Bolt and an Alternative Approach to Genomic EPDs

Bruce L. Golden, Ph.D., CEO, Theta Solutions; Rohan Fernando, Professor, Iowa State University; and Dorian J. Garrick, Professor, Iowa State University

Introduction

Misztal et al. (2009) and Aguilar et al. (2010) introduced approaches for single-step analyses that combined genotyped and non-genotyped animals in the same analysis on the basis of pedigree and genomic relationship matrices and their inverses. Fernando et al. (2014) and Garrick et al. (2014) presented an alternative computing algorithm for the same model that gives identical genomic enhanced EPDs but has different computational properties. That single-step approach, called the Hybrid Model, provides solutions for the marker effects and the imputation errors for non-genotyped animals, rather than directly providing the EBVs. It has computational advantages over that of Misztal et al. (2009) in that it does not require any large matrix inverse, and it has the ability to implement marker selection methods such as Bayes C (or other forms of the Bayesian alphabet). Implementing a marker selection approach resulted in a substantial increase in the accuracy of the predictions from the same amount of genotype data. Our aim in this paper was to present an alternative formulation of the mixed model equations (MME) from those presented in Fernando, et al. (2014) and Garrick, et al. (2014). Putatively called the Super Hybrid Model (SHM; Fernando and Garrick, unpublished), the new MME are even easier to assemble and solve then those of the Hybrid Model.

Additionally, we present here results from implementing these models in the Biometric Open Language Tools software package (Bolt) available from ThetaSolutions LLC. Designed specifically to work with commodity class General Purpose Graphics Processing Units (GPU), we have solved and sampled very large, complex multiple trait SHM that include maternal effects. In addition to the original Hybrid Model, and the SHM, we have implemented the original single-step approach (SSGBLUP) of Misztal et al. (2009) in Bolt. Breeding companies and organizations in several countries are currently converting their routine evaluations to use Bolt, including the Pan American Cattle Evaluation (PACE) of Hereford cattle and the multibreed International Genetic Solutions (IGS) evaluation of Angus, Red Angus, Gelbvieh, Limousin, Maine Anjou, Shorthorn, and Simmental pure- and crossbred cattle from US and Canada. Here we discuss implementation of Bolt for IGS for their twelve participating beef breed associations.

The Super Hybrid Model

The basic form of the SHM includes the usual fixed effects; marker effects, α; breeding value effects for animals that are not genotyped, u_n; and residual effects, e. The model equation is

$$y = Xb + Z_n u_n + Z_g M_g \alpha + e$$

where X is an incidence matrix of fixed effects, b, on observations in y; $Z = [Z_n \quad Z_g]$, is an incidence matrix of animals with observations in y; g and n subscripts refer to animals who were genotyped and animals who were not genotyped respectively; M is the matrix of marker values

and α are the random additive marker effects; u_n are the random breeding values for animals who were not genotyped and e are the residual effects on y.

In large problems the only substantial amount of work is the formation of the diagonal block for the marker effects. However, Bolt has optimized routines that have achieved over 7TFlops on inexpensive enthusiast class hardware when performing this computation making it highly tractable. Additionally, the computation of the diagonal blocks is performed only once, in parallel, during assembly of the MME.

We have developed optimized asynchronous parallel methods for high performance sampling of the dependent variables (Golden et al., 2014). Using Gibbs sampling results in high quality estimates of the prediction error variances including the variance of functions of the EPD such as economic indexes. Sampling also permits the implementation of marker selection models which results in substantial increases in accuracy over SSGBLUP which always fits all markers.

Other features of the SHM formulation of the MME include no large inverse matrices as are required in SSGBLUP or solutions involving the forward/backward substitution solve (or other solve) at each round of sampling as was required for efficient implementation of Fernando et al. (2014).

Expanding the SHM to include extra polygenic effects is trivial and extending it to maternal effects and multiple traits is straight forward. Including an extra polygenic effect is important when the markers do not describe all the additive genetic variance of the traits. This has been shown to be the case with current marker information (Saatchi and Garrick, 2016). Failure to do so results in widely-used sires with high accuracy EPD having slightly different genomic prediction estimates compared to those from traditional pedigree analyses.

The Bolt Software

Bolt is a collection of over one hundred software tools implemented as a set of commands used to manipulate data and the matrices involved in statistical problem assembly and solution. Combining Bolt with an environment such as the Born Again Shell (bash) or other computer language like environments (e.g., Python) provides a full featured language the professional analyst can use for many classes of statistical analysis of very large data sets.

Bolt is supported in the Linux environment and is designed to use low-cost computer workstation hardware with at least one general purpose CUDA class graphics processing unit (GPU). GPU computing has become a standard method in scientific computing for achieving very high performance computations at a relatively low cost (Owens, et al., 2008). Originally developed to process data for video editing and the computer gaming industry, GPU were adapted to provide general purpose computing for numerically intensive problems. Two widely used programming environments available for GPU computing include OpenCL and CUDA. The CUDA environment is available only for use with GPU designed by the NVidia Corporation while OpenCL can be used on other manufacturers' GPU (e.g., Advanced Micro Devices, Inc.) as well as Nvidia's GPU. However, the CUDA programming environment developed by

NVidia, is more fully featured and exceeds most performance benchmarks compared to other manufacturers' GPU. Particularly, the sparse matrix libraries and basic linear algebra libraries in the CUDA environment are highly optimized. CUDA is freely available and is well supported. NVidia sells both enterprise and enthusiast class GPU. We have found that lower cost enthusiast class GPU actually perform faster than the enterprise class GPU and are a fraction of the cost. Although Bolt supports both types of processors, we recommend its use with low cost consumer class workstations using enthusiast class GPU in the CUDA environment and Linux.

Bolt is designed not only to use GPU but to maximize the parallel execution capability of multiple core CPU, often achieving full so-called embarrassingly parallel execution. Additionally, Bolt is designed to take advantage of systems with multiple-GPU installed. Bolt's design makes it easy for professional analysts to make decisions about applying CPU cores and GPU to a single analysis or splitting the CPU cores and GPU among different problems. The complexity of CPU and GPU control is largely abstracted from the analyst so that the best analytical methods to apply to a problem can be focused on.

The IGS International Genetic Evaluation

Lead by a consortium of beef breed associations, the International Genetic Solutions organization has worked with Theta Solutions, LLC to implement a multi-breed genetic evaluation including data from twelve different beef breed associations from North America.

The first prototype analysis included thirteen traits' representing threshold and continuous observations from 6,987,238 pedigree observations including 45,176 observations on animals with genotypes, and 5,663,965 animals with performance observations. Traits were run in meaningful multiple trait combinations. For example, EPDs for birth weight, weaning weight, milk, and total maternal were solved together. The model included extra polygenic effects for birth additive direct, weaning additive direct and weaning additive maternal effects.

The analysis was performed on a computer built on an ASRock X99 Extreme11 motherboard with an Intel Xeon E5-2643 V3 (6 core at 3.4GhHz) processor. It had 64G of ECC DDR4 memory and four Titan X GPU. No overclocking of the CPU or GPU was performed. Our previous work (Golden, et al., 2015) has shown that substantial benefit from overclocking can be obtained. However, the E5-2643 cannot be overclocked.

The timings given here are for the so-called MSRP (Saatchi and Garrick, 2016) subset of genetic markers. The strategy implemented for the IGS analysis is to use a Bayes C0 analysis for an informative subset of markers identified from a Bayes C analysis (with pi=.95) applied to higher denity (e.g. 70k markers) periodically performed to refresh and validate the subset list. Our as yet unpublished studies have shown that this results in equivalent accuracies of the Bayes C analysis predictions' and are substantially more accurate than Bayes C0 of larger marker sets (e.g., BovSNP50). Another advantage is these analyses using relevant subsets of markers complete relatively quickly, allowing for new analyses to be performed when new genotype data are received. This way, IGS can turn around results to their members and customers as frequently as daily. The wall-clock time to assemble the SHM for this analysis was 50 minutes

and 25 seconds. Once assembled the time to solve the equations to obtain the EPDs using a PCG solver was 9 minutes and 39 seconds.

A Bayes C0 sampling strategy with four parallel chains of ten thousand samples each after being seeded with the PCG solver solutions was used to obtain prediction error variances. The wall clock time to obtain the prediction error variances was 5 hours and 44 minutes.

Citations

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta and T. J. Lawlor. 2010. Hot topic: a unified approach to utilize phenotype, full pedigree and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93:743-752.

Fernando, R. L., J. C. M. Dekkers and D. J. Garrick. 2014. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analysis. Genetics Selection Evolution, 46:50.

Garrick, D. J., J. C. M. Dekkers, B. L. Golden and R. L. Fernando. 2014. Bayesian prediction combining genotyped and non-genotyped individuals. Proc. 10th World Congress of Genetics Applied to Livestock Production, https://www.asas.org/docs/default-source/wcgalp-proceedings-oral/053_paper_10311_manuscript_1300_0.pdf?sfvrsn=2.

Golden, B. L., R. L. Fernando and D. J. Garrick. 2015. High performance Gibbs Sampler for mixed density general linear systems. Proc. GTC 2015, http://on-demand.gputechconf.com/gtc/2015/posters/GTC_2015_Life___Material_Science_06_P5265_WEB.pdf.

Misztal, I., A. Legarra and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J. Dairy Sci. 92:4648-4655.

Owens, J. D., M. Houston, D. Luebke, S. Green, J. E. Stone, and J. C. Phillips. GPU Computing. Proceedings of the IEEE, 96(5), pages 879–899, May 2008.

Saatchi, M., and D. J. Garrick. 2016. Developing an efficient reduced panel for low-cost genotyping in beef cattle. Proc. Plant and Anim. Genome Meeting, https://pag.confex.com/pag/xxiv/webprogram/Paper22335.html.