SELECCIÓN GENÓMICA MEDIANTE MODELOS PARAMÉTRICOS

Regresión de Ridge y Lasso Bayesiano

Carolina Saavedra Diaz[1]

[1]
Pontificia Universidad Javeriana, Cali

Septiembre 20, 2019

Plan

Introducción

Cebada (*Hordeum vulgare*)

- Es una planta anual perteneciente a la familia de las poáceas (gramíneas)
- Un cereal de gran importancia tanto para animales como para humanos y es el quinto cereal más cultivado en el mundo (53 millones de hectáreas o 132 millones de acres).

Introducción

La fenología de las plantas caracteriza los eventos del ciclo de vida del desarrollo de las plantas y cómo estos eventos están influenciados por las variaciones estacionales e interanuales del clima, así como los factores del hábitat.

- En la cebada, diferentes etapas de desarrollo, como la iniciación de la espiguilla y la duración del desarrollo del grano, pueden influir seriamente en el rendimiento y la calidad.

Introducción

EL GRANO

- Forma ahusada, más grueso en el centro y disminuyendo hacia los extremos. La cáscara lo protege contra depredadores y es de utilidad en los procesos de malteado y cervecería ; representa un 13% del peso del grano.
- Se usa ampliamente para la alimentación humana y animal, así mismo en la producción de almidón y la industria química.

Introducción

ALTURA DE PLANTA

- Es un aspecto de gran importancia, ya que si esta es muy baja se dificulta su cosecha mecánica. Cuando los materiales son de porte muy alto, se ocasiona una mayor susceptibilidad al acame que puede afectar de manera gradual el rendimiento de grano final.
- Es afectada por las condiciones climáticas (falta de agua)-> Portes de planta diferentes entre cada ciclo.

Introducción

Heading date o inicio de floración (aparición de espiguillas)

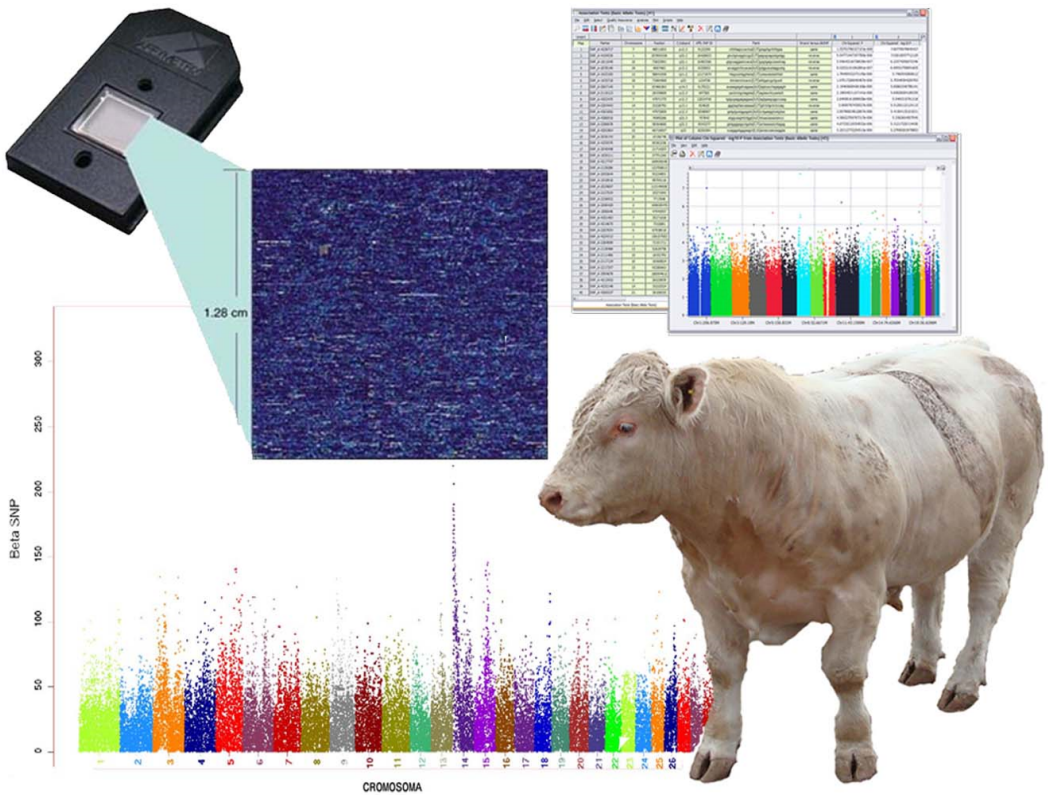
- Es un rasgo complejo en la cebada que tiene un impacto directo en el rendimiento y la calidad del grano y también forma la base de la adaptación evolutiva al clima cambiante.

Introducción

- El rendimiento y la calidad son rasgos complejos importantes en cualquier programa de reproducción. La mejora de estos rasgos es muy difícil de lograr debido a su complejidad genética, fisiológica y física.

- Por lo cual, el mejoramiento genético en plantas tiene como propósito la obtención de germoplasma con características de mayor rendimiento, mayor calidad comercial y mayor resistencia a factores bióticos y abióticos adversos al cultivo. En otras palabras, tiene por finalidad la generación de germoplasma más eficientes, producir productos aprovechables por el hombre como alimento, como materias primas para la industria y como forraje para los animales.

Introducción



Introducción

REGRESIÓN DE RIDGE

- El método de ridge tiende a contraer los coeficientes de regresión al incluir el término de penalización en la función objetivo.
  Aproxima a cero los coeficientes de los predictors pero sin llegar a excluir ninguno.

Inicialmente : eludir los efectos adversos del problemas de colinealidad

Introducción

λ parámetro de penalización, determina la fuerza de penalización

- λ grande algunos coeficientes son practicamente nulos
- λ pequeño los coeficientes crecer
- λ=0 no hay penalización

Introducción

La regresión Ridge consiste en utilizar como estimador de β, el siguiente :

$$\beta = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{\rho} \beta_j x_j \right)^2 + \lambda \sum_{j=1}^{\rho} \beta_j^2$$

$$\beta = (X'X + \lambda I)^{-1} X'y$$

Donde λ es una constante pequeña arbitraria. Cuando todos los predictores están estandarizados, se tiene que XX es la matriz de correlaciones, con unos en la diagonal.

Introducción

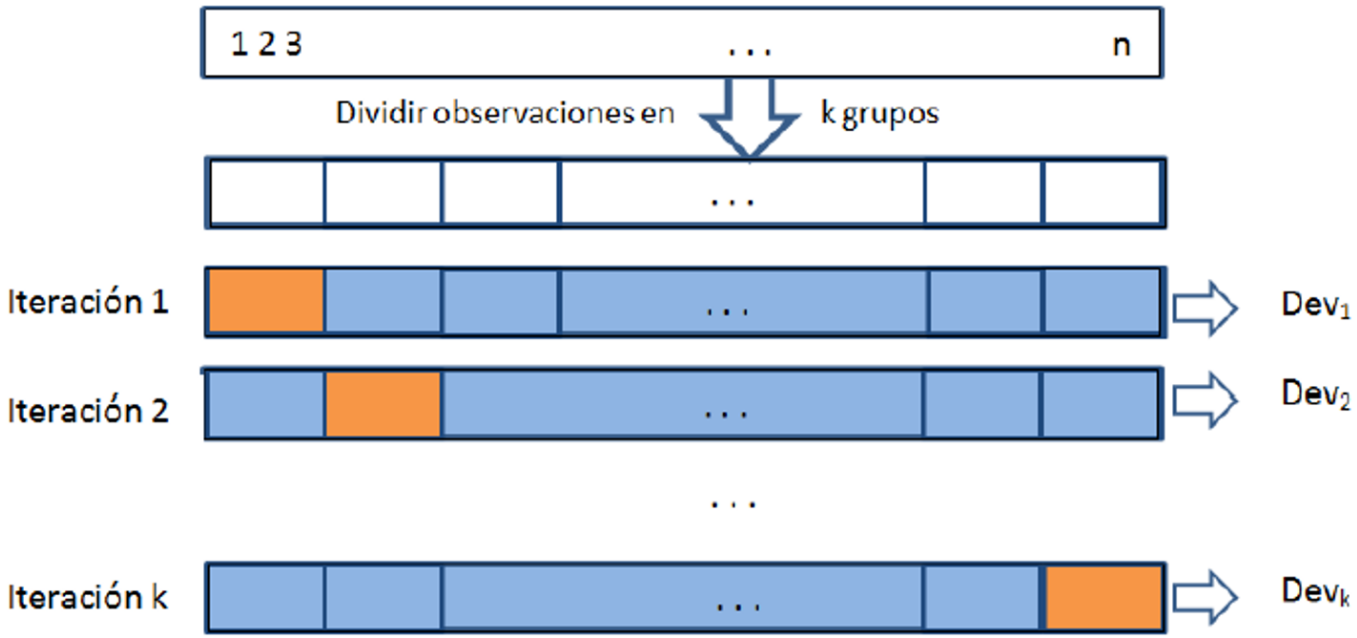
En genética, este método asume que los efectos de los marcadores g son aleatorios con una varianza común. Ridge puede aplicarse en selección genómica de la siguiente manera :

$$\mathbf{g} = (\mathrm{X}'\mathrm{X} + \lambda \mathrm{I})^{-1} \mathrm{X}\mathrm{y}$$

Introducción

VALIDACIÓN CRUZADA

El método hold-out y/o k-fold



Introducción

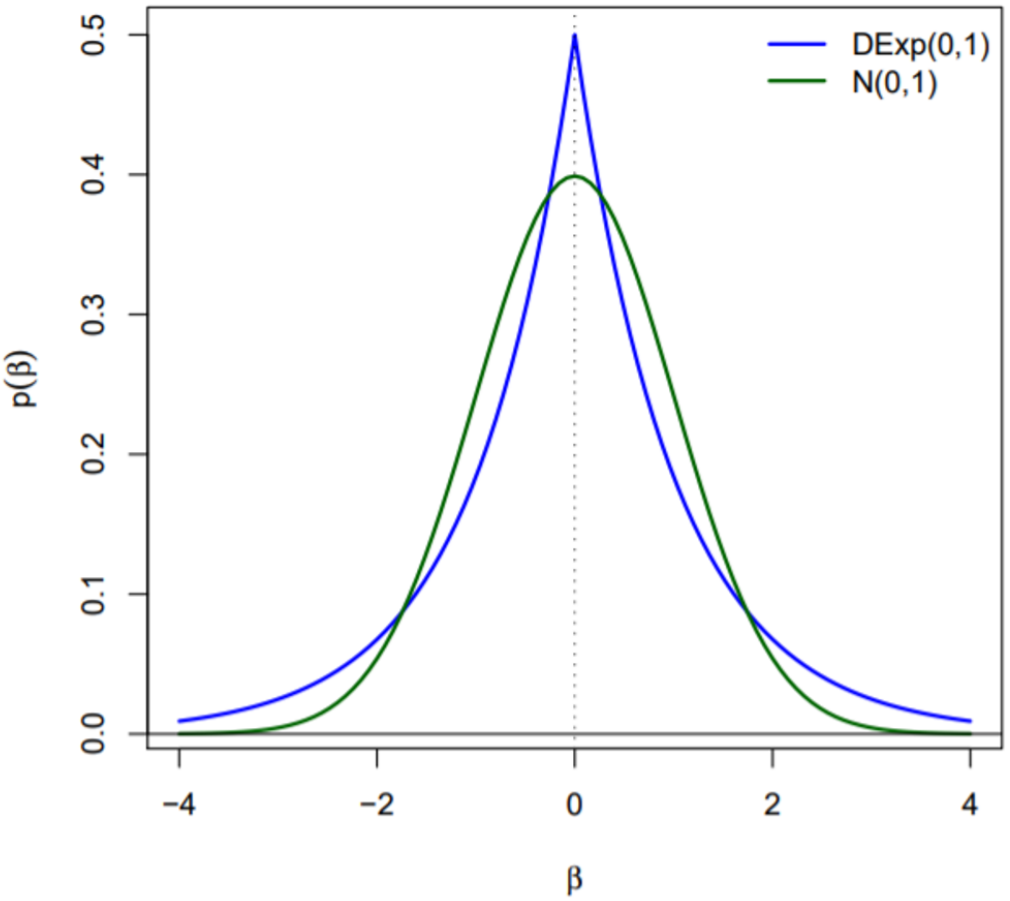
LASSO BAYESIANO

El estimador β_{lasso} puede presentarse como la moda de la distribución posteriori de β, es decir :

$$\beta_{\text{lasso}} = \text{argmax}_{\beta} f\left(\beta|y, \sigma^2, \lambda\right)$$

cuando los parámetros de regresión tienen a priori distribución idéntica e independiente Laplace (doble exponencial).

Introducción



OBJETIVOS

- Comprender el funcionamiento de dos modelos paramétricos : regresión de Ridge (RR) y Regresión Lasso Bayesiano, como modelos utilizados en selección genómica.
- Comparar dos métodos paramétricos usados en selección genómica.
- Ejemplificar el potencial de selección genómica de dos modelos paramétricos mediante el uso del programa R, en datos genéticos con marcadores tipo SNPs y tres rasgos fenotípicos.

MATERIALES Y MÉTODOS

Datos estudiados

Caracteres fenotipicos :

- Producción de grano
- Altura
- Inicio de floración

Marcadores

- 1176 SNPs en 96 observaciones, los SNPs se codificaron como 1 (homocigoto parental 1), 0 (heterocigotos) y -1 (homocigotos parental 2).

MATERIALES Y MÉTODOS

- REGRESIÓN RIGDE : GMNET
- LASSO BAYESIANO : función BGLR de la librería BGLR

Utiliza el algoritmo de Monte Carlo vía Cadenas de Markov denominado Gibbs Sampler que muestrea repetidamente y calcula estadísticas resúmenes de las distribuciones a posteriori.
200.000 iteraciones para cada una de las poblaciones y caracteres fenotípicos, descartando las primeras 1.000 iteraciones.

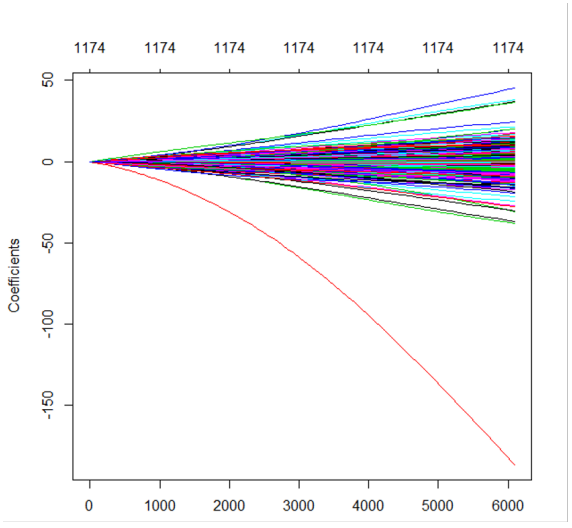
Los datos faltantes (NA), fueron removidos mediante el comando A.mat() presente en la librería rrBLUP.

	Rendimiento del grado	Altura	Inicio de floración
Bestlam	95623.57	305.4752	27.5468

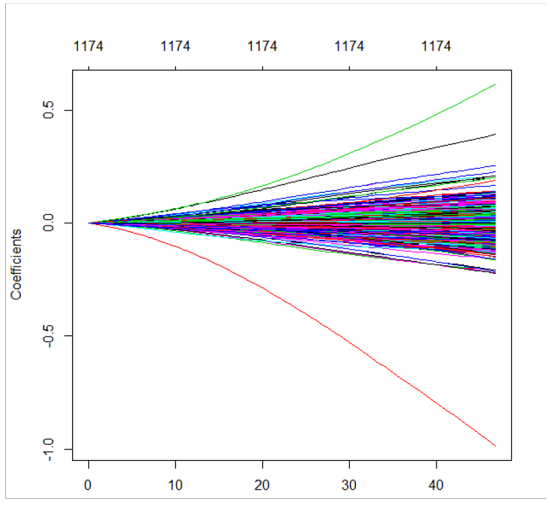
El valor del bestlam corresponde al lambda en el cual se presenta el menor error de predicción por validación cruzada.

Valores grandes de λ, algunos coeficientes son prácticamente nulos, mientras que a medida que se impone una penalización menor los coeficientes crecen, mostrando un efecto del marcador mayor

Trayectoria de la estimación de los coeficientes según el valor del logaritmo de cada λ

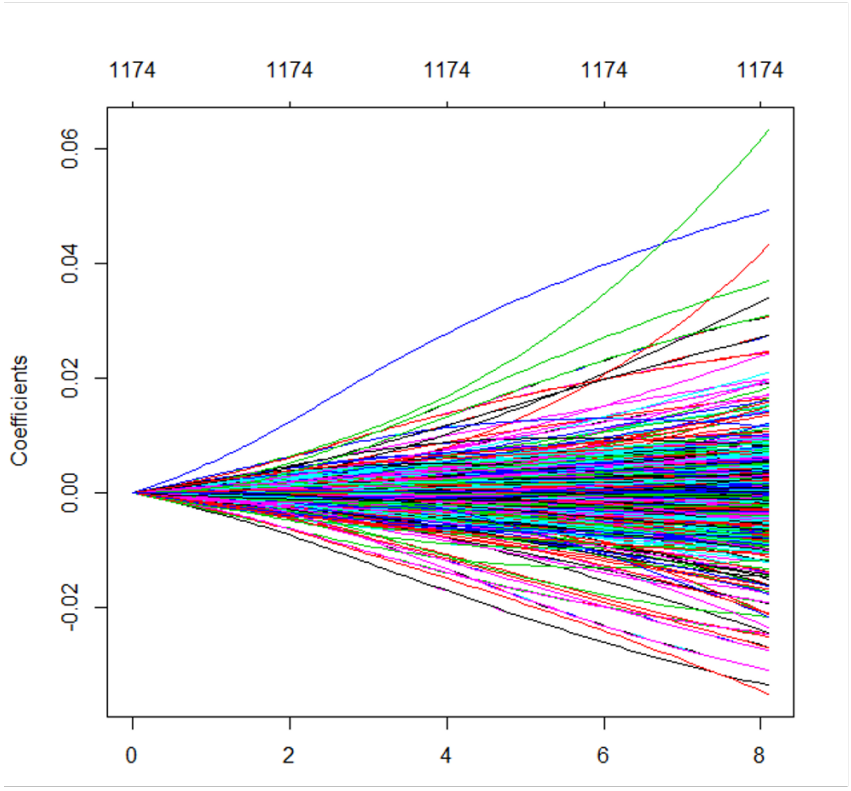


Rendimiento del grano



Altura de la planta

Trayectoria de la estimación de los coeficientes según el valor del logaritmo de cada λ



La función (ridge.coef !=0), se ratifica que el número de coeficientes obtenidos mediante este regresión, iguales a cero es nulo, ya que, aunque se pueden presentar coeficientes muy pequeños (cercanos a cero), no se puede considerar que el modelo sea disperso, por lo tanto no se pueden determinar variables relacionadas con los rasgos.
No permite la selección de variables, ya que este método consigue minimizar la influencia sobre el método de los predictores menos relacionados con la variable de respuesta, pero estos continúan en él.

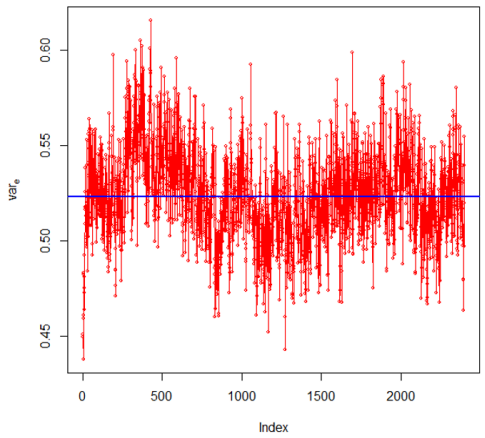
Puede ayudar a determinar o detectar que marcadores presentan un efecto mayor sobre las variables lo cual permitirá reducir el número de marcadores y direccionar un poco la investigación, lo cual puede generar la inclusión de nuevas variables y/o la reducción en costo computacional.

LASSO BAYESIANO

Fenotipo	preAbi.	Rank	Bias	Best10	λ BL	λ clásico
Rendimiento del grado	0.72	0.68	2.37	5747.6	45.8	220.90
Altura	0.74	0.68	1.79	82.2	1.11	0.9772
inicio de floración	0.55	0.42	0.26	59.1	43.4	0.1116

Puede mostrar que los marcadores evaluados pueden estar asociados con estos rasgos fenotipos de forma directa, es decir, implicado en un gen especifico para cada rasgo o de forma indirecta en la casada de genes implicados en estos rasgos.

TRAZA VARIANZA RESIDUAL



Rendimiento del grano

Con un gran número de marcadores se necesitan largas cadenas para inferir los parámetros de regularización con precisión.

Fenotipo	λ BL	λ clásico
Rendimiento del grado	45.8	220.90
Altura	1.11	0.9772
inicio de floración	43.4	0.1116

Al comprar los valores del parámetro, se percibe la influencia del método para la determinación del mismo, ya que el parámetro se determino mediante validación cruzada en Lasso clásico y el en Lasso bayesiano va a depender de la densidad previa.

No se encontraron variables con cuyo coeficientes es igual a cero ; por lo tanto se puede inferir que las variables evaluadas pueden tener una influencia sobre los rasgos estudiados, la cual puede ser aditiva, dominante o epistatica.
Sin embargo para determinar la influencia de estas sobre los rasgos, se deben realizar trabajos evaluando tanto en la descendencia como los parentales los componentes de la heredabilidad (dominancia, aditiva, epistatica), ya que de acuerdo a los encontrado en este trabajo, existe una relación (genética) de los marcadores son los rasgos.

Como se ha reportado diferentes genes influyen dentro de las características aquí evaluadas :

- Genes de semienanismo han sido ampliamente estudiados en programas de mejoramiento de cebada para reducir la altura de la planta y su resistencia al encamado.
- Genes implicados en la vernalización : condición natural física a periodos variables de frío para que se produzca la apertura de sus flores

Por lo tanto es comprensible que los marcadores tipo SNPs, pueden estar asociados con las variables, sin embargo no se logra determinar una asociación exacta, ya que no se cuentan con los datos de ubicación dentro del genoma del los marcadores, por lo que no se logra determinar su real influencia.

CONCLUSIONES

- Se logró determinar la regresión de Rigde como un método de regularización, sin embargo solo permite un primer acercamiento para la selección de variables, ya que al no obtener coeficientes nulo (iguales a cero), no se puede discriminar que marcado puede influir en los rasgos evaluados.
- Comparación de un metodología clásica (Rigde y Lasso) con una con un enfoque bayesiano, permitiendo comprender la influencia que presentan la incoporación de distribución a priori ; así mismo se mostró la capacidad del método Lasso bayesiano para determinar la capacidad predictiva del marcador así como el sesgo del modelo.

CONCLUSIONES

- De acuerdo a la estimación Lasso (clásica y bayesiana), se determinó que los SNPs estudiados pueden tener una influencia sobre los rasgos evaluados, los cuales son de gran importancia económica.

BIBLIOGRAFIA

{1} Allasian, M.B. Branco, M.E Quaglino, M.B. 2016. Regresión lasso bayesiana. Ajuste de modelos lineales penalizados mediante la asignación de priores normales con mezcla de escala. Vigesimoprimeras Jornadas "Investigaciones en la Facultad" de Ciencias Económicas y Estadística.

{2} Collard, B. C. Y., Jahufer, M. Z. Z., Brouwer, J. B., Pang, E. C. K. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker ?assisted selection for crop improvement : the basic concepts. Euphytica, 142(1 ?2), 169 ?196.

{3} MacLeod IM, Hayes BJ, Savin KW, Chamberlain a. J, McPartlan HC and Goddard ME (2010) Power of a genome scan to detect and locate quantitative trait loci in cattle using dense single nucleotide polymorphisms. J Anim Breed Genet 127 :133 ?142.

{4} Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redona, E., Atlin, G., Jannink, J.-L., y McCouch, S. R. (2015). Genomic selection and association mapping in rice (oryza sativa) : Effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLoS Genetics, 11(2) :1 ? 25.

{5} Zhang H, Wang Z, Wang S and Li H (2012) Progress of genome wide association study in domestic animals. J Anim Sci Biotechnol 3 :26.