Lasso bayesiano mejorado para seleccion genomica

Andres Legarra, Christèle Robert-Granié, Pascal Croiseau, Guillaume François, Sébastien Fritz

XV Reunion Nacional de Mejora Genetica Animal, 16-18 juin 2010, Vigo, Espagne

Lasso Bayesiano Mejorado para selección genómica

Andrés Legarra[1]*[#], Christèle Robert-Granié[1], Pascal Croiseau[2], François Guillaume[3] y Sébastien Fritz[4]

[1]INRA, UR 631 SAGA, F-31326 Castanet-Tolosan, France
[2] INRA, UMR1313 GABI, F-78352 Jouy en Josas, France
[3] Institut de l'Elevage, F-75595 Paris, France
[4] UNCEIA, F-75595 Paris, France

[#]This work has been financed by ANR project AMASGEN (2009-2011). Project partly supported by Toulouse Midi-Pyrénées bioinformatic platform. We thank G. de los Campos for carefully explaining to us the Bayesian Lasso and for his initial code in R.

Resumen. Las experiencias empíricas en selección genómica sugieren que la distribución de los efectos de los SNPs es no-normal. Una alternativa, que evita el uso de información a priori arbitraria o parámetros *ad hoc*, es el Lasso Bayesiano (BL). El BL normal utiliza un mismo parámetro de varianza para los efectos residuales o de los SNPs (BL1Var). Proponemos un BL con diferentes varianzas para los residuos y los SNPs (BL2Var), que es equivalente a la formulación original del Lasso de Tibshirani. El parámetro λ del Lasso está relacionado con la varianza genética de la población. Asimismo, proponemos la estimación previa de las varianzas individuales de cada efecto SNP mediante BL2Var, varianzas que se usan después en un modelo mixto (HetVar-GBLUP).

Los diferentes modelos se probaron en caracteres de producción lechera mediante validación cruzada, considerando 1.756 toros Holstein Francia; 1.216 fueron usados para el aprendizaje, y el resto para verificación. El número de SNP considerado fue de 51.325. Otros métodos utilizados fueron modelos mixtos para los efectos SNP, sea con varianzas estimadas de los datos, sea con varianzas fijadas a priori a partir de parámetros genéticos de la valoración.

Las estimas de variación genética en la población fueron similares a las estimas genealógicas para BL2Var pero no para BL1Var. BL1Var regresó demasiado poco los valores de cría, debido a la varianza común para residuos y efectos de SNPs. BL2Var fue en general el método más preciso, así como adecuado para la presencia de genes mayores como DGAT1, sobre todo para porcentaje de grasa. BL1Var fue el método menos preciso. HetVar-GBLUP fue casi tan preciso como BL2Var, y es fácil de calcular y de extender. **Palabras clave:** Selección genómica, todo el genoma, evaluación genética, BLUP.

1 Introducción

En selección genómica, los modelos mixtos bajo normalidad (también conocidos como "ridge regression"- o GBLUP o GWBLUP) son sencillos de entender, calcular, y extender metodológicamente [8, 1]. Sin embargo, en algunos caracteres parece haber una no-normalidad de los efectos de los SNPs, lo que resulta (y se observa en) una mayor precisión de métodos que utilizan distribuciones *a priori* más sofisticadas para los efectos de los marcadores; éstas incluyen regresión no lineal o BayesA,B [4, 9]. Al día de hoy, el conocimiento biológico de la arquitectura genética no permite concebir fácilmente estas distribuciones *a priori*.

Una de las posibles distribuciones de los efectos SNPs es el llamado "Lasso" o Lasso bayesiano (BL, de "Bayesian Lasso"; [6, 5, 2]). Este estimador permite grandes desviaciones de la normalidad, es decir, SNPs con un gran efecto sobre el carácter, así como una fuerte regresión o "shrinkage", que mejora la capacidad predictiva. El BL es muy sencillo de parametrizar (a diferencia de alternativas como el BayesB) y de calcular mediante el muestreo de Gibbs, gracias a la introducción de variables adicionales τ_i^2, que se pueden interpretar como componentes individuales de varianza del efecto de cada SNP, a_i, en el siguiente modelo jerárquico:

$$\begin{aligned} &\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e};\ p(\mathbf{a} \mid \boldsymbol{\tau}) \sim N(\mathbf{0}, \mathbf{D}\sigma^2); diag(\mathbf{D}) = (\tau_1^2 \ldots \tau_n^2); \\ &p(\boldsymbol{\tau} \mid \lambda) = \prod_i \frac{\lambda^2}{2} \exp(-\lambda^2 \tau_i^2 / 2); p(\mathbf{e} \mid \sigma^2) \sim N(\mathbf{0}, \mathbf{I}\sigma^2) \end{aligned} \qquad (1)$$

El modelo de BL en (1) –descrito en [5, 2], y que llamaremos Lasso Bayesiano de 1 Varianza (BL1Var)- tiene un inconveniente, y es que se usa un mismo parámetro, σ^2, para modelizar los efectos **a** de los SNP así como la varianza residual en **e**. Es decir, se considera de la misma manera la variación ambiental (producto de un manejo o un tipo de control lechero) que la variación de los efectos SNP, que es una característica de la población. De acuerdo a los principios aceptados de partición de la varianza en componentes ambientales y genéticos, proponemos un modelo similar:

$$\begin{aligned} &\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e};\ p(\mathbf{a} \mid \boldsymbol{\tau}) \sim N(\mathbf{0}, \mathbf{D}\sigma_a^2); diag(\mathbf{D}) = (\tau_1^2 \ldots \tau_n^2); \\ &p(\boldsymbol{\tau} \mid \lambda) = \prod_i \frac{\lambda^2}{2} \exp(-\lambda^2 \tau_i^2 / 2); p(\mathbf{e} \mid \sigma_e^2) \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2) \end{aligned} \qquad (2)$$

Con componentes específicos de varianza de los SNPs σ^2_a y residual, σ^2_e. Sin embargo, en ausencia de otra información, los parámetros λ y σ^2_a están confundidos, ya que $\mathbf{a} \mid \lambda, \sigma_a^2 \sim \prod_i \frac{\lambda}{2\sigma_a} \exp\left(\frac{-\lambda|a_i|}{\sigma_a}\right)$ y por tanto se puede fijar uno de los dos (en este trabajo se fijó $\sigma^2_a = 1$). Haciendo eso, (2) es idéntico al Lasso original de Tibshirani [6]; sin embargo nosotros retendremos la formulación jerárquica en (2). Así, λ se puede calcular tanto bayesianamente como por máxima verosimilitud. Llamaremos a este modelo Lasso Bayesiano de 2 Varianzas (BL2Var).

Atención: si conociéramos $\mathbf{D}\sigma^2$, la matriz de covarianzas de los efectos SNP, la construcción de un modelo lineal sería extremadamente sencilla, sea resolviendo explícitamente el estimador BLUP $(\mathbf{Z'Z} + \mathbf{D}\sigma_e^2 / \sigma_a^2)\hat{\mathbf{a}} = \mathbf{Z'y}$, sea con un modelo equivalente [8] usando una matriz "genómica" de covarianzas entre individuos $\mathbf{G} = \mathbf{ZDZ'}\sigma_a^2$. Ambos modelos son extensiones del llamado GBLUP.

Otro aspecto poco estudiado es la interpretación de los parámetros del Lasso o BL en un contexto de selección genómica. Asumiendo Hardy-Weinberg y equilibrio de ligamiento (es decir, en una población ideal) la varianza genética es [3]:

$$\sigma_u^2 = 2\sum_i p_i(1-p_i)\frac{2\sigma^2}{\lambda^2} \text{ para (1) y } \sigma_u^2 = 2\sum_i p_i(1-p_i)\frac{2}{\lambda^2} \text{ en (2)}. \qquad (3)$$

Por tanto, el parámetro λ^2 es inversamente proporcional a la variación genética en una población. En este trabajo:

1. Comparamos BL1Var, BL2Var, y otros métodos relacionados (GBLUP), tanto en capacidad predictiva en un juego de datos real como en interpretación de los resultados como variación genética.
2. Proponemos el uso del BL2Var para obtener varianzas individuales de SNPs, que se puedan usar luego en un estimador tipo GBLUP.

2 Material y métodos

Datos. Se analizó un conjunto de toros Holstein franceses, consistiendo en 1216 y 540 toros de "aprendizaje" y "verificación" respectivamente, genotipados para 51325 SNPs polimórficos. Los caracteres analizados fueron cantidad de leche, grasa y proteína (KL, KG, KP) y porcentajes de grasa y proteína (PG, PP), en forma de DYDs (fenotipos corregidos de sus hijas) ponderados por su precisión.

Análisis. Se usaron los modelos BL1Var y BL2Var (descritos más arriba), así como modelos mixtos con distribuciones normales de los efectos SNPs, tanto asumiendo varianzas conocidas (GBLUP; mediante la equivalencia en [3]) como integradas mediante MCMC (MCMC-GBLUP). Asimismo, utilizando BL2Var, se estimó la matriz de varianzas de SNPs, **D**, y se usó para la predicción el estimador $\left(\mathbf{Z}'\mathbf{Z}+\mathbf{D}\sigma_e^2/\sigma_a^2\right)\hat{\mathbf{a}}=\mathbf{Z}'\mathbf{y}$ (HetVar-GBLUP). La matriz **Z** se parametrizó según [8]. Las estimas de efectos SNPs en el juego de datos de aprendizaje se usaron para predecir el valor genético de los toros de validación; la estima con los DYDs de estos últimos sirvió para calcular la capacidad predictiva. En general, todos los modelos se resolvieron por MCMC; BL2Var se calculó asimismo por máxima verosimilitud marginal por MCEM.

Asimismo, se estimó la varianza genética poblacional mediante (3) y se comparó con los valores actuales en la valoración de rutina y a estimas REML basadas en genealogía con el mismo conjunto de datos.

3 Resultados

Estimas. La tabla 1 muestra las estimas de los diferentes modelos para el parámetro λ. Las estimas de los modelos BL1Var y BL2Var son muy diferentes, mientras que los dos estimadores (bayesiano o máxima verosimilitud por EM) en BL2Var son prácticamente idénticos. Curiosamente, el mayor valor de λ en BL1Var no implica mayor regresión, como se puede ver en la Figura 1, ya que la varianza σ^2 también juega un papel. A partir de los estimadores de λ y/o las diferentes varianzas según el modelo, las estimaciones de la varianza genética poblacional se presentan en la Tabla 2. Las estimaciones del modelo BL1Var son muy diferentes de cualquiera de las otras, lo que las hace inadecuadas. La razón es que se la modelización de la variación de los efectos SNP considera la varianza residual. Por tanto, a más varianza residual, más inflación de la estima de varianza genética poblacional. Sin embargo,

las estimaciones de BL2Var y MCMC-GBLUP son muy similares entre sí y razonablemente similares a las usadas actualmente o las obtenidas mediante genealogía.

TABLA 1. Estimas del parámetro λ (± error estándar)

Carácter	BL1Var	BL2Var	BL2Var - EM
KL	17.06 ±0.05	0.26 ±0.01	0.26
KG	20.60 ±0.05	6.56 ±0.20	6.51
KP	19.92 ±0.05	8.43 ±0.23	8.41
PG	15.06 ±0.05	57.20 ±1.64	55.61
PP	16.32 ±0.06	135.82 ±4.40	134.01

TABLA 2. Estimas de la varianza genética poblacional σ^2_u (± error estándar)

Carácter	BL1Var	BL2Var	MCMC-GBLUP	Pedigree REML	Valores de rutina
KL[a]	1260 ±50	448 ±27	451 ±26	570	635
KG	1876 ±84	710 ±44	710 ±39	893	973
KP	1127 ±50	429 ±24	428 ±20	473	520
PG	27.6 ±1.09	9.32 ±0.54	11.60 ±0.60	14.90	8.80
PP	5.51 ±0.03	1.66 ±0.10	1.60 ±0.12	2.56	2.19

[a] dividido por 1000

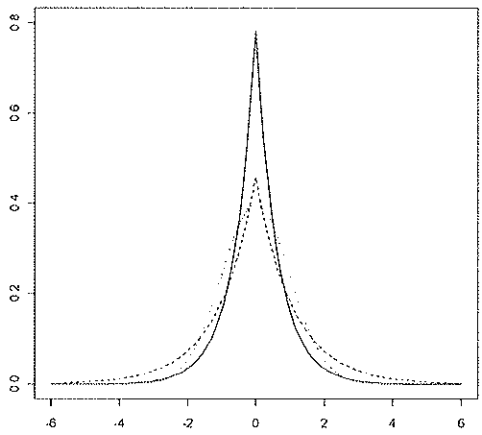

Figura 1. Distribución teórica de efectos SNP para PG según las estimas de σ^2_e , σ^2_a y λ en BL2Var (línea continua), BL1Var (línea discontinua gris), y MCMC-GBLUP (línea de puntos negra).

Capacidad predictiva. En cuanto a la capacidad predictiva, se presenta en la Tabla 3. Es sistemáticamente mejor para BL2Var, especialmente para los caracteres (porcentajes) controlados por genes mayores como DGAT1, es decir donde hay grandes desviaciones de la normalidad. BL1Var es el peor de los métodos genómicos,

debido a la restricción que impone la varianza residual en la distribución de SNPs. El método HetVar-GBLUP, que precalcula las varianzas de los SNPs, es casi tan preciso como BL2Var.

TABLA 3. Precisiones: correlación entre valores genéticos estimados y 2DYDs en los datos de validación

Carácter	BL1Var	BL2Var	GBLUP	MCMC-GBLUP	HetVar-GBLUP
KL	0.28	0.41	0.42	0.40	0.41
KG	0.35	0.37	0.34	0.37	0.36
KP	0.27	0.30	0.31	0.30	0.30
PG	0.53	0.73	0.59	0.61	0.71
PP	0.36	0.48	0.44	0.46	0.47

4 Discusión

El parámetro λ ha sido poco discutido, así como su estimación. Las estimas que obtenemos por métodos Bayesianos o de máxima verosimilitud son muy similares y precisas. Además, su interpretación como inversamente proporcional a la varianza genética poblacional facilita su comprensión. Sin embargo, la restricción en BL1Var lo fuerza a valores que no tienen mucho sentido como varianza genética.

En cuanto a la varianza genética poblacional, los valores obtenidos por el resto de métodos son razonables y comparables entre sí, incluso si las poblaciones base comparadas son diferentes: fundadores no emparentados para las estimas con genealogía, y una población con SNPs en Hardy-Weinberg y equilibrio de ligamiento en las estimas con SNPs (BL y MCMC-GBLUP).

Las precisiones obtenidas son casi siempre óptimas mediante BL2Var; sin embargo otros autores han encontrado que los métodos lineales son a veces mejores [9, 4]. Esto se puede deber al número de caracteres limitado en este estudio. Otra razón –pero esto es más especulativo- es que en este trabajo no se ha usado información a priori importante, a diferencia de métodos como BayesB, que requieren parámetros cuya información no desaparece asintóticamente [3]. En el BL, es difícil postular un "mal" prior y de hecho el parámetro λ se puede obtener por máxima verosimilitud con gran sencillez computacional. Una alternativa es el cálculo de parámetros por validación cruzada [7], pero que es laboriosa y muy dependiente de la constitución de los ficheros de aprendizaje y validación. En cuanto a BL1Var, la restricción impuesta por el modelo hace que prediga mal y no sea recomendable en la práctica.

La alternativa de precalcular las varianzas de los efectos SNP, y después usar un modelo lineal (HetVar-GBLUP), dio una buena capacidad predictiva. Esta estrategia es muy similar al uso habitual del REML + BLUP para predicción de valores de cría; si los valores de las varianzas son "estables", permitiría calcular varianzas en un subconjunto y usarlas en una población mas grande; también permite, en principio, usar una matriz de parentesco genómico con diferentes pesos para cada SNP en la valoración en un solo paso [1], con los beneficios de los métodos no lineales y usando toda la información disponible.

5 Conclusión

El Lasso Bayesiano con 2 varianzas (BL2Var), equivalente al Lasso original, es mejor para selección genómica que el BL1Var [5, 2] que no se recomienda. Los parámetros del Lasso se relacionan con la varianza genética poblacional. Un estimador lineal en dos etapas (HetVar-GBLUP) es interesante para su uso en rutina.

Bibliografía

[1] I. Aguilar, I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of holstein final score. *J Dairy Sci*, 93(2):743–752, Feb 2010.

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

[3] D. Gianola, G. de los Campos, W. G Hill, E. Manfredi, and R. Fernando. Additive genetic variability and the bayesian alphabet. *Genetics*, 183(1):347–363, Sep 2009.

[4] B. J. Hayes, P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. Invited review: Genomic selection in dairy cattle: Progress and challenges. *J Dairy Sci*, 92(2):433–443, Feb 2009.

[5] T. Park and G. Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[6] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[7] M. Graziano Usai, Mike E Goddard, and Ben J Hayes. Lasso with cross-validation for genomic selection. *Genet Res*, 91(6):427–436, Dec 2009.

[8] P. M. VanRaden. Efficient Methods to Compute Genomic Predictions. *J. Dairy Sci.*, 91(11):4414–4423, 2008.

[9] P. M. VanRaden, C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. Invited review: reliability of genomic predictions for north american holstein bulls. *J Dairy Sci*, 92(1):16–24, Jan 2009.