FORMAÇÃO DE GRUPOS PRODUTIVOS EM VACAS LEITEIRAS POR MEIO DE COMPONENTES PRINCIPAIS

Eucymara França Nunes SANTOS[1]
Kleber Régis SANTORO[1]
Rinaldo Luiz Caraciolo FERREIRA[2]
Eufrázio de Souza SANTOS[1]
Gladston Rafael de Arruda SANTOS[3]

- RESUMO: Os objetivos deste trabalho foram utilizar dados referentes à produção de leite de três grupos geneticamente divergentes, com o intuito de visualizar a separação destes grupos por meio de gráficos, e eliminar as variáveis menos importante sem muita perda de informação, através da análise de componentes principais. A identificação dos grupos auxilia na forma como o manejo é aplicado aos animais, tal como o tipo de ração que recebem, a ordenha, as instalações, o tipo de reprodução. Animais de um mesmo grupo necessitam de manejo semelhante. Na identificação dos grupos, as variáveis características para a produção de leite analisadas foram: grupo genético, peso do leite (kg) produzido no dia do controle, peso do leite (kg) produzido na primeira, segunda e terceira ordenhas, idade da vaca (dias) ao parto, idade da vaca (dias) na data do controle leiteiro e intervalo de partos (dias). As análises foram realizadas com os dados estandardizados, devido as diferentes medidas experimentais. Tal estudo proporcionou a obtenção de três componentes com a explicação de 92,84% da variabilidade dos dados. A técnica permitiu uma redução de cinco variáveis não significativas na dimensionalidade dos dados e apresentou o melhor gráfico de separação dos grupos genéticos.

- PALAVRAS-CHAVE: Componentes principais; grupos genéticos; vacas leiteiras.

1 Introdução

A estatística multivariada é um conjunto de técnicas exploratórias de sintetização (ou simplificação) da estrutura de variabilidade dos dados, consistindo em estudar diversas variáveis simultaneamente. Dentre as técnicas de análise multivariada, está a de componentes principais (Principal Components Analysis – PCA), que consiste em transformar um conjunto de variáveis originais em um pequeno número de combinações lineares, os chamados componentes principais, de dimensões equivalentes. O objetivo é

1 Universidade Federal Rural de Pernambuco – UFRPE, Departamento de Estatística e Informática, CEP: 52.171-900, Recife, Pernambuco, Brasil. E-mail: *eucymara@gmail.com / krsantoro@uag.ufrpe.br*, *eufrazio@deinfo.ufrpe.br*
2 Universidade Federal Rural de Pernambuco - UFRPE, Departamento de Ciência Florestal, CEP: 52.171-900, Recife, Pernambuco, Brasil. E-mail: *rinaldo@dcfl.ufrpe.br*
3 Empresa Pernambucana de Pesquisa Agropecuária – IPA, CEP:56.600-000 , Sertânia, Pernambuco, Brasil. E-*mail: gladstonrafael@ipa.br*

obter variáveis que retenham o máximo possível de informações e expliquem a maior parte da variabilidade total, revelando que tipo de relacionamento existe entre eles (REIS, 2001).

Em se tratando de um conjunto de variáveis, a análise univariada pode ser incompleta, só a multivariada avalia simultaneamente um conjunto de características, levando em consideração as correlações existentes. Esta é mais eficiente e proporciona enriquecimento das informações extraídas de um conjunto de dados experimentais.

Este método é utilizado para identificar o fator dimensão dos dados: a redução da dimensão fornece gráficos para um estudo por meio dos escores dos componentes, nos quais se pode observar a formação de grupos. Portanto a representação gráfica dos componentes principais são ferramentas valiosas na explanação da análise de dados (DILLON e GOLDSTEIN, 1984).

Em estudos de divergência genética, o estudo de componentes principais contribui na interpretação das relações existentes entre as variáveis, propõe relativa economia de tempo e custo em experimentos futuros descartando variáveis redundantes sem perda considerável de informação: auxilia na tomada de decisões e pode levar a aceleração de um programa de melhoramento genético (MORRISON, 1976).

A utilização desta técnica para identificar grupos produtivos geneticamente divergentes, está diretamente associada à produção animal, ao que os profissionais chamam de "manejo", forma com que os animais são tratados. Conhecer as variáveis que afetam a produção e conseguir separar os grupos produtivos auxilia na forma como o manejo é aplicado aos animais, terão tratamento diferenciado, de acordo com as qualidades de cada grupo.

Os objetivos deste trabalho foram observar a possibilidade de formação de diferentes grupos de animais, por meio da investigação de componentes principais, utilizando os dados referentes à produção de leite de três variados grupos, geneticamente diferentes, verificarem a possibilidade de redução no número de variáveis envolvidas na análise e melhorar a compreensão da dispersão dos dados.

2 Análise de componentes principais

A técnica de componentes principais foi originalmente desenvolvida por Karl Pearson (1901) e, posteriormente, aplicada por Harold Hotteling (1933) em diversas áreas da ciência (CHATFIELD e COLLINS ,1980).

As medidas das variáveis originais são correlacionadas entre si, indicando que algumas informações contidas em uma variável também estão em outra. Então o objetivo da análise de componentes principais é transformar a quantidade de variáveis correlacionadas em uma quantidade de variáveis não correlacionadas, ou seja, os componentes principais (Cruz e Regazzi, 1997). Os melhores resultados são obtidos quando as variáveis originais são altamente correlacionadas, positiva ou negativamente.

Seja um conjunto de variáveis observadas X_1, X_2, ... , X_j o componente principal é obtido pela combinação linear.

$$Y_i = a_{i1} X_1 + a_{i2} X_2 + ... + a_{ij} X_j, \quad (1)$$

a partir da matriz de correlação (R) ou variância-covariância (S). A solução é obtida através da equação característica

$$|R - \lambda I| = 0 \quad (2)$$

sendo I a matriz identidade.

Para cada autovalor (λ) determina-se um autovetor a_i:

$$[R - \lambda I]a_i = \phi \quad (3)$$

onde φ é um vetor nulo, de dimensão i x 1.

Cada autovalor representa a variância de cada componente. A interpretação relativa de cada componente principal Y_i é então avaliada pela porcentagem da variância total que ele explica. A variância total é definida pela soma das variâncias de cada uma das variáveis. As variâncias individuais constituem os elementos da diagonal principal (traço da matriz), somando-os encontra-se-á a variância total.

Os autovetores são determinados pelos coeficientes das equações lineares de cada componente principal. Substituindo-se as variáveis originais nas equações dos componentes, encontra-se-ão os escores de cada componente.

A importância da variância dos componentes está na sua organização em ordem decrescente. Os últimos componentes são responsáveis por uma fração muito pequena da variabilidade dos dados. Segundo Johnson e Wichern (1998), 80 ou 90% da variabilidade total pode ser explicada pelos primeiros componentes, que poderão ser usados no lugar das variáveis originais sem perder muita informação.

O critério para descarte de variáveis, segundo Jolliffe (1972), baseado em dados simulados e reais, recomenda que quando a análise de componentes principais utiliza a matriz de correlação, estabelece-se que o número de variáveis descartadas deve ser igual ao número de componentes cuja variância (autovalor) é inferior a 0,7.

Os primeiros componentes conta com grandes proporções do total da variância: então os últimos componentes podem ser desconsiderados por correspoderem a pouca variância dos dados. O descarte sugerido por Mardia et al. (1979) e Morrison (1976), indica que a variável que possuir maior coeficiente no componente principal de menor autovalor (variância), deve ser menos importante para explicar a variância total, portanto, passível de descarte.

3 Materiais e métodos

Neste trabalho foram utilizados dados provenientes de cruzamentos entre animais Holandês (HO) e Gir (GL) de vacas leiteiras de três diferentes grupos genéticos: 1/2 HO, GL; 3/4 HO, GL e 7/8 HO,GL, com 326 animais por categoria. Os dados foram coletados semanalmente, no período de dezembro/2000 a dezembro/2006 numa fazenda com agropecuária semi-intensiva de leite, na região de Ribeirão, Zona da Mata de Pernambuco.

Foram analisadas as seguintes características da produção de leite: grupo genético (GS), peso do leite (Kg) produzido no dia do controle (PL), peso do leite (Kg) produzido na primeira ordenha (PL1), peso do leite (Kg) produzido na segunda ordenha (PL2), peso do leite (Kg) produzido na terceira ordenha (PL3), idade da vaca (dias) na data do controle (IPL), idade da vaca (dias) ao parto (IP) e intervalo de partos (IEP), com 13.643, 13.643, 13.600, 12.128, 9.643, 13.643, 13.643 e 3.119 observações por variável, respectivamente.

Os dados foram examinados por meio da análise de componentes principais. Como

as variáveis observadas possuíam diferentes unidades de medidas, então as componentes principais podem ser acompanhadas de unidade de medidas sem sentido, neste caso é necessário trabalhar com as variáveis estandardizadas, eliminando as diferenças nas dimensões e na variação das variáveis.

Para análise dos componentes principais utilizou-se o procedimento PRINCOMP, do software SAS (SAS INSTITUTE, 2002).

4 Resultados e discussão

Pode-se observar a matriz de correlação obtida através da análise de componentes principais na Tabela 1, que houve alta correlação positiva entre as variáveis peso do leite no dia do controle em relação ao peso do leite da primeira, segunda e terceira ordenhas, e entre a idade da vaca no dia do controle e a idade da vaca no dia do parto. A correlação entre a variável grupo genético em relação as demais variáveis foi baixa, sendo que o intervalo de partos foi a que obteve uma melhor correlação dentre elas. A variável intervalo de partos obteve baixa correlação com as características do peso do leite. McManus et al. (2002) encontraram médias/altas correlações entre a variável fertilidade real, com os pesos da vaca e alta com intervalo de partos. Para explicar a distribuição dos grupos, não será necessário um grande número de componentes, por possuir altas correlações entre as variáveis, podendo variar de acordo com a população estudada (MANLY, 2004).

Tabela 1 - Matriz de correlação entre as variáveis utilizadas na análise de componentes principais

Variável	Variável							
	GS	PL	PL1	PL2	PL3	IPL	IP	IEP
GS	1,00							
PL	0,22	1,00						
PL1	0,04	0,72	1,00					
PL2	0,14	0,90	0,81	1.00				
PL3	0,15	0,91	0,81	0,88	1,00			
IPL	-0,14	0,03	0,02	0,01	-0,00	1,00		
IP	-0,16	0,10	0,10	0,10	0,08	0,98	1,00	
IEP	-0,11	0,24	0,20	0,22	0,21	0,88	0,89	1,00

Conforme pode-se observar na Tabela 2, os três primeiros componentes principais explicaram 92% da variação total, cinco componentes (62,5%) apresentaram variância (autovalor) inferior a 0,7. Assim, as cinco variáveis podem ser descartadas de acordo com o critério de Jolliffe (1972).

Barbosa et al. (2006) sugeriram a redução de dez para três componentes principais por apresentarem variância inferior a 0,7, sugerido por Jolliffe (1972). Essas contribuíram com 60,65% da variação total das características da qualidade da carne de suínos.

Leite et al. (2009), ao avaliarem as características de carcaça de codorna, os quatro dos onze componentes principais explicaram 75% da variação total, e encontraram sete variáveis redundantes, podendo ser descartadas, segundo critério de Morrison (1976).

Tabela 2 - Autovalores, proporção individual e acumulada da variação dos dados por meio da análise dos componentes principais

Componente	Autovalor	Proporção Individual	Proporção Acumulada
1	3.74	0.46	0.46
2	2.74	0.34	0.81
3	0.94	0.11	0.92
4	0.27	0.03	0.96
5	0.11	0.01	0.97
6	0.11	0.01	0.99
7	0.05	0.00	0.99
8	0.01	0.00	1,00

A Figura 1 mostra que a partir do terceiro componente a variância dos próximos componentes torna-se insignificante.

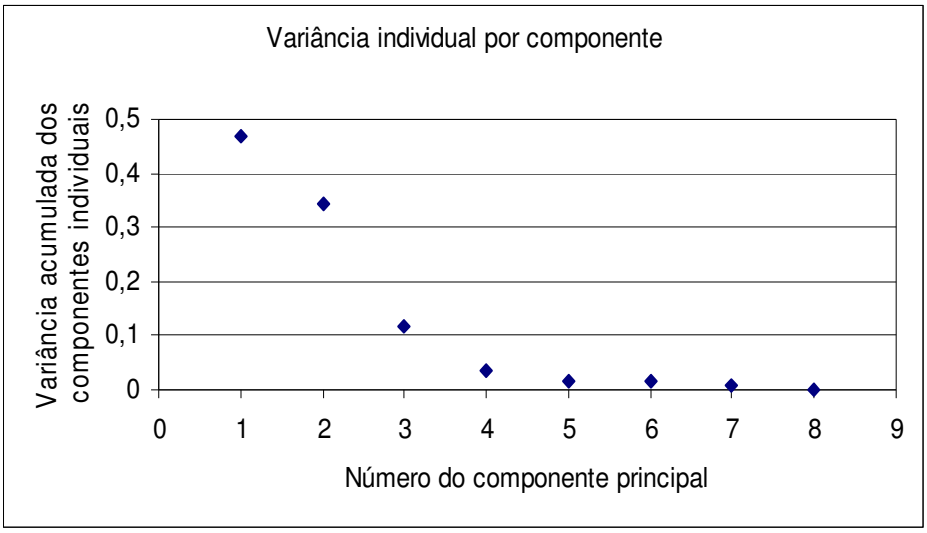

Figura 1 - Variância individual por componente.

As variáveis passíveis de descarte podem ser vistas na Tabela 3, aquelas que possuem maior correlação dos últimos cinco componentes, como são o caso das variáveis em ordem crescente de importância: PL1, IEP, PL2, PL e IP: as demais, devem ser mantidas. Com menos variáveis a serem analisadas, poupa-se relativa economia de tempo e custo, na tomada de novas medidas e em análises futuras.

As variáveis passíveis de descarte apresentam correlação linear simples significativa: (PL1, PL2 e PL) com a variável selecionada PL3 e (IEP e IP) com a variável selecionada IPL, isto é podem ser responsáveis por algum tipo de informação exclusiva.

Tabela 3 - Autovetores (coeficiente de ponderação) e suas correlações (em percentagem)

	Prin4	Prin5	Prin6	Prin7	Prin8
GS	0,16(8,82)	-0,00(-0,31)	-0,01(-0,36)	0,04(1,03)	-0,02(0,24)
PL	-0,47(-24,94)	0,08(2,75)	0,12(4,09)	**-0,70(-7,09)**	-0,00(-0,09)
PL1	**0,82(43,17)**	-0,04(-1,37)	0,04(1,45)	-0,24(-5,92)	0,01(0,20)
PL2	-0,14(-7,36)	0,27(9,33)	**-0,72(-24,25)**	0,35(8,60)	0,03(0,33)
PL3	-0,17(-8,97)	-0,16(-5,50)	0,60(20,44)	0,55(13,55)	0,03(0,33)
IPL	0,04(2,15)	0,36(12,51)	0,12(4,12)	0,00(0,14)	0,70(7,63)
IP	0,05(2,64)	0,37(12,70)	0,11(4,00)	0,04(1,13)	**-0,70(-7,66)**
IEP	-0,06(-3,31)	**-0,78(-26,96)**	-0,25(-8,41)	-0,02(-0,64)	-0,00(-0,04)

* Valor da correlação entre parênteses

Se os dois primeiros componentes principais explicam uma grande proporção da variação total, entre 70 a 80%, então isto será útil para plotar os valores dos escores dos componentes para cada individuo, em outras palavras possibilita plotar os dados em duas dimensões, então pode-se observar formação de grupos.

Observando os três primeiros componentes principais, o gráfico de dispersão mostrado na Figura 2, foi obtido por intermédio dos escores dos componentes principais 2 e 3. Foi a que melhor representou a separação dos grupos genéticos produtivos.

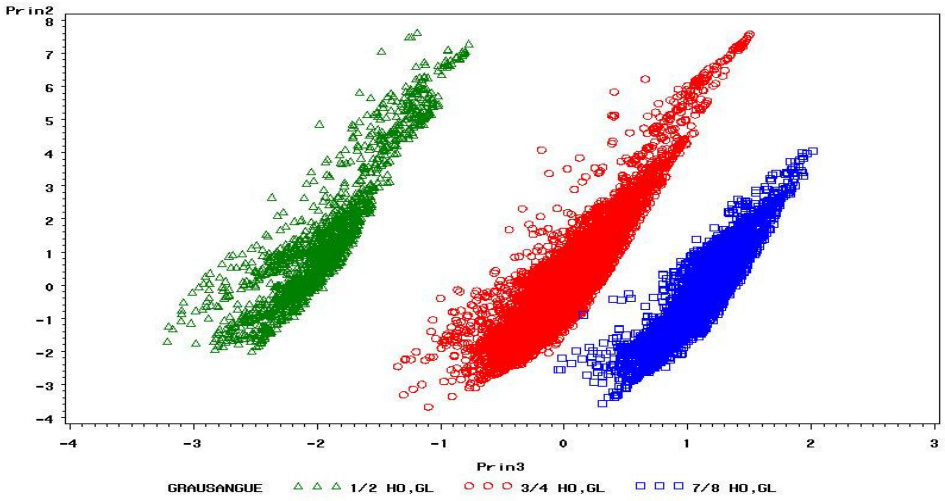

Figura 2 - Separação entre os três grupos genéticos produtivos através da dispersão entre os escores do segundo (Prin2) e do terceiro (Prin3) componentes principais.

Conclusões

A investigação de componentes principais permitiu um melhor entendimento dos dados, tornando-se útil na utilização de técnicas subsequentes, assim como a análise de agrupamentos e a discriminante.

É recomendado estender o estudo para mais propriedades, para verificar a formação dos mesmos ou de outros grupos. Isto pode levar a considerar que animais de um determinado tipo podem ter o mesmo manejo. Este processo ajudaria na diminuição de custos, otimização de mão-de-obra e recursos de infra-estrutura, enfim, ajuda a todo planejamento da propriedade e também do manejo dos animais.

Agradecimentos

À UFRPE (Propesquisador 2003/2005) por proporcionar boas condições de trabalho, e ao proprietário da fazenda, pela colaboração e cessão dos dados.

SANTOS, E. F. N.; SANTORO, K. R.; FERREIRA, R. L. C.; SANTOS, E. S.; SANTOS, G. R. Formation of productive genetic groups in dairy cows through principal components. *Rev. Bras. Biom.*, São Paulo, v.28, n.3, p.15-22, 2010.

- *ABSTRACT: This work objectives went to utilize data regarding production to the of milkfrom three genetically divergents groups, with the intention of visualizing the separation of these groups through graphs and eliminating less important variables without a lot of loss of information ,through the principal components analysis The identification of the groups in the form as the management it is applied the animals, just as the ration type that it received, it milks, the installation, the reproduction type. Animals of a same group need similar management. The characteristic variables for the production of milk analyzed were: group genetic, weigh of the milk (kg) produced in the day of the control, weight of the milk (kg) produced in the first it milks, weigh of the milk (kg) produced in the second it milks, weigh of the milk (kg) produced in the third it milks, age of the cow (days) in the date of the control, age of the cow to the childbirth and interval of childbirths. The analysis were accomplished with the original data of collection and with the standardized data, due the different measures experimentals. The analysis provided the aquisition of three components with the explanation of 92,84% of the variability of the data. The technique permited theelimination of five non significant variables and presented the best separation graph of the genetic group, with the second and third scores of the principal components.*

- *KEYWORDS: Principal components; genetic groups; cows mil.*

Referências

BARBOSA, L.; LOPES. P. S.; REGAZZI, A. J.; GUIMARÃES, S. E. F.; TORRES, R. de. A. Avaliação de características de qualidade da carne de suínos por meio de componentes principais. *Rev. Bras. Zootec.*, Viçosa, v. 35, n. 4, p. 1639 – 1645, 2006 (supl.)

CHATFIELD, A.; COLLINS, A. J. *Introduction to multivariate analysis*. London: Chapman & Hall/CRC, 1980. 246p.

CRUZ, C. D.; REGAZZI, A. J. *Modelos biométricos aplicados ao melhoramento genético*. 2.ed. Viçosa: Editora UFV, 1997. 390p.

DILLON, W. R.; GOLDSTEIN, M. *Multivariate analysis:* Methods and applications. New York: John Wiley e Sons, 1984. 608p.

JOHNSON, R. A.; WICHERN, D. W. *Applied multivariate statistical analysis*. 4 ed. New Jersey: Pratice Hall, 1998. 816p.

JOLLIFFE, I. T. Discarding variables in a principal component analysis. I. Artificial data. *J. R. Stat., Soc., Serie C: Appl. Stat.*, London, v.21, n.2, p.160-173, 1972.

LEITE, C. D. S.; CORRÊA, G. S. S.; BARBOSA, L.; MELO, A. L.; YAMAKI, M.; SILVA, M. A.; TORRES, R. A. Avaliação de características de desempenho e de carcaça de codornas de corte por meio de análise de componentes principais. *Rev. Bras. Med. Vet. Zootec.*, Belo Horizonte, v.61, n. 2, p.498-503, 2009.

MACMANUS, C.; SAUERESSING, M. G.; FALCÃO, R. A.; SERRANO, G.; MARCELINO, K. R. A.; PALUDO, G. R. Componentes reprodutivos no rebanho de corte da Embrapa Cerrados. *Ver. Bras. Zootec.*, Viçosa, v.31, n.2, p. 648-657, 2002.

MANLY, B. F. J. *Multivariate statistical methods a primer*. 3 ed. London: Chapman & Hall/CRC, 2004. 208p.

MARDIA, A. K. V.; KENT, J. T.; BIBBLY. J. M. *Multivariate analysis*. London: Academic Press, 1997. 518p.

MORRISON, D. F. *Multivariate statistical methods*. 2.ed. New York: McGraw-Hill Company, 1976. 415p.

REIS, E. *Estatística multivariada aplicada*. 2.ed. Lisboa: Sílabo, 2001. 253 p.

SAS INSTITUTE. *SAS Stat user's guide. Version 9*. Carry: SAS Institute, 2002. 1 CD-ROM.

Recebido em 20.01.2010.

Aprovado após revisão 02.08.2010.