mathbf{K}^{-1} \boldsymbol{\mu}_2 \\
&= \mathbf{y}' \mathbf{K}^{-1} \mathbf{y} - 2\mathbf{y}' \mathbf{K}^{-1} \boldsymbol{\mu}_2 + \boldsymbol{\mu}_2' \mathbf{K}^{-1} \boldsymbol{\mu}_2
\end{aligned}$$

Una vez desarrollado las formas cuadráticas anteriores, eliminando términos y empleando la propiedad $E(\mathbf{y}'\mathbf{K}\mathbf{y}) = (E(\mathbf{y}))'\mathbf{K}(E(\mathbf{y})) + \text{traza}(\mathbf{K}\mathbf{K}')$, la divergencia de Kullback se resume en

$$KL(F_{\mathbf{Y}_o}, F_{\mathbf{Y}_s}) = -\log(z) + \frac{1}{2} [(\boldsymbol{\mu}_s - \boldsymbol{\mu}_2)'\mathbf{K}^{-1}(\boldsymbol{\mu}_s - \boldsymbol{\mu}_2) - (\boldsymbol{\mu}_s - \boldsymbol{\mu}_1)'\mathbf{K}^{-1}(\boldsymbol{\mu}_s - \boldsymbol{\mu}_1)],$$

con $z = (2\pi)^{t/2} |\mathbf{K}|^{-1/2} \int_{\mathbf{y}_c}^{\infty} \exp\left\{\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}_1)'\mathbf{K}^{-1}(\mathbf{y} - \boldsymbol{\mu}_1)\right\} d\mathbf{y}$.

A2. Derivación de la función CRPS cuando la distribución predictiva es normal

Antes de proceder con la derivación de la expresión presentada en la ecuación 3.6, procedemos con presentar algunos resultados de la distribución normal plegada (*folded normal distribution*).

La distribución normal plegada corresponde a la distribución del valor absoluto de una variable aleatoria distribuida normal. Suponga que $Y \sim N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$ y $\sigma > 0$. Sea $W = |Y|$, $W \sim N^f(\mu, \sigma^2)$.

La función de densidad de probabilidad f de W se expresa como

$$\begin{aligned} f(w) &= \frac{1}{\sigma} \left[\phi\left(\frac{w - \mu}{\sigma}\right) + \phi\left(\frac{w + \mu}{\sigma}\right) \right] \\ &= \frac{1}{\sigma\sqrt{2\pi}} \left\{ \exp\left(-\frac{1}{2}\left(\frac{w + \mu}{\sigma}\right)^2\right) + \exp\left(-\frac{1}{2}\left(\frac{w - \mu}{\sigma}\right)^2\right) \right\}, \quad w \in [0, \infty). \end{aligned}$$

Los dos primeros momentos de W están dados por

- $E(W) = \mu [1 - 2\Phi(-\frac{\mu}{\sigma})] + \sigma\sqrt{2/\pi} \exp(-\mu^2/2\sigma^2),$
- $E(W^2) = \mu^2 + \sigma^2$

Detalles de la distribución normal plegada pueden consultarse en [Tsagris et al. \(2014\)](#). Enseguida calculamos la esperanzas de la expresión presentada enseguida

$$CRPS(F_{Y_o}, \mu_s) = E_F|Y_o - \mu_s| - \frac{1}{2}E_F|Y_o - Y'_o|$$

Primero note $F_{|Y_o - Y'_o|}(\cdot) = N^f(0, \sigma^2)$ y $F_{|Y_o - \mu_s|}(\cdot) = N^f(\mu_2 - \mu_s, \sigma^2)$, entonces, utilizando los resultados del valor esperado de la normal plegada, se tiene que

$$E|Y_o - Y'_o| = \frac{2\sigma}{\sqrt{\pi}},$$

$$\begin{aligned} E|Y_o - \mu_s| &= (\mu_2 - \mu_s) \left[1 - 2\Phi\left(-\frac{\mu_2 - \mu_s}{\sigma}\right) \right] + \sigma\sqrt{\frac{2}{\pi}} \exp\left(-\frac{1}{2\sigma^2}(\mu_2 - \mu_s)^2\right) \\ &= (\mu_2 - \mu_s) \left[1 - 2\Phi\left(\frac{\mu_s - \mu_2}{\sigma}\right) \right] + 2\sigma\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\mu_2 - \mu_s}{\sigma}\right)^2\right) \\ &= (\mu_2 - \mu_s) \left[1 - 2\Phi\left(\frac{\mu_s - \mu_2}{\sigma}\right) \right] + 2\sigma\phi\left(\frac{\mu_2 - \mu_s}{\sigma}\right). \end{aligned}$$

Simplificando

$$\begin{aligned} CRPS(F_{Y_o}, \mu_s) &= E_F|Y_o - \mu_s| - \frac{1}{2}E_F|Y_o - Y'_o| \\ &= (\mu_2 - \mu_s) \left[1 - 2\Phi\left(\frac{\mu_s - \mu_2}{\sigma}\right) \right] + 2\sigma\phi\left(\frac{\mu_2 - \mu_s}{\sigma}\right) - \frac{1}{2}\frac{2\sigma}{\sqrt{\pi}} \\ &= \sigma \left[\frac{\mu_s - \mu_2}{\sigma} \left(2\Phi\left(\frac{\mu_s - \mu_2}{\sigma}\right) - 1 \right) + 2\phi\left(\frac{\mu_s - \mu_2}{\sigma}\right) - \frac{1}{\sqrt{\pi}} \right] \\ &= -\sigma \left[\frac{1}{\sqrt{\pi}} - 2\phi\left(\frac{\mu_s - \mu_2}{\sigma}\right) - \left(\frac{\mu_s - \mu_2}{\sigma}\right) \left(2\Phi\left(\frac{\mu_s - \mu_2}{\sigma}\right) - 1 \right) \right]. \end{aligned}$$

A3. Código R de las funciones de pérdida propuestas

Las FPs univariadas propuestas en esta investigación utilizan muestras MCMC de las distribuciones marginales *a posteriori* de cada cantidad y/o parámetro involucrado. Las muestras que alimentan las funciones siguientes, se consideran realizaciones de las distribuciones estacionarias, después del periodo de calentamiento. En el caso de las funciones de pérdida univariadas, las muestras MCMC

se originan al ajustar los modelos planteados mediante el paquete BGLR; mientras que el caso multivariado, dichas muestras provienen del paquete MTM.

A3.1. Funciones univariadas

Los argumentos de las funciones escritas en R son:

- Xb: matriz de muestras MCMC para la media de las líneas candidatas para los fenotipos no observados. En las filas las líneas o individuos, en columnas las muestras MCMC (μ_2).
- MU: vector de muestras MCMC para la media poblacional (μ_1).
- sigma: vector de muestras MCMC para la desviación estándar poblacional ($\sigma_1 = \sqrt{\sigma_1^2}$).
- y_c: valor de truncamiento (escalar).
- alpha: parámetro de asimetría (escalar) en la FP LinLin que es igual a 1-p, donde p corresponde a la proporción de líneas a seleccionar del conjunto de candidatas a la selección.
- Nsel: número de líneas a seleccionar.

La salida de la función corresponde a los `id` de las líneas seleccionadas, cuyas pérdidas esperadas *a posteriori* resultaron menores.

```
1 ## Selection through Standard Method (STD)
2 selStd <- function(Xb, y_c, Nsel, MU, sigma) {
3   yHat <- apply(Xb, 1, mean, na.rm = TRUE)
4   selected <- order(yHat, decreasing = TRUE)[1:Nsel]
5   return(selected)
6 }
7
8 ## Selection through Kullback-Leibler (KL)
9 selKL <- function(Xb, y_c, Nsel, MU, sigma) {
10  n.mcmc <- ncol(Xb)
11  n.lines <- nrow(Xb)
12  losses <- matrix(0, n.lines, n.mcmc)
13  for(i in 1:n.mcmc) {
14    mu1 <- MU[i]
15    mu2 <- Xb[,i]
16    sd <- sigma[i]
```

```

17     yz <- (y_c - mu1) / sd           # standardizing
18     Z <- 1 - pnorm(yz)              # normalization factor
19     dens <- dnorm(yz, mean = 0, sd = 1) # density
20     prob <- pnorm(yz, mean = 0, sd = 1) # probability
21     muT <- mu1 + sd * (dens / (1 - prob))
22     # posterior predictive function at each iter.
23     yppdf <- sapply(mu2, function(ypred) {
24         rnorm(1, mean = ypred, sd=sd)})
25     losses[,i] <- -log(Z) + ((muT - yppdf)^2 +
26         (muT - mu1)^2) / (2 * sd^2)
27     }
28     e.loss <- apply(losses, 1, mean, na.rm = TRUE)
29     selected <- order(e.loss, decreasing = FALSE)[1:Nsel]
30     return(selected)
31 }
32
33 ## Selection through Continuous Ranked Probability Score (CRPS)
34 selCRPS <- function(Xb, y_c, Nsel, MU, sigma) {
35     n.mcmc <- ncol(Xb)
36     n.lines <- nrow(Xb)
37     losses <- matrix(0, n.lines, n.mcmc)
38     for(i in 1:n.mcmc){
39         mu1 <- MU[i]
40         mu2 <- Xb[,i]
41         sd <- sigma[i]
42         zz <- (y_c-mu1)/sd
43         muT <- mu1 + sd*(dnorm(zz)/(1-pnorm(zz)))
44         # posterior predictive function at each iter.
45         yppdf <- sapply(mu2, function(ypred) {
46             rnorm(1, mean = ypred, sd=sd)})
47         yz <- (muT-yppdf) / sd           # standardizing
48         dens <- dnorm(yz, mean = 0, sd = 1) # density
49         prob <- pnorm(yz, mean = 0, sd = 1) # probability
50         losses[,i] <- -1 * (sd * (1 / sqrt(pi) +
51             2 * dens - yz * (2 * prob - 1)))
52     }
53     e.loss <- apply(losses, 1, mean, na.rm = TRUE)
54     selected <- order(e.loss, decreasing = FALSE)[1:Nsel]
55     return(selected)
56 }
57
58 ## Selection through Linear-Linear (LinLin)
59 selLin <- function(Xb, y_c, Nsel, MU, sigma, alpha) {
60     n.mcmc <- ncol(Xb)
61     n.lines <- nrow(Xb)

```

```
62     losses <- matrix(0, n.lines, n.mcmc)
63     for(i in 1:n.mcmc){
64         mu1 <- MU[i]
65         mu2 <- Xb[,i]
66         sd <- sigma[i]
67         z = (y_c-mu1)/sd
68         prob <- 1-pnorm(z)
69         dens <- dnorm(z)
70         bias <- sd*dens/prob
71         muT <- mu1 + bias
72         # posterior predictive function at each iter.
73         yppdf <- sapply(mu2, function(ypred) {
74             rnorm(1, mean = ypred, sd=sd)})
75         losses[,i] <- ((alpha-ifelse(yppdf < muT, 1, 0))*(yppdf-muT))
76     }
77     e.loss <- apply(losses, 1, mean, na.rm = TRUE)
78     selected <- order(e.loss, decreasing = FALSE)[1:Nsel]
79     return(selected)
80 }
```

A3.2. Funciones multivariadas

Los argumentos de las funciones escritas en R son:

- **Xb**: matriz que contiene las muestras MCMC para la media predicha de las líneas candidatas a la selección (μ_2). Valores en columnas corresponden a las medias para cada rasgo, en las filas cada realización MCMC.
- **MU**: matriz que contiene las muestras MCMC para la media de la población base (μ_1). Valores en columnas corresponden a las líneas, en las filas cada realización MCMC.
- **K**: matriz de muestras MCMC para la matriz de varianzas y covarianzas. Valores en columnas cada componente de varianzas o covarianzas, en filas cada realización MCMC.
- **y_c**: vector con los puntos de truncamiento para cada uno de los rasgos.
- **tau**: vector parámetro de asimetría, de longitud igual al número de rasgos.
- **Nsel**: número de líneas a seleccionar.

ANEXOS

La salida de la función corresponde a los `id` de las líneas seleccionadas, cuyas pérdidas esperadas *a posteriori* resultaron menores.

```
1  ## Selection through Multivariate Kullback-Leibler (KL)
2  selKL <- function(Xb, MU, K, y_c, Nsel) {
3      Xb <- as.matrix(Xb); MU <- as.matrix(MU); K <- as.matrix(K);
4      n.traits <- ncol(MU)
5      n.lines <- ncol(Xb)/n.traits
6      n.mcmc <- nrow(Xb)
7      losses <- matrix(0, nrow = n.lines, ncol = n.mcmc)
8      Kall <- apply(K, 1, function(V) as.matrix(nearPD(xpnd(unlist(V)))$mat))
9      Kallinv <- llply(Kall, solve)
10     Xb <- apply(Xb, 1, split, f=rep(1:n.traits, each=n.lines))
11     for(i in 1:n.mcmc) {
12         mu1 <- as.vector(MU[i,])
13         mu2 <- do.call(cbind, Xb[[i]])
14         K <- Kall[[i]]
15         Kinv <- Kallinv[[i]]
16         muS <- as.vector(mtmvnorm(mean = mu1, sigma = K, lower = y_c,
17                                 upper = rep(Inf, length(y_c)),
18                                 doComputeVariance=FALSE)$tmean)
19         Z = pmvnorm(lower=y_c, upper=Inf, mean=mu1, sigma=K)
20         # posterior predictive distribution at each iteration.
21         yppdf <- as.matrix(t(apply(mu2, 1, function(ypred) {
22             rmvnorm(1, mean = ypred, sigma=K)})))
23         S <- muS-mu1 # selection differential
24         UKU <- as.numeric(t(S)%*%Kinv%*%S)
25         muSmu2 <- sweep(yppdf,2,muS, '-') # muS - mu2
26         losses[,i] <- as.vector(apply(muSmu2, 1, function(x) {
27             0.5*(t(x)%*%Kinv%*%x-UKU)-log(Z) }))
28     }
29     e.loss <- apply(losses, 1, mean, na.rm = TRUE)
30     selected <- order(e.loss, decreasing = FALSE)[1:Nsel]
31     return(selected)
32 }
33
34 ## Selection through Energy Score (EnergyS)
35 selEnergyS <- function(Xb, MU, K, y_c, Nsel) {
36     Xb <- as.matrix(Xb); MU <- as.matrix(MU); K <- as.matrix(K);
37     n.traits <- ncol(MU)
38     n.lines <- ncol(Xb)/n.traits
39     n.mcmc <- nrow(Xb)
40     losses <- matrix(0, nrow = n.lines, ncol = n.mcmc)
```



```

41 Kall <- alply(K, 1, function(V) as.matrix(nearPD(xpnd(unlist(V)))$mat) )
42 Xb <- apply(Xb, 1, split, f=rep(1:n.traits, each=n.lines))
43 for(i in 1:n.mcmc) {
44   mu1 <- as.vector(MU[i,])
45   mu2 <- do.call(cbind, Xb[[i]])
46   K <- Kall[[i]]
47   muS <- as.vector(mtmvnorm(mean = mu1, sigma = K, lower = y_c,
48     upper = rep(Inf, length(y_c)),
49     doComputeVariance=FALSE)$tmean)
50   # posterior predictive distribution at each iteration.
51   yppdf <- as.matrix(t(apply(mu2, 1, function(ypred) {
52     rmvnorm(1, mean = ypred, sigma=K)})))
53   yppdf_p <- as.matrix(t(apply(mu2, 1, function(ypred) {
54     rmvnorm(1, mean = ypred, sigma=K)})))
55   muSmu2 <- sweep(yppdf, 2, muS, '-') # X-y
56   mu2mu2 <- yppdf-yppdf_p # X-X'
57   xj_y <- as.vector(apply(muSmu2, 1, function(x) sqrt(sum(x^2)) ))
58   xj_xjp <- as.vector(apply(mu2mu2, 1, function(x) 0.5*sqrt(sum(x^2))))
59   losses[,i] <- xj_y - xj_xjp
60 }
61 e.loss <- apply(losses, 1, mean, na.rm = TRUE)
62 selected <- order(e.loss, decreasing = FALSE)[1:Nsel]
63 return(selected)
64 }
65
66 ## Selection through Multivariate Asymmetric Loss Function (MALF)
67 selMALF <- function(Xb, MU, K, y_c, Nsel, tau) {
68   Xb <- as.matrix(Xb); MU <- as.matrix(MU); K <- as.matrix(K);
69   n.traits <- ncol(MU)
70   n.lines <- ncol(Xb)/n.traits
71   n.mcmc <- nrow(Xb)
72   losses <- matrix(0, nrow = n.lines, ncol = n.mcmc)
73   Kall <- alply(K, 1, function(V) as.matrix(nearPD(xpnd(unlist(V)))$mat) )
74   Xb <- apply(Xb, 1, split, f=rep(1:n.traits, each=n.lines))
75   for(i in 1:n.mcmc) {
76     mu1 <- as.vector(MU[i,])
77     mu2 <- do.call(cbind, Xb[[i]])
78     K <- Kall[[i]]
79     muS <- as.vector(mtmvnorm(mean = mu1, sigma = K, lower = y_c,
80       upper = rep(Inf, length(y_c)),
81       doComputeVariance=FALSE)$tmean)
82     # posterior predictive distribution at each iteration.
83     yppdf <- as.matrix(t(apply(mu2, 1, function(ypred) {
84       rmvnorm(1, mean = ypred, sigma=K)})))
85     error <- -1*sweep(yppdf, 2, muS, '-') # mu2 - muS

```

```
86     abs.error <- abs(error)
87     sum.e <- rowSums(abs.error)
88     score <- sum.e + t(error %*% tau)
89     losses[,i] <- score
90   }
91   e.loss <- apply(losses, 1, mean, na.rm = TRUE)
92   selected <- order(e.loss, decreasing = FALSE)[1:Nsel]
93   return(selected)
94 }
```
