

METHODOLOGY

GUIA PRÁTICO PARA ANÁLISE DE DADOS SEGUNDO MODELO ANIMAL EM DFREML COM MATRIZES ESPARSAS. I. ANÁLISES UNIVARIADAS (Practical Guidelines for Analyzing Data Under an Animal Model with DFREML Using Sparse Matrix Solver. I. Univariate Analysis)

J.B.S. Ferraz*

ABSTRACT

The best estimation of variance components and genetic parameters in the unbalanced data usually used in animal breeding is that obtained by restricted maximum likelihood (REML) procedures under an animal model that consider all the pedigree information available; a model that is as close as possible to the biological situation. The use of a derivative-free algorithm, where the inversion of the big matrices generated in animal models is not needed and the use of a sparse matrix solver - SPARSPAK- resulted in a software that solves the complex mixed model equations in a shorter time and using less computer resources than the original version of DFREML. This paper presents the practical guidelines for animal breeders that need to use this package but are not familiar with FORTRAN or the previous version of DFREML.

INTRODUÇÃO

A melhor predição linear não viciada (BLUP, Henderson, 1949; Henderson *et al.*, 1959) sob modelo animal, tem se tornado rapidamente no método de escolha pelos responsáveis pela avaliação de animais (Henderson, 1988). A estimação de componentes

Departamento de Produção Animal, Faculdade de Medicina Veterinária e Zootecnia, USP, Caixa Postal 23, 13630 Pirassununga, SP, Brasil.

* Present address: University of Nebraska, Dept. of Animal Science, Lincoln, NE 68583-0908, USA.

de variância e covariância por procedimentos de máxima verossimilhança restrita (REML) é geralmente considerado o melhor método para os dados não balanceados geralmente utilizados em melhoramento animal (Boldamn e Van Vleck, 1991).

O termo "modelo animal" foi provavelmente criado por Quaas e Pollak (1980) e representa atualmente uma série de muitos modelos diferentes que têm em comum que todos os animais, com ou sem registros de produção, animais com repetidos registros de produção, efeitos de ninhada, efeitos ambientais fixos ou aleatórios são avaliados conjuntamente na população de interesse. Os modelos animais utilizam, de forma direta ou indireta, a matriz de parentesco completa, com todas as relações aditivas entre todos os possíveis animais com informações disponíveis e seu objetivo principal chegar-se a um modelo que descreva a real situação biológica da maneira mais próxima possível, para predição, em melhoramento animal, dos valores genéticos aditivos (breeding values) e estimação de parâmetros genéticos como herdabilidade (Kennedy *et al.*, 1988).

Importantes revisões e discussões sobre as bases teóricas e genéticas do modelo animal ocorreram em um simpósio realizado em Edmonton, Alberta, Canada, cujos Proceedings foram publicados pelo Journal of Dairy Science em 1988 (Henderson, 1988; Kennedy *et al.*, 1988; Meyer, 1988a entre outros).

Os problemas computacionais que ocorrem com o uso de estimações por métodos REML em melhoramento animal estão ligados à inversão de matrizes e primeiras e segundas derivadas necessárias à avaliação da maximização (ou minimização) de funções, o que leva à utilização de enormes recursos computacionais (tempo e memória). Um algoritmo livre de derivadas foi proposto por Graser *et al.* (1987) para análise univariada em modelo animal onde o animal era o único efeito aleatório. Este algoritmo baseou-se na repetida utilização da eliminação Gaussiana nas equações de modelo misto em conjunção com técnicas de matrizes esparsas para avaliar o log da função de verossimilhança. Esta aproximação livre de derivadas foi utilizada por Meyer (1988a,b, 1989 e 1991) para desenvolver DFREML, um sistema de programas, originalmente escrito em FORTRAN, especialmente projetado para estimar componentes de variância em modelos mistos.

Utilizando-se de diferentes técnicas de solução para matrizes esparsas, o software SPARSPAK (George *et al.*, 1980) e fatoração de Cholesky, Boldman e Van Vleck (1991) adaptaram a versão original de DFREML de tal forma que este novo software passou a necessitar de muito menos recursos computacionais (este novo sistema é de 100 a 500 vezes mais rápido e utiliza cerca de 1/5 dos recursos necessários à versão original).

A maioria das informações necessárias para utilizar-se do sistema DFREML e de sua versão com soluções por SPARSPAK estão contidas na documentação interna dos próprios programas e nos trabalhos publicados pelos autores (Meyer, 1988 a,b e 1989; Boldman e Van Vleck, 1991), mas este trabalho foi elaborado para os usuários de

DFREML não familiarizados com FORTRAN e teve como objetivo tentar esclarecer as etapas necessárias para analisar dados com este poderoso sistema. Os passos descritos foram desenvolvidos em computador de grande porte, IBM modelo 4381, operando sob sistema operacional CMS (Copyright IBM Corporation).

MATERIAIS E MÉTODOS

Qualquer tipo de dados usualmente manipulados pelos geneticistas animais, seja de múltiparos ou uníparos, pode ser analisado sob modelo animal por DFREML, sem ser afetado pela estrutura dos acasalamentos (se inteiramente casualizados, hierarquizados ou de classificação cruzada), pois a matriz de parentesco e relações genéticas aditivas considera tais estruturas. Embora este sistema tenha sido desenvolvido para estimar componentes de variância, ele pode ser utilizado, desde que modificado para informar o vetor de soluções, para estimar efeitos fixos ou covariáveis.

São necessários dois tipos de arquivo para rodar DFREML: o **arquivo de pedigree**, onde encontram-se as informações sobre a genealogia dos animais e o **arquivo de registros**. Eventualmente o **arquivo de pedigree** pode fazer parte do **arquivo de registros**, mas deve ser lido separadamente.

As características mais importantes do **arquivo de pedigree** são:

- a ordem de codificação do arquivo de pedigree deve ser animal, pai e mãe;
- todos os animais, com ou sem registros, devem constar deste arquivo. Adicione tantas informações quantas sejam possível, pois as mesmas tornarão mais completa a matriz de parentesco e relações aditivas;
- todos os animais devem ser codificados com números maiores que o de seus pais;
- se possível, separe as informações com espaços vazios (brancos), caso contrário o arquivo dever ser lido como formatado;
- é possível, logo após o campo de identificação da mãe, colocar-se uma identificação de efeito fixo (como raça ou sexo), de tal forma que o programa possa calcular coeficientes de endogamia (F) médios para cada subclasse daquele efeito fixo.

As características mais importantes do **arquivo de registros**, que contém todas as informações sobre a identificação do animal que tem registros (tal identificação, numérica, deve ser a mesma contida no **arquivo de pedigree**), efeitos fixos, covariáveis e os registros observados para todas as características e todos os animais são:

- Se possível, separe as informações com espaços vazios (brancos), caso contrário o arquivo deverá ser lido como formatado;
- a sequência correta para entrada de dados é:

- 1-Identificação do animal, pai e mãe;
- 2-Identificação dos efeitos fixos;
- 3-Identificação do efeito de ambiente permanente (animal, ninhada ou mãe);
- 4- Covariáveis e
- 5- As observações dos animais, ou seja, as variáveis dependentes.

DFREML, em sua versão conjunta com SPARSPAK, utiliza para a maioria dos problemas os seguintes programas em FORTRAN:

- BINTRAN
- DFNRM
- DFCODE
- DFMME1
- DFUNIS, subdividido em DFMAIN, DFSUBSPO e IDIAGST
- SPARSPAK

As subrotinas usadas por estes programas estão adicionadas ao fim de cada um, de tal forma que são compiladas conjuntamente.

Adicionalmente, são necessários alguns arquivos executáveis pelo sistema operacional, que informam ao computador as definições dos arquivos de entrada e saída de cada programa, compilam, carregam e rodam os programas em FORTRAN. Estes arquivos executáveis são escritos na linguagem do sistema operacional do computador e são:

- BINTRANF
- DFCODEF
- DFMME1F
- DFSPARSF ou DFUNIS (na versão original de DFREML)

Quando a análise é feita em BATCH, mais um arquivo é necessário, DF5 DATA, que responde às perguntas interativas de DFUNIS. Detalhes adicionais deste arquivo são dados adiante.

1) Como construir a matriz A-1 (matriz de parentesco e relações aditivas)

1.a) Rodando BINTRAN

BINTRAN é um programa em FORTRAN que lê os arquivos e os reescreve em linguagem binária, que é acessada mais rapidamente pelos computadores. É um programa

muito simples e poderoso, que pode editar dados, ler arquivos formatados ou em formato livre.

Em sua versão BINDUP, ele elimina duplicatas. Isto é muito útil quando não se sabe ao certo se existem animais com registros em duplicata. Neste caso, o arquivo deve estar ordenado segundo a variável que vai ser verificada para eliminação de duplicatas. Os Anexos 1 e 2 apresentam, BINTRANF EXEC, BINTRAN FORTRAN, BINDUP e BINTRAN adaptado para não ler observações iguais a zero, que pode ser usado em casos de observações perdidas.

Para rodar BINTRAN:

- edite o arquivo executável, BINTRANF EXEC para ter certeza dos nomes dos arquivos de entrada e saída;

- compile BINTRAN FORTRAN, utilizando-se do comando FORTVS BINTRAN (opt(3), que dá otimização nível 3 compilação;

- rode o arquivo executável (BINTRANF), depois de compilar BINTRAN; verifique a definição dos arquivos de entrada e saída, através do comando QFI (query files). O arquivo de entrada deve ter o mesmo nome do **arquivo de pedigree**.

- carregue e rode Bintran no computador, digitando o comando "LOAD BINTRAN (CLEAR START " (sem fechar o parênteses). Responda às questões interativas. Responda "3" à pergunta sobre o número de variáveis "integer"(inteiras), pois refere-se ao animal, pai e mãe. Caso algum efeito fixo tenha sido especificado e o nível de endogamia dentro de cada um seja desejado, digite "4" para o número de variáveis inteiras. Todas as variáveis de identificação são definidas como variáveis inteiras, sem decimais. Digite "0" para informar que nenhuma variável "real" (variável fracionária, com decimais) vai ser lida. BINTRAN mostrará na tela do computador as primeiras 5 e a última observações. Verifique se a leitura está correta.

1.b) Rodando DFNRM:

Depois de ler o **arquivo de pedigree** na forma binária com BINTRAN, o programa DFNRM vai montar a inversa da matriz de parentesco e relações aditivas (A-1). Este programa pode ser modificado para imprimir os elementos da diagonal da matriz, que informam o valor $1 + F$ (coeficiente de endogamia) para cada animal, inclusive os "animais base", aqueles sobre os quais não se têm informações de genealogia.

Os parâmetros definidos pelo programa e suas subrotinas têm que estar compatíveis com os dados, estando definidos com valores iguais ou superiores aos observados nos dados. Os níveis definidos para estes parâmetros no programa devem ser repetidos nas subrotinas. Tais parâmetros são:

MAXROW= número máximo de animais

MAXAN = MAXROW

MAXNRM= número máximo de elementos não iguais a zero (NZE).

Caso o número definido não seja suficiente, o programa será abortado e uma mensagem de erro será impressa.

Para rodar DFNRM:

- rode DFNRMF EXEC, o arquivo executável, que definirá os arquivos de entrada (que foi o arquivo de saída de BINTRAN) e de saída, compilar, carregar e rodar DFNRM; (Anexo 3);

- responda às questões interativas. Para a questão sobre o animal com maior número, responda com quantos "9" quantos forem o número de dígitos da identificação do animal, pai ou mãe. Por exemplo, se seus animais forem identificados por 6 dígitos, o maior animal será "999999". Responda às outras questões, segundo seu interesse;

- imprima DF66 DATA, um dos arquivos de saída, pois ele contém informações que serão necessárias nas próximas etapas, tais como o número total de animais, incluindo os "animais base".

Uma modificação que pode ser feita em DFNRM, mostrada no Anexo 3, destina-se a gerar o arquivo DF88 DATA, que contém a identificação real e recodificada de cada animal. Este arquivo pode ser útil no caso de desejar o valor genético aditivo de cada animal.

Ao fim desta etapa, a matriz de parentesco e relações genéticas aditivas está montada.

2) Como entrar com o arquivo de registros

ATENÇÃO: Antes de rodar esta parte do programa, esteja certo de que DF66 DATA da etapa ou da análise anterior foi impresso, pois um dos arquivos de saída desta etapa também é DF66 DATA, que vai apagar o anterior.

Nesta etapa é necessário ler-se o **arquivo de registros** na forma binária, através do segundo uso do programa BINTRAN.

2.a) Rodando BINTRAN:

- não é necessário compilar-se BINTRAN novamente, a menos que alguma alteração, como por exemplo, não ler variáveis com valor zero, tenha sido feita. Se a alteração foi feita, compile BINTRAN com o comando FORTVS BINTRAN (opt(3);

- altere BINTRANF EXEC, que agora deve definir como arquivo de entrada o **arquivo de registros**. Rode BINTRANF e verifique, com o comando QFI, se os arquivos estão definidos corretamente;

- carregue e rode BINTRAN, com o comando "Load BINTRAN (clear start" (não feche o parênteses);

- responda às questões interativas:

- a primeira pergunta refere-se ao número de variáveis inteiras (integer) que vai ser lida. Nesta classe de variáveis estão incluídos a identificação do animal, pai e mãe, os níveis de efeitos fixos, e o efeito de ambiente permanente (efeito aleatório adicional).

Conte quantas classes existem nestas condições e informe o número;

- a segunda pergunta refere-se ao número de variáveis reais (real), ou seja, variáveis que têm decimais. Neste caso estão incluídas as covariáveis, e todos os registros de produção dos animais, **mesmo que as medidas originais não contenham decimais**. Conte o número de classes existentes nestas condições e informe o número.

BINTRAN mostrará na tela as primeiras 4 e a última observação e informará o total de observações lidas. Verifique se a leitura está correta.

Depois que o **arquivo de registros** está em forma binária, o programa DFCODE vai recodificar as variáveis na ordem necessária, preparando o arquivo para os programas seguintes. Este programa pode recodificar um segundo efeito aleatório, como o efeito materno e pode ler registros múltiplos de um mesmo animal.

2.b) Rodando DFCODE:

- rode DFCODEF EXEC, o arquivo executável que vai definir os arquivos de entrada e saída, compilar, carregar e rodar DFCODE;

- responda às questões interativas:

- a primeira questão deve ser respondida como "análise univariada", pois esta versão do programa não executa análises multivariadas;

- a segunda questão refere-se ao número de variáveis dependentes que vai ser analisado. Se neste arquivo existe apenas uma variável, responda "1", caso contrário, informe o número;

- a terceira pergunta refere-se ao número de covariáveis que vão ser consideradas na análise;

- a quarta pergunta trata do número de classes de efeitos fixos que a análise vai considerar;

- a quinta pergunta refere-se ao número de efeitos aleatórios "adicionais" ao animal. Caso se deseje na análise o efeito materno e o efeito de ambiente permanente, informe "2" nesta pergunta. Caso apenas o efeito materno, ou apenas o efeito de ambiente permanente seja desejado, informe "1";

- a sexta questão é relativa à necessidade de se recodificar o "segundo efeito animal", que pode ser o animal, seu pai ou sua mãe. Caso não se deseje estimar os valores genéticos aditivos para outro efeito que não seja o animal propriamente dito, responda "0" (não) a esta pergunta; caso contrário, responda "1";

- a sétima questão é ligada à sexta e pergunta qual é o "segundo efeito animal";

- a oitava questão dá a opção de se transformar os dados para a escala logarítmica ou não. Escolha sua opção. Caso seja necessário um outro tipo de transformação, o arquivo já deve conter as variáveis transformadas.

Após responderem-se às questões interativas, DFCODE apresenta na tela as primeiras 3 observações e informa o número total de observações lidas. Verifique se a leitura está correta.

3) Como montar a parte de mínimos quadrados das equações de modelos mistos- DFMME1

ATENÇÃO: Antes de rodar esta parte do programa, esteja certo de que DF66 DATA da etapa ou da análise anterior foi impresso, pois um dos arquivos de saída desta etapa também é DF66 DATA, que vai apagar o anterior.

O programa DFMME1 monta a parte de mínimos quadrados das equações de modelo misto, aumentada pelos elementos independentes (right hand side-RHS) e pela soma total de quadrados (YY'), armazenando os elementos diferentes de zero (NZE) na parte inferior da armazenagem esparsa.

3.a) Rodando DFMME1:

- para rodar DFMME1, execute DFMME1F EXEC, o arquivo executável que vai definir os arquivos de entrada e saída, compilar, carregar e executar DFMME1.

- responda às perguntas do sistema interativo de entrada:

- a primeira espera pelos comentários (até 6 linhas), que deverão ser impressos em todas as análises que se fizerem com este modelo. Para terminar os comentários, digite "*" na primeira coluna da primeira linha após seu último comentário.

- as próximas perguntas referem-se às variáveis dependentes que serão analisadas. Após informar o número de variáveis independentes, informe o nome de cada uma, na sequência em que se encontram no **arquivo de registros**.

- a sequência seguinte refere-se à parte fixa do modelo. Inicialmente pergunta quantas covariáveis existem no modelo. Informe o número, o nome de cada uma delas e a ordem da regressão desejada para o efeito (1 para linear, 2 para quadrática, 3 para cúbica,

etc). Em seguida, informe o número de efeitos fixos, seus nomes e o número de níveis para cada um deles.

- ao fim da parte fixa do modelo, o programa vai perguntar se existe alguma dependência conhecida entre os efeitos fixos. Se nenhuma dependência for conhecida, responda "0" e ele vai informar uma mensagem com o número de equações consideradas como zero e sua localização. A solução para aquele nível será zero e para os outros níveis, dentro daquele efeito, as soluções serão diferenças em relação a zero.

- seguem-se perguntas relativas à parte aleatória do modelo. A primeira questão refere-se ao número total de animais (main random effects). Este número foi informado no arquivo de saída de DFNRM e refere-se a todos os animais, incluindo pais, mães e "animais base". A próxima pergunta trata do efeito aleatório "adicional", como por exemplo efeito permanente de ambiente. Informe se o modelo considera este efeito, seu nome e o número de níveis. Caso este número não seja conhecido, podem-se utilizar programas como COUNT FORTRAN (Anexo 4), que fará esta contagem. Este efeito deve estar codificado no **arquivo de registros**, de forma sequencial. A última série de perguntas refere-se ao segundo efeito animal aleatório, usualmente o efeito materno. Responda se o modelo considera este efeito, a estrutura de covariância que este efeito tem com o principal efeito aleatório (para detalhes, veja Meyer, 1988b, Table II) e o nome do efeito.

- após todas as perguntas terem sido respondidas, DFMME1 informará qual o modelo de análise, quantas observações foram lidas, o número de equações e o número de elementos das matrizes que não são iguais a zero (NZE).

4) Como obter as soluções - valores genéticos aditivos, estimativas de componentes de variância e herdabilidade = DFUNIS, SPARSPAK version.

ATENÇÃO: Antes de rodar esta parte do programa, esteja certo de que DF66 DATA da etapa ou da análise anterior foi impresso, pois um dos arquivos de saída desta etapa também é DF66 DATA, que vai apagar o anterior.

DFUNIS é o programa que resolve, através de sistema interativo, as equações e dá as estimativas dos componentes de variância. Em sua versão original, ele não reordena a matriz, não dá as soluções para os efeitos analisados e exige maiores recursos computacionais. A versão que usa SPARSPAK para solucionar o sistema (Boldman e Van Vleck, 1991) além de ser mais rápida e exigir menores recursos de computador, pode, se modificada para tal, mostrar as soluções para os efeitos fixos, covariáveis e os valores genéticos aditivos para cada animal e para o segundo efeito animal, se este foi analisado no modelo.

Antes de rodar este programa, vários subprogramas e subrotinas têm que ser compilados. Compile usando o comand FORTVS e a opção de otimização 3 (opt (3)). Os programas que devem ser compilados são DFMAIN, DFSUBSPO, IDIAGST e SPARSPAK.

Este programa pode ser executado de modo interativo (como se faz quando é rodado em microcomputadores), mas como a solução do sistema pode levar horas em processamento, dependendo do modelo, estrutura e quantidade de dados, geralmente ele é rodado em BATCH. Para tal, é necessário um outro arquivo, o DF5 DATA, que responde às questões interativas de entrada, necessárias para rodar DFUNIS. Este arquivo é mostrado no Anexo 5. Ele informa:

- se a rodada é a primeira (opção 0) ou se este modelo já foi analisado previamente (opção 1). Esta informação é a chamada "run option". Caso esta seja uma rodada posterior, nos casos onde a convergência ainda não foi atingida, copie os arquivos de saída da última rodada DF57 DATA (que tem a matriz reordenada), DF58 DATA (que tem os dados, como eles são lidos pelo programa) e DF54 DATA, este último com o nome de DF53 DATA (que contém as estimativas iniciais -"priors" para a próxima rodada de interações).

- as estimativas iniciais para os parâmetros genéticos h^2 , m^2 e c^2 , respectivamente herdabilidade direta, materna e efeito de ambiente permanente e de covariância entre os parâmetros direto e materno. Esta configuração de estimativas iniciais é válida quando se usa um modelo com todos estes parâmetros estimáveis. Quando não se usa o efeito de ambiente permanente e/ou o efeito materno ou correlação entre direto e materno, omita estas estimativas iniciais. Nas rodadas futuras, quando as estimativas iniciais serão dadas por DF53 DATA, omita estas informações.

- o número da variável que será analisada nesta rodada. Este número representa a sequência original da variável dependente no arquivo de registros. Se apenas uma variável for analisada, delete ou comente esta linha.

- o critério de convergência, ou seja a variância máxima permitida para a estimativa de log da função de verosimilhança. Usualmente se usam valores de 1.d-9 a 1.d-12, onde "d" significa uma variável de dupla precisão e -9 ou -12 a potência de 10.

- o número de interações, segundo o procedimento Simplex, permitida. Alguns conjuntos de dados demandam de grande tempo de processamento para reordenação da matriz e por isso é indicado numa primeira rodada especificar-se um pequeno número de simplexes (2 ou 3), deixando um número maior, usualmente de 200 a 500, para rodadas posteriores, quando a matriz já foi reordenada.

- se a matriz já foi reordenada (opção 1, usada quando de rodadas posteriores à primeira) ou não (opção 0, usada em primeiras rodadas).

Assim, DF5 DATA vai ser completo na primeira rodada, mas conterà apenas a "run option", critério de convergência, número de simplexes permitido e a informação sobre a reordenação da matriz. (Anexo 5).

Uma vez verificado DF5 DATA, a rodada deve ser submetida ao computador, através da submissão do arquivo executável DFSPARSF EXEC (Anexo 6), que define os arquivos de entrada e saída, carrega os programas e subrotinas, define DF5 DATA como entrada nas telas interativas e finalmente roda o sistema. Para submeter o sistema ao computador, use o comando "VMBATCH SUBMIT DFSPARSF EXEC (CL I NAME _____)". Este comando submete o "job" denominado _____ ao sistema de análise em "batch", na classe i. O sistema de classes, seus limites e sua utilização no momento devem ser verificados em cada mainframe antes de se submeter o "job".

DFSUBSPO pode ser modificado para obterem-se as soluções para os efeitos, após a convergência ser atingida. Uma destas modificações é mostrada no Anexo 6. Com esta modificação, o vetor de soluções será impresso no arquivo DF99 DATA, na seguinte ordem:

- as primeiras linhas referem-se aos coeficientes de regressão para as covariáveis, tantos quantos forem as covariáveis x ordem da regressão. Assim, se o modelo considerou 2 covariáveis, uma com efeito quadrático e outra com efeito cúbico, as primeiras cinco linhas apresentarão estes coeficientes;

- as próximas linhas mostram as soluções para todos os níveis do primeiro efeito fixo;

- após as soluções para o primeiro efeito fixo, são mostradas as (n-1) soluções para cada um dos subseqüentes efeitos fixos;

- após as soluções para efeitos fixos, são apresentadas as soluções para cada um dos efeitos aleatórios não correlacionados, geralmente os efeitos permanentes de ambiente;

- finalmente são apresentadas as estimativas BLUP de valor genético aditivo, direto numa linha e materno na seguinte (se este foi analisado no modelo) para cada um dos animais recodificados em DFNRM, incluindo os "animais base" e os sem registro. O número total de estimativas de valor genético aditivo será dado portanto pelo número total de animais (informado em DFNRM) se o modelo considerou apenas o efeito direto e este número multiplicado por 2 se o efeito materno foi solicitado.

A última linha de DF99 DATA é o valor genético aditivo materno do último animal, a penúltima é seu valor genético aditivo direto e assim por diante.

Para relacionarem-se os arquivos DF88 DATA, com todos os animais em seus códigos originais e recodificações e DF99 DATA, o arquivo de saída após a convergência ter sido atingida com o fim de obterem-se em um só arquivo o número do animal e suas

estimativas de valor genético aditivo direto e materno, um pequeno programa em SAS foi escrito, CODEBV, que trata estes arquivos com este fim. Tal programa é apresentado no Anexo 7. Neste caso, antes de se rodar o programa, há que se modificar a informação "FIRSTOBS=___", substituindo-se ___ pelo número da primeira linha que contém o valor genético aditivo direto do primeiro animal. Esta primeira linha é obtida quer somando-se o número de linhas que traz os outros efeitos, quer subtraindo-se do total de linhas de DF99 DATA o número total de animais x 2 (caso haja efeito materno no modelo). O arquivo de saída de CODEBV EBVCODED DATA, que pode ser tratado por SAS ou qualquer outro programa caso se deseje classificar, ordenar ou analisar os valores genéticos aditivos.

5) *Algumas mensagens de erros e seus significados:*

5.A) "Required id not found". Isto significa que o animal indicado ("Id given was _____") foi encontrado no arquivo de registros, mas não está no arquivo de pedigree.

5.b) "Error in Sparspak, ierr = 53"

Esta mensagem vem impressa em DF66 DATA, arquivo de saída de DFUNIS. Este erro aparece quando a matriz não é "full rank", ou seja uma ou mais subclasses foi perdida. Isto ocorre quando se lê o arquivo de dados em BINTRAN não se considerando os zeros e algumas classes são perdidas. Retorne ao arquivo de dados e recodifique os efeitos perdidos (usualmente uma ou mais mães ou ninhadas, quando estas caracterizam o efeito de ambiente permanente).

5.c) "Error in Sparspak, ierr = 31"

Este erro também vem impresso em DF66 DATA após rodar-se DFUNIS. Neste caso, embora DF5 DATA tenha informado que a matrix foi reordenada, o programa não reconheceu DF57 DATA. Ou tal não foi copiado, ou sofreu algum dano. Caso ele não tenha sido copiado, copie o arquivo para seu setor de trabalho e repita a operação. Caso contrário, rode a análise novamente, especificando "0" na opção que indica se esta rodada é a primeira ou não.

5.d) "Value given implies correlation gt or eq 1"

Neste caso, o erro é impresso no arquivo J8281CON, saída de VMBATCH e indica que os valores dados como estimativas iniciais para o processo interativo são

absurdos, dando uma correlação entre h^2 e m^2 maior que a unidade. É portanto um erro em DF5 DATA. Assim,

$$\frac{\text{Cov}(a,m)}{h \cdot m} > 1.0, \quad \text{onde:}$$

$\text{Cov}(a,m)$ = covariância entre herdabilidade direta e materna, informada como estimativa inicial,

h = a raiz quadrada do coeficiente de herdabilidade (direta) informado,

m = a raiz quadrada do coeficiente de herdabilidade (materna) informado.

5.e) "End of Data Set, File _____"

Neste erro, também de DF5 DATA, foi informado que a matriz foi reordenada, mas o programa não encontra DF57 DATA ou o programa não encontra DF58 DATA, com o arquivo na forma em que é lido pelo programa. Rode a análise novamente, iniciando com DFMME1.

5.f) "Illegal decimal point ..."

É um outro erro de DF5 DATA. Provavelmente o modelo não considera efeito materno, ou a correlação do efeito direto com o materno, mas DF5 DATA informa tais estimativas iniciais.

RESULTADOS

Os resultados finais das análises feitas em DFREML, SPARSPAK version dependem não só do modelo, mas também das adaptações feitas no programa. Caso seja usado um modelo que estime efeito de ambiente permanente e os valores aditivos direto e materno, para cada animal, ambos correlacionados -(modelo 8)- e as adaptações para gerar DF88 DATA e DF99 DATA tenham sido feitas, os resultados serão impressos nos seguintes arquivos:

1)DFNRM:

Este programa gera arquivos de saída para a próxima etapa, na forma binária, sem interesse do ponto de vista de resultados. Gera também DF88 DATA, que contém

três colunas: a primeira com a palavra "ID", a segunda com a identificação recodificada do animal e a terceira com a identificação original do animal. Não imprima este arquivo, que somente será utilizado caso se deseje um arquivo com os valores genéticos aditivos de todos os animais. DF66 DATA contém informações a respeito da quantidade de dados lida, do número de animais, do número total de animais e de "animais base", a média do coeficiente de endogamia, o número de animais endogâmicos e sua média de endogamia. Imprima DF66 DATA.

2)DFCODE:

DFCODE gera arquivos intermediários para a próxima etapa e também DF66 DATA. Este arquivo apresenta as seguintes estatísticas a respeito das variáveis dependentes: média, desvio-padrão, Coeficiente de variação, mínimo, máximo, mínimo-desvio (em desvios-padrão) em relação à média e máximo-desvio (em desvios-padrão) em relação à média. Imprima este arquivo para checar seus dados.

3)DFMME1:

Além dos arquivos com a parte de mínimos quadrados das equações de modelos mistos, em forma binária e que serão processados por DFUNIS, DFMME1 gera DF66 DATA, que traz a descrição dos dados e do modelo, além do número de equações e de elementos não iguais a zero. Imprima este arquivo, pois todas as características do modelo lá estão.

4)DFUNIS:

Ao fim da rodada em batch, o sistema envia três arquivos, informando os tempos de processamento, eventuais erros e o mapa de alocação de recursos computacionais. É interessante verificar tais arquivos antes de imprimirem-se os outros, à procura de eventuais erros. Os arquivos de saída DF54 DATA, DF57 DATA e DF58 DATA só terão serventia caso vá se rodar novas interações com a mesma variável ou então novas análises com o mesmo conjunto de dados.

Já DF99 DATA traz todas as soluções, conforme descrito e precisa ser cuidadosamente analisado, pois um engano de apenas uma linha altera completamente a interpretação dos resultados. Este arquivo tem tantas linhas quantas forem as equações, usualmente milhares e só deve ser impresso quando estritamente necessário. Caso o interesse seja nas soluções para covariáveis e efeitos fixos, imprima só as primeiras linhas, onde estão estas informações. Caso o interesse seja pelos valores genéticos aditivos, trabalhe este arquivo através de CODEBV SAS e imprima EBVCODED DATA.

DF66 DATA é o arquivo de saída mais importante, onde estão não só as informações sobre o modelo, como também tempo de processamento, as variâncias da função log da verosimilhança, a variância final encontrada para aquela função, os componentes de variância, as estimativas dos parâmetros genéticos e finalmente o tempo total de processamento. As informações contidas neste arquivo são muito claras e dispensam maiores detalhes.

CONCLUSÕES

DFREML, em sua versão que utiliza SPARSPAK para resolver o sistema de equações com matrizes esparsas, é uma poderosa ferramenta para análise de dados de melhoramento animal e estimação de parâmetros genéticos ou valor genético aditivo. No entanto, as modificações introduzidas na versão original o transformaram em um pacote aplicável a quaisquer tipos de análises onde modelos mistos sejam aplicados, já que as soluções apresentadas são BLUE (best linear unbiased estimates) ou BLUP (best linear unbiased predictors) das funções estimáveis que dão respostas aos problemas analisados. Este guia torna este sistema de programas acessível à maioria dos pesquisadores.

AGRADECIMENTOS

Aos Drs. Dale Van Vleck, Keith G. Boldman, Karin Meyer, Rafael Nuñez-Dominguez e Octávio Martinez, pelo acesso a seus programas e pelas informações; aos Drs. Gustavo Maria-Levrino e Raysildo B. Lobo pelas sugestões; a University of Nebraska, Department of Animal Science, Lincoln, Ne, USA e ao Dr. Rodger K. Johnson, pela oportunidade de trabalhar com o sistema; ao CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico pelo apoio financeiro; ao Departamento de Produção Animal da Fac. de Med. Veterinária e Zootecnia da Universidade de São Paulo, Instituição de origem do autor.

Publicação subvencionada pela FAPESP.

RESUMO

A melhor estimação de componentes de variância e parâmetros genéticos nos dados não balanceados usualmente utilizados em melhoramento genético animal é aquela obtida por máxima verosimilhança restrita (REML), utilizada sob um modelo animal, que considera todas as informações disponíveis sobre a genealogia do animal, aproximando-se ao máximo da situação biológica real. O uso de algoritmos livres de derivadas, onde não é necessária a inversão de grandes matrizes, geradas em análises de modelo animal e o uso de sistemas especializados em solução de matrizes esparsas - SPARSPAK - resultou em um sistema de programas, originalmente escritos em FORTRAN, que resolve as complexas equações de modelos mistos em tempo menor e com menor utilização de recursos de computador que a versão original de DFREML. Este trabalho apresenta um guia prático destinado aos melhoristas que precisam utilizar-se destes programas mas não estão familiarizados com FORTRAN ou com a versão prévia de DFREML.

ANEXO 1

BINTRANF EXEC

```

/* an exec to define filedefs for bintran */
trace results
/* erase old copies of output files */
"ERASE DF33 DATA A"
"FILEDEF * CLEAR"
/* define input files */
"FILEDEF 1 DISK IW73 DATA A1"
/* define output files */
"FILEDEF 2 DISK DF33 DATA A"

```

BINTRAN, adaptado para ler arquivos não formatados, formatados ou não considerar variáveis com valor=zero

```

C-----
          PROGRAM BINTRAN
C-----
C      PURPOSE:   Program to read unformatted data and transform
C                  to binary as input for KM'S programs
C      WRITTEN:   By KM; modified by KGB 6-22-89
C-----
          implicit double precision (a-h,o-z)
          dimension ivec(20),rvec(20)
          WRITE(6,*)'no. of integer variables ?'
          READ(5,*)ni
          WRITE(6,*)'no. of real variables ?'
          READ(5,*)nr
          WRITE(6,*)' '
          WRITE(6,*)'unit 1 is input; unit 2 is output '
          WRITE(6,*)' '
C50      READ(1,5,END=99)(IVEC(L),L=1,NI),(RVEC(J),J=1,NR)
C-----
C50      FORMAT TO READ RABBIT DATA FILE=REPREML, FERRAZ, 1991
C 5      FORMAT (I7,1X,I6,1X,I6,1X,I1,1X,I1,1X,I2,1X,I2,1X,I3,1X,F2.0,1X,

```

```

C      CF7.5,1X,F3.0,1X,F4.0,1X,F4.0,1X,F3.0,1X,F3.0,1X,F3.0,1X,F6.4)
C----- STATEMENT TO READ NON FORMATED FILES (ALL VARIABLES):
50    READ(1,*,END=99)(IVEC(L),L=1,NI),(RVEC(J),J=1,NR)
C----- MODIFICATION TO SKIP VARIABLES =0, IN THIS CASE THE 3RD
C----- REAL VARIABLE. IT CAN BE DEFINED BY THE FORMAT STATEMENT
C      K=3
C      IF (RVEC(K).EQ.0) GOTO 50
      NREC=NREC+1
C      WRITE(6,*)NREC
      if(nrec.lt.5) WRITE(6,*)(ivec(l),l=1,ni),(rvec(j),j=1,nr)
      WRITE(2)(IVEC(L),L=1,NI),(RVEC(J),J=1,NR)
      go to 50
99    WRITE(6,*)(ivec(l),l=1,ni),(rvec(j),j=1,nr)
      WRITE(6,*)' '
      WRITE (6,*)'NREC = ',NREC
      stop
      end

```

ANEXO 2

BINDUP = BINTRAN adaptado para descartar duplicatas

```

C-----
PROGRAM BINDUP
C-----
C      PURPOSE:   Program to read unformatted data and transform
C                to binary as input for KM'S programs
C      WRITTEN:   By KM; modified by KGB 6-22-89
C-----
      implicit double precision (a-h,o-z)
      dimension ivec(20),rvec(20)
      WRITE(6,*)'no. of integer variables ?'
      READ(5,*)ni
      WRITE(6,*)'no. of real variables ?'
      READ(5,*)nr
      WRITE(6,*)' '
      WRITE(6,*)'NO. OF INTEGER VARIABLE TO DELETE DUPLICATE?'

```

```

READ(5,*)ND
WRITE(6,*)' '
WRITE(6,*)'unit 1 is input; unit 2 is output '
WRITE(6,*)' '
IDOLD=0
50  read(1,*,end=99)(ivec(l),l=1,ni),(rvec(j),j=1,nr)
    NREC=NREC+1
    if(nrec.lt.5) WRITE(6,*)(ivec(l),l=1,ni),(rvec(j),j=1,nr)
    IF(IVEC(ND).EQ.IDOLD)THEN
        NDUP=NDUP+1
        GOTO 50
    ELSE
        IDOLD=IVEC(ND)
    END IF
    write(2)(ivec(l),l=1,ni),(rvec(j),j=1,nr)
    go to 50
99  WRITE(6,*)(ivec(l),l=1,ni),(rvec(j),j=1,nr)
    WRITE(6,*)' '
    WRITE (6,*)'NREC = ',NREC
    WRITE(6,*)' '
    WRITE (6,*)'NDUP = ',NDUP
    stop
    end

```

ANEXO 3

DFNRMF EXEC = o arquivo executável para rodar DFNRM

```

/* an exec to define filedefs for dfnrm */
trace results
/* erase old copies of output files */
"FILEDEF * CLEAR"
"ERASE DF11 DATA A"
"ERASE DF44 DATA A"
"ERASE DF66 DATA A"

```

```
"ERASE DF88 DATA A"
/* define input files */
"FILEDEF 33 DISK DF33 DATA A"
/* define output files */
"FILEDEF 11 DISK DF11 DATA a1"
"FILEDEF 22 DISK DF22 DATA a1"
"FILEDEF 23 DISK DF23 DATA a1"
"FILEDEF 44 DISK DF44 DATA a1"
"FILEDEF 66 DISK DF66 DATA a1"
"FILEDEF 88 DISK DF88 DATA A1"
"LOAD DFNRM(CLEAR START"
```

DFNRM, modificação para gerar DF88 DATA

```
if(idebug.eq.1)then
    write(iun99,*)nrec2,mrow2,ifac,max
    write(iun99,*)(idvec(i),i=1,nrec2)
    write(iun99,*)(kfirst(i),i=1,mrow2)
```

```
end if
```

```
END IF
```

```
DO 9989 KK=1,NREC2
```

```
    WRITE(IUN88,*)'ID',KK,IDVEC(KK)
```

```
9989    CONTINUE
```

ANEXO 4

COUNT FORTRAN = um programa destinado a contar o número de subclasses diferentes dentro de uma classe.

```
C *****
C *THIS PROGRAM COUNTS REPEATED OBSERVATIONS OF ONE VARIABLE
C *WITHIN ANOTHER VARIABLE, I.E., REPEATED COWS WITHIN BREED.
C *DATA MUST BE SORTED BY BREED AND COW.
C *WRITEN AT UNL, 09/05/91 BY RAFAEL NUNEZ DOMINGUEZ, OCTAVIO
C *MARTINEZ AND KEITH BOLDMAN.
C *****
```

```

      INTEGER INC(140)
C     MAXIMUM CODE FOR IB=136
      DO 20 I=1,140
        INC(I)=0
20    CONTINUE
      IOLD=0
      NREC=0
      NCOW=0
50    READ(1, 10, END=99)IB, IC
10    FORMAT(I3, I6)
      NREC=NREC+1
C     IF(NREC.LT.5)WRITE(6,*)IB,IC
      IF(IOLD.NE.IC)THEN
        IOLD=IC
        INC(IB)=INC(IB)+1
        NCOW=NCOW+1
      END IF
      GOTO 50
99    CONTINUE
C     WRITE(6,*)IB,IC
      DO 199 I=1,140
        IF(INC(I).GT.0)THEN
C     THIS IS IMPORTANT SINCE THE SIZE OF THE INC VECTOR IS GREATER
C     THAN THE NUMBER OF BREEDS
          WRITE(16,*)'BREED CODE=',I,' NO. COWS =',INC(I)
        END IF
199   CONTINUE
      WRITE(16,*)' '
      WRITE(16,*)'RECS. READ=',NREC
      WRITE(16,*)'TOTAL COWS=',NCOW
      STOP
      END

```

ANEXO 5

DFSUBSPO = modificações para gerar DF99 DATA

```

CALL IDIAGST(LSTART)
REWIND(99)
DO 171 I=1,(NEQ-NQ-NRZERO)
    DIAGL=S(I+LSTART-1)
c    DET=DET+DLOG(DIAGL*DIAGL)
    DET=DET+DLOG(DIAGL)
WRITE(99,924)'SOLUTION FOR ',I,' ', S(I)
171    SSMOD=SSMOD+(S(I)*RHS(I))
    det=det*2.d0
924    FORMAT (A12,3X,I6,2X,A1,3X,F15.7)

```

DF5 DATA = arquivo que informa parâmetros iniciais para rodar DFUNIS, versão SPARSPAK, em batch. Modelo para primeira rodada.

```

1      RUN OPTION: 0 FIRST; 1 LATER
0.10D0 H2{ FOLLOWS RUN OPTION IF 0 }
0.05D0 C2
0.01D0 M2
-.03D0 COV(A,M)
4      TRAIT NUMBER{FOLLOWS PRIORS IF NQ1 AND RUN OPTION=0}
1.D-9  CONVERGENCE
2000   NUMBER OF SIMPLEXES
1      MATRIX PREVIOUSLY REORDERED: 0 - NO; 1 - YES

```

```

0.10D0 H2 { FOLLOWS RUN OPTION IF 0 }
0.05D0 C2
0.01D0 M2
-.03D0 COV(A,M)
4      TRAIT NUMBER{FOLLOWS PRIORS IF NQ1 AND RUN OPTION=0}

```

----- GENERAL FORM:

```

0      run option: 0 first; 1 later
0.4d0  h2
0.1d0  c2

```

```

0.15d0  m2
-.03D0  COV(A,M)
1       trait number
1.d-12  convergence
2       NUMBER OF ROUNDS (200 AFTER REORDER)

```

GENERAL FORM FOR SECOND ROUND:

```

1       RUN OPTION: 0 FIRST; 1 LATER
1.d-9   convergence
200     NUMBER OF SIMPLEXES
1       MATRIX PREVIOUSLY REORDERED: 0 - NO; 1 - YES

```

DF5 DATA = arquivo que informa parâmetros iniciais para rodar DFUNIS, versão SPARSPAK, em batch. Modelo para segunda rodada.

```

1       RUN OPTION: 0 FIRST; 1 LATER
1.D-9   CONVERGENCE
2000    NUMBER OF SIMPLEXES
1       MATRIX PREVIOUSLY REORDERED: 0 - NO; 1 - YES

```

ANEXO 6

DFSPARSF EXEC = o arquivo executável para DFUNIS na versão SPARSPAK.

```

&TRACE ALL
SET BLIP OFF
** delete old copies of files to be written */
ERASE DF66 DATA A
FILEDEF * CLEAR
FILEDEF 51 DISK DF51 DATA
FILEDEF 52 DISK DF52 DATA
FILEDEF 44 DISK DF44 DATA
FILEDEF 5 DISK DF5 DATA
** D-1 file: row,col,coef */
FILEDEF 45 DISK DF45 DATA
** input for later rounds */

```

FILEDEF 53 DISK DF53 DATA

** define output files */

** YPY, row number-I, # elem-NIR */

** column #s-IR(NIR) */

** values-VALUES(NIR) */

** RHS(NEQ-NQ-NRZERO) */

FILEDEF 57 DISK DF57 DATA

** S matrix */

FILEDEF 58 DISK DF58 DATA

** output for use in later rounds */

FILEDEF 54 DISK DF54 DATA

FILEDEF 66 DISK DF66 DATA(RECFM F LRECL 132

FILEDEF 99 DISK DF99 DATA(RECFM F LRECL 132

LOAD DFMAIN DFSUBSPO SPARSPAK IDIAGST(CLEAR START

DFUNIS EXEC = o arquivo executável para rodar DFUNIS

/* an exec to run DFUNIS interactively */

TRACE ALL

"FORTVS DFUNIS(OPT(3))"

if rc^=0 then

do

say"FAILED COMPILE"

say"JOB TERMINATING"

exit

end

/* delete old copies of files to be written */

"ERASE DF66 DATA A"

/* define input files */

/* non-interactive input */

/* "FILEDEF 5 DISK PK5 DATA A" */

"FILEDEF 44 DISK DF44 DATA A"

"FILEDEF 51 DISK DF51 DATA A"

"FILEDEF 52 DISK DF52 DATA A"

"FILEDEF 53 DISK DF53 DATA A"

/* D-1 file: row,col,coef */

"FILEDEF 45 DISK DF45 DATA A"

```

/* define output files */
"FILEDEF 54 DISK DF54 DATA A"
"FILEDEF 66 DISK DF66 DATA A(RECFM F LRECL 132)"
"LOAD DFUNIS (CLEAR START)"

```

ANEXO 7

CODEBV SAS - programa em SAS para relacionar DF88 DATA e DF9 DATA e imprimir os valores genéticos aditivos direto e materno para cada animal (número real e recodificado).

CODEBV

THIS IS A SAS PROGRAM TO RELATE TWO SPECIAL OUPUTS OF DFREML/ SPARSPAK VERSION(K.Meyer, 1988, DALE Van Vleck & KEITH G.Boldman 1991. 1) DF88, OUTPUT OF DFNRM, THAT HAS THE ANIMALS ID'S (CODED AND REAL) AND 2) DF99, OUTPUT FORMATED OF DFSUBSPO, THAT HAS THE SOLUTIONS FOR DIRECT AND MATERNAL BREEDING VALUES, PLUS CO-VARIATES (REGRESSIONS), AND OTHER EFFECTS. YOU NEED FIRST TO LOCATE IN DF99 DATA WERE IS THE FIRST DIRECT BREEDING VALUE, THAT YOU ARE GOING TO USE AS FIRSTOBS IN THE PROGRAM AND THEN RUN THIS SAS PROGRAM. PROGRAMMED BY JOSE BENTO STERMAN FERRAZ, UNL, AUGUST 1991

JBSF

```

*/
DATA A;
INFILE 'DF99 DATA A1' FIRSTOBS=144;
INPUT DIRECT 32-42 ú2 MATERNAL 32-42;
CODE= _N_;
DATA B;
INFILE 'DF88 DATA';
INPUT CODE 11-14 ANIMAL 21-26;
PROC SORT;
BY CODE;
DATA ALL;

```

MERGE A B;
BY CODE;
FILE EBVCODED;
PUT (CODE ANIMAL DIRECT MATERNAL)(5. 8. 13.8 13.8);

REFERENCIAS

- Boldman, K.G. and Van Vleck, L.D. Derivative-free Restricted Maximum Likelihood estimation in animal models with a sparse matrix solver. *J. Dairy Sci.* (in press).
- George, A., Liu, J. and Ng, E. (1980). *User guide for SPARSPAK: Waterloo sparse linear equations package.* CS-78-30, Depto. Computer Science, Univ. of Waterloo, Ontario, Canada.
- Graser, H.U., Smith, S.P. and Tier, B. (1987). A derivative-free approach for estimating variance components in animal models by Restricted Maximum Likelihood. *J. Anim. Sci.* 64: 1362-1370.
- Henderson, C.R. (1949). Estimation of changes in herd environment. *J. Dairy Sci.* 32: 706.
- Henderson, C.R. (1988). Theoretical basis and computacional methods for a number of different animal models. *J. Dairy Sci.* 71 (Suppl.): 1-16.
- Henderson, C.R., Kempthorne, O., Searle, S.R. and Von Krosigk, C.M. (1959). The estimation of genetic and environmental trends from records subject to culling. *Biometrics* 15: 192-218.
- Kennedy, B.W., Schaeffer, L.R. and Sorensen, D.A. (1988). Genetic properties of animal models. *J. Dairy Sci.* 71 (Suppl.): 17-26.
- Meyer, K. (1988a). DFREML - A set of programs to estimate variance components under an individual Animal Model. *J. Dairy Sci.* 71 (Suppl.): 33-34.
- Meyer, K. (1988b). *DFREML - Programs to Estimate Variance Components for Individual Animal Models by Restricted Maximum Likelihood (REML).*- USER NOTES. University of Edinburgh.
- Meyer, K. (1989). Estimation of variance components for individual Animal Models. I. Univariate analyses. *Genet. Sel. Evol.* 21: 317-340.
- Meyer, K. (1991). Estimation of variance components for individual Animal Models. II. Multivariate analyses. *Genet. Sel. Evol.* 23: 67- 83.
- Quaas, R.L. and Pollak, E.J. (1980). Mixed model methodology for farm and ranch beef cattle testing programs. *J. Anim. Sci.* 51: 1277-1287.
- Van Vleck, L.D. and Nuñez-Dominguez, R. (1990). Genetic evaluation of Dairy herds bulls and cows with an animal model. In: *International Dairy Seminar. Proceedings.* Centro de Ganaderia, Colegio de Postgraduados, Chapingo, Mexico.