Opportunities and Challenges for a New Approach to Genomic Prediction

Dorian Garrick
dorian@iastate.edu

Prediction of Merit

- Philosophical concept embodied in the "model" that is the basis for prediction
- Statistical method used to estimate effects and perhaps other parameters in the model
- Computing algorithm(s) to implement the statistical method

Philosophical Concept

- A Model describes cause and effect - the underlying process believed to result in the observations

  Performance = Breeding + Feeding

  Phenotype = Genotype + Environment

- The model (or a simplification of the model) is the basis for prediction

Model Equation

$$y = Xb + Zu + e$$

- y: Vector of phenotypes (performance)
- Xb: Vector of non-genetic effects herd-year age of dam date-of-birth (fixed effects)
- Zu: Vector of additive genetic (random) effects (EPD)
- e: Vector of leftover parts we do not know how to model and cannot explain (residuals)

This represents a "mixed" model (as it contains fixed and random effects)

Model

- The model is not completely specified with the model equation
  - Must also define information about;
    - the locations (means) of effects
    - the dispersion (variance-covariance) of effects
      - Based on pedigree-relationships for true EPD
    - sometimes the distributional assumptions of effects
      - eg normality of genetic and residual effects
  - Heritabilities, phenotypic standard deviations, genetic and phenotypic correlations are derived from these parameters

Statistical Method

- Preferred method is known as "BLUP"
  - Best meaning it minimizes the variance of prediction errors
  - Linear meaning EPD are computed from weighted sums and differences of observations
  - Unbiased meaning that estimates are equally likely to increase or decrease when more information is obtained
  - Prediction refers to estimates of random effects

Computing Algorithm(s)

- Henderson invented an efficient strategy to predict EPD based on mixed model equations

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \lambda A^{-1} \end{bmatrix} \begin{bmatrix} b \\ u \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

Scalar Variance Ratio

$$\lambda = (1-h^2)/h^2$$

Inverse of pedigree-based relationship matrix

Implementation

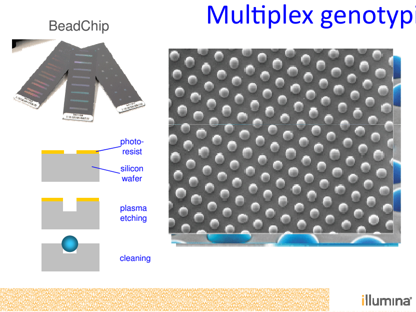
- Brute force formation of the sparse elements of the MME
  - Form the sparse inverse relationship matrix from pedigree
  - Accumulate and store only non-zero values
- Iteratively solve to obtain EPD
  - Start with some values for every effect
  - Iteratively refine the values to get a solution
    - Gauss-Seidel was the initial method of choice
    - Preconditioned Conjugate Gradient (PCG) was later adopted
- Some methods avoid forming the MME (IOD)

Implementation

- Solving the MME gives the EPD, but the prediction error variances needed to obtain EPD accuracies or reliabilities require the inverse of the left-hand side of MME
  - Too big to obtain with national data
  - Various approximations were developed
- Whole analysis is so much work it is often run 2-3x per year with regular interim solutions

BeadArray™ Technology: Array format generation

Multiplex genotyping



Genomic Technology

- This lead to some suggested philosophical changes in the model
  - Nejati-Javaremi (1997) imagined replacing the pedigree-based relationship matrix by relationships assessed using genomic information
  - Meuwissen, Hayes and Goddard (2001) extended Falconer's definition of a breeding value as the sum of gene effects to predict the EPD as the sum of estimated SNP effects
- These two approaches are actually equivalent and give the same EPDs
  - Stranden and Garrick (2009)

Breeding Value Model

- Use genotypes to obtain some kind of genomic relationship matrix
  - Using this instead of the pedigree-based relationship matrix is known as GBLUP
  - Minor modifications required to old software
- Now the relationship matrix and its inverse (if it exists) are dense, not sparse, requiring more computing effort
- Now the approximations for prediction error variances are not as good

Marker-Effects Model

- Use Henderson's MME to predict marker effects rather than breeding values
  - Use the marker effects to obtain EPD
- These models have been a major focus of researchers at Iowa State University over the last 6 years
  - New software has been developed (GenSel)
    - BayesC, BayesCpi, more efficient BayesA, Bayes B
      - Categorical data, dominance effects, etc

New Computing Strategies

- Markov chain Monte Carlo (MCMC) has become a popular strategy for model fitting
  - Not just a Bayesian technique
  - Alternative to methods for iterative solution like Gauss-Seidel and PCG
- MCMC provides plausible values for each of the effects in the model, not just the estimates of effects that solve the equations
  - This gives you the EPD and the PEV, accuracy etc

Not everyone genotyped

- Now we have two different models – one for genotyped and another for non genotyped

Only some animals genotyped (1)

- First Approach: Breeding Value Model for all
  - Same model equation (use EPDs)
  - Single Step HBLUP strategy
    - Various publications (Misztal, Legarra, Aguilar)
  - Assumed variance-covariance (H) is based
    - primarily on pedigree relationships for non-genotyped
    - primarily on genomic relationships for genotyped
  - Use its inverse in conventional software
  - Limit on about 100,000 genotyped animals

Single Step HBLUP

- First Attempt to model covariance

$$var\begin{bmatrix} u_n \\ u_g \end{bmatrix} = \begin{bmatrix} A_{nn} & A_{ng} \\ A_{gn} & G_{gg} \end{bmatrix} \sigma_a^2$$

Misztal et al (2009)

- Second Attempt to model covariance

$$H = var\begin{bmatrix} u_n \\ u_g \end{bmatrix} \sigma_a^{-2} = \begin{bmatrix} A_{nn} + A_{ng}A_{gg}^{-1}G_{gg}A_{gg}^{-1}A_{gn} & A_{ng}A_{gg}^{-1}G_{gg} \\ G_{gg}A_{gg}^{-1}A_{gn} & G_{gg} \end{bmatrix}$$

Legarra et al (2009)

$$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G_{gg}^{-1} - A_{gg}^{-1} \end{bmatrix}$$

Aguilar et al (2010)

Only some animals genotyped (2)

- Second Approach: Hybrid Model
  - Impute genotypes for non genotyped from their genotyped relatives
  - Estimate marker effects from all animals
  - Fit a residual breeding value effect for non genotyped animals to account for imputation errors

Fernando, Dekkers and Garrick (2014) GSE

Let's revisit the basic idea

$$\begin{bmatrix} y_n \\ y_g \end{bmatrix} = \begin{bmatrix} X_n \\ X_g \end{bmatrix} b + \begin{bmatrix} Z_n & 0 \\ 0 & Z_g \end{bmatrix} \begin{bmatrix} u_n \\ u_g \end{bmatrix} + \begin{bmatrix} e_n \\ e_g \end{bmatrix}$$

$with\ u_g = M_g \alpha\ for\ genotyped\ individuals$ (MHG 2001)

$$whereas\ u_n = \widehat{u_n}/u_g + (u_n - \widehat{u_n}/u_g) = \widehat{u_n}/u_g + \varepsilon_n$$

$$with\ \widehat{u_n}/u_g = A_{ng} A_{gg}^{-1} u_g$$

$$so\ u_n = A_{ng} A_{gg}^{-1} u_g + (u_n - A_{ng} A_{gg}^{-1} u_g) = (A_{ng} A_{gg}^{-1} M_g)\alpha + \varepsilon_n$$

Fernando, Dekkers and Garrick (2014) GSE

With "Hybrid" Mixed Model Equations

$$\begin{bmatrix} X'X & X'ZM & X_n'Z_n \\ M'Z'X & M'Z'ZM + \phi & M_n'Z_n'Z_n \\ Z_n'X_n & Z_n'Z_nM_n & Z_n'Z_n + A^{nn}\lambda \end{bmatrix} \begin{bmatrix} b \\ \alpha \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} X'y \\ M'Z'y \\ Z_n'y_n \end{bmatrix}$$

$$where\ X = \begin{bmatrix} X_n \\ X_g \end{bmatrix}, Z = \begin{bmatrix} Z_n \\ Z_g \end{bmatrix}, M = \begin{bmatrix} M_n \\ M_g \end{bmatrix} = \begin{bmatrix} A_{ng} A_{gg}^{-1} M_g \\ M_g \end{bmatrix}, y = \begin{bmatrix} y_n \\ y_g \end{bmatrix}$$

$with\ EBV\ given\ by$

$$\widehat{u_g} = M_g \widehat{\alpha}$$

$$\widehat{u_n} = M_n \widehat{\alpha} + \widehat{\varepsilon_n}$$

NB Single-Step GBLUP
is a special case of the above
(but in this equivalent model no inversion is needed)

$$M_n = A_{ng} A_{gg}^{-1} M_g$$

Fernando, Dekkers and Garrick (2014) GSE

If everyone is genotyped

$$\begin{bmatrix} X'X & X'ZM & X_n'Z_n \\ M'Z'X & M'Z'ZM + \phi & M_n'Z_n'Z_n \\ Z_n'X_n & Z_n'Z_nM_n & Z_n'Z_n + A^{nn}\lambda \end{bmatrix} \begin{bmatrix} b \\ \alpha \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} X'y \\ M'Z'y \\ Z_n'y_n \end{bmatrix}$$

These are the MME that form the basis of BayesA, BayesB, BayesC etc

If no one is genotyped

$$\begin{bmatrix} X'X & X'ZM & X_n'Z_n \\ M'Z'X & M'Z'ZM + \phi & M_n'Z_n'Z_n \\ Z_n'X_n & Z_n'Z_nM_n & Z_n'Z_n + A^{nn}\lambda \end{bmatrix} \begin{bmatrix} b \\ \alpha \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} X'y \\ M'Z'y \\ Z_n'y_n \end{bmatrix}$$

These MME form the basis of traditional pedigree-based BLUP

Single Step HBLUP special case

$$\phi = diagonal\ \sigma_e^2/\sigma_{\alpha i}^2\ (general\ locus\ specific)$$

$$\lambda = \sigma_e^2/\sigma_g^2 = (1-h^2)/h^2$$

$$Suppose\ \phi = \lambda/2\overline{pq}k\ for\ k\ loci\ (one\ special\ choice)$$

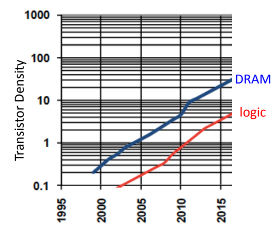
- In this special case, the hybrid model gives the same EPDs for genotyped and non-genotyped animals as does single step HBLUP but without needing any matrix inversions or needing the genomic relationship matrix to be full rank

Computing Strategy

- These hybrid MME are
  - straightforward to form and solve using conventional approaches for pedigrees < 1 million animals
  - straightforward to form and fit using MCMC methods to obtain EPD and accuracies for pedigrees < 1 million animals
  - require advanced computing techniques to be efficiently used for >10 million animals

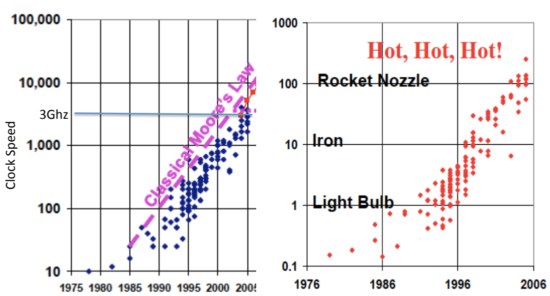
High-Performance Computing Trends

- Moore's Law "the number of transistors that can be integrated on a die would double every 18-14 months"
  - Memory density will increase 4x every 3 years



Latest trend is 3d wafer processing

High-Performance Computing Trends



www.top500.com Based on 500 most powerful computers

Parallel Computing

Multiple cores on the standard CPU (eg 4, 6, 8, 16)
Multiple cores on graphics cards (eg >2,000, >4,000)

Computing Since 2004

- No increases in clock speed (often decrease)
- Increase the number of computer cores to increase whole machine power
- Less memory available per core
- Less electricity produced per core
- Liquid cooling of cores
- Hybrid computing using graphics cards

Genetic Evaluation changes since 2004

- Buy bigger computers with more cores and more memory
- Use just one core while all the others do nothing
  - Using 1/6, 1/8, or 1/16 computer power available

Parallel Computing

- Need new software
- Need new computing approaches
- Need big problems
  - Have not been able to speed up GenSel using multiple processors or graphics cards unless we have many more genotyped animals
- Single Step using our hybrid model is a perfect example of a problem suited to parallel computing

Challenges

- Few individuals who understand the animal breeding aspects and the computing aspects
- Few individuals who have used MCMC approaches on large-scale problems
- So far unsuccessful in obtaining federal funding for these initiatives – they are seen as "development" rather than "research", "education" or "extension"
- Market not big enough for venture capital

Summary of New Approach

- Opportunities
  - New algorithms
  - New hardware
  - Technically sound approach without approximations
- Challenges
  - Funding for initial and ongoing developments
    - Developing new approaches along with ongoing research
  - Identifying expertise to assist in development
  - Overcoming the "can't be done" attitude
  - Streamlining interface(s) to association databases