Revista da Estatística UFOP, Vol III(3), 2014, ISSN 2237-8111
Edição especial: 59ª Reunião Anual da Regional Brasileira da Sociedade Internacional de Biometria - RBRAS
Trabalhos apresentados na versão poster.

Capacidade preditiva como critério para determinação do número de componentes principais em Seleção Genômica Ampla

Filipe Ribeiro Formiga Teixeira[13]
Mayra Marques Bandeira[13]
Moysés Nascimento[1]
Ana Carolina Campana Nascimento[1]
Camila Ferreira Azevedo[1]
Fabyano Fonseca e Silva[2]
Paulo Sávio Lopes[2]
Simone E. F. Guimarães[2]

1 Introdução

A seleção genômica ampla (Genome Wide Selection - GWS) foi proposta por Meuwissen et al. (2001) visando aumentar a eficiência na seleção e acelerar o melhoramento genético (RESENDE et al., 2010) por meio da utilização de um grande número de marcadores amplamente distribuídos no genoma (AZEVEDO, 2012). Este método de seleção prioriza a predição simultânea dos efeitos genéticos de milhares de marcadores de DNA.

Em geral, visto que nos estudos de seleção genômica o número de SNP's é geralmente maior que o número de indivíduos genotipados e fenotipados, surgem dois problemas na estimação dos parâmetros: a alta dimensionalidade das variáveis e a multicolinearidade entre elas.

Para lidar com esse problema, métodos de redução de dimensionalidade são requeridos (AZEVEDO, 2012). Dentre eles se destacam o PCR (Regressão via componentes Componentes Principais) e o da Regressão via Quadrados Mínimos Parciais (Partial Least Squares Regression - PLSR) (SOLBERG et al., 2009). Tais métodos visam reduzir a quantidade de variáveis explicativas no modelo por meio de componentes, de modo que estes representem uma quantidade aceitável da variabilidade total contida nos marcadores.

Entretanto, para a aplicação dos referidos métodos, é necessária a definição do número de componentes a serem utilizados. A literatura apresenta alguns critérios para determinar o número "ótimo" de componentes principais e, dentre elas estes se destacam, pela facilidade de interpretação,

[1] DET - UFV. Email: filipeformiga1@live.com
[2] DZO- UFV.
[3] Agradecimentos: Fapemig, Capes e CNPq pelo apoio financeiro.

o *scree-plot* (FERREIRA, 2011) e o percentual acumulado de explicação da variabilidade dos dados.

Tais critérios se mostram eficientes no que se trata da redução de dados e explicação da variância total, porém na seleção genômica ampla, visto que temos como objetivo predizer valores a partir de um modelo de regressão, tais critérios podem não ser os mais recomendados.

Diante do exposto, este trabalho tem teve como objetivo avaliar a capacidade preditiva (correlação entre o valor real e o valor estimado pelo modelo) como critério para definição do número de componentes principais, tendo como variável resposta o peso ao abate de suínos.

2 Materiais e métodos

Os dados utilizados são provenientes da Granja de Melhoramento de Suínos do Departamento de Zootecnia da Universidade Federal de Viçosa no período de novembro de 1998 a julho de 2011, que contém uma amostra F2 de 335 suínos, onde foram coletadas informações sobre o sexo, lote, presença ou ausência de halotano, peso ao abate (variável dependente resposta) e 237 marcadores SNP's em todo o seu DNA.

Os dados fenotípicos foram corrigidos para efeitos fixos, ou seja, a variável utilizada no ajuste é dada pelo resíduo do ajuste da regressão tendo como variável resposta o peso ao abate dos animais, e como variáveis independentes o sexo, lote e a presença ou ausência de halotano.

Posteriormente, como critério para avaliação para o número de componentes principais, utilizaremos a correlação entre os valores reais corrigidos e os valores estimados (capacidade preditiva). Uma alta correlação nos indica que a redução no número de variáveis é viável, visto uma vez que indicará uma maior semelhança entre esses os valores observados e estimados. Como são 237 marcadores, serão realizadas estimadas 237 análises de regressão, tomando-se de 1 a 237 componentes e calculadas as correlações de acordo com o número de componentes.

A regressão via componentes principais considera cada componente encontrado como variável independente, e a variável resposta será é estimada de acordo com esses componentes. O modelo é dado por:

$$= Z\alpha + e$$

Em que $Z = XP, \alpha = P^T \beta e P^T X^T XP = \Lambda^1$.

As colunas de Z, que definem o novo conjunto de regressores ortogonais, são referidas como componentes principais. O estimador de mínimos quadrados de α é dado por:

$$\hat{\alpha} = (Z^T Z)^{-1} Z^T \quad = \Lambda^{-1} Z^T$$

E a matriz de covariância é dada por:

$Cov(\hat{\alpha}) = \sigma^2 (Z^T Z)^{-1} = \sigma^2 \Lambda^{-1}$

Assim, pequenos valores de λ_j indicam que a variância do coeficiente de regressão ortogonal associado será grande. Da estatística multivariada sabemos que a variância associada a cada componente é dada pelo autovalor λ_j.

O software utilizado na análise foi o R, onde foram criadas rotinas para realização a estimação dos modelos de regressões regressão, e para encontrarmos as correlações e as variâncias explicadas de acordo com cada número de componentes. Os pacotes utilizados foram "MASS" e "pls".

3 Resultados e Discussão

Utilizando a abordagem usual, isto é, considerando o número de componentes encontrado necessários para explicar pelo menos 70% da variabilidade total dos dados, poderíamos trabalhar com 37 componentes ao invés de 237 marcadores, o que seria uma redução considerável da dimensionalidade dos dados (Tabela 1).

Para avaliar a viabilidade da metodologia proposta, foram ajustados 237 modelos de regressão, de acordo com o número de componentes utilizados. Para cada ajuste calculou-se a correlação entre os valores observados e os valores estimados (capacidade preditiva), no intuito de verificar se esta correlação é um bom critério para determinação do número de componentes principais.

Tabela 1. Variabilidade explicada e correlação de acordo com o número de componentes

Variabilidade Explicada (%)	Número de Componentes	Corr $(\hat{Y}, Y)$
70	37	0,35
80	54	0,41
90	86	0,55
100	237	0,85

Podemos notar que ao utilizando utilizar os 37 componentes principais, encontramos um valor baixo para a correlação (0,35), o que indica que os valores estimados com base no modelo ajustado não estão próximos aos valores observados. Observa-se também que quanto maior o número de componentes, maior será essa correlação, ou seja, para a predição do peso ao abate é mais interessante usarmos todas as variáveis, visto que o número de componentes é diretamente proporcional à capacidade preditiva.

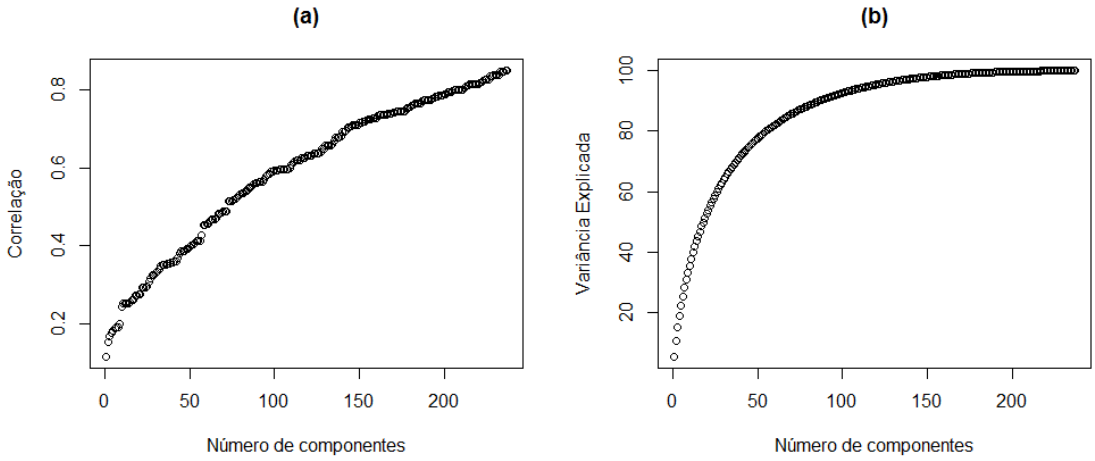

Figura 1: (a) correlação de acordo com o número de componentes; (b) variabilidade total explicada de acordo com o número de componentes.

Graficamente, podemos identificar um ponto onde a variância explicada se estabiliza por volta de 75 componentes, de acordo com o aumento do número de componentes principais (Figura 1b). Usando como critério a capacidade preditiva, observa-se um aumento quase linear, indicando que não é possível identificar um ponto "ótimo", onde a correlação se estabilize (Figura 1a).

4 Conclusão

A capacidade preditiva não se apresenta como um critério para avaliar o número de componentes retidos no modelo. Porém, novas pesquisas e estudos devem ser realizados para comprovar esse resultado de uma maneira mais abrangente, visto que neste estudo foi utilizada apenas uma variável (PA).

5 Referências Bibliográficas

[1] RESENDE, M. D. V.; JUNIOR, M. F. R. R.; AGUIAR, A. M.; ABAD, J. I. M.; MISSIAGGIA, A. A.; SANSALONI, C.; PETROLI, C. GRATTAPAGLIA, D.; **Computação da Seleção Genômica Ampla (GWS)**, Colombo-PRdezembro de 2010.

[2] MEUWISSEN, T.H.E. et al. **Prediction of total genetic value using genome wide dense marker maps**. Genetics, v.157, p.1819-1829, 2001.

[3] AZEVEDO, C. F., **Métodos de redução de dimensionalidade aplicados na seleção genômica para característica de carcaça em suínos**. 2012. 50 p. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa, Viçosa.

[4] SOLBERG, T. R.; SONESSON A. K.; WOOLLIAMS, J. A.; MEUWISSEN, T. H. E. Reducing dimensionality for prediction of genome-wide breeding values. **Genetics Selections Evolution**, v. 41, n. 29, 2009.

[5] FERREIRA, D. F., **Estatística Multivariada**. 2ª edição. Lavras: Ed. UFLA, 2011.