COLEGIO DE POSTGRADUADOS

INSTITUCIÓN DE ENSEÑANZA E INVESTIGACIÓN
EN CIENCIAS AGRÍCOLAS

CAMPUS MONTECILLO

SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA
ESTADÍSTICA

UN MÉTODO DE REGRESIÓN BAYESIANA PARA SELECCIÓN GENÓMICA

ESPERANZA NICOLÁS POPOCA

T E S I S

PRESENTADA COMO REQUISITO PARCIAL PARA
OBTENER EL GRADO DE:

MAESTRA EN CIENCIAS

MONTECILLO,TEXCOCO, EDO. DE MÉXICO, SEPTIEMBRE 2013

La presente tesis titulada: **UN MÉTODO DE REGRESIÓN BAYESIANA PARA SELECCIÓN GENÓMICA**, realizada por el alumno: **ESPERANZA NICOLÁS POPOCA**, bajo la dirección del Consejo Particular indicado ha sido aprobada por el mismo y aceptada como requisito parcial para obtener el grado de:

MAESTRA EN CIENCIAS

SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA
ESTADÍSTICA

CONSEJO PARTICULAR

CONSEJERO

Dr. Sergio Pérez Elizalde

DIRECTOR

ASESOR

Dr. Francisco Julián Ariza Hernández

ASESOR

Dr. José Crossa Hiriart

ASESOR

Dr. Flaviano Godínez Jaimes

Montecillo, Texcoco, México, Septiembre de 2013

AGRADECIMIENTOS

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por su apoyo económico brindado durante mis estudios de posgrado.

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

Al Dr. Sergio Pérez Elizalde,por su paciencia , y excelente direción sin la cual no hubiera sido posible concluir esta investigación ...

A los Doctores:

Dr. Francisco Julián Ariza Hernández
Dr. Flaviano Godínez Jaimes
Dr. José Crossa Hiriart

Por su tiempo y por sus valiosas sugerencias en la revisión a este escrito.

A mis profesores y compañeros de clases...
A mis amigos...

A todo el personal administrativo del Colegio y en especial a nuestro angel E.O.S. siempre estarás en nuestro corazón

Gracias a mi esposo Jorge por su amor, apoyo, comprensión y ayuda en el desarrollo y conclusión de este importante proyecto.

DEDICATORIA

A JORGE

Índice

Índice de tablas

Índice de figuras

Capítulo 1

INTRODUCCIÓN

Desde hace miles de años el hombre ha seleccionado y mejorado diversas especies vegetales y animales basándose solamente en sus cualidades observables útiles para el ser humano; esto es, el fenotipo. Por otro lado el mejoramiento genético es posible debido a la variabilidad genética, a la heredabilidad del carácter que se quiere aislar; sin embargo, muchos aspectos son desconocidos, como el número y efecto de los genes implicados en la expresión de un carácter, la localización de estos genes, y su función fisiológica. Recientemente, por medio de marcadores moleculares, es posible localizar genes con características de intéres dentro de un cromosoma o genoma completo; esto es, los marcadores se usan en el mapeo genético para encontrar la posición e identidad de un gen. Cuando varios marcadores moleculares se asocian inequívocamente con un rasgo genético, se dice que forman un QTL (*loci* de rasgos cuantitativos o cuantificables). Los efectos genéticos de QTL y los valores fenotípicos de un carácter pueden relacionarse estadísticamente por medio de un modelo de regresión lineal múltiple, sin embargo los QTL contienen un gran número de marcadores que en teoría tiene poco efecto individual sobre el fenotipo, es decir, el **número de observaciones es mucho**

menor que el número de covariables ($p >> n$), por lo que la estimación de los parámetros de dicho modelo no puede realizarse por medio de los métodos de estimación estándar como Máxima Verosimilitud y Mínimos Cuadrados Ordinarios no proveen soluciones únicas ya que $\boldsymbol{X}^t\boldsymbol{X}$ es singular. Un problema relacionado con que $p >> n$ es la multicolinealidad, que puede considerarse la existencia de dependencias lineales altas entre las covariables, lo que tiene como consecuencia que los estimadores de los coeficientes de regresión posean varianzas muy grandes, por lo que no son de utilidad practica. Para enfrentar el problema antes descrito existen métodos alternativos, conocidos como métodos de regresión penalizada. Desde el punto de vista de estadística clásica la solución del problema anterior está dada por métodos como la Regresión Ridge (RR), que puede ser visto como una aplicación de la regularización de Tikhonov. Este problema también se clasifica como un problema inverso lineal discreto, donde debido a los datos conocidos y un operador lineal queremos inferir acerca de los parámetros de regresión desconocidos. Para la interpretación Bayesiana de los métodos de regularización en regresión, se supone una distribución a priori para los parámetros de regresión; sin embargo, las inferencias posteriores son muy sensibles al cambio de la distribución *a priori* y a la información limitada de los datos comparados con la cantidad de información previa. Esta sensibilidad se refleja tanto en el sobreajuste como en el poco poder predictivo del modelo ajustado.

EL objetivo principal de este trabajo es proponer una distribución *a priori* mínimo informativa para el vector $\boldsymbol{\beta}$ basada en la divergencia de Kullback-Leibler (KL), obtener la distribución *a posteriori* correspondiente y evaluar el poder predictivo del método propuesto. Este

documento está estructurado de la siguiente manera: en el Capítulo 1, se describe de forma general el problema bajo estudio y los métodos alternativos de solución; en el Capítulo 2, se revisa el modelo de regresión lineal múltiple, supuestos y métodos de regresión penalizada; en el Capítulo 3, se hace una breve revisión de las distribuciones mínimo informativas propuestas por Zellner y Siow (1980), Jeffreys (1961), las distribuciones de referencia; se describe la propuesta de García-Donato (2003) y el resultado de Pérez (2005) para la contrucción de la distribución *a priori* convencional basada en la divergencia de Kullback-Leibler; en el Capítulo 4, se obtiene la distribución inicial mínimo informativa o convencional para el modelo lineal normal, de rango completo e incompleto; en el Capítulo 5, se obtiene la distribución *a posteriori* correspondiente a la distribución inicial propuesta, dado que no es sencillo obtener las distribuciones marginales de los parámetros, se describen métodos de simulación como el mestreador de Gibbs para muestrear de la distribución conjunta; además, se describe el método de validación cruzada utilizado para comparar el poder de predicción del método propuesto; se presentan los resultados de la simulación, discusión y Por último se presentan las conclusiones finales y algunas recomendaciones. se describe el algoritmo de simulación empleado en esta investigación y por último, se presentan los resultados obtenidos y conclusiones finales.

Capítulo 2

REGRESIÓN LINEAL MÚLTIPLE

La predicción de valores genéticos mediante la implementación de métodos de regresión utilizando el genoma ha contribuido con al mejoramiento de plantas y animales (de los Campos y Calus, 2012a, Heffner *et al.*, 2009, Heslot *et al.*, 2012, Lorenz *et al.*, 2011, Meuwissen *et al.*, 2001). Además son útiles para el estudio la genética humana (de los Campos *et al.*, 2010a, de los Campos y Allison, 2012b, Makowsky *et al.*, 2011, Ober *et al.*, 2012, Vazquez *et al.*, 2012). El modelo paramétrico básico para selección genómica fue propuesto por Meuwissen *et al.* (2001)

$$y = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + e_i \tag{2.1}$$

o

$$\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.2}$$

en forma matricial, donde $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\boldsymbol{I})$, $\boldsymbol{y} = (y_1, ..., y_n)^t$ es el vector de valores del genotipo, $\boldsymbol{X}$ es la matriz diseño de orden $n \times p + 1$ de los marcadores del genotipo, codificados como -1, 0, 1, para aa, Aa , y

AA respectivamente, donde (aa, AA) son homocigotos, es decir, poseen 2 copias(alelos) idénticas de un gen específico y Aa es un heterocigoto ya que sus alelos son diferentes y $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^t$ es el vector de parámetros desconocidos(efectos de los marcadores).

En genómica se manejan grandes bases de datos donde el número de observaciones del fenotipo, n, es mucho menor que el número de marcadores, p, $(n << p)$, lo que implica que los métodos convencionales de estimación como Máxima Verosimilitud (MV) y Mínimos Cuadrados Ordinarios (MCO) no son viables debido que la inversa de $\boldsymbol{X}^t\boldsymbol{X}$ puede no existir ya que es muy probable que haya combinaciones lineales entre las columnas de $\boldsymbol{X}$, es decir, existen dependencias lineales cercanas conocidas como colinealidad. En tales circunstancias se utilizan métodos de regularización.

2.1. Métodos de regresión penalizada

Estos métodos se basan en la minimización de la suma de cuadrados del error (SCE) sujeta a restricciones sobre los posibles valores de los estimadores para reducir su varianza, obteniéndose así predicciones más precisas. A continuación se describen brevemente algunos de estos métodos.

2.1.1. Regresión Ridge

El estimador tipo Ridge reduce el error cuadrático medio penalizando la SCE con la norma L_2. Así, minimizando

$$\text{mín}_\beta \left\{ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + k \|\boldsymbol{\beta}^t \boldsymbol{\beta}\|^2 \right\} \tag{2.3}$$

se tiene que el estimador Ridge es:

$$\hat{\boldsymbol{\beta}}_R = \left(\boldsymbol{X}^t \boldsymbol{X} + kI \right)^{-1} \boldsymbol{X}^t \boldsymbol{y} \tag{2.4}$$

donde k es el parámetro de Ridge, el cual no tiene un valor específico, algunas propuestas para determinar su valor (le Cessie y van Houwelingen, 1992, Lee y Silvapulle, 1988, Schaefer *et al.*, 1984) son :

1. $k_0 =$ el valor de $\boldsymbol{\beta}_R$ donde se estabiliza el gráfico conocido como la traza de Ridge.
2. $k_1 = \dfrac{1}{\boldsymbol{\beta}^t \boldsymbol{\beta}}$
3. $k_2 = \dfrac{p+1}{\boldsymbol{\beta}^t \boldsymbol{\beta}}$
4. $k_3 = \dfrac{(\lambda_1 - 100\lambda_p)}{99}$; donde λ_1 y λ_2 son el mayor y menor valor propio de $\boldsymbol{X}^t \boldsymbol{X}$

donde $\|\boldsymbol{\beta}^t \boldsymbol{\beta}\|^2$ es la norma L_2 de $\boldsymbol{\beta}$ que tiene el efecto de contraer los coeficientes de regresión estimados hacia cero, lo que introduce un sesgo pero al mismo tiempo reduce la varianza de los estimadores (Hoerl y Kennard, 1970, Zou y Hastie, 2005). El estimador Ridge incluye todas las covariables en el modelo, por lo que no es un método apropiado cuando el objetivo es la selección de variables. Otra desventaja del

método es que k depende de $\boldsymbol{\beta}$, que es desconocido, y que penaliza a todos los $\boldsymbol{\beta}_i$ estén o no en la relación de colinealidad.

2.1.2. Regresión LASSO

la regresión LASSO(Least Absolute Shrinkage and Selection Operator) fue propuesta por Tibshirani (1996), donde el estimador LASSO se obtiene como solución de la minimizar:

$$\text{mín}_{\beta} \left\{ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\| \right\} \tag{2.5}$$

para algún $\lambda > 0$, donde $\|\boldsymbol{\beta}\|$ es la norma L_1. En la solución algunos coeficientes se hacen cero, lo que hace al modelo parsimonioso. Efron *et al.* (2004) propusieron el algoritmo LARS (Least Angle Regression) que produce modelos estimados de forma semejante a los procedimientos de selección de variables hacia adelante y hacia atrás, pero que también puede obtener estimaciones LASSO.

Así como la regresión Ridge, el método LASSO también presenta algunas desventajas, (Zou y Hastie, 2005). Así, cuando $p > n$, LASSO selecciona a lo más n variables; en presencia de altas correlaciones por pares en subconjuntos de variables, elige una en cada subconjunto sin importar cual de ellas es seleccionada, y si todos los predictores estan correlacionados la regresión LASSO es superada por la regresión Ridge (Kyung *et al.*, 2010, Tibshirani, 1996, Zou y Hastie, 2005).

2.1.3. Elastic-Net (E-N)

La regresión Elastic-Net es una generalización de la regresión LASSO y la regresión Ridge, donde la penalización de la suma de cuadrados del

error es una combinación de las normas L_1 y L_2. Esto es:

$$\text{mín}_\beta \left\{ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda_1 \|\boldsymbol{\beta}\| + \lambda_2 \|\boldsymbol{\beta}^t \boldsymbol{\beta}\|^2 \right\}, \quad \lambda_1, \lambda_2 > 0 \quad (2.6)$$

De forma equivalente, para las constantes no negativas λ_1 y λ_2 se tiene:

$$\text{mín}_\beta \left\{ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \right\}$$

sujeto a la restricción:

$$(1 - \alpha) \sum_{j=1}^{p} |\beta_j| + \alpha \sum_{j=1}^{p} \beta_j^2 \leq t$$

donde $t > 0$ y $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$, para $\alpha \in (0, 1)$ tiene como solución el estimador Elastic-Net. Para encontrar el estimador del Elastic Net de una manera eficiente, Zou y Hastie (2005) proponen el algoritmo LARS-EN basado en el algoritmo LARS de Efron *et al.* (2004).

2.1.4. BLUP

Los modelos mixtos son un extensión de los modelos lineales, en el cual algunos de los parámetros son efectos aleatorios. Así el modelo lineal mixto es:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e} \quad (2.7)$$

donde $\boldsymbol{X}$ y $\boldsymbol{Z}$ son matrices conocidas, $\boldsymbol{\beta}$ es el vector de efectos fijos y $\boldsymbol{u}$ es el vector de efectos aleatorios, además

$$\text{E} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{e} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix} \qquad \text{y} \qquad \text{Cov} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{e} \end{bmatrix} = \begin{bmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{bmatrix}$$

Para estimar $\boldsymbol{\beta}$ y predecir $\boldsymbol{u}$, dadas las matrices de covarianzas $\boldsymbol{G}$ y $\boldsymbol{R}$, el modelo (2.7) se reescribe como sigue:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\xi}, \tag{2.8}$$

donde $\boldsymbol{\xi} = \boldsymbol{Zu} + \boldsymbol{e}$, se tiene que $\text{Cov}(\boldsymbol{\xi}) \equiv \boldsymbol{V} = \boldsymbol{ZVZ}^t = \boldsymbol{R}$, es decir, estan correlacionados. Bajo ciertas condiciones de regularidad Henderson (1953) desarrolla las ecuaciones del modelo mixto(MME, por sus siglas en inglés), que simultaneamente produce el mejor estimador lineal insesgado(BLUE, siglas en inglés) de $\boldsymbol{X\beta}$ (o para cualquier función estimable $\boldsymbol{K^t\beta}$) y el mejor predictor lineal insesgado(BLUP) de $\boldsymbol{u}$(o de cualquier vector $\boldsymbol{w} = \boldsymbol{K^t\beta} + \boldsymbol{L^tu}$)

$$\begin{pmatrix} \boldsymbol{X}^t\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}^t\boldsymbol{R}^{-1}\boldsymbol{X} \\ \boldsymbol{Z}^t\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}^t\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{R}^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ \boldsymbol{u} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^t\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}^t\boldsymbol{R}^{-1}\boldsymbol{y} \end{pmatrix} \tag{2.9}$$

Para más detalles ver (Christensen, 2001, Gianola, 2013, Gianola, Ruppert *et al.*, 2003, Witkovsky, 2013), entre otros.

2.1.5. Regresión RKHS

La regresión RKHS (Reproducing kernel Hilbert space) es un método de regresión semiparamétrica que ha sido utilizado para la predicción genómica (Crossa *et al.*, 2010, Gianola *et al.*, 2006, Gianola y van Kaam, 2008). En este método se utiliza una función llamada *kernel* mediante la cual se transforma el conjunto de datos (marcadores) en un conjunto de medidas de similitud, por ejemplo distancias, entre pares de observaciones, lo que da como resultado una matriz cuadrada que es utilizada como matriz diseño en un modelo lineal normal. Dado que la regresión RKHS no asume linealidad, se considera que se podría captar mucho

mejor los efectos no aditivos. El modelo puede ser formulado como sigue (Heslot *et al.*, 2012):

$$Y = W\boldsymbol{\mu} + \boldsymbol{K}_h\alpha + \varepsilon \tag{2.10}$$

donde

1) $\boldsymbol{\mu}$ es un vector de efectos fijos,

2) Se asume que $\varepsilon \sim N(0, I\sigma^2_\varepsilon)$ y que, de forma independiente, *a priori* $\alpha \sim N(0, \boldsymbol{K}_h\sigma^2_\alpha)$.

3) $\boldsymbol{K}_h$: Es una matriz de incidencias, que depende de la elección del kernel y del parámetro h, que de forma semejante a una matriz de correlaciones mide la similitud genómica entre los genotipos.

Un ejemplo de Kernel es el Gaussiano:

$$\boldsymbol{K}_h(x_i, x_j) = exp(-hd_{ij}) = exp\left(-\frac{\theta}{k}d_{ij}\right)$$

donde d_{ij} es la distancia de los marcadores del individuo i con el j, algunas de estas distancias pueden ser; la distancia Euclideana y la distancia de Manhattan. Para más detalles de la regresión RKHS aplicada a la predicción basada en el genoma ver: Crossa *et al.* (2010), de los Campos *et al.* (2010b, 2009), Gianola y van Kaam (2008), Schaid (2010).

2.2. Enfoque Bayesiano

Bajo éste enfoque se tiene que el parámetro o vector de parámetros poblacional θ sobre el cual se desea hacer inferencia es considerado co-

mo una cantidad desconocida a la que se asigna una distribución *a priori* para representar la incertidumbre sobre su valor verdadero. Así, la inferencia bayesiana se basa en $\pi(\theta \mid x)$; esto es, en la distribución del parámetro dados los datos. Los métodos Bayesianos de inferencia reciben este nombre por que son capaces de sintetizar la información muestral y la llamada información *a priori* (no muestral) utilizando el Teorema de Bayes. Así, dado que el objetivo es relacionar probabilísticamente a un parámetro θ con los datos, el teorema de Bayes puede presentarse en términos de densidades:

$$\pi(\theta \mid x) = \frac{\pi(\theta, x)}{\pi(x)} = \frac{\pi(x \mid \theta)\pi(\theta)}{\pi(x)} \tag{2.11}$$

donde $\pi(x)$ puede ser obtenido, dependiendo de si θ es continuo o discreto, de la siguiente manera:

$$\pi(x) = \begin{cases} \sum_\theta \pi(x \mid \theta)\pi(\theta) & \text{si} \quad \theta \quad \text{es discreto} \\ \int_\theta \pi(x \mid \theta) f(\theta) d\theta & \text{si} \quad \theta \quad \text{es continuo} \end{cases} \tag{2.12}$$

En la expresión (2.11):

a) $f(\theta)$ representa, en términos probabilísticos, lo que es conocido de θ antes de recolectar los datos y es llamada distribución *a priori* de θ.

b) $\pi(\theta \mid x)$ representa lo que se conoce de θ después de recolectar los datos y es llamada la distribución *a posterior* de θ dado x y

c) $\pi(x \mid \theta)$ es la distribución fundamental que incorpora al modelo la información proporcionada por los datos

Dado que x es conocido y θ no, $\pi(x \mid \theta)$ puede ser reconocido como una función de θ dado un valor fijo de x, a la cual se le denomina la función de verosimilitud de θ dado x que se denota usualmente por $L(\theta \mid x)$.

Una forma equivalente de presentar $\pi(\theta \mid x)$ es omitiendo el factor $\pi(x)$ ya que no depende de θ y, al ser x fijo, puede ser considerado como una constante. Entonces:

$$\pi(\theta \mid x) \propto L(x \mid \theta)\pi(\theta) \propto L(\theta \mid x)\pi(\theta)$$

donde el lado derecho, es la distribución *a posteriori* no normalizada que tiene aplicaciones en regresión bayesiana.

2.2.1. Regresión Ridge Bayesiana (RRB)

La RRB se obtiene al asignar una distribución *a priori* al vector de parametros $\boldsymbol{\beta}$ que es el producto de distribuciones normales independientes dada por (2.13) , donde σ^2_β es la varianza *a priori* común de los efectos.

$$\pi\left(\boldsymbol{\beta}\right) = \left(\frac{1}{2\pi\sigma^2_\beta}\right)^{\frac{p}{2}} \exp\left\{-\frac{1}{2\sigma^2_\beta}\boldsymbol{\beta}^t\boldsymbol{\beta}\right\} \tag{2.13}$$

Si se asume que $\boldsymbol{\epsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2\boldsymbol{I})$ y la distribución *a priori* no informativa $\pi(\sigma^2) \propto 1/\sigma^2$, se tiene que la distribución *a posteriori* de $\boldsymbol{\beta}$ es (2.16), cuya media *a posteriori* dada por (2.19) es un estimador bayesiano de $\boldsymbol{\beta}$. Esto es:

$$\begin{aligned} \pi(\boldsymbol{\beta}|\sigma^2_\beta, \sigma^2, \boldsymbol{y}) \;\; &\propto \;\; exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right\} \\ &\text{x} \;\; exp\left\{-\frac{1}{2\sigma^2_\beta}\boldsymbol{\beta}^t\boldsymbol{\beta}\right\}\pi(\sigma^2) \end{aligned} \tag{2.14}$$

desarrollando y factorizando (2.14)

$$\pi\left(\boldsymbol{\beta}|\sigma^2,\sigma^2_\beta,\boldsymbol{y}\right) \propto exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\boldsymbol{X\beta}\right)^t\left(\boldsymbol{y}-\boldsymbol{X\beta}\right)\right\} exp\left\{-\frac{1}{2\sigma^2_\beta}\boldsymbol{\beta}^t\boldsymbol{\beta}\right\}$$
$$\propto exp\left\{-\frac{1}{2\sigma^2}\left[-2\boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{y}+\boldsymbol{\beta}^t\left(\boldsymbol{X}^t\boldsymbol{X}+\frac{\sigma^2}{\sigma^2_\beta}I\right)\boldsymbol{\beta}\right]\right\} \tag{2.15}$$

Completando la forma cuadrática en (2.15) se tiene :

$$\pi\left(\boldsymbol{\beta}|\sigma^2_\beta,\sigma^2,Y\right) \propto exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{a})^t\,\sigma^*\,(\boldsymbol{\beta}-\boldsymbol{a})\right\} \tag{2.16}$$

Por lo anterior, la distribución final en (2.16) es una distribución normal, es decir, $\boldsymbol{\beta} \mid \boldsymbol{X} \sim N(\boldsymbol{a},\boldsymbol{\sigma}^*)$ con parámetros

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

y

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

Una vez obtenida la distribución final, se procede a calcular el estimador del vector de parámetros desconocidos. Aplicando la definición del estimador Bayes se tiene:

$$T = arg\,\text{mín}_t E[l(\boldsymbol{\beta},t)] \tag{2.17}$$

donde $l(\boldsymbol{\beta},t)$ es una función de pérdida que minimiza el estimador de Bayes. Algunas funciones de pérdida son:"todo o nada", "cuadrática" y la "absoluta".

En particular, tomando la pérdida cuadrática $l(t, \boldsymbol{\beta}) = (t-\boldsymbol{\beta})^2$, se tiene:

$$T = E[l(t, \beta)] = \int_{\Re} (t - \boldsymbol{\beta})^2 \pi \left(\boldsymbol{\beta} | \sigma_\beta^2, \sigma^2, \boldsymbol{y}\right) d\boldsymbol{\beta} \tag{2.18}$$

Dado que la distribución final es normal, se obtiene:

$$T = \hat{\beta} = E[\boldsymbol{\beta}|\boldsymbol{y}] = \left(\boldsymbol{X}^t\boldsymbol{X} + k\boldsymbol{I}\right)^{-1} \boldsymbol{X}^t\boldsymbol{y} \tag{2.19}$$

es el estimador de Ridge ordinario o clásico, donde $k = \sigma^2/\sigma_\beta^2$.

2.2.2. LASSO Bayesiano (LB)

Park y Casella (2008) proponen un enfoque bayesiano para la regresión LASSO, considerando *a priori* para cada elemento de $\boldsymbol{\beta}$ una distribución condicional Laplace (Doble Exponencial) tal que:

$$\pi(\boldsymbol{\beta}|\sigma) = \prod_{j=1}^{p} \left(\frac{\lambda}{2\sigma}\right) exp\left\{-\frac{\lambda|\beta_j|}{\sigma}\right\} \tag{2.20}$$

y una *a priori* no informativa para σ^2 dada por :

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2} \tag{2.21}$$

de modo que la distribución final está dada por:

$$\pi(\boldsymbol{\beta} \mid \sigma, \lambda) = \prod_{j=1}^{p} \pi(\boldsymbol{\beta}_j|\sigma) \prod_{j=1}^{n} \mathrm{N}(y_i \mid \boldsymbol{x}_i^t\beta, \sigma^2)\pi(\sigma^2) \tag{2.22}$$

Dado que no es posible obtener de forma analítica la distribución *a posteriori*, se utiliza el muestreador de Gibbs en un modelo Jerarquico para más detalles ver Park y Casella (2008). Al respecto, Pérez *et al.* (2010) desarrollaron un paquete para R (R Development Core Team,

2011) llamado BLR (Bayesian Linear Regression), el cual implementa de forma eficiente un muestreo de Gibbs para los métodos de RRB y LB.

2.2.3. Elastic-Net Bayesiano (EN-B)

En el enfoque bayesiano de Elastic-Net Li y Lin (2010) demuestran que obtener el estimador de (E-N) es equivalente a encontrar la moda marginal *a posteriori* de $\boldsymbol{\beta}|\boldsymbol{y}$ cuando la distribución *a priori* de $\boldsymbol{\beta}$ es:

$$\pi(\boldsymbol{\beta}|\sigma) \propto exp\left\{-\frac{1}{2\sigma}\left(\lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=1}^{p}\beta_j^2\right)\right\} \tag{2.23}$$

y la distribución *a priori* para σ^2 es (2.21).

Existen otros métodos de predicción genómica agrupados en lo que Gianola (2013), Gianola *et al.* (2009) llama el *alfabeto bayesiano*, además de los basados en redes neuronales González-Camacho *et al.* (2012).

2.3. Discusión

De acuerdo con lo anterior, las distribuciones *a priori* más utilizadas para la regresión lineal cuando $p >> n$ son las informativas; sin embargo, en selección genómica se asume que los valores de los efectos de los marcadores son muy cercanos a cero, por lo que es razonable suponer distribuciones que tengan mayor densidad de masa alrededor del cero, pero el incoveniente es que producen estimaciones de los efectos demasiado contraídas hacía dicho valor. Esto es, las distribuciones a priori informativas, tienen el inconveniente de que pueden conducir a un sobreajuste del modelo, dado que aportan demasiada información en relación a la proporcionada por los datos. Por el contrario, las distribuciones no informativas o convencionales típicamente tienen colas más densas por lo que podrían producir una menor contracción de las estimaciones de los marcadores con efectos considerables; ademas, éstas distribuciones aportan mínima información, lo que que evita el sobreajuste del modelo. No obstante, en general, las distribuciones no informativas para el problema de interés son impropias y conducen a distribuciones finales impropias. En el siguiente capítulo se realiza una revisión de algunas distribuciones *a priori* no informativas para el modelo de regresión encontradas en la literatura; en particular se revisan las propuestas de Berger y Bernardo (1989, 1992a,b), Jeffreys (1961), Zellner (1986), Zellner y Siow (1980). Especial atención recibe la propuesta de García-Donato (2003) que, en el contexto de selección de modelos, propone una distribución a priori basada en la discrepancia de Jeffreys.

Capítulo 3

DISTRIBUCIONES INICIALES NO INFORMATIVAS

En ciertos problemas de investigación, el conocimiento inicial sobre el verdadero valor del parámetro θ de un modelo $p(x \mid \theta)$ puede ser muy vago o nulo, lo que ha llevado a generar un tipo de distribuciones iniciales llamadas distribuciones *a priori* no informativas, las cuales reflejan un estado de ignorancia inicial, con el propósito de dejar que los datos hablen por sí mismos.

En el enfoque bayesiano se combina la información sobre θ contenida en x con la información disponible antes de que se realize el experimento. Esta información*a priori* se resume en la distribución inicial $\pi(\theta)$. Algunas distribuciones inciales son impropias, esto es,

$$\int_{\theta} \pi(\theta) d\theta = \infty$$

No obstante, siempre que $\pi(x) < \infty$ en (2.12) se tiene que la distribución *a posteriori* $\pi(\theta \mid x)$ es propia. Así, surge la pregunta ¿Cómo asignar distribuciones iniciales?, por lo que se han implementado di-

versos métodos para obtener distribuciones no informativas, donde el procedimiento más antiguo es el de (Laplace, 1951) Laplace o 'principio de razón insuficiente" que consiste en asignar una distribución uniforme a θ. A continuación se expone un método propuesto por Jeffreys (1961) que está basado en la información esperada de Fisher.

3.1. Distribuciones de Jeffreys

La regla de Jeffreys nos permite obtener distribuciones *a priori* invariantes ante transformaciones de θ y se aplica cuando no se tiene información sobre θ. En la literatura se ha encontrado que éstas distribuciones han sido utilizadas para el problema de contraste de hipótesis o equivalentemente para la selección de covariables (modelos competidores).
La distribución *a priori* de Jeffreys para θ esta dada por

$$f(\theta) \propto I(\theta)^{\frac{1}{2}} \tag{3.1}$$

donde $I(\theta)$ es la matriz de información de Fisher

$$I(\theta) = -E\left[\frac{\partial^2 log f(x|\theta)}{\partial \theta^2}\right]$$

Cuando se tiene más de un parámetro desconocido la distribución *a priori* se define como sigue:

$$f(\boldsymbol{\theta}) \propto \sqrt{|I(\boldsymbol{\theta})|} \tag{3.2}$$

donde $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)^t$ y $I(\boldsymbol{\theta})$ es la matriz de derivadas parciales

$$I(\boldsymbol{\theta}) = -E\left[\frac{\partial^2 log f(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j}\right]$$

Otras distribuciones que estan ligadas con la distribución de Jefreys son las de Zellner.

3.2. Distribuciones de Zellner

Considere el problema de seleción de modelos de regresión lineal, con vector de respuesta $\boldsymbol{y} = (y_1, ..., y_n)^t \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{I}_n/\phi)$, donde ϕ es el parámetro de precisión y la matriz identidad $\boldsymbol{I}_n$ de $n \times n$. Dado un conjunto de posibles variables predictoras $\boldsymbol{X}_1, ..., \boldsymbol{X}_p$. Se asume que el vector de medias $\boldsymbol{\mu}$ se expande a $\mathbf{1}_n, \boldsymbol{X}_1, ..., \boldsymbol{X}_p$, donde $\mathbf{1}_n$ es un vector de unos de tamaño n. El problema de elección de modelos consiste en seleccionar un subconjunto de variables predictoras con ciertas restricciones adicionales sobre un subespacio que contiene la media. Bajo un modelo $\boldsymbol{\mathcal{M}}_\gamma$, $\boldsymbol{\mu}$ se puede expresar en forma vectorial como:

$$\boldsymbol{\mathcal{M}_\gamma} : \mathbf{1}_n\alpha + \boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma \tag{3.3}$$

note que, $\boldsymbol{\gamma}$ es un vector de variables indicadoras, esto es, si $\gamma_j = 1$, $\boldsymbol{X}_\gamma$ es incluida en el conjunto de variables predictoras ó, si $\gamma_j = 0$, $\boldsymbol{X}_\gamma$ es excluida, α es el intercepto común en todos los modelos, $\boldsymbol{X}_\gamma$ es la matriz diseño de $n \times p_\gamma$ bajo el modelo $\boldsymbol{\mathcal{M}}_\gamma$, $\boldsymbol{\beta}_\gamma$ es el vector de dimensión p_γ de coeficientes de regresión distintos de cero. para la selección de modelos en el enfoque bayesiano es necesario especificar distribuciones *a priori* para el vector de parámetros desconocidos $\boldsymbol{\theta}_\gamma = (\alpha, \boldsymbol{\beta_\gamma}, \phi) \in \boldsymbol{\Theta_\gamma}$.

(Zellner, 1986), bajo el modelo lineal normal propone distribuciones *a priori* basadas en la familia conjugada Normal-Gamma, llamadas g-*prior*, que se dan a continuación

$$p(\phi) \propto \frac{1}{\phi}, \qquad \boldsymbol{\beta}|\phi \sim N\left(\boldsymbol{\beta}_\alpha, \frac{g}{\phi}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\right)$$

donde

1) $\boldsymbol{\beta}_\alpha$ es la media *a priori*

2) La matriz de varianzas y covarianzas es un múltiplo escalar g de la matriz de informacion de Fisher

Estas distribuciones son muy utilizadas en regresión lineal para la selección de variables mediante factores Bayes.

Otra distribución inicial a utilizarse cuando se tiene poca o ninguna información es la dada por Zellner y Siow (1980), quienes proponen una distribución *a priori* Cauchy multivariada para los coeficientes de regresión, definida como sigue:

$$\pi(\boldsymbol{\beta}_\gamma|\phi) \propto \frac{\Gamma(p_\gamma/2)}{\pi^{p_\gamma/2}} \left|\frac{\boldsymbol{X}^t\boldsymbol{X}}{n/\phi}\right|^{\frac{1}{2}} \left(1 + \frac{\boldsymbol{\beta}_\gamma^t \boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta}_\gamma}{n/\phi}\right)^{-p_\gamma/2} \tag{3.4}$$

es una distribución Cauchy multivariada centrada en $\boldsymbol{\beta}_\gamma = \mathbf{0}$ y la matriz de precisión tiene la forma de la "matriz de información de Fisher". Note que la distribución Cauchy se puede expresar como una mezcla de distribuciones normales. Esto es, la distribución de (Zellner y Siow, 1980) se puede representar como una mezcla, donde la distribución *a priori* de g es $\text{Inv} - \text{Gamma}(1/2, n/2)$, de este modo:

$$\pi(\boldsymbol{\beta}_\gamma \mid \phi) \propto \int \mathrm{N}\left(\boldsymbol{\beta}_\gamma \mid \mathbf{0}, \frac{g}{\phi}(\boldsymbol{X}_\gamma^t \boldsymbol{X}_\gamma)^{-1}\right) \pi(g)\mathrm{d}g \tag{3.5}$$

con $\pi(g) = \dfrac{(n/2)^{\frac{1}{2}}}{\Gamma(1/2)} g^{-3/2} e^{-n/(2g)}$, obteniendo de esta manera grandes ventajas computacionales.

3.3. Distribuciones de referencia

Las distribuciones de referencia son una representación matemática de la falta de conocimiento a priori sobre la magnitud de interés y que parecen superar algunas de las dificultades a las que se enfrentan otras definiciones de distribuciones iniciales objetivas. Las distribuciones de referencia pueden ser impropias y, en tales casos, sólo son herramientas matemáticas que se utilizan para hacer inferencias bayesianas basadas únicamente en los datos y el modelo. En tal situación se usa el término. Las funciones iniciales de referencia propuestas por Bernardo (1979), Berger y Bernardo (1989) , Berger y Bernardo (1992b) y Berger y Bernardo (1992a) son distribuciones a priori que, dado un modelo paramétrico, permiten hacer inferencias sobre un parámetro de interés basadas sólo en el modelo y los datos disponibles.

La idea detrás de la función inicial de referencia es que, ésta aproxima a una inicial $\pi(\theta)$, para el parámetro de interés θ con respecto a la cual es maximizado el valor esperado de la información desconocida sobre θ.

Definición 3.1 ***Distribuciones de referencia unidimensionales.***

Sea y el resultado de un experimento e, que consiste en una observación de $p(y|\theta)$, $y \in \mathcal{Y}$, $\theta \in \Theta \subseteq \mathbb{R}$; sea $\boldsymbol{z}_k = \boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_k$ el resultado de k repeticiones independientes de e La distribucion de referencia unidimensional se define por

$$f_k^*(\theta) = exp\left\{\int p(\boldsymbol{z}_k|\theta) \log p^*(\theta|\boldsymbol{z}_k) d\boldsymbol{z}_k\right\}, \tag{3.6}$$

donde

$$p^*(\theta|\boldsymbol{z}_k) = \frac{\prod_{i=1}^k p(\boldsymbol{y}_i|\theta)}{\int \prod_{i=1}^k p(\boldsymbol{y}_i|\theta) d\theta}.$$

Así la distribución de θ después de que se observó $\boldsymbol{y}$ esta dada por $\pi(\theta|\boldsymbol{y})$, de modo que

$$E[\delta\{\pi_k(\theta|\boldsymbol{y}), \pi(\theta|\boldsymbol{y})\} \to 0], \text{cuando } k \to \infty.$$

donde $\delta(g, h) = \text{mín}\{K(g|h), K(h|g)\}$, es conocida como la discrepancia intrinseca entre g y h, que es una simetrización de la divergencia de Kullback-Leiber, y

$$\pi_k(\theta|\boldsymbol{y}) = c_k(\boldsymbol{y}) p(\boldsymbol{y}|\theta) f_k^*(\theta);$$

donde c_k son constantes de normalización. Así para $\pi(\theta)$ positiva, $c(\boldsymbol{y}) > 0$ y $\forall \theta \in \Theta$

$$\pi(\theta|\boldsymbol{y}) = c(\boldsymbol{y}) p(\boldsymbol{y}|\theta) \pi(\theta) \tag{3.7}$$

Es la FIR de θ para el experimento e. La función $p^*(\theta \mid \boldsymbol{z}_k)$ que aparece en la definición es igual a la densidad final para θ, si se utiliza una inicial uniforme y son observados los datos $\boldsymbol{z}_k$. No es necesario utilizar esta forma de $p^*(\theta \mid \boldsymbol{z}_k)$, ya que, como Bernardo y Smith (1994) enfatizan, se puede utilizar cualquier aproximación asintótica de la distribución final de θ.

Definición 3.2 ***Distribuciones de referencia y un parámetro de ruido***.
Bernardo y Smith (1994) Sea y el resultado de un experimento e que consiste en una observación del modelo probabilístico $p(\boldsymbol{y}|\phi,\lambda)$, $\boldsymbol{y}\in Y$, $\phi\in\Phi\subseteq\mathbb{R}$, $\lambda\in\Lambda\subseteq\mathbb{R}$.
La distribución final de referencia $\pi(\phi|\boldsymbol{y})$ para ϕ, relacionada con el experimento e y la sucesión creciente de subconjuntos de Λ, $\{\Lambda_i(\phi)\}$, $\quad\phi\in\Phi$; $\quad\cup_i\Lambda_i(\phi)=\Lambda$, se define como el resultado del siguiente proceso:

i) Aplicando la definición 3.1 a $p(\boldsymbol{y}|\phi,\lambda)$, para un ϕ fijo, se obtiene la FIR condicional, $\pi(\lambda|\phi)$, para λ

ii) Se normaliza $\pi(\lambda|\phi)$ en cada Λ_i para obtener una sucesión de distribuciones iniciales propias $\pi_i(\lambda|\phi)$;

iii) Utilizar éstas para obtener una sucesión de modelos integrados.

iv) Con estos se obtiene la sucesión de funciones iniciales de referencia

$$\pi_i(\phi)=c\lim_{k\to\infty}\frac{f_k^*(\phi)}{f_k^*(\phi_0)},$$

con

$$f_k^*(\theta)=exp\left\{\int p_i(\boldsymbol{z}_i|\phi)\log p^*(\phi|\boldsymbol{z}_k)d\boldsymbol{z}_k\right\},$$

donde

$\boldsymbol{z}_k$ es una repetición k-dimensional de e de una observación del modelo integrado $p_i(\boldsymbol{y}|\phi)$. Para los datos $\boldsymbol{y}$ se obtienen las distribuciones finales correspondientes

$$\pi_i(\phi|\boldsymbol{y})\propto\pi(\phi)\int_{\Lambda_i(\phi)}p(\boldsymbol{y}|\phi,\lambda)\pi(\lambda|\phi)d\lambda;$$

v) Por otro lado se define $\pi(\phi|\boldsymbol{y})$ tal que $\delta\{\pi_i(\phi|\boldsymbol{y}),\pi(\phi|\boldsymbol{y})\}\to 0$ cuando $i\to\infty$.

La FIR de la parametrización ordenada (ϕ, λ) es una función positiva $\pi(\phi, \lambda)$ tal que

$$\pi(\phi|\boldsymbol{y}) \propto \int p(\boldsymbol{y}|\phi, \lambda)\pi(\phi, \lambda)d\lambda.$$

Esta se obtiene simplemente como

$$\pi(\phi, \lambda) = \lim_{i\to\infty} \frac{\pi_i(\phi)\pi_i(\lambda|\phi)}{\pi_i(\phi_0)\pi_i(\lambda_0|\phi_0)}$$

Claramente este algoritmo requiere que el investigador seleccione de manera apropiada la sucesión, $\Lambda_i(\phi)$; en general, la FIR resultante dependerá de dicha selección.

Proposición 3.1 ***Función inicial de referencia bajo normalidad asintótica***

Considere el experimento e_n, que consiste en la observación de una muestra aleatoria $\boldsymbol{y} = y_1, \ldots, y_n$ de $p(y|\phi, \lambda)$, $(\phi, \lambda) \in \Phi \times \Lambda \subseteq \mathbb{R} \times \mathbb{R}$ y sea $\{\Lambda_i(\phi)\}$ una sucesión apropiada de subconjuntos de Λ, como en la Definición 3.2. Suponga que la distribución final conjunta asintótica de (ϕ, λ), dada una repetición k-dimensional de e_n, es una normal multivariada con matriz de precisión $kn\boldsymbol{H}(\hat{\phi}_{kn}, \hat{\lambda}_{kn})$, donde

- $(\hat{\phi}_{kn}, \hat{\lambda}_{kn})$ es un estimador consistente de (ϕ, λ)
- $\hat{h}_{ij} = h_{ij}(\hat{\phi}_{kn}, \hat{\lambda}_{kn})$, $i = 1, 2$, $j = 1, 2$ es la partición de $\boldsymbol{H}$ correspondiente a ϕ, λ.

Entonces

$$\pi(\lambda|\phi) \;\propto\; \{h_{22}(\phi,\lambda)\}^{1/2}, \quad \text{y}$$

$$\pi(\phi,\lambda) \;=\; \pi(\lambda|\phi)\lim_{i\to\infty}\left\{\frac{\pi_i(\phi)c_i(\phi)}{\pi_i(\phi_0)c_i(\phi_0)}\right\}, \quad \phi_0 \in \Phi$$

definen la FIR correspondiene a la parametrización ordenada (ϕ,λ) donde

$$\pi_i(\phi) \;\propto\; exp\left\{\int_{\Lambda_i(\phi)} \pi_i(\lambda|\phi) \quad \log\left(\{h_\phi(\phi,\lambda)\}^{1/2}\right) d\lambda\right\}, \quad \text{con}$$

$$\pi_i(\lambda|\phi) \;=\; c_i(\phi)\pi(\lambda|\phi) = \frac{\pi(\lambda|\phi)}{\int_{\Lambda_i(\phi)}\pi(\lambda|\phi)d\lambda} \quad \text{y}$$

$$h_\phi \;=\; (h_{11} - h_{12}h_{22}h_{21})$$

En algunos casos $\{\Lambda_i\}$ no depende de ϕ y $\{h_{22}(\phi,\lambda)\}$; $\{h_\phi(\phi,\lambda)\}$ se pueden factorizar en funciones separadas de ϕ y λ, obteniendo la FIR más simple.

3.3.1. Función inicial de referencia multiparámetrica

El procedimiento anterior contempla sólo un parámetro de ruido y se basa en la parametrización ordenada (ϕ,λ), con la primer componente llamada parámetro de interés y de segunda el parámetro ruido. La función de referencia $\pi(\phi,\lambda)$ para (ϕ,λ) se construyo de tal manera

que se pudiera expresar de la siguiente forma

$$\pi(\phi, \lambda) = \pi(\lambda|\phi)\pi(\phi).$$

Ahora considere que el vector de parámetros $\boldsymbol{\theta}$ tiene mas de 2 componentes, donde $\boldsymbol{\theta}$ es la siguiente parametrización ordenada $(\theta_1, \ldots, \theta_m)$ y generando por condicionamiento sucesivo una FIR para cada parametrización, esto es:

$$\pi(\boldsymbol{\theta}) = \pi(\theta_m|\theta, \ldots, \theta_{m-1}) \ldots \pi(\theta_2|\theta_1)\pi(\theta_1).$$

Proposición 3.2 *FIR bajo normalidad asintótica*

Bajo condiciones de regularidad Bernardo y Smith (1994) generalizan la proposición 3.1, la FIR $\pi(\boldsymbol{\theta})$, correspondiente a la parametrización ordenada $(\theta_1, \ldots, \theta_m)$, está dada por

$$\pi(\boldsymbol{\theta}) = \lim_{i \to \infty} \frac{\pi^l(\boldsymbol{\theta})}{\pi^l(\boldsymbol{\theta}^*)}, \quad \text{para algún} \quad \boldsymbol{\theta}^* \in \Theta$$

donde $\pi^l(\boldsymbol{\theta})$ está definida en la siguiente recursión:

i) Para $j = m$ y $\theta_m \in \Theta_m^l$,

$$\pi_m^l\left(\boldsymbol{\theta}_{[m-1]}|\boldsymbol{\theta}^{[m-1]}\right) = \pi_m^l(\theta_m|\theta_1, \ldots, \theta_{m-1}) = \frac{\{h_m(\boldsymbol{\theta})\}^{1/2}}{\int_{\Theta_m^l}\{h_m(\boldsymbol{\theta})\}^{1/2}d\theta_m}$$

ii) Para $j = m-1, m-2, \ldots, 2$ y $\theta_j \in \Theta_j^l$,

$$\pi_j\left(\boldsymbol{\theta}_{[j-1]}|\boldsymbol{\theta}^{[j-1]}\right) = \pi_{j+1}(\boldsymbol{\theta}_{[j]}|\boldsymbol{\theta}^{[j]})\left[\frac{exp\{\frac{1}{2}E_j^l[\log h_j(\boldsymbol{\theta})]\}}{\int_{\Theta_j^l} exp\{\frac{1}{2}E_j^l[\log h_j(\boldsymbol{\theta})]\}d\theta_j}\right],$$

donde

$$E_j^l[\log h_j(\boldsymbol{\theta})] = \int_{\Theta_{[j]}^l} \log h_j(\boldsymbol{\theta}) \pi_{j+1}^l \left(\boldsymbol{\theta}_{[j]} | \boldsymbol{\theta}^{[j]}\right) d\boldsymbol{\theta}_{[j]}.$$

iii) Para $j = 1$, $\boldsymbol{\theta}_{[0]} = \boldsymbol{\theta}$, con $\boldsymbol{\theta}^{[0]}$ vacio y

$$\pi(\boldsymbol{\theta}) = \pi_1^l \left(\boldsymbol{\theta}_{[0]} | \boldsymbol{\theta}^{[0]}\right)$$

Para seguir con más detalle como se demuestra esta proposición véase (Berger y Bernardo, 1992a,b).

En la siguiente sección se muestra un procedimiento para obtener una distribución *a priori* no informativa para selección de modelos.

3.4. Selección convencional de modelos lineales

García-Donato (2003) cita, como base la distribución de Jeffreys para desarrollar un método no informativo en la selección de modelos, que consiste en uns distribución convencional, obtenida a partir de la divergencia de Kullback-Leibler (Kullback, 1997).

3.4.1. Divergencia de Kullback-Leibler (DKL)

La divergencia de Kullback-Leibler es una medida de la similitud entre dos funciones de distribución de probabilidad $f(\boldsymbol{y} \mid \boldsymbol{\theta}_0, \boldsymbol{\eta})$ y $f(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\eta})$. La distribución $f(\boldsymbol{y} \mid \boldsymbol{\theta}_0, \boldsymbol{\eta})$ representa la distribución "verdadera" y $f(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\eta})$ representa una teoría, un modelo, descripción o aproximación de $f(\boldsymbol{y} \mid \boldsymbol{\theta}_0, \boldsymbol{\eta})$.

La DKL es el promedio de la diferencia entre los logartimos de las probabilidades $f(\boldsymbol{y} \mid \boldsymbol{\theta}_0, \boldsymbol{\eta})$ y $f(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\eta})$, donde el promedio se toma usando las probabilidades $f(\boldsymbol{y} \mid \boldsymbol{\theta}_0, \boldsymbol{\eta})$.

Definición 3.3 *Dados dos modelos* $f(\boldsymbol{y} \mid \boldsymbol{\theta}_0, \boldsymbol{\eta})$ *y* $f(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\eta})$, *se tiene que la divergencia entre* $\boldsymbol{\theta}_0$ *y* $\boldsymbol{\theta}$ *es:*

$$KL(\boldsymbol{\theta_0}, \boldsymbol{\eta} \mid \boldsymbol{\theta}, \boldsymbol{\eta}) = \int log \frac{f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\eta})}{f(\boldsymbol{y}|\boldsymbol{\theta}_0, \boldsymbol{\eta})} f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\eta}) d\boldsymbol{y} \tag{3.8}$$

Note que la divergencia es una medida no simétrica de la similitud o diferencia entre dos modelos que puede ser simetrizada. Bayarri y Garcia-Donato (2006) proponen lo siguiente:

$$D^S(\boldsymbol{\theta}, \boldsymbol{\theta_0}|\boldsymbol{\eta}) = KL(\boldsymbol{\theta}, \boldsymbol{\eta} \mid \boldsymbol{\theta_0}, \boldsymbol{\eta}) + KL(\boldsymbol{\theta_0}, \boldsymbol{\eta} \mid \boldsymbol{\theta}, \boldsymbol{\eta}) \tag{3.9}$$

$$D^M(\boldsymbol{\theta}, \boldsymbol{\theta_0}|\boldsymbol{\eta}) = 2\text{mín}\{KL(\boldsymbol{\theta}, \boldsymbol{\eta} \mid \boldsymbol{\theta_0}, \boldsymbol{\eta}), KL(\boldsymbol{\theta_0}, \boldsymbol{\eta} \mid \boldsymbol{\theta}, \boldsymbol{\eta})\} \tag{3.10}$$

donde D^S y D^M son la suma y el mínimo de las divergencias respectivamente. Una de las aplicaciones de esta medida es en pruebas de hipótesis, donde dados los datos $\boldsymbol{y}$, con densidad $f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\eta})$, probar las hipótesis:

$$H_1 : \boldsymbol{\theta} = \boldsymbol{\theta_0} \quad vs \quad H_2 : \boldsymbol{\theta} \neq \boldsymbol{\theta_0} \tag{3.11}$$

es equivalente al problema de selección entre el modelo 1 (M_1) y el modelo 2 (M_2):

$$M_1 : f_1(\boldsymbol{y}|\boldsymbol{\eta_1}) = f(\boldsymbol{y}|\boldsymbol{\theta_0}, \boldsymbol{\eta_1}) \; Vs \; M_2 : f_2(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\eta_2}) = f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\eta_2}) \tag{3.12}$$

donde $\boldsymbol{\eta_1}$ y $\boldsymbol{\eta_2}$ son parámetros de ruido que pueden representar cantidades diferentes en cada modelo. García-Donato (2003) propone las siguientes distribuciónes *a priori* para la selección de modelos:

1) Lla distribución de Jeffreys para M_1: $\pi_1(\boldsymbol{\theta}_0, \boldsymbol{\eta}) = h(\boldsymbol{\eta})$, y

2) Para M_2: $\pi_2(\boldsymbol{\theta}, \boldsymbol{\eta}) = h(\boldsymbol{\eta})\pi_q(\boldsymbol{\theta}|\boldsymbol{\eta})$, donde

$$\pi_q(\boldsymbol{\theta}|\boldsymbol{\eta}) \propto \left[1 + \frac{D(\boldsymbol{\theta}, \boldsymbol{\eta})}{n}\right]^{-q} \tag{3.13}$$

donde $q \in (0, \infty)$ y $D(\boldsymbol{\theta}, \boldsymbol{\eta})$ es la discrepancia de Jeffreys, que se define a continuación:

Definición 3.4 ***Discrepancia de Jeffreys***
Dados cualesquiera dos modelos $f_i(\boldsymbol{y}|\boldsymbol{\theta}_i)$, $i = 1, 2$, se define la discrepancia entre f_1 y f_2 como

$$D = \int log\left[\frac{f_1(\boldsymbol{y}|\boldsymbol{\theta}_1)}{f_2(\boldsymbol{y}|\boldsymbol{\theta}_2)}\right]\{f_1(\boldsymbol{y}|\boldsymbol{\theta}_1) - f_2(\boldsymbol{y}|\boldsymbol{\theta}_2)\}d\boldsymbol{y} \tag{3.14}$$

se puede verificar que:

$$D = KL[f_1(\boldsymbol{y}|\boldsymbol{\theta}_1), f_2(\boldsymbol{y}|\boldsymbol{\theta}_2)] + KL[f_2(\boldsymbol{y}|\boldsymbol{\theta}_2), f_1(\boldsymbol{y}|\boldsymbol{\theta}_1)]$$

Jeffreys demuestra que D es invariante ante transformaciones no singulares de $\boldsymbol{y}$ y ante transformaciones en los parámetros.

Por otro lado, Bayarri y Garcia-Donato (2006) modificaron $\pi_q(\boldsymbol{\theta}|\boldsymbol{\eta})$ para el caso en que se tienen parámetros de ruido en el modelo. Así, es necesario obtener las siguientes distribuciones las siguiente distribuciones *a priori*, $\pi_1(\boldsymbol{\eta})$ para M_1 y $\pi_2(\boldsymbol{\theta}, \boldsymbol{\eta})$ para M_2, Jeffrey asigna la misma distribución *a priori* objetiva para los parámetros comunes $\boldsymbol{\eta}$ en M_1 y una distribución *a priori* condicional para el parámetro nuevo $\boldsymbol{\theta} \mid \boldsymbol{\eta}$ para M_2, así,

$$\pi^D(\boldsymbol{\theta}|\boldsymbol{\eta}) = c(q*, \boldsymbol{\eta})^{-1}(1 + \bar{D}[(\boldsymbol{\theta}, \boldsymbol{\theta}_0)|\boldsymbol{\eta}])^{-q*}\pi^N(\boldsymbol{\theta}|\boldsymbol{\eta}) \tag{3.15}$$

donde:

1) $c(q, \boldsymbol{\eta}) = \int (1 + \bar{D}[(\boldsymbol{\theta}, \boldsymbol{\theta}_0 | \boldsymbol{\eta})])^{-q} \pi^N(\boldsymbol{\theta}|\boldsymbol{\eta}) d\boldsymbol{\theta}$,

2) $\pi^N(\boldsymbol{\theta}|\boldsymbol{\eta})$ es una distribución objetiva (de Jeffrey o de referencia) para M_2

3) $\underline{q} = inf\{q \geq 0 : c(q, \boldsymbol{\eta} < \infty)\}$, $q* = \underline{q} + \frac{1}{2}$.

Sí $\underline{q} < \infty$, entonces para M_1: $\pi_1^D(\boldsymbol{\eta}) = \pi^N(\boldsymbol{\eta})$, y dado que $\boldsymbol{\theta}_0$ es un valor conocido, sólo es necesario calcular $\pi_2^D(\boldsymbol{\theta}, \boldsymbol{\eta})$ bajo M_2:

$$\pi_2^D(\boldsymbol{\theta}, \boldsymbol{\eta}) = \pi^D(\boldsymbol{\theta}|\boldsymbol{\eta}) \pi^N(\boldsymbol{\eta}) \tag{3.16}$$

Con respecto a las medidas de divergencia en presencia de parámetros de ruido, Pérez (2005) utiliza argumentos de geometría diferencial para demostrar que $KL(\boldsymbol{\theta}_0, \boldsymbol{\eta} \mid \boldsymbol{\theta}, \boldsymbol{\eta})$ puede intepretarse como la medida de divergencia entre M_1 y M_2 sólo para el vector de parámetros de interés $\boldsymbol{\theta}$ cuando éste y $\boldsymbol{\eta}$ son ortogonales de acuerdo con el criterio de Cox y Reid (1987).

Resultado 3.1 *Supóngase el modelo parámetrico* $S = \{\pi(\boldsymbol{X} \mid \boldsymbol{\theta}, \boldsymbol{\lambda})\}$. *Sea* $(\boldsymbol{\theta}, \boldsymbol{\eta}(\boldsymbol{\theta}, \boldsymbol{\nu}))$ *una tranformación ortogonal, entonces:*

$$KL(\boldsymbol{\theta}_2, \boldsymbol{\eta}_2 \mid \boldsymbol{\theta}_1, \boldsymbol{\eta}_1) = KL(\boldsymbol{\theta}_1, \boldsymbol{\eta}_2 \mid \boldsymbol{\theta}_1, \boldsymbol{\eta}_1) + KL(\boldsymbol{\theta}_2, \boldsymbol{\eta}_2 \mid \boldsymbol{\theta}_1, \boldsymbol{\eta}_2)$$
$$KL(\boldsymbol{\theta}_1, \boldsymbol{\eta}_1 \mid \boldsymbol{\theta}_2, \boldsymbol{\eta}_2) = KL(\boldsymbol{\theta}_2, \boldsymbol{\eta}_1 \mid \boldsymbol{\theta}_2, \boldsymbol{\eta}_2) + KL(\boldsymbol{\theta}_1, \boldsymbol{\eta}_1 \mid \boldsymbol{\theta}_2, \boldsymbol{\eta}_1)$$

por lo que la discrepancia íntrinseca ortogonal(DIO) entre $\pi(\boldsymbol{X} \mid \boldsymbol{\theta}_1, \boldsymbol{\eta})$ y $\pi(\boldsymbol{X} \mid \boldsymbol{\theta}_2, \boldsymbol{\eta})$ es:

$$\delta(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\eta}) = 2\text{mín}\{KL(\boldsymbol{\theta}_2, \boldsymbol{\eta}_2 \mid \boldsymbol{\theta}1, \boldsymbol{\eta}_2), KL(\boldsymbol{\theta}_1, \boldsymbol{\eta}_1 \mid \boldsymbol{\theta}_2, \boldsymbol{\eta}_1)\} \tag{3.17}$$

Así, la distribución *a priori* convencional,bajo ortogonalidad entre $\boldsymbol{\theta}$ y $\boldsymbol{\eta}$, está dada por:

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{\eta}) \propto \left[1 + \frac{\delta(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\eta})}{n}\right]^{-q} \tag{3.18}$$

donde $\pi(\boldsymbol{\eta})$ puede ser la distribución *a priori* de Jeffreys y $q > 0$.

3.5. Discusión

En el capítulo anterior se mostraron algunas métodos para obtener distribuciones *a priori* no informativas, las cuales, son de utilidad en problemas en los que no es factible o no se quiere utilizar una distribución inicial subjetiva.

Uno de los problemas en los que estas distribuciones son usadas, es en selección de modelos (Zellner, 1986, Zellner y Siow, 1980) y ambos procedimientos están ligados con la propuesta de (Jeffreys, 1961); bajo el modelo lineal normal. Por otro lado, se describe la propuesta de (García-Donato, 2003) generalizando la propuesta de Jeffreys y a la vez aplicando la definición de DKL obtiene lo que él llama *Discrepancia de Jeffreys* denotada por D, (García-Donato, 2003, Cap. 6) nos dice que la medida D del parámetro de interés θ con el parámetro de ruido η, esto es, $D(\theta, \eta)$ es un caso particular de la llamada $\phi-$divergencia (Amari, 1985, 1990) que se define como:

$$D_\phi(\theta, \theta_0) = \int \phi\left(\frac{f_1(y \mid \theta_0)}{f_2(y \mid \theta_0, \theta)}\right) f_2(y \mid \theta_0, \theta) dy$$

donde ϕ es una función real convexa en $[0, \infty]$. Las ϕ-divergencias tambien se han utiizado en contextos bayesianos de selección de modelos, usualmente analizando su distribución *a posteriori* para más detalles ver (Bayarri, 1987, Bernardo, 1985, Esteban *et al.*, 2000), entre otros. Otra aplicación de esta medida es que se propone como medio para la contruir la distribución *a priori* convencional π para el parámetro de interés; ya que tiene la ventaja que se puede aplicar independientemente de la dimensión del parámetro $\boldsymbol{\theta}$, por lo que es una propuesta automatica de previas convencionales, es propia y esta basada en la de Jeffreys (García-Donato, 2003, sección 6.5).

Capítulo 4

REGRESIÓN BAYESIANA CONVENCIONAL (RBC)

En este trabajo de investigación se retoma la definición de la divergencia de Kullback-Leibler para obtener la distribución *a priori* para la predicción en selección genómica ya que, considerando que π_q se puede aplicar independientemente de la dimension de $\boldsymbol{\beta}$ y que además es invariante ante transformaciones.

4.1. Modelo Lineal General

4.1.1. Modelo de rango completo

Considere el modelo:

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

donde la matriz diseño $\boldsymbol{X}$ de $n \times p$ es de rango completo por lo que existe $(\boldsymbol{X}^t\boldsymbol{X})^{-1}$, además $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \sigma^2)$ y $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$

Por lo que la verosimilitud se define por:

$$L(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}) = (\sigma^2)^{-\frac{n}{2}} exp\left\{\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right\}$$

la cual se puede expresar de la siguiente forma:

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}) = (\sigma^2)^{-\frac{n}{2}} exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{y}^t\boldsymbol{M}\boldsymbol{y}\right\} \\ \text{x} \quad exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^t\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right\} \end{aligned} \tag{4.1}$$

donde, $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$ es el estimador de Máxima Verosimilitud y $M = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t$ es la matriz de proyección en el espacio columna de $\boldsymbol{X}$.

Así, aplicando la definición (3.8) de la divergencia de Kullback-Leibler en (4.1)

$$\begin{aligned} KL(\boldsymbol{\beta}_0, \sigma_0^2 \mid \boldsymbol{\beta}, \sigma^2) &= \int log\left[\frac{L(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y})}{L(\boldsymbol{\beta}_0, \sigma_0^2 \mid \boldsymbol{y})}\right] p(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2) d\boldsymbol{y} \\ &= \frac{n}{2} log\left(\frac{\sigma_0^2}{\sigma^2}\right) + \frac{(n-p)}{2}\left(\frac{\sigma^2}{\sigma_0^2} - 1\right) + \frac{n}{2}\left(\frac{\sigma^2}{\sigma_0^2} - 1\right) \\ &+ \frac{(\boldsymbol{\beta}_0 - \boldsymbol{\beta})^t\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{\beta}_0 - \boldsymbol{\beta})}{2\sigma_0^2} \end{aligned} \tag{4.2}$$

Similarmente:

$$\begin{aligned} KL(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{\beta}_0, \sigma_0^2) &= \int log\left[\frac{L(\boldsymbol{\beta}_0, \sigma_0^2 | \boldsymbol{y})}{L(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y})}\right] p(\boldsymbol{y} | \boldsymbol{\beta}_0, \sigma_0^2) d\boldsymbol{y} \\ &= \frac{n}{2} log\left(\frac{\sigma^2}{\sigma_0^2}\right) + \frac{(n-p)}{2}\left(\frac{\sigma_0^2}{\sigma^2} - 1\right) + \frac{n}{2}\left(\frac{\sigma_0^2}{\sigma^2} - 1\right) \\ &+ \frac{(\boldsymbol{\beta}_0 - \boldsymbol{\beta})^t\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{\beta}_0 - \boldsymbol{\beta})}{2\sigma^2} \end{aligned} \tag{4.3}$$

Note que si, $\boldsymbol{Z} \sim \mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ y $\boldsymbol{A}$ es una matriz simétrica no negativa definida, entonces $E(\boldsymbol{Z}^t\boldsymbol{A}\boldsymbol{Z}) = tr(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^t\boldsymbol{A}\boldsymbol{\mu}$. Además $\hat{\boldsymbol{\beta}} \sim \mathrm{N}\left(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1}\right)$. En este caso se puede verificar directamente que $\boldsymbol{\beta}$ es ortogonal a σ^2, este es un caso particular del resultado discutido por (Barndorff-Nielsen, 1978, 1983) , y establece que en modelos lineales generalizados el parámetro esperanza y el de dispersión son ortogonales.

Por tanto, la divergencia de Kullback-Leibler, Para $\boldsymbol{\beta}_0 = \mathbf{0}$ y $\sigma_0^2 = \sigma^2$ es:

$$KL(\boldsymbol{\beta}_0, \sigma_0^2 \mid \boldsymbol{\beta}, \sigma^2) = \int log\left[\frac{L(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y})}{L(\mathbf{0}, \sigma_0^2|\boldsymbol{y})}\right] p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)d\boldsymbol{y} = \frac{\boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta}}{2\sigma^2}$$

de manera similar

$$KL(\boldsymbol{\beta}, \sigma^2|\boldsymbol{\beta}_0, \sigma_0^2) = \int log\left[\frac{L(\mathbf{0}, \sigma_0^2|\boldsymbol{y})}{L(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y})}\right] p(\boldsymbol{y}|\mathbf{0}, \sigma^2)d\boldsymbol{y} = \frac{\boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta}}{2\sigma^2}$$

entonces en la ecuación (3.17) se tiene :

$$D = 2min\{KL(\boldsymbol{\beta}_0, \sigma_0^2 \mid \boldsymbol{\beta}, \sigma^2), KL(\boldsymbol{\beta}, \sigma^2|\boldsymbol{\beta}_0, \sigma_0^2)\} = \frac{\boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta}}{\sigma^2}$$

por lo que la distribución inicial no informativa de $\boldsymbol{\beta}$ para el modelo de rango completo es:

$$\pi_q(\boldsymbol{\beta}|\sigma^2) \propto \left[1 + \frac{D(\boldsymbol{\beta}, \sigma^2)}{n}\right]^{-q} \propto \left[1 + \frac{\boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta}}{n\sigma^2}\right]^{-q}$$

para, $q = \frac{dim(\boldsymbol{\beta})+1}{2}$, (García-Donato, 2003)

$$\pi_q(\boldsymbol{\beta}|\sigma^2) = Ca_{dim(\boldsymbol{\beta})}\left[\boldsymbol{\beta}|\mathbf{0}, n\sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1}\right]$$

Antes de continuar con el modelo lineal de rango y aplicar la definición de la divergencia de Kullback-Leibler para construir la distribución *a priori* no informativa, hagamos un paréntesis para definir el criterio de ortogonalidad de Cox y Reid.

Dos parámetros $\boldsymbol{\theta}$ y $\boldsymbol{\eta}$ son ortoganales con respecto a la información de Fisher Cox y Reid (1987) cuando

$$\frac{1}{n}\left[-\frac{\partial^2 l(\boldsymbol{\theta},\boldsymbol{\eta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\eta}} \mid \boldsymbol{\theta},\boldsymbol{\eta}\right] = \mathbf{0} \tag{4.4}$$

donde $l(\boldsymbol{\theta},\boldsymbol{\eta})$ es la log-verosimilitud de los parámetros $(\boldsymbol{\theta},\boldsymbol{\eta})$ y n es el tamaño de muestra en el que se basa la verosimilitud. Si (4.4) se cumple para todo $(\boldsymbol{\theta},\boldsymbol{\eta})$ en el espacio paramétrico, entonces se le llama *ortogonalidad global*; si esto sólo ocurre para un valor $(\boldsymbol{\theta}^0,\boldsymbol{\eta}^0)$, entonces los vectores $\boldsymbol{\theta}$ y $\boldsymbol{\eta}$ son *localmente ortogonales* en $(\boldsymbol{\theta}^0,\boldsymbol{\eta}^0)$. La interpretación estadística más directa es que los componentes relevantes del estadístico *score*, el gradiente de la log-verosimilitud, no están correlacionados, esto es la matriz de información en diagonal por bloques.

4.1.2. Modelo de rango incompleto

En los modelos de rango incompleto el rango de la matriz diseño $\boldsymbol{X}$ es menor que el numero de columnas $(r \leq p)$ por lo que $(\boldsymbol{X}^t\boldsymbol{X})^{-1}$ no existe, no obstante, esto no es un inconveniente ya que los modelos sobreparametrizados son ampliamente utilizados.

Supóngase que $\boldsymbol{y}$ y $\boldsymbol{X}$ estáan relacionados por el modelo

$$\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4.5}$$

En lo que sigue se utilizarán algunos resultados del álgebra matricial de los modelos lineales. Como se mencionó antes, $(\boldsymbol{X}^t\boldsymbol{X})^{-1}$ no existe, de modo que $\hat{\boldsymbol{\beta}}$ tampoco existe y la función de verosimilitud no puede ser expresada como en (4.1). Sin embargo, se puede utilizar una inversa generalizada $\boldsymbol{G}$ de $\boldsymbol{X}^t\boldsymbol{X}$ tal que

$$\begin{aligned} \boldsymbol{X}^t\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^t\boldsymbol{X} = \boldsymbol{X}^t\boldsymbol{X}\boldsymbol{G}^t\boldsymbol{X}^t\boldsymbol{X} = \boldsymbol{X}^t\boldsymbol{X}, \\ \boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^t\boldsymbol{X} = \boldsymbol{X}, \quad \text{y} \\ \boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^t \qquad \text{es invariante a la selección} \quad \boldsymbol{G} \end{aligned} \tag{4.6}$$

entonces de (4.6)

$$\dot{\boldsymbol{\beta}} = \boldsymbol{G}\boldsymbol{X}^t\boldsymbol{y} \tag{4.7}$$

es una solución del sistema de ecuaciones lineales

$$\boldsymbol{X}^t\boldsymbol{X}\dot{\boldsymbol{\beta}} = \boldsymbol{X}^t\boldsymbol{y} \tag{4.8}$$

Note que $\dot{\boldsymbol{\beta}}$ no es una solución única debido a que depende de la selección de $\boldsymbol{G}$, si embargo existen combinaciones lineales $\boldsymbol{\lambda}^t\boldsymbol{\beta}$ tales que $\boldsymbol{\lambda}^t\dot{\boldsymbol{\beta}}$ es invariante a la selección de $\boldsymbol{G}$. Bajo el enfoque clásico de modelos lineales éstas combinaciones lineales reciben el nombre de funciones estimables. Se dice que una función $\boldsymbol{\lambda}^t\boldsymbol{\beta}$ es estimable si y sólo si

$$E(\boldsymbol{\lambda}^t\dot{\boldsymbol{\beta}}) = \boldsymbol{T}^tE(\boldsymbol{y}) = \boldsymbol{T}^t\boldsymbol{X}\boldsymbol{\beta} \tag{4.9}$$

donde $\boldsymbol{T}^t$ es una matriz $r \times n$ con rango $r \leq q$, por lo que $r(\boldsymbol{\lambda}) = r$. Entonces de lo anterior, se verifica que $\boldsymbol{\lambda}$ satisface

$$\boldsymbol{\lambda}^t = \boldsymbol{T}^t\boldsymbol{X}, \quad \text{y} \tag{4.10}$$

$$\boldsymbol{\lambda}^t\boldsymbol{G}\boldsymbol{X}^t\boldsymbol{X} = \boldsymbol{\lambda}^t \tag{4.11}$$

la condición (4.10) es consecuencia directa de (4.9) y (4.11) que se

debe a la segunda igualdad en (4.6). Entonces, si $\boldsymbol{\lambda}^t\boldsymbol{\beta}$ es una función estimable se tiene que

$$\begin{aligned} E(\boldsymbol{\lambda}^t\dot{\boldsymbol{\beta}}) &= \boldsymbol{\lambda}^t\boldsymbol{G}\boldsymbol{X}^t E(\boldsymbol{y}) = \boldsymbol{T}^t\boldsymbol{X}^t\boldsymbol{G}\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\lambda}\boldsymbol{\beta} \\ var(\boldsymbol{\lambda}^t\dot{\boldsymbol{\beta}}) &= \boldsymbol{\lambda}^t\boldsymbol{G}\boldsymbol{X}^t var(\boldsymbol{y})\boldsymbol{X}\boldsymbol{G}^t\boldsymbol{\lambda} = \sigma^2\boldsymbol{\lambda}^t\boldsymbol{G}\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{G}^t\boldsymbol{X}^t\boldsymbol{T} = \sigma^2\boldsymbol{\lambda}^t\boldsymbol{G}\boldsymbol{\lambda}; \end{aligned}$$

además, $\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$

$$\boldsymbol{\lambda}^t\dot{\boldsymbol{\beta}} \sim \mathrm{N}(\boldsymbol{\lambda}^t\boldsymbol{\beta}, \sigma^2\boldsymbol{\lambda}^t\boldsymbol{G}\boldsymbol{\lambda}) \tag{4.12}$$

Por otro lado, la verosimilitud correspondiente al modelo (4.5) es

$$L(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}) = (\sigma^2)^{-\frac{n}{2}} exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right\}$$

donde

$$(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{y}-\boldsymbol{X}\dot{\boldsymbol{\beta}})^t(\boldsymbol{y}-\boldsymbol{X}\dot{\boldsymbol{\beta}}) + (\boldsymbol{X}\dot{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{X}\dot{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\beta})$$

Note que el primer sumando del lado derecho de la expresión anterior puede escribirse como

$$\begin{aligned} (\boldsymbol{y}-\boldsymbol{X}\dot{\boldsymbol{\beta}})^t(\boldsymbol{y}-\boldsymbol{X}\dot{\boldsymbol{\beta}}) &= (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^t\boldsymbol{X}\dot{\boldsymbol{\beta}})^t(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^t\boldsymbol{X}\dot{\boldsymbol{\beta}}) \\ &= \boldsymbol{y}^t\boldsymbol{y} - \boldsymbol{y}^t\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^t\boldsymbol{X}\dot{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{G}^t\boldsymbol{X}^t\boldsymbol{y} \\ &\quad + \dot{\boldsymbol{\beta}}\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{G}^t\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^t\boldsymbol{X}\dot{\boldsymbol{\beta}} \\ &= \boldsymbol{y}^t\boldsymbol{y} - \boldsymbol{y}^t\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^t\boldsymbol{y} = \boldsymbol{y}^t(\boldsymbol{I}-\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^t)\boldsymbol{y} = \boldsymbol{y}^t\boldsymbol{M}\boldsymbol{y}; \end{aligned} \tag{4.13}$$

donde $\boldsymbol{M} = (\boldsymbol{I}-\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^t)$ y sea $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}(\boldsymbol{\lambda}^t)^{-1}\boldsymbol{\lambda}^t\boldsymbol{\beta}$ por lo que

$$
\begin{aligned}
(\boldsymbol{X}\dot{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{X}\dot{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}) &= (\boldsymbol{X}(\boldsymbol{\lambda}^t)^{-1}\boldsymbol{\lambda}^t\dot{\boldsymbol{\beta}} - \boldsymbol{X}(\boldsymbol{\lambda}^t)^{-1}\boldsymbol{\lambda}^t\boldsymbol{\beta})^t \\
&\quad \text{x} \quad (\boldsymbol{X}(\boldsymbol{\lambda}^t)^{-1}\boldsymbol{\lambda}^t\dot{\boldsymbol{\beta}} - \boldsymbol{X}(\boldsymbol{\lambda}^t)^{-1}\boldsymbol{\lambda}^t\boldsymbol{\beta}) \\
&= \dot{\boldsymbol{\beta}}^t\boldsymbol{\lambda}(\boldsymbol{\lambda})^{-1}\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{\lambda}^t)^{-1}\boldsymbol{\lambda}^t\dot{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}}^t\boldsymbol{\lambda}(\boldsymbol{\lambda})^{-1}\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{\lambda}^t)^{-1}\boldsymbol{\lambda}^t\boldsymbol{\beta} \\
&\quad -\boldsymbol{\beta}^t\boldsymbol{\lambda}(\boldsymbol{\lambda})^{-1}\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{\lambda}^t)^{-1}\boldsymbol{\lambda}^t\dot{\boldsymbol{\beta}} + \boldsymbol{\beta}^t\boldsymbol{\lambda}(\boldsymbol{\lambda})^{-1}\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{\lambda}^t)^{-1}\boldsymbol{\lambda}^t\boldsymbol{\beta} \\
&= (\boldsymbol{\lambda}^t\boldsymbol{\beta} - \boldsymbol{\lambda}^t\dot{\boldsymbol{\beta}})\boldsymbol{\lambda}^{-1}\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{\lambda}^t)^{-1}(\boldsymbol{\lambda}^t\boldsymbol{\beta} - \boldsymbol{\lambda}^t\dot{\boldsymbol{\beta}})
\end{aligned} \tag{4.14}
$$

Por lo tanto, de (4.13) y (4.14) se tiene que la función de verosimilitud de $(\boldsymbol{\lambda}^t\boldsymbol{\beta}, \sigma^2)$ está dada por

$$
\begin{aligned}
L(\boldsymbol{\lambda}^t\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}) &= (\sigma^2)^{-\frac{n}{2}} exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{y}^t\boldsymbol{M}\boldsymbol{y}\right\} \\
&\quad \text{x} \quad exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\lambda}^t\dot{\boldsymbol{\beta}} - \boldsymbol{\lambda}^t\boldsymbol{\beta})\boldsymbol{\lambda}^{-1}\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{\lambda}^t)^{-1}(\boldsymbol{\lambda}^t\dot{\boldsymbol{\beta}} - \boldsymbol{\lambda}^t\boldsymbol{\beta})\right\}
\end{aligned} \tag{4.15}
$$

Una vez más aplicando (3.8) y desarrollando el logaritmo:

$$
\begin{aligned}
log\left[\frac{L(\boldsymbol{\lambda}^t\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y})}{L(\boldsymbol{\lambda}^t\boldsymbol{\beta}_0, \sigma^2 \mid \boldsymbol{y})}\right] &= -\frac{1}{2\sigma^2}(\boldsymbol{\lambda}^t\dot{\boldsymbol{\beta}} - \boldsymbol{\lambda}^t\boldsymbol{\beta})^t\boldsymbol{\lambda}^{-1}\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{\lambda}^t)^{-1}(\boldsymbol{\lambda}^t\dot{\boldsymbol{\beta}} - \boldsymbol{\lambda}^t\boldsymbol{\beta}) \\
&\quad + \frac{1}{2\sigma^2}(\boldsymbol{\lambda}^t\boldsymbol{\beta}_0 - \boldsymbol{\lambda}^t\dot{\boldsymbol{\beta}})^t\boldsymbol{\lambda}^{-1}\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{\lambda}^t)^{-1}(\boldsymbol{\lambda}^t\boldsymbol{\beta}_0 - \boldsymbol{\lambda}^t\dot{\boldsymbol{\beta}})
\end{aligned}
$$

Note que la DKL es un promedio, así :

$$
\begin{aligned}
& E\left[\frac{L(\boldsymbol{\lambda}^t\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y})}{L(\boldsymbol{\lambda}^t\boldsymbol{\beta}_0, \sigma^2 \mid \boldsymbol{y})} \mid \boldsymbol{\beta}, \sigma^2\right] = -\frac{1}{2\sigma^2} \\
\text{x} \quad & E[(\boldsymbol{\lambda}^t\dot{\boldsymbol{\beta}} - \boldsymbol{\lambda}^t\boldsymbol{\beta})^t\boldsymbol{\lambda}^{-1}\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{\lambda}^t)^{-1}(\boldsymbol{\lambda}^t\dot{\boldsymbol{\beta}} - \boldsymbol{\lambda}^t\boldsymbol{\beta})] \\
& + \frac{1}{2\sigma^2}E[\dot{\boldsymbol{\beta}}^t\boldsymbol{\lambda}(\boldsymbol{\lambda}^{-1}\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{\lambda}^t)^{-1})\boldsymbol{\lambda}^t\dot{\boldsymbol{\beta}}]\} \\
& = -\frac{1}{2\sigma^2}\left\{tr(\boldsymbol{\lambda}^{-1}\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{\lambda}^t)^{-1}\boldsymbol{\lambda}^t\boldsymbol{G}\boldsymbol{\lambda})\right\} \qquad (4.16) \\
& - \frac{1}{2\sigma^2}\left\{tr(\boldsymbol{\lambda}^{-1}\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{\lambda}^t)^{-1}\boldsymbol{\lambda}^t\boldsymbol{G}\boldsymbol{\lambda})\right\} \\
& + \frac{1}{2\sigma^2}\left\{\boldsymbol{0} + (\boldsymbol{\lambda}^t\boldsymbol{\beta}_0 - \boldsymbol{\lambda}^t\boldsymbol{\beta})^t\boldsymbol{\lambda}^{-1}\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{\lambda}^t)^{-1}(\boldsymbol{\lambda}^t\boldsymbol{\beta}_0 - \boldsymbol{\lambda}^t\boldsymbol{\beta})\right\}
\end{aligned}
$$

Por lo tanto la divergencia de Kullback-Leibler para $\boldsymbol{\beta}_0 = \boldsymbol{0}$ y $\sigma_0^2 = \sigma^2$ es:

$$
KL(\boldsymbol{\lambda}^t\boldsymbol{\beta}_0, \sigma_0^2 \mid \boldsymbol{\beta}, \sigma^2) = \int log\left[\frac{L(\boldsymbol{\lambda}^t\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y})}{L(\boldsymbol{0}, \sigma^2 \mid \boldsymbol{y})}\right] p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)d\boldsymbol{y}
$$

$$
= \frac{\boldsymbol{\beta}^t\boldsymbol{\lambda}\boldsymbol{\lambda}^{-1}\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{\lambda}^t)^{-1}\boldsymbol{\lambda}^t\boldsymbol{\beta}}{2\sigma^2}
$$

o bien

$$
KL(\boldsymbol{\lambda}^t\boldsymbol{\beta}, \sigma^2|\boldsymbol{\beta}_0, \sigma_0^2) = \int log\left[\frac{L(\boldsymbol{0}, \sigma_0^2 \mid \boldsymbol{y})}{L(\boldsymbol{\lambda}^t\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y})}\right] p(\boldsymbol{y}|\boldsymbol{0}, \sigma^2)d\boldsymbol{y}
$$

$$
= \frac{\boldsymbol{\beta}^t\boldsymbol{\lambda}\boldsymbol{\lambda}^{-1}\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{\lambda}^t)^{-1}\boldsymbol{\lambda}^t\boldsymbol{\beta}}{2\sigma^2}
$$

Ahora, note que, si $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)$, que cumple el criterio de ortogonalidad de Cox y Reid (1987) , esto es, $\boldsymbol{\lambda}_1 \perp \boldsymbol{\lambda}_2$, y $\boldsymbol{X}^t\boldsymbol{X} = \boldsymbol{\lambda}_1\boldsymbol{D}\boldsymbol{\lambda}_1^t$ es la descomposición del valor singular, donde $\boldsymbol{D}$ es una matriz diagonal, se

tiene :

$$\begin{aligned}\boldsymbol{\beta}^t\boldsymbol{\lambda}\boldsymbol{\lambda}^{-1}\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{\lambda}^t)^{-1}\boldsymbol{\lambda}^t\boldsymbol{\beta} &= \boldsymbol{\beta}^t\boldsymbol{\lambda}\boldsymbol{\lambda}^t\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\lambda}\boldsymbol{\lambda}^t\boldsymbol{\beta} \\ &= \boldsymbol{\beta}^t\begin{bmatrix}\boldsymbol{\lambda}_1 & \boldsymbol{\lambda}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\lambda}_1^t \\ \boldsymbol{\lambda}_2^t\end{bmatrix}\boldsymbol{X}^t\boldsymbol{X}\begin{bmatrix}\boldsymbol{\lambda}_1 & \boldsymbol{\lambda}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\lambda}_1^t \\ \boldsymbol{\lambda}_2^t\end{bmatrix}\boldsymbol{\beta} \\ &\text{desarrollando y simplificando} \\ &= \boldsymbol{\beta}^t\boldsymbol{\lambda}_1(\boldsymbol{\lambda}_1^t\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\lambda}_1)\boldsymbol{\lambda}_1^t\boldsymbol{\beta} = \boldsymbol{\beta}^t\boldsymbol{\lambda}(\boldsymbol{\lambda}^t\boldsymbol{\lambda}_1\boldsymbol{D}\boldsymbol{\lambda}_1^t\boldsymbol{\lambda})\boldsymbol{\lambda}^t\boldsymbol{\beta}\end{aligned}$$

Por otro lado

$$\begin{aligned}\boldsymbol{\lambda}_1^t\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_1^t\boldsymbol{\lambda}_1\boldsymbol{D}\boldsymbol{\lambda}_1^t\boldsymbol{\lambda}_1 &= [\boldsymbol{\lambda}_1^{-1}(\boldsymbol{\lambda}_1^t)^{-1}\boldsymbol{D}^{-1}\boldsymbol{\lambda}_1^{-1}(\boldsymbol{\lambda}_1^t)^{-1}]^{-1} \\ &= [\boldsymbol{\lambda}_1^{-1}(\boldsymbol{\lambda}_1\boldsymbol{D}^{-1}\boldsymbol{\lambda}_1^t)\boldsymbol{\lambda}_1]^{-1} \\ &= [\boldsymbol{\lambda}_1^t(\boldsymbol{\lambda}_1\boldsymbol{D}^{-1}\boldsymbol{\lambda}_1^t)\boldsymbol{\lambda}_1]^{-1} = (\boldsymbol{\lambda}_1^t\boldsymbol{G}\boldsymbol{\lambda}_1)^{-1}\end{aligned}$$

por tanto, de acuerdo con (García-Donato, 2003), la distribución convencional para $\boldsymbol{\lambda}_1^t\boldsymbol{\beta}$ es:

$$\pi(\boldsymbol{\lambda}_1^t\boldsymbol{\beta}|\sigma^2) \propto \left[1 + \frac{\boldsymbol{\beta}^t\boldsymbol{\lambda}_1(\boldsymbol{\lambda}_1^t\boldsymbol{G}\boldsymbol{\lambda}_1)^{-1}\boldsymbol{\lambda}_1{}^t\boldsymbol{\beta}}{n\sigma^2}\right]^{-q} \tag{4.17}$$

Con $q = \frac{r+1}{2}$

$$\pi(\boldsymbol{\lambda}_1^t\boldsymbol{\beta}|\sigma^2) = Ca_r(\boldsymbol{\lambda}_1^t\boldsymbol{\beta}|\mathbf{0}, \boldsymbol{\lambda}_1^t\boldsymbol{G}\boldsymbol{\lambda}_1) \tag{4.18}$$

Sea $\boldsymbol{\lambda}_2^t$ tal que

$$\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)$$

es de rango completo, entonces (García-Donato, 2003)

$$\pi(\boldsymbol{\beta}|\sigma^2) \propto \left[1 + \frac{\boldsymbol{\beta}^t\boldsymbol{\lambda}_1(\boldsymbol{\lambda}_1^t\boldsymbol{G}\boldsymbol{\lambda}_1)^{-1}\boldsymbol{\lambda}_1^t\boldsymbol{\beta}}{n\sigma^2}\right]^{-q} f_{p-r}(\boldsymbol{\lambda}_2^t\boldsymbol{\beta})|\boldsymbol{\lambda}| \tag{4.19}$$

esto es,

$$\pi(\boldsymbol{\beta}|\sigma^2) = PIC_r(\boldsymbol{\beta}|(\boldsymbol{\lambda}_1^t\boldsymbol{G}\boldsymbol{\lambda}_1), \boldsymbol{\lambda})$$

donde

1. $f_{p-r}(\boldsymbol{\lambda}_2^t\boldsymbol{\beta})$ es una densidad $p-r$ variante.
2. PIC_r es una densidad Cauchy parcial informativa

Acontinuación se definirán éstos conceptos más a fondo y también pueden ver Bayarri y Garcia-Donato (2007)

Definición 4.1 *Distribución Cauchy parcial informativa*

Sea $\boldsymbol{y} \in R^n$. Diremos que $\boldsymbol{y}$ se distribuye PIC de parámetros $(\boldsymbol{A}, \boldsymbol{C})$ denotado $PIC_n(\boldsymbol{A}, \boldsymbol{C})$ con $\boldsymbol{A}$: nxn simétrica semidefinida positiva y $\boldsymbol{C}$: nxn no singular, si la desidad conjunta (posiblemente impropia) de $\boldsymbol{y}$ es

$$p(\boldsymbol{y}|\boldsymbol{A}, \boldsymbol{C}) = \int GPIN_n\left(\boldsymbol{y}|\frac{\boldsymbol{A}}{t}, \boldsymbol{C}\right) IGa\left(t|\frac{1}{2}, \frac{1}{2}\right) dt \tag{4.20}$$

donde la Normal generalizada parcialmente informativa

$$GPIN(\boldsymbol{y}|\boldsymbol{A}, \boldsymbol{C}) = det_+\left[\frac{\boldsymbol{A}}{2\pi}\right]^{1/2} |det(\boldsymbol{C})| \exp\left\{\frac{1}{2} y^t(\boldsymbol{C}^t\boldsymbol{A}\boldsymbol{C})\right\} \tag{4.21}$$

y $det_+\left[\frac{\boldsymbol{A}}{2\pi}\right]^{1/2}$ es el producto de los valores propios de $\boldsymbol{A}$

Regresando a la ecuación (4.19) y si hacemos $\boldsymbol{G} = \boldsymbol{U}_1\boldsymbol{D}^{-1}\boldsymbol{U}_1^t$, donde $\boldsymbol{G}$ es la inversa generalizada de Moore-Penrose de $\boldsymbol{X^tX}$ y $\boldsymbol{\lambda}_1^t = \boldsymbol{U}_1^t$, se tiene:

$$\pi(\boldsymbol{\beta}|\sigma^2) \propto \left[1 + \frac{\boldsymbol{\beta}^t\boldsymbol{G}\boldsymbol{\beta}}{n\sigma^2}\right]^{-q} f(\boldsymbol{U}_2^t\boldsymbol{\beta})|\boldsymbol{U}| \tag{4.22}$$

así,

$$\pi(\boldsymbol{\beta}|\sigma^2) = PIC_r\left[\boldsymbol{\beta}|\boldsymbol{G}, \boldsymbol{I}\right] \tag{4.23}$$

Ya se tiene la distribución *a priori* convencional para $\boldsymbol{\beta}$ para el modelo de rango incompleto. No obstante, si la reparametrización se hace directamente en el modelo (4.5), es más directo obtener la divergencia de Kullback-Leibler para poder construir la distribución *a priori* para una función de $\boldsymbol{\beta}$.

Por tanto, sea la reparametrización, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^t, \boldsymbol{\theta}_2^t)^t = \left((\boldsymbol{U}_1^t\boldsymbol{\beta})^t, (\boldsymbol{U}_2^t\boldsymbol{\beta})^t\right)^t$ y sea $\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{U}_1^t$, donde $\boldsymbol{U} = [\boldsymbol{U}_1 \mid \boldsymbol{U}_2]$ y $\boldsymbol{V}$ son matrices ortogonales de la descomposición del valor singular (SVD) de $\boldsymbol{X}$ y $\boldsymbol{D}^{\frac{1}{2}}$ es una matriz diagonal de orden $r \times r$ de los valores singulares. Considerando que $\boldsymbol{U} = [\boldsymbol{U}_{1_{n\times p}} \mid \boldsymbol{U}_{2_{p\times(p-n)}}]$, donde $\boldsymbol{U}_2$ es el complemento ortogonal de $\boldsymbol{U}_1$ y note que $\boldsymbol{\theta} = \boldsymbol{U}^t\boldsymbol{\beta}$, entonces:

$$\boldsymbol{\theta} = \boldsymbol{U}^t\boldsymbol{\beta} \Rightarrow (\boldsymbol{U}^t)^{-1}\boldsymbol{\beta} = (\boldsymbol{U}^t)^{-1}\boldsymbol{U}^{-1}\boldsymbol{\theta} \Rightarrow (\boldsymbol{U}^t)^{-1}\boldsymbol{\theta} = \boldsymbol{\beta} \Rightarrow \boldsymbol{U}\boldsymbol{\theta} = \boldsymbol{\beta}$$

Sustituyendo en (4.5) se tiene:

$$\begin{aligned} \boldsymbol{y} &= \mu\mathbf{1} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ &= \mu\mathbf{1} + \boldsymbol{V}\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{U}_1^t(\boldsymbol{U}_1\boldsymbol{\theta}_1 + \boldsymbol{U}_2\boldsymbol{\theta}_2) + \boldsymbol{\epsilon} \\ &= \mu\mathbf{1} + \boldsymbol{V}\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{\theta}_1 + \boldsymbol{\epsilon} \end{aligned} \tag{4.24}$$

Note que el modelo (4.24) es regresión en componentes principales; así, sustituyendo la media de $\boldsymbol{y}$ en la verosimilitud se tiene:

$$\begin{aligned} L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\sigma^2 \mid \boldsymbol{y}) &= (\sigma^2)^{-\frac{n}{2}} \\ &\text{x} \quad exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}^* - \boldsymbol{V}\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{\theta}_1)^t(\boldsymbol{y}^* - \boldsymbol{V}\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{\theta}_1)\right\} \end{aligned} \tag{4.25}$$

si, $\boldsymbol{y}^* = \boldsymbol{y} - \mu\mathbf{1}$ y desarrollando la forma cuadrática en la exponencial

de (4.25) se observa:

$$\begin{aligned}(\boldsymbol{y}^* - \boldsymbol{V}\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{\theta}_1)^t(\boldsymbol{y}^* - \boldsymbol{V}\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{\theta}_1) &= \boldsymbol{y}^{*^t}\boldsymbol{y}^* - 2\boldsymbol{\theta}_1^t\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{V}^t\boldsymbol{y}^* + \boldsymbol{\theta}_1^t\boldsymbol{D}\boldsymbol{\theta}_1 \\ &= (\boldsymbol{\theta}_1 - \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{V}^t\boldsymbol{y}^*)^t\boldsymbol{D}(\boldsymbol{\theta}_1 - \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{V}^t\boldsymbol{y}^*) \\ &+ \boldsymbol{y}*^t\boldsymbol{y}^* - \boldsymbol{y}^{*^t}\boldsymbol{V}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{D}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{V}^t\boldsymbol{y}^*\end{aligned} \tag{4.26}$$

por lo que la función de verosimilitud es:

$$\begin{aligned}L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\sigma^2 \mid \boldsymbol{y}) &\propto (\sigma^2)^{-\frac{n}{2}} \\ \text{x} \quad exp&\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\theta}_1 - \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{V}^t\boldsymbol{y}^*)^t\boldsymbol{D}(\boldsymbol{\theta}_1 - \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{V}^t\boldsymbol{y}^*)\right\}\end{aligned} \tag{4.27}$$

Entonces, $\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{V}^t\boldsymbol{y}^* \sim \mathrm{N}(\boldsymbol{\theta}_1, \sigma^2\boldsymbol{D}^{-1})$. La media y varianza de $\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{V}^t\boldsymbol{y}^*$ son:

$$\begin{aligned}E[\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{V}^t\boldsymbol{y}^*] &= \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{V}^t E[\boldsymbol{y}^*] = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{V}^t\boldsymbol{V}\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{\theta}_1 = \boldsymbol{\theta}_1 \\ var[\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{V}^t\boldsymbol{y}^*] &= \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{V}^t var(\boldsymbol{y}^*)\boldsymbol{V}\boldsymbol{D}^{-\frac{1}{2}} = \sigma^2\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{V}^t\boldsymbol{V}\boldsymbol{D}^{-\frac{1}{2}} = \sigma^2\boldsymbol{D}^{-1}\end{aligned}$$

De (4.26) y (4.27) se tiene que la DKL esta dada por:

$$\begin{aligned}&KL(\boldsymbol{\theta}_{1_0}, \boldsymbol{\theta}_{2_0}, \sigma_0^2 \mid \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma^2) = E\left[\left(\frac{L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma^2 \mid \boldsymbol{y})}{L(\boldsymbol{\theta}_{1_0}, \boldsymbol{\theta}_{2_0}, \sigma_0^2 \mid \boldsymbol{y})}\right) \mid \boldsymbol{\theta}, \sigma^2\right] \\ &= -\frac{1}{2\sigma^2}E\left[(\boldsymbol{\theta}_1 - \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{V}^t\boldsymbol{y}^*)^t\boldsymbol{D}(\boldsymbol{\theta}_1 - \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{V}^t\boldsymbol{y}^*)\right] \\ &+ \frac{1}{2\sigma^2}E\left[(\boldsymbol{\theta}_{1_0} - \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{V}^t\boldsymbol{y}^*)^t\boldsymbol{D}(\boldsymbol{\theta}_{1_0} - \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{V}^t\boldsymbol{y}^*)\right] \\ &= \frac{n}{2}log\left(\frac{\sigma_0^2}{\sigma^2}\right) + \frac{n}{2}\left(\frac{\sigma^2}{\sigma_0^2} - 1\right) \\ &+ \frac{1}{2\sigma_0^2}(\boldsymbol{\theta}_{1_0} - \boldsymbol{\theta}_1)^t\boldsymbol{D}(\boldsymbol{\theta}_{1_0} - \boldsymbol{\theta}_1)\end{aligned} \tag{4.28}$$

de manera similar: aaaaaaaa

$$
\begin{aligned}
&KL(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma^2 \mid \boldsymbol{\theta}_{1_0}, \boldsymbol{\theta}_{2_0}, \sigma_0^2) \\
&= E\left[\left(\frac{L(\boldsymbol{\theta}_{1_0}, \boldsymbol{\theta}_{2_0}, \sigma_0^2 \mid \boldsymbol{y})}{L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma^2 \mid \boldsymbol{y})}\right) \mid \boldsymbol{\theta}_{1_0}, \boldsymbol{\theta}_{2_0}, \sigma_0^2\right] \\
&= \frac{1}{2\sigma^2} E\left[(\boldsymbol{\theta}_1 - \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{V}^t\boldsymbol{y}^*)^t \boldsymbol{D}(\boldsymbol{\theta}_1 - \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{V}^t\boldsymbol{y}^*)\right] \\
&- \frac{1}{2\sigma_0^2} E\left[(\boldsymbol{\theta}_{1_0} - \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{V}^t\boldsymbol{y}^*)^t \boldsymbol{D}(\boldsymbol{\theta}_{1_0} - \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{V}^t\boldsymbol{y}^*)\right] \qquad (4.29) \\
&= \frac{n}{2} log\left(\frac{\sigma^2}{\sigma_0^2}\right) + \frac{n}{2}\left(\frac{\sigma^2}{\sigma_0^2} - 1\right) + \frac{1}{2\sigma^2}(\boldsymbol{\theta}_{1_0} - \boldsymbol{\theta}_1)^t \boldsymbol{D}(\boldsymbol{\theta}_{1_0} - \boldsymbol{\theta}_1),
\end{aligned}
$$

nótese que $KL(\boldsymbol{\theta}_{1_0}, \boldsymbol{\theta}_{2_0}, \sigma_0^2 \mid \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma^2)$ y $KL(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma^2 \mid \boldsymbol{\theta}_{1_0}, \boldsymbol{\theta}_{2_0}, \sigma_0^2)$ no dependen de $\boldsymbol{\theta}_2$. Se puede verificar:

$$
\begin{aligned}
KL(\boldsymbol{\theta}_{1_0}, \sigma_0^2 \mid \boldsymbol{\theta}_1, \sigma^2) &= KL(\boldsymbol{\theta}_1, \sigma_0^2 \mid \boldsymbol{\theta}_1, \sigma^2) \\
&+ KL(\boldsymbol{\theta}_{1_0}, \sigma_0^2 \mid \boldsymbol{\theta}_1, \sigma_0^2) \qquad (4.30)
\end{aligned}
$$

$$
\begin{aligned}
KL(\boldsymbol{\theta}_1, \sigma^2 \mid \boldsymbol{\theta}_{1_0}, \sigma_0^2) &= KL(\boldsymbol{\theta}_{1_0}, \sigma^2 \mid \boldsymbol{\theta}_{1_0}, \sigma_0^2) \\
&+ KL(\boldsymbol{\theta}_1, \sigma^2 \mid \boldsymbol{\theta}_{1_0}, \sigma^2) \qquad (4.31)
\end{aligned}
$$

Por tanto, de acuerdo al resultado (3.1), la DIO para $\boldsymbol{\theta}$ es:

$$
\begin{aligned}
\delta(\boldsymbol{\theta}_1, \sigma^2) &= 2\,\text{mín}\{KL(\boldsymbol{\theta}_{1_0}, \sigma_0^2 \mid \boldsymbol{\theta}_1, \sigma_0^2), KL(\boldsymbol{\theta}_1, \sigma^2 \mid \boldsymbol{\theta}_{1_0}, \sigma^2)\} \\
&= \frac{n}{2} log\left(\frac{\sigma^2}{\sigma_0^2}\right) + \frac{n}{2}\left(\frac{\sigma^2}{\sigma_0^2} - 1\right) \\
&+ \frac{1}{2\sigma^2}(\boldsymbol{\theta}_{1_0} - \boldsymbol{\theta}_1)^t \boldsymbol{D}(\boldsymbol{\theta}_{1_0} - \boldsymbol{\theta}_1) \qquad (4.32)
\end{aligned}
$$

Para $\boldsymbol{\theta}_{1_0} = \mathbf{0}$ y $\sigma_0^2 = \sigma^2$, se tiene:

$$
KL(\boldsymbol{\theta}_{1_0}, \sigma_0^2 \mid \boldsymbol{\theta}_1, \sigma^2) = KL(\boldsymbol{\theta}_1, \sigma^2 \mid \boldsymbol{\theta}_{1_0}, \sigma_0^2) = \frac{1}{2\sigma^2}\boldsymbol{\theta}_1^t \boldsymbol{D}\boldsymbol{\theta}_1
$$

Así, la distribución convencional de $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^t, \boldsymbol{\theta}_2^t)^t$ es:

$$\pi(\boldsymbol{\theta} \mid \sigma^2) \propto \left[1 + \frac{\boldsymbol{\theta}_1^t \boldsymbol{D} \boldsymbol{\theta}_1}{n\sigma^2}\right]^{-q} f_{p-r}(\boldsymbol{\theta}_2^t) \tag{4.33}$$

donde $f_{p-r}(\boldsymbol{\theta}_2^t)$ es una distribución *a priori* propia para $\boldsymbol{\theta}_2$.
Note que $\pi(\boldsymbol{\theta} \mid \sigma^2)$, con $q = \frac{r+1}{2}$ puede expresarse como una mezcla Normal-Gamma:

$$\pi(\boldsymbol{\theta} \mid \sigma^2, \gamma) = \mathrm{N}(\boldsymbol{\theta} \mid \mathbf{0}, n\sigma^2\gamma^{-1}\boldsymbol{D}^{-1}) \qquad \text{y} \tag{4.34}$$

$$\pi(\gamma) = \mathrm{Ga}\left(\gamma \mid \frac{1}{2}, \frac{1}{2}\right) \tag{4.35}$$

así la distribución a priori conjunta es:

$$\begin{aligned}
\pi(\boldsymbol{\theta} \mid \sigma^2, \gamma) &= \int_{\mathbb{R}^+} \pi(\boldsymbol{\theta} \mid \sigma^2, \gamma)\pi(\gamma) d\gamma \\
&= \int_{\mathbb{R}^+} \mathrm{N}(\boldsymbol{\theta} \mid \mathbf{0}, n\sigma^2\gamma^{-1}\boldsymbol{D}^{-1}) \mathrm{Ga}\left(\gamma \mid \frac{1}{2}, \frac{1}{2}\right) f_{p-r}(\boldsymbol{\theta}_2^t) d\gamma \\
&\propto \int_{\mathbb{R}^+} \gamma^{\frac{1}{2}-1} \left(\frac{\gamma}{\sigma^2}\right)^{\frac{r}{2}} exp\left\{-\frac{\gamma}{2} - \frac{\gamma \boldsymbol{\theta}^t \boldsymbol{D} \boldsymbol{\theta}}{2n\sigma^2}\right\} f_{p-r}(\boldsymbol{\theta}_2^t) d\gamma \\
&\propto \int_{\mathbb{R}^+} \gamma^{\frac{r+1}{2}-1} exp\left\{-\frac{\gamma}{2}\left(1 + \frac{\boldsymbol{\theta}^t \boldsymbol{D} \boldsymbol{\theta}}{n\sigma^2}\right)\right\} f_{p-r}(\boldsymbol{\theta}_2^t) d\gamma \\
&\propto \left(1 + \frac{\boldsymbol{\theta}^t \boldsymbol{D} \boldsymbol{\theta}}{n\sigma^2}\right)^{-\frac{r+1}{2}} \\
&\quad \mathrm{x} \int_{\mathbb{R}^+} \frac{\left(1 + \frac{\boldsymbol{\theta}^t \boldsymbol{D} \boldsymbol{\theta}}{n\sigma^2}\right)^{\frac{r+1}{2}}}{\Gamma\left(\frac{r+1}{2}\right)} \gamma^{\frac{r+1}{2}-1} exp\left\{-\frac{\gamma}{2}\left(1 + \frac{\boldsymbol{\theta}^t \boldsymbol{D} \boldsymbol{\theta}}{n\sigma^2}\right)\right\} f_{p-r}(\boldsymbol{\theta}_2^t) d\gamma \\
&\propto \left(1 + \frac{\boldsymbol{\theta}^t \boldsymbol{D} \boldsymbol{\theta}}{n\sigma^2}\right)^{-\frac{r+1}{2}} f_{p-r}(\boldsymbol{\theta}_2^t) d\gamma
\end{aligned} \tag{4.36}$$

Capítulo 5

DISTRIBUCIÓN *A POSTERIORI*

Para obtener la distribución *a posteriori* o distribución final, asúmase las distribuciones *a priori* de Jeffrey para σ^2, dada por:

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2} \tag{5.1}$$

y para $\boldsymbol{\mu}$ se asigna una distribución *a priori* no informativa $\pi(\boldsymbol{\mu}) \propto \mathbf{1}$.

De (4.24) la distribución condicional de $\boldsymbol{y}$ está dada por:

$$\pi(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}) = N_n(\boldsymbol{y} \mid \mu\mathbf{1} + \boldsymbol{V}\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{\theta}_1, \sigma^2\boldsymbol{I}) \tag{5.2}$$

De acuerdo al teorema de Bayes la distribución conjunta *a posteriori* de $(\boldsymbol{\theta}, \sigma^2, \gamma)$ es:

$$\begin{aligned} \pi(\boldsymbol{\theta}, \sigma^2, \gamma \mid \boldsymbol{y}) \propto \mathrm{N_n}(\boldsymbol{y} \mid \mu\mathbf{1} + \boldsymbol{V}\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{\theta}_1, \sigma^2\boldsymbol{I}) \\ \text{x} \quad \mathrm{N}(\boldsymbol{\theta} \mid \mathbf{0}, n\sigma^2\gamma^{-1}\boldsymbol{D}^{-1})\mathrm{Ga}\left(\gamma \mid \frac{1}{2}, \frac{1}{2}\right) f_{p-r}(\boldsymbol{\theta}_2^t)\pi(\sigma^2) \end{aligned} \tag{5.3}$$

Dado que la distribución *a posteriori* en (5.3) no tiene una forma cerrada es necesario implementar un algoritmo de simulación con la finalidad de obtener las distribuciones marginales y otras cantidades de interés; por ejemplo, estimaciones puntuales de $\boldsymbol{\theta}$ y σ^2. También es de interés en este caso generar predicciones dados los datos observados por medio de la distribución predictiva *a posteriori*. Por todo lo anterior, se sugiere simular de la distribución *a posteriori* conjunta por medio de un algoritmo MCMC, siendo en este caso el muestreador de Gibbs la mejor opción dado que es posible simular de las distribuciones condicionales.

Para poder aplicar el algoritmo del muestreador de Gibbs es necesario obtener las distribuciones condicionales de cada uno de los parámétros involucrados.
Las distribuciones condicionales bajo la reparametrización en (4.24) son las siguientes: para $\boldsymbol{\theta_1}$ es:

$$\begin{aligned}\pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2,\sigma^2,\mu,\gamma,\boldsymbol{y}) &= \pi(\boldsymbol{\theta}_1|\sigma^2,\boldsymbol{\mu},\gamma,\boldsymbol{y})\pi(\boldsymbol{y}|\boldsymbol{\theta}_1,\sigma^2,\mu) \\ &\propto \mathrm{N}(\boldsymbol{y}|\mu\mathbf{1}+\boldsymbol{V}\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{\theta}_1,\sigma^2\boldsymbol{I})\mathrm{N}\left(\boldsymbol{\theta}|\mathbf{0},\sigma^2\gamma^{-1}\boldsymbol{D}^{-1}\right)\end{aligned} \tag{5.4}$$

la distribución de σ^2 es:

$$\begin{aligned}\pi(\sigma^2|\boldsymbol{\theta_1},\theta_2,\boldsymbol{\mu},\boldsymbol{y}) \propto \mathrm{N_n}(\boldsymbol{y} \mid \boldsymbol{V}\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{\theta}_1,\sigma^2\boldsymbol{I}) \\ \text{x} \quad \mathrm{N}\left(\boldsymbol{\theta}|\mathbf{0},\sigma^2\gamma^{-1}\mathbf{D}^{-1}\right)\pi(\sigma^2)\end{aligned} \tag{5.5}$$

para $\boldsymbol{\theta}_2$, la distribución condicional es:

$$\pi(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1,\sigma^2,\boldsymbol{\mu},\boldsymbol{y},\gamma) \propto f_{p-r}(\boldsymbol{\theta}_2) \tag{5.6}$$

la distribución condicional de $\boldsymbol{\mu}$ es

$$\pi(\boldsymbol{\mu} \mid \sigma^2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{y}) = \mathrm{N}\left(\boldsymbol{m}^t(\mathbf{1}/n), \sigma^2/n\right), \tag{5.7}$$

donde $\boldsymbol{m} = \boldsymbol{y} - \boldsymbol{V}\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{\theta}_1$, y por último la distribución condicional del parámetro γ es:

$$\begin{aligned} \pi(\gamma \mid \boldsymbol{\theta}, \sigma^2\boldsymbol{y}) &= \pi(\boldsymbol{\theta}|\sigma^2, \gamma)\pi(\gamma) \\ &\propto \mathrm{N}\left(\boldsymbol{\theta}|\mathbf{0}, \sigma^2\gamma^{-1}\boldsymbol{D}^{-1}\right) \mathrm{Ga}\left(\gamma|1/2, n/2\right) \end{aligned} \tag{5.8}$$

5.1. Algoritmos de simulación

En esta sección describiremos dos métodos computacionales que permiten muestrear de distribuciones mediante un proceso de cadenas de Markov Monte Carlo (MCMC) que son: Metropolis-Hasting y muestreador de gibbs

5.1.1. Metropolis-Hasting

En éste método se trata de simular de un parámetro del que conocemos la densidad final $f(\boldsymbol{\theta}|\boldsymbol{x})$ salvo, quizás, por la constante de integración. Este algoritmo es semejante al algoritmo del rechazo para simular variables aleatorias. En este caso, se busca una densidad condicional $g(\cdot|\boldsymbol{\theta})$ llamada distribución de cubrimiento de la que sea fácil muestrear. Después se generan observaciones de esta distribución de cubrimiento para decidir si pertenecen a la distribución de $\boldsymbol{\theta}|\boldsymbol{x}$ mediante un sorteo. El algoritmo se puede resumir de la siguiente forma :

1. . Definir un valor inicial $\boldsymbol{\theta}^{(0)}$,
2. $t = 0$,
3. Generar $\boldsymbol{\phi} \sim g(\cdot \mid \boldsymbol{\theta}^{(t)})$,
4. Definir
$$\alpha(\boldsymbol{\theta}^{(t)}, \phi) = \text{mín}\left[1, \frac{f(\boldsymbol{\phi} \mid \boldsymbol{x}) g(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\phi})}{f(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{x}) g(\boldsymbol{\phi} \mid \boldsymbol{\theta}^{(t)})}\right]$$
5. Tomar
$$\boldsymbol{\theta}^{(t+1)} = \begin{cases} \boldsymbol{\phi} & \text{con probabilidad} \quad \alpha \\ \\ \boldsymbol{\theta}^{(t)} & \text{en caso contrario} \end{cases}$$
6. $t = t + 1$. Ir a 3

Se puede utilizar cualquier distribución $g(\boldsymbol{\phi} \mid \boldsymbol{\theta}^{(t)})$, lo importante es que sea fácil de muestrear, sin embargo la convergencia del algoritmo se facilita de $g \approx f$.

5.1.2. Gibbs Sampler

El muestreador de Gibbs es una versión del metodo de Metropolis-Hastings donde se toma como distribución propuesta la distribución condicionada de la distribución original de donde se quiere simular el parametro $\boldsymbol{\theta}_i$, es decir,

$$g_i(\cdot \mid \boldsymbol{\theta}_i, \boldsymbol{\theta}_{-i}) = f(\cdot \mid \boldsymbol{\theta}_{-i}, \boldsymbol{x}) \tag{5.9}$$

Dada (5.9), se puede demostrar que la probabilidad de aceptar ϕ es siempre 1, así, los valores propuestos siempre son aceptados. Aunque,

en este caso, las distribuciones condicionadas deben ser todas fáciles de simular. El algoritmo para el muestreador de Gibbs es el siguiente:

1. Valores iniciales arbitrarios $\boldsymbol{\theta}^{(0)}$
2. Generar $\boldsymbol{\theta}_1^{(t+1)} \sim f(\boldsymbol{\theta}_1 \mid \boldsymbol{x}, \boldsymbol{\theta}_2^{(t)}, \ldots, \boldsymbol{\theta}_k^{(t)})$,
3. Generar $\boldsymbol{\theta}_2^{(t+1)} \sim f(\boldsymbol{\theta}_2 \mid \boldsymbol{x}, \boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_3^{(t)}, \ldots, \boldsymbol{\theta}_k^{(t)})$,
   ...
   Generar $\boldsymbol{\theta}_k^{(t+1)} \sim f(\boldsymbol{\theta}_k \mid \boldsymbol{x}, \boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \ldots, \boldsymbol{\theta}_{k-1}^{(t+1)})$
4. Ir a 2.

Así repetir 2-4 hasta tener un tamaño de muestra n.

5.1.3. Validación cruzada (VC)

Para medir la capacidad predictiva del modelo propuesto se utilizó validación cruzada (VC) y se compararon los resultados, con los obtenidos utilizando modelos alternativos (Crossa *et al.*, 2010). La idea básica de este método consiste en dividir los casos (datos) aleatoriamente en dos conjuntos disjuntos, llamados muestra de entrenamiento y muestra de validación. De acuerdo con Zhang (1993) la idea de la VC múltiple o en k- subgrupos apareció por primera vez en Geisser (1975), que en lugar de eliminar una observación VC simple (dejar uno fuera), se eliminan más de una. Suponga un tamaño de muestra n que se puede escribir como $n = k \times d$, donde k y d son números enteros. En lugar de sumar sobre todos los posibles subconjuntos de tamaño d, se divide $1, ..., n$ en k subgrupos $s_1, ..., s_k$, que se excluyen mutuamente. Por tanto sin pérdida de generalidad, supongamos que la división es la siguiente:

$$\underbrace{1, ..., d}_{s_1}, \underbrace{d+1, ..., 2k}_{s_2}, ..., \underbrace{(k-1)d, ..., kd}_{s_k} \tag{5.10}$$

Por ejemplo, supóngase que se utlizan $k = 10$-subgrupos. Así, los datos son divididos en 10 conjuntos de tamaño $k = [n/10]$ de la siguiente forma:

$$\boldsymbol{y}^* = \underbrace{y_{[1]}, ..., y_{[k]}}_{\text{conjunto} \quad 1}, \underbrace{y_{[k+1]}, ..., y_{[2k]}}_{\text{conjunto} \quad 2}, ..., \underbrace{y_{[9k+1]}, ..., y_{[10k]}}_{\text{conjunto} \quad 10}$$

donde $y[i]$ es cualquier elemento de $\boldsymbol{y}$ ordenado en la posición [i]. La matriz $\boldsymbol{X}$ también se ordena de tal manera que sus n hileras correspondan a los n elementos de $\boldsymbol{y}^*$, resultando $\boldsymbol{X}^*$. El método procede de la siguiente manera:

1. Se ajusta el modelo $\boldsymbol{y}_{-1} = \boldsymbol{X}_{-1}\boldsymbol{\beta}$y se obtienen los valores $\hat{\boldsymbol{\beta}}_{-1}$, donde $\boldsymbol{y}_{-1}$ es el vector $\boldsymbol{y}^*$ de donde se suprimieron las k observaciones correspondientes al conjunto 1 y $\boldsymbol{X}_{-1}$ es la matriz $\boldsymbol{X}^*$de la cual se eliminaron las hileras correspondientes al conjunto 1.

2. Con el vector de cofcientes estimados $\hat{\boldsymbol{\beta}}_{-1}$, se predicen los k valores eliminados:

$$\hat{\boldsymbol{y}}_1 = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_{-1} + \sigma_\epsilon z \tag{5.11}$$

   donde $z \sim \mathrm{N}(0, 1)$, $\hat{\boldsymbol{y}}_1 = (\hat{y}_{[1]}, ..., \hat{y}_{[k]})^t$ y $\boldsymbol{X}_1$ es la matriz que solo contiene las k hileras correspondientes al conjunto 1.

3. Se repiten los pasos 1 y 2 para cada uno de los conjuntos restantes.

4. Una vez que se tiene el vector de valores predichos para los 10

conjuntos,

$$\hat{\boldsymbol{y}}^* = (\hat{y}_{[1]}, ..., \hat{y}_{[k]}, ..., \hat{y}_{[k+1]}, ..., \hat{y}_{[2k]}, ..., \hat{y}_{[9k+1]}, ..., \hat{y}_{[10k]})^t,$$

se calcula su correlación con el vector de observaciones $\boldsymbol{y}^*$. Entre más grande es tal correlación, se tiene un modelo con mayor capacidad de predicción.

5.2. Simulación

Recuérdese que la inferencia bayesiana está basada en la distribución *a posteriori* conjunta. Como se puede verificar en (5.3), no es posible obtener de forma analítica la distribución conjunta de $\left(\boldsymbol{\theta}_1, \sigma^2\right)$, por lo que se debe utilizar un método para aproximarla. En este trabajo se utilizó un método de simulación de cadenas de Markov (MCMC, por sus siglas en inglés) para obtener una muestra de la distribución conjunta *a posteriori* a partir de la cual se estiman cantidades de interés; por ejemplo, los momentos *a posteriori*. En particular, se utiliza el muestreo de Gibbs generando muestras de las distribuciones condicionales en el siguiente orden:

1. Asignar valores iniciales para $\boldsymbol{\mu}$, $\boldsymbol{\theta}_1$, σ^2 y γ, es decir, $\boldsymbol{\mu}^{(0)}$,$\boldsymbol{\theta}_1^{(0)}$, $\sigma^{2^{(0)}}$ y $\gamma^{(0)}$
2. Hacer $t = 1$
3. Muestrear de la distribución condicional de $\boldsymbol{\theta}_1 \mid \boldsymbol{y}, \boldsymbol{\mu}^{(0)}, \sigma^{2^{(0)}}, \gamma^{(0)}$ dada en (5.4) y asi obtener los $\boldsymbol{\theta}_1^{(t)}$.
4. Muestrear de la distribución condicionada de $\sigma^2 \mid \boldsymbol{y}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\theta}^{(t)}, \gamma^{(0)}$ dada en (5.5), hasta obtener $\sigma^{(t)}$.

5. Muestrear de la distribución de $\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{\beta}^{(t)}, \sigma^{(t)}, \gamma^{(0)}$ de la ecuación (5.6) y obtener $\boldsymbol{\mu}^{(t)}$.

6. Muestrear de la distribución de $\gamma \mid \boldsymbol{y}, \boldsymbol{\beta}^{(t)}, \sigma^{(t)}, \boldsymbol{\mu}^{(t)}$ de la ecuación (5.8), hasta obtener $\gamma^{(t)}$

7. Por último hacer $\boldsymbol{\mu}^{(0)} = \boldsymbol{\mu}^{(t)}$, $\boldsymbol{\beta}^{(0)} = \boldsymbol{\beta}^{(t)}$, $\gamma^{(0)} = \gamma^{(t)}$ y $\sigma^{(0)} = \sigma^{(t)}$ para obtener $t = t + 1$

8. Repetir los pasos de 3-6 hasta obtener el tamaño de muestra s deseado.

Para evaluar la capacidad predictiva del procedimiento propuesto se llevó a cabo una validación cruzada, para la cual el vector de observaciones del fenotipo rendimiento de grano(GY) fue dividido aleatoriamente en 2 subgrupos, uno de entrenamiento $\boldsymbol{y}_{ent}$ y otro de datos a predecir $\boldsymbol{y}_{pred}$; así, $\boldsymbol{y} = \left[\boldsymbol{y}_{ent} \mid \boldsymbol{y}_{pred}\right]^t$ de modo que la distribución *a posteriori* conjunta de los parámetros y los datos predichos esta dada por:

$$\pi(\boldsymbol{\theta}, \sigma^2, \boldsymbol{\mu}, \gamma, \boldsymbol{Y}_{pred} \mid \boldsymbol{y}_{ent}) \propto \pi(\boldsymbol{y}_{ent} \mid \boldsymbol{\theta}, \sigma^2, \boldsymbol{Y}_{pred}) \pi(\boldsymbol{Y}_{pred} \mid \boldsymbol{\theta}, \sigma^2) \pi(\boldsymbol{\theta}, \boldsymbol{\mu}, \sigma^2)$$

así, la distribución marginal predictiva de $\boldsymbol{Y}_{pred} \mid \boldsymbol{y}_{ent}$ se obtiene con:

$$\pi(\boldsymbol{Y}_{pred} \mid \boldsymbol{y}_{ent}) = \int_{\boldsymbol{\Theta}^{\mathcal{R}}} \int_{\boldsymbol{\Sigma}^{\boldsymbol{\mathcal{R}}^+}} \int_{\boldsymbol{\mu}^{\mathcal{R}}} \int_{\gamma^{\mathcal{R}^+}} \pi(\boldsymbol{\theta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{Y}_{pred} \mid \boldsymbol{y}_{ent}) d\gamma d\boldsymbol{\mu} d\sigma^2 d\boldsymbol{\theta}$$

Una medida de la capacidad predictiva del modelo ajustado es la correlación entre el vector de valores predichos, dado por el valor esperado

de la distribución predictiva, y el vector de valores a predecir; esto es:

$$\text{Cor}\left(\boldsymbol{y}_{pred}, E[\boldsymbol{Y}_{pred} \mid \boldsymbol{y}_{ent}]\right)$$

una estimación de $E[\boldsymbol{Y}_{pred} \mid \boldsymbol{y}_{ent}]$ puede obtenerse, mediante simulación MCMC, como el promedio de los valores simulados $\boldsymbol{y}_{pred}^{(s)}$ de la distribución predictiva en cada iteración del muestreo de Gibbs:

$$\hat{E}[\boldsymbol{Y}_{pred} \mid \boldsymbol{y}_{ent}] = \sum_{i=1}^{s} \frac{\boldsymbol{y}_{pred}^{(s)}}{s}$$

Para estimar la distribución conjunta *a posteriori* se generó una muestra MCMC de tamaño 30,000, se utilizó un periodo de calentamiento de 15000 iteraciones. Los datos utilizados como ejemplo son de una colección de 599 variedades de trigo del CIMMYT (Centro Internacional para el Mejoramiento del Maíz y el Trigo) evaluados en cuatro ambientes (A1,...,A4). El fenotipo evaluado fue el rendimiento del grano (GY). Estos datos (wheat) están disponibles en la libreria BLR de R, con la siguiente estructura: la matriz $\boldsymbol{Y}$ que contiene el rendimiento promedio de grano para los cuatro mega-ambientes, la matriz $\boldsymbol{A}$ contiene las relaciones aditivas obtenidas del pedigrí y la matriz $\boldsymbol{X}$ tiene la información de los marcadores.

El método presentado para la predicción de efectos genéticos bajo el modelo de rango incompleto con una distribución a priori no informativa (RBC), se compara con los métodos BLUP (Best Linear Unbiased Predictor), RKHS (Reproducing Kernel Hilbert Space) y BL (Bayesian Lasso). Estos modelos fueron ajustados para el fenotipo rendimiento de grano (GY) (Tabla 5.1) en (Crossa *et al.*, 2010); P es modelo del pedigrí, (M-RKHS) es la regresión RKHS en la que se incluye sólo los

marcadores y (PM-RKHS) con Pedigrí-Marcadores, al igual para BL con marcadores (M-BL) y con Marcadores-Pedigrí (PM-BL); mediante un proceso de validación cruzada descrito anteriormente.

5.3. RESULTADOS

En la figura 5.1 puede observarse que la regresión Ridge bayesiana (RRB) produce estimaciones de los efectos de los marcadores muy contraidas hacia el cero; en comparación, después de transformar las estimaciones, el método RBC produce estimaciones menos encogidas al cero. En la figura 5.2 se presentan los valores ajustados de la RRB y de RBC, donde se puede observar que la RBC presenta un mejor ajuste.

En la Tabla 5.1 se muestran las correlaciones de la VC para los 6 modelos comparados en los 4 ambientes diferentes. Se observa que la RBC es la que posee las correlaciones más altas en los ambientes A2, A3 y A4; éstas son 0.606, 0.482 y 0.522 respectivamente , y en el ambiente A1, RBC es superada por la regresión RKHS sólo con marcadores, obteniendo una correlación de 0.608, otro modelo que tambien presenta alas correlaciones es el PM-BL respecto a M-BL. Además, es posible observar que el cambio relativo en relación al efecto del pedigrí (P) es positivo, para la regresión RKHS con marcadores y marcadores-pedigrí se tiene un incremento del 35.7 % y un 34.2 % respectivamente en en (A1), mientras que la RBC tiene los más altos incrementos de capacidad preditiva en A2 y A3 de 45.32 % y 15.59 % respectivamente.

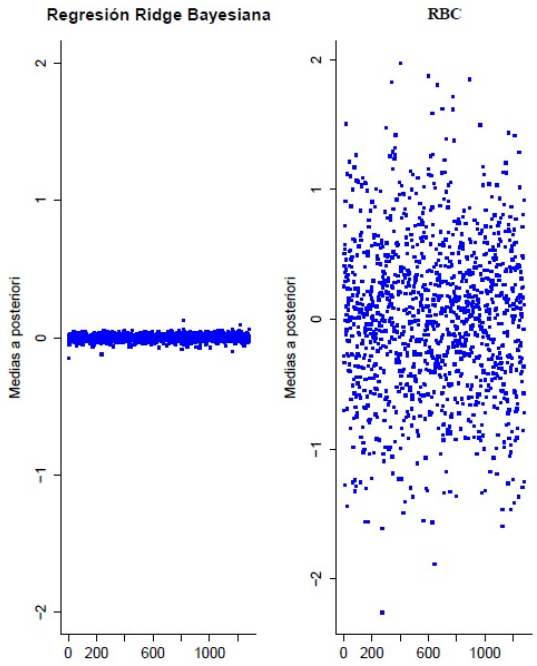

Figura 5.1: Medias *a posteriori* en el ambiente A1

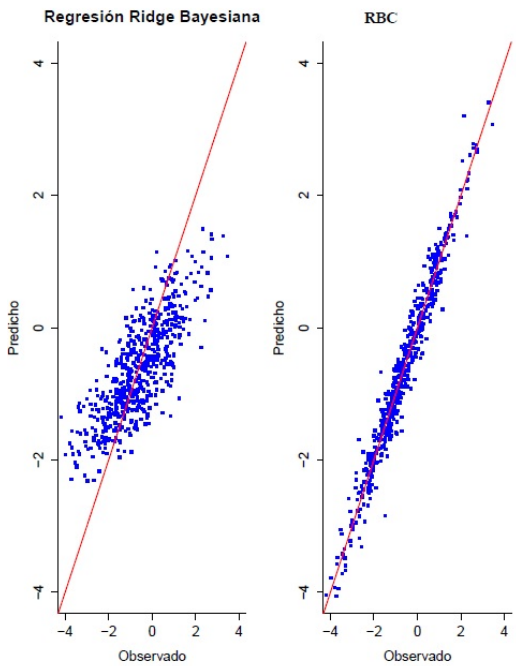

Figura 5.2: Valores ajustados en el ambiente A1

Tabla 5.1: Correlaciones obtenidas para diferentes modelos aplicando validación cruzada de 10 folds con 30000 iteraciones

	Modelos						
FEN-AMB	P	PM-RKHS	M-RKHS	M-BL	PM-BL	BLUP	REG.CONV
				Correlación			
GY-A1	0.448	0.601	0.608	0.518	0.542	0.480	0.584
GY-A2	0.417	0.494	0.497	0.493	0.501	0.488	0.606
GY-A3	0.417	0.445	0.478	0.403	0.449	0.355	0.482
GY-A4	0.449	0.524	0.524	0.457	0.495	0.464	0.522
				% cambio (en relación P)			
GY-A1	–	34.2	35.7	15.6	21.0	7.1	30.36
GY-A2	–	18.5	19.2	18.2	20.1	17.0	45.32
GY-A3	–	6.7	14.6	-3.4	7.7	-14.9	15.59
GY-A4	–	16.7	16.7	1.8	10.2	3.3	16.26

P: es el efecto del pedrigrí

5.4. Discusión

Anteriormente se mencionó que las distribuciones informativas tienden a contraer los valores de los parámetros hacia el cero; al graficar las medias *a posteriori*; esto es, el efecto promedio estimado de los marcadores genéticos bajo la RRB (figura 5.1) observamos que la contracción al cero es muy marcada lo que puede conducir a eliminar marcadores importantes para explicar la variación en la característica bajo estudio (en este caso producción de grano) si se toma como critero de selección la magnitud de los efectos. Recuérdese en que la RRB se asigna cada efecto de los marcadores una distribución *a priori* con media cero; así, dado que $p >> n$ la información aportada por la distribución inicial es muy relevante en relación con la aportada por los datos, lo que se refleja en la contracción de la media *a posteriori* hacia la media *a priori*. Al respecto, la RBC asigna una distribución *a priori* no informativa a no más de n parámetros de regresión y una distribución normal con media cero a los $n-p$ restantes, para los que no hay información en la muestra y son eliminados mediante marginalización. De este modo, la inferencia sobre los parámetros de interés está basada sólo en la información contenida en la muestra.

Referente al proceso de validación cruzada, se observa (Tabla 5.1) que los modelos que presentan altas correlaciones son los que incluyen marcadores-pedigrí (PM-BL y PM-RKHS), tales correlaciones van de 0.44 a 0.608 y con un incremento de la capacidad predictiva del 7.7 % al 35.7 % en relación al pedigrí; No obstante bajo ciertas condiciones BL y RKHS presenta algunos problemas, por ejemplo, cuando $p > n$ BL selecciona a lo más n variables y al seleccionar de cada subgrupo sólo toma una sin importar cual sea, lo que provoca perdida de infroma-

ción relevante en el factor de estudio, mientras que la regresión RKHS es capaz de capturar características más complejas, los resultados son sensibles a la elección del Kerne; por tanto, debido a estos problemas en este trabajo se retoma la propuesta de (Bayarri y Garcia-Donato, 2006, García-Donato, 2003) de obtener una distribución *a priori* no informativa basada en la DKL y aplicando el resultado de (Pérez, 2005) cuando se presentan parámetros de ruido en las divergencias, obteniendo la distribución *a priori* dada en (4.33) y se observó que se puede expresar como una mezcla Normal-Gamma (Zellner y Siow, 1980), la cual se aplicó en el modelo de regresión en componentes principales (RBC) presentando mayor capacidad predictiva y un incremento en relación al efecto del pedigrí de el 16.26 % al 45.32 % en los 4 ambientes.

Capítulo 6

CONCLUSIONES Y RECOMENDACIONES

1. Con los resultados obtenidos de la metodología desarrollada en esta investigación es posible concluir que la RBC tuvo un mejor desempeño en la capacidad predictiva cuando se presenta el problema de $(n << p)$, comparando los valores ajustados con los observados de algunos modelos más utilizados en selección genómica como la regresión RKHS y BL.

2. Se comprobó que al utilizar una distribución *a apriori* no informativa se tiene como resultado un menor encogimiento en los valores de los efectos de los marcadores hacia el cero, obteniendo de esta forma predicciones más precisas ya que no se eliminan covariables importantes para explicar la variación en la característica de interés.

3. La distribución *a priori* convencional expresada como una mezcla Normal-Gama (Zellner y Siow, 1980) tuvó muy buenos resultados para el modelo de regresión usando todos los componentes principales cuando se tiene el problema de $n << p$, sin embargo, se recomienda que en investigaciones futuras el ajuste del modelo sea con un número determinado de componentes para ver si se obtiene un buen ajuste.

4. Los marcadores moleculares tienen un gran impacto en las decisiones de selección genómica; sin embargo, si la información del pedigrí es agregada al modelo se ha observado que la capacidad predictiva aumenta considerablemente; por lo que se sugiere que en trabajos futuros se incluya el efecto del pedigrí a la RBC a través de la distribución *a priori*.

5. Recordemos que las distribuciones *a priori* de (García-Donato, 2003, Zellner y Siow, 1980) son muy aplicadas en selección de variables, por tanto una extensión de esta investigación es aplicar la RBC como un método indirecto de selección variables,aplicando algunos métodos como el factor Bayes.

REFERENCIAS

Amari, S. (1985). *Differential-Geometrical Metods in Statistics*. tomo 28. Springer Verlag, New York.

Amari, S. (1990). *Differential-Geometrical Metods in Statistics*. tomo 28. Springer, Berlin.

Barndorff-Nielsen, O. E. (1978). *Information and exponential families in statistical theory*. New York.

Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, 70, 343–365.

Bayarri, M. J. (1987). Comentario de Testing Precise Hypotheses de Berger, J. O. y Delampady, M. *Statistical Science*, 2, 317–352.

Bayarri, M. J. y Garcia-Donato, G. (2006). Divergence Based Priors for Bayesian Hypothesis testing. *Journal of the Royal Statistical Society*, 70, 299–318.

Bayarri, M. J. y Garcia-Donato, G. (2007). Extending Conventional priors for Testing General Hypotheses in Linear Model. *Biometrika*, 94, 1, 135–152.

Berger, J. O. y Bernardo, J. M. (1989). Estimating a Product of Means: Bayesian Analysis with Reference Priors. *Journal of the American Statistical Association*, 84, 405, 200–207.

Berger, O., James y Bernardo, J. M. (1992a). On the development of reference prior. *Oxford University Press. (with discussion)*, 35–60.

Berger, O., James y Bernardo, J. M. (1992b). Ordered group reference prior with application to the multinomial problem. *Biometrika*, 1, 79, 25–37.

Bernardo, J. M. (1979). Reference posterior distributions for bayesian inference. *The Journal of the Royal Statistics Society B*, 41, 2, 113–147.

Bernardo, J. M. (1985). Análisis Bayesiano de los contrastes de hipótesis. *Trabajos de Estadística*, 36, 45–54.

Bernardo, J. M. y Smith, A. F. M. (1994). *Bayesian Theory*. Wiley and Sons.

Christensen, R. (2001). *Plane Answers to Complex Questions: The Theory of Linear Models*. tercera edición.

Cox, D. R. y Reid, N. (1987). Parameter Orthogonality and Approximate Conditional Inference. *Journal of the Royal Statistical Society*, 49, 1, 1–39.

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

de los Campos, G., D., G. y Allison, D. B. (2010a). Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nature Reviews Genetics*, 11, 880–886.

de los Campos, G., D, G., M, R. G. J., Weigel, K. y Crossa, J. (2010b). Semiparametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genetics*, 92, 295–308.

de los Campos, G., Gianola, D. y Rosa, G. J. (2009). Reproducing Kernel Hilbert Spaces Regression: a General Framework for Genetic Evaluation. *Animal Science*, 87, 1883–1887.

de los Campos, J. M. H. R. P.-W. H. D. . D., G. y Calus, M. P. L. (2012a). Whole genome regression and prediction methods applied to plant an animal breeding. *Genetics*.

de los Campos, Y. C. K. A. I. V., G. y Allison, D. B. (2012b). Prediction of expected years of life using whole-genome markers. *PLoS One*, 7, 1–7.

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

Esteban, M. D., M., M. A., Morales, D. y Morales, J. (2000). Some New Statistics for Testing Point Null Hypotheses with Prior Information. *Statistical Planning and Inference*, 87, 251–271.

García-Donato, G. (2003). *Factores bayes y factores bayes convencionales: Algunos aspectos relevantes*. Tesis Doctoral, Universidad de Valencia departamento de estadistica.

Geisser, S. (1975). The Predictive Sample Reuse Method with Applications. *Journal of the American Statistical Association*, 70, 350, 320–328.

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

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

Gianola, D. y van Kaam, J. B. C. H. M. . (2008). Reproducing Kernel Hilbert Spaces Regression Methods for Genomic Assisted Prediction of Quantitative Traits. *Genetics*, 178, 2289–2303.

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

Gianola, G. (????). Los Métodos Estadísticos en el Mejoramiento Genético.

González-Camacho, J. M., de los Campos, G., Gianola, D., Cairns, J. E., Mahuku, G., Babu, R. y Crossa, J. (2012). Genome-enabled Prediction of Genetic Values Using Radial Function Neural Networks. *Theorical and Applied Genetics*, 125, 759–771.

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

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

Heslot, N., Yang, H. P., Sorrells, E. M. y Jannink, J. L. (2012). Genomic Selection In plant breeding: A comparison of models. *Crop Science Society of America*, 52, 146–160.

Hoerl, A. y Kennard, R. (1970). Ridge regression: biased estimation for non-orthogonal problems. *Technometrics*, 12, 55–67.

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press.

Kullback, S. (1997). *Information theory and Statistics*. Dover Publication Inc.

Kyung, M., Gilly, J., Ghoshz, M. y G., C. (2010). Penalized Regression, Standard Errors, and Bayesian Lassos. *Bayesian Analysis*, 2, 5, 369–412.

Laplace, P. S. (1951). *Essai philosophique sur les probabilités*. Dover, New York.

le Cessie, S. y van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Appl. Statistics*, 41, 1, 191–201.

Lee, A. H. y Silvapulle, M. J. (1988). Ridge estimation in logistic regression. *Comm. in Statistics-Theory and Methods*, 17, 4, 1231–1257.

Li, Q. y Lin, N. (2010). The bayesian Elastic Net. *Bayesian analysis*, 5, 151–170.

Lorenz, A., Chao, S., Asoro, F., ner, E. H., Hayashi, T., H. Iwata, K. S., Sorrells, M. y Jannink, J.-L. (2011). Genomic selection in plant breeding: Knowledge and prospects. *Adv. Agron.*, 110, 77–123.

Makowsky, R., C., P. N. M. K. Y., I., V. A., W., D. C. y et al. (2011). Beyond Missing Heritability: Prediction of Complex Traits. *PLoS Genet*.

Meuwissen, T. H. E., Hayes, B. J. y Goddard, M. E. (2001). Prediction of total genetic value using genomewide dense marker maps. *Genetics*, 159, 1819–1829.

Ober, U., Ayroles, F. J., Stone, A. E., Richards, S., Zhu, D., Gibbs, A. R., Stricker, C., Gianola, D., Schlather, M., Mackay, F. C. T. y Simianer, H. (2012). Using Whole-Genome Sequence Data to Predict Quantitative Trait Phenotypes in Drosophila melanogaster. *PLoS Genetics*.

Park, T. y Casella, G. (2008). The bayesian lasso. *American Statistical Association*, 103, 681–686.

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

Pérez, S. (2005). *Métodos bayesianos objetivos de comparación de medias*. Tesis Doctoral, Universidad de Valencia: Departamento de Estadística.

Ruppert, D., Wand, M. P. y Carroll, R. J. (2003). *Semiparametric Regression*. United States of America by Cambridge University Press, New York.

Schaefer, R. L., Roi, L. D. y Wolfe., R. A. (1984). A Ridge logistic estimator. *Comm. Statistics-Theory and Methods*, 13, 1, 99–113.

Schaid, D. J. (2010). Genomic Similarity and Kernel Methods I: Advancements by Building on Mathematical and Statistical Foundations. *Human Heredity*, 70, 2, 109–131.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statistics soc*, 58, 267–288.

Vazquez, I. A., de los Campos, G., Klimentidis, C. Y., Rosa, M. J. G., Gianola, d., Yi, N. y Allison, B. D. (2012). A Comprehensive Genetic Approach for Improving Prediction of Skin Cancer Risk in Humans. *Genetics*.

Witkovsky, V. (2013). Estimation, Testing, and Prediction Regions of the Fixed and Random Effects by Solving the Henderson's Mixed Model Equations.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6, 233–243.

Zellner, A. y Siow, A. (1980). Posterior odds ratio for selected regression hypotheses. *Bayesian Statistics*, 585–603.

Zhang, P. (1993). Model Selection Via Multifold Cross Validation. *The Annals of Statistics*, 21, 1, 299–313.

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