Extension of single-step ssGBLUP to many genotyped individuals

I. Misztal

1 University of Georgia, Department of Animal and Dairy Science, 30602 Athens, GA, USA
ignacy@uga.edu (Corresponding Author)

Summary

Single-step methodology (SS) became a method of choice for genomic computations. As originally introduced, SS required quadratic memory and cubic computations with the number of genotyped animals. Those computational costs would become prohibitive for more than 100k animals, but more than 1 million Holsteins have been genotyped. Various methods have been proposed to reduce these costs. Methods that involved unsymmetric equations failed to scale up when the inversion was not needed. Hybrid methods based on breeding values (BV) estimation for nongenotyped animals and SNP estimation for genotyped animals also failed to scale up. A method called SS Bayesian Regression "imputes" genotypes for nongenotyped animals, eliminating the need for relationship matrices in the equations. While the imputation is computationally expensive, graphical processors facilitate computations, but complex models are hard to implement. The dimensionality of the genomic information is limited by the effective population size and varies from 5k to 15k in livestock species. An inverse of genomic relationship matrix (G) called APY exploits those limits, is sparse, and has almost linear memory and computational costs for large populations. In tests, national evaluations with more than 700k genotyped animals took less than a day. The inverse can also be derived implicitly, for use in PCG type iteration, by a singular value decomposition (SVD) of gene content; such a decomposition for a 720k × 60k matrix took less than a day. An exact product of the inverse of blended G by a vector can be derived based on a Woodbury formula, where the only matrix to invert is that of SNP BLUP, and memory requirement scale up linearly. Tests reported good convergence in multibreed populations. Most SS methods need an inverse of the pedigree relationship matrix for genotyped animals. Such matrix may be dense, especially with long pedigrees, but can be accommodated in PCG based iteration programs at a low cost implicitly. Computations with SS methodology are now applicable to very large genotyped populations.

Keywords: single-step, large data, APY

Introduction

Single-step methodology (SS) combines phenotypic, pedigree and genomic information jointly (Aguilar et al., 2010; Christensen and Lund, 2010). This method avoids problems with de-regressed proofs or blending of pedigree information. Single-step also accounts for preselection of young genotyped animals (Legarra et al., 2014).

As originally introduced, SS required quadratic memory and cubic computations based on the number of genotyped animals. While such requirements are reasonable for at most 100k genotyped animals, more than 2 million Holsteins have been genotyped. This study focuses on extensions of SS that have reduced computing requirements and expands on developments described in Misztal & Legarra (2016).
**Material and methods**

**Original single-step**

Let the breeding values (BV) of nongenotyped animals be \( u_1 \) and of genotyped animals be \( u_2 \). Assume only pedigree relationships and additive variance of 1 for simplicity. The variance of BV is:

\[
H = \text{combines pedigree and genomic relationships and is based properties of the conditional distribution (Legarra et al., 2009; Christensen & Lund, 2010).}
\]

Assuming \( \text{var}(u_2) = G \) and \( \text{var}(u) = H \) leads to:

After rearrangement, the matrix \( H \) is

\[
H = A + \begin{bmatrix}
A_{12}A_{21}^{-1} & 0 \\
0 & I
\end{bmatrix}
\begin{bmatrix}
I & 0 \\
0 & I
\end{bmatrix}
\begin{bmatrix}
A_{22}^{-1}A_{21} & 0 \\
0 & I
\end{bmatrix},
\]

and its inverse is (Aguilar et al., 2010; Christensen and Lund, 2010):

\[
H^{-1} = A^{-1} + \begin{bmatrix}
0 & 0 \\
0 & G^{-1} - A_{22}^{-1}
\end{bmatrix}.
\]

In the original implementation (Aguilar et al., 2011), \( G \) and \( A_{22} \) matrices were created explicitly and inverted. With 50k genotyped animals and optimized programs, computing requirements were about 40 Gb memory and 1 h computing, with quadratic increases for memory and cubic increases for computations as the number of genotyped animals increased. Scaled to 1 million animals, the requirements would be 32 Tb memory and more than 1 month of computing, too computationally intensive for all but the largest computers.

**Initial attempts at reducing costs**

As the number of genotyped animals increased, there was an interest in decreasing the cost of computations of both \( G^{-1} \) and \( A_{22}^{-1} \). Faux et al. (2012) attempted to mimic procedures for the efficient creation of \( A^{-1} \) and to create \( G^{-1} \) by recursion on a small number of relatives; however, the method was inaccurate. They also tried to create an efficient inverse of \( A_{22}^{-1} \) by selective absorption (Faux et al., 2013). Consequently, they found \( A_{22}^{-1} \) was no longer sparse with many nongenotyped ancestors, and both storage and computations were demanding.

Legarra & Ducrocq (2012) developed a method where \( G \) was not inverted, but the method did not scale up. Liu et al. (2014) proposed a method involving BV for nongenotyped animals and model with SNP effects estimated for genotyped animals; the model did not converge. Reducing the computational burden of SS required further research.

**Indirect computations of \( A_{22}^{-1} \)**

Stranden et al. (2014) found that the formula for \( A_{22}^{-1} \) involves sparse matrices:
When only a product of $A_{2t}^{-1}$ and a vector is required in the iteration process as in the PCG algorithm, that product can be calculated sequentially every round as follows (Masuda et al., 2016): \( \text{where the product } ; \text{ is computed as a solution to:} \\
Masuda et al. (2017) found that for a U.S. Holstein population, this algorithm required 2 min setup time and less than 1 s per round.

**Single-step Bayesian Regression**

If 50k SNP are enough for predictions, an alternative idea was to impute genotypes of nongenotyped animals, resulting in the same 50k SNP effects to estimate regardless of the number of genotyped animals. Let \( u_2 = Z a \), where \( a \) is a vector of SNP effects. Applying the previous formula:

where \( T \) can be called an imputation matrix for nongenotyped animals and \( \epsilon \) can be called an imputation error. Then, the BV in an animal model can be replaced by:

Regardless of the number of animals, the number of unknowns is equal to the number of SNP, although there is an additional uncorrelated effect \( \epsilon \). As the imputation was expensive, a solution by BOLT used graphical processing units (GPU; Golden et al., 2016). The method, single-step Bayesian Regression (SS-BR), allows for implementation of Bayesian methods for estimation of SNP but is complicated for complex models.

**APY algorithm**

Misztal et al. (2014) noticed that in dairy nearly all genomic information was derived from genotyped proven bulls. He suggested an algorithm for proven (p) and young (y) (APY) where BV of young animals were linear function of BV for proven animals as:

and where the matrix \( P \) was derived from \( G \). The resulting matrix was:

where \( M \) was a diagonal matrix with elements:

and the method had a linear cost (computations and memory) for young animals.

Fragomeni et al. (2015) tested the APY algorithm and found that accuracy was dependent on the number, not the type, of animals in the recursion: recursions on 20k bulls, cows or random animals gave similar accuracy. Ostersen et al. (2016) found the choice of animals in the recursion (called core) critical. In simulation, Bradford et al. (2017) found that choice not critical for accuracy, provided that the pedigrees were complete. However, the choice influenced convergence, and overall the best choice was random.

The theory for the APY algorithm was given by Misztal (2016). He hypothesized that because of a limited dimensionality of the genomic information, say \( k \), \( k \) animals (excluding extreme cases) carry the same information. Assume a vector \( t \) of length \( k \) contains all additive information about the population. Any number of animals greater than \( k \) (excluding extreme cases) will contain a linear combination of \( t \). Call those animals core (c) and the remaining animals noncore (n). Then:
where \( P \) is as before, and \( \varepsilon \) is a small error.

Pocrnic et al. (2016a) simulated multiple populations with different effective population sizes (Ne). He found that: 1) 95% (98%) of the largest eigenvalues approximately equaled \( 2NeL \) (4NeL; \( L \) = genome length) and 2) the accuracy of GEBV peaked when the number of core animals equaled the number of 98% largest eigenvalues. The accuracy was only slightly less for fewer core animals (95% of eigenvalues). Pocrnic et al. (2016b) also analyzed commercial data sets for swine, broilers, beef, and dairy and found similar results. Dimensionality ranged from 5k (pigs and broilers) to 15k (Holsteins). The dimensionality of the genomic information (4NeL) may be equal to the number of distinct haplotypes. Thus, Ne haplotypes would exist for each \( L/4 \) segment.

Inverse by singular value decomposition

The inverse of \( G \) can be derived from the eigenvalue decomposition:

\[
G^{-1} = U D^{-1} U^T
\]

where \( U \) is a matrix of eigenvectors and \( D \) is a matrix of eigenvalues. If all eigenvalues are positive, the inverse of \( G \) is

\[
G^{-1} = \frac{1}{\lambda} U D^{-1} U^T
\]

Following Pocrnic et al. (2016a), \( G \) has a dimensionality less than 20k regardless of the number of animals and SNP. Let \( D_i \) indicate a fraction of \( D \) with non-negligible eigenvalues, and let \( U_i \) be corresponding eigenvectors. Then:

For full rank and stability, small diagonal elements can be added to the inverse, e.g.:

If \( G^{-1} \) is to be created explicitly, the computing cost is cubic with quadratic storage. If \( G^{-1} \) is used by a PCG algorithm as:

and the number of eigenvalues in \( D_i \) is small, \( G^{-1} q \) can be calculated at a linear cost.

While eigenvalue decomposition of \( G \) requires creating \( G \) explicitly and can be very expensive, a less expensive alternative is singular value decomposition (SVD) of gene content \( (Z; \ G = ZZ/\varepsilon) \) when there are more genotyped animals than SNP. Masuda (2016, unpublished) found that the SVD decomposition for a matrix of 720k animals by 60k SNP took less than a day.

Inverse by Woodbury formula

Mätysaari et al. (2017) proposed an inverse of:

\[
Z'Z \varepsilon + \varepsilon I
\]

based on the Woodbury formula:

\[
Z'Z \varepsilon + \varepsilon I
\]

where \( Z'Z \) is the design matrix of SNP BLUP. The formula is an exact inversion but is based on the arbitrary value of \( \varepsilon \), without which \( G \) could not full rank. The "Woodbury" \( G^{-1} \) is dense and is used only for PCG systems in which only a product of this matrix by a vector is desired, reformulated as:
Matrix $S$ has dimensions equal to the number of animals by the number of SNP. In practice, the SNP BLUP design-matrix $Z'Z$ is not full rank, and one dimension can be reduced to the actual rank (5k to 15k for one breed) by truncating $U$ and $D$ to eliminate small eigenvalues:

$$U = \begin{pmatrix} A & B \\ B' & D \\ \end{pmatrix},$$

$$D = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \\ \end{pmatrix}$$

While the rearrangement can substantially reduce the computations and perhaps improve accuracy, a lossless correspondence between $G$ and $G^{-1}$ is no longer possible.

**Estimation of variance components**

Variance components are usually estimated using either REML or Bayesian methods. Genomic information results in additional dense blocks in the MME. In REML, Masuda et al. (2015) showed that the old sparse-matrix package was inefficient and unstable but a new package called YAMS was much more efficient and stable. YAMS recognizes dense blocks in the MME, rearranges computations accordingly, and allows for parallel computations for large dense blocks. Efforts to reduce REML computations by using APY were unsuccessful (unpublished) because the sparse inverse was relatively dense as used for the computation of traces. With Bayesian methods, the use of APY can drastically reduce storage and computations for large genotyped populations. An important concern could be density of, which is nearly dense with very long pedigrees. Cutting pedigrees to a point where is sufficiently sparse needs to be researched to show whether REML estimates are biased.

**Additional Discussion**

**Meaning of reduced dimensionality of genomic information**

Stam (1980) defined the dimensionality of genomic information as $4NeL$, where $Ne$ is effective population size and $L$ is genome length in Morgans. The formula was based on the expected number of junctions in the genome, and the paper is hard to read. The formula can be understood in two not necessarily exclusive interpretations: 1) as $4NeL$ consecutive genomic segments and 2) as $Ne$ haplotypes within each 0.25 M segment. Assume $Ne = 100$, $L = 30$ M and 3 Gb genome length. The first interpretation suggests 12k blocks of about 250 Kb each. These blocks would limit GWAS resolution to large intervals. The second interpretation suggests 100 haplotypes within each 25 Mb interval. As accuracies with APY were only slightly reduced assuming the dimensionality of $2NeL$ or even $Ne$ (Pocrnic et al, 2016b), most haplotypes have either small effect or small frequency. If a single QTL exists within an interval, the dimensionality in that interval is reduced from $Ne$ to 1. This agrees with findings of Fragomeni et al (2017a), where the dimensionality was less when $G$ was weighted by variances obtained from GWAS. A comprehensive study on the topic would provide better insight into the limits of genomic selection.

**APY and persistence of genomic evaluation**

When the number of core animals was approximately $4 NeL$, using core animals four generations before the validation animals reduced the accuracy less than 1% comparing to using core animals from the previous generation (Bradford et al, 2017). For smaller number of animals ($\sim NeL$), the loss was bigger (2%) but still small. This suggests that when the
number of genotyped animals with information is large enough (equivalent to > NeL animals with 0.99 accuracy), the genomic predictions are persistent for many generations. Also, when the number of genotyped animals with information exceeds that equivalent to 4 NeL animals with 0.99 accuracy, there is no further increase of accuracy of genomic prediction. Such findings have been only partially verified with field data (Masuda et al., 2016).

**Single-step with major SNP or causative SNP**

Major SNP or causative SNP can be incorporated in single-step by weighted G or heterogeneous SNP variances for single-trait models. SS-BR can lead directly to Bayes-B type procedures but with a large cost for large data sets. If all causative SNP and their variances are known, Fragomeni et al. (2017a) showed that ssGBLUP has a low cost (dimensionality of genomic information close to the number of causative SNP) if G is blended by a small fraction of the identity matrix. They also have shown that weighting G reduces its rank. Studies suggest that the impact of weighting is smaller with larger populations (Karaman et al., 2016). In a livestock population, Fragomeni et al (2017b) found no improvement with weighting in ssGBLUP (which had the greatest accuracy) but an improvement for GBLUP using de-regressed proofs, especially with homogenous residuals. This suggests that some weighting could reflect problems with de-regression instead of QTL.

In multiple-trait models, weighting G differently by trait in SS is complicated. The extra accuracy through weighted G in single-trait models may not compensate the additional gain of multiple-trait models with unweighted G.

**Scaling**

To avoid inflation and biases, G and A22 need to be compatible (Christensen and Lund, 2010; Vitezica et al., 2011). In practice, this also requires uniform pedigrees, with nonzero inbreeding for unknown parents (Legarra et al., 2015; Misztal et al., 2017; Tsuruta et al., 2017). Matching G and A22 is simpler for methods where explicit G and A22 are available, at least partially.

**Multibreed SS analyses**

Multibreed analyzes can be accommodated in SS as separate relationship matrices for each purebred line with phasing for the crossbreds (e.g., Christensen et al., 2014), or as a single relationship matrix for all breeds and breed combinations. The second choice is simpler but could reduce accuracy. For instance, with 60k SNPs and 10 breeds, only 6 k SNP are available on average per breed.

**Conclusion**

Single-step methods are now available to handle the largest data sets. Modifications have been developed to handle the inversion of G and A22 making SS computationally feasible for the increasing numbers of genotyped animals. The methods are based, implicitly or explicitly, on limited dimensionality of the genomic information and on the sparsity of A.
List of References


