Introduction to Single-step GBLUP

Austin Putz

October 30, 2016

Contents

1 Introduction

Motivation: Single-step GBLUP (ssGBLUP) can combine pedigree, genotypes, and phenotypes for *all* animals into one step.

Meuwissen [13] introduced dense SNP panels to animal breeding in a Bayesian paradigm. They introduced BayesA and BayesB still used today. They both allow SNP specific marker variances to get at a Student's t-distribution (fatter tails) than the normal distribution. However, it's mathematically challenging to obtain marker effects with the t-distribution so it's assumed each SNP effect comes from a normal distribution, but there is a SNP specific variance σ^2_g for each. Therefore if σ^2_g is large then $\hat{g}$, the allele substitution effect, will be large. BayesA allows all markers to have an effect while BayesB sets a π value as the proportion of markers with no effect, which creates a mixture distribution.

For Bayesian methods you need a genotype and a phenotype (or pseudophenotype) on all animals. In many circumstances we do not have genotypes on all animals, but many relatives have phenotypes we would like to include. Or we have genotypes and no phenotypes (think Dairy bulls). It's has been expensive to genotype all animals, so generally only a subset gets genotyped or genotyped with a lower density panel. Previous to single-step GBLUP the procedure looked something like:

1. Run basic pedigree based BLUP
2. Calculate "pseudo-phenotypes" or adjusted phenotypes

   Daughter yield deviations (DYDs) or de-regressed proofs (dEBVs) (see Garrick [7])
3. Use these genotyped and "phenotyped" individuals to calculate marker effects
4. Use to calculate DGVs
5. Combine DGV with pedigree based evaluation to get final GEBV

Note: You can train (estimate SNP effects) without DYDs or dEBVs by using the phenotype, but they are usually pre-adjusted. Most software for Bayesian models do not handle fixed effects so it's actually on the residual of a model after accounting for fixed effects (see de los Campos [5]).

Single-step GBLUP is a way to combine genotyped and non-genotyped animals into one covariance matrix to substitute for the numerator relationship matrix (**A**) in MME equations. The new covariance matrix is refered to as the **H** matrix. Then one can simply substitute the new **H** matrix for the **A** matrix. This is what makes ssGBLUP so easy to implement.

2 Relationship matrices

We will start by building matrices from the beginning with **A**, the numerator relationship matrix. This was followed by the total allelic relationship matrix (**TA**) that became more conceptual. Then the **G** matrix, genomic relationship matrix, was introduced using dense SNP panels. Finally, the **H** matrix was utilized to combine the **A** and **G** matrices.

2.1 A matrix

Henderson 1976 [8] used the numerator relationship matrix (**A**) matrix to be used to link relatives together (i.e. covariances). The **A** matrix uses the expected values based on the pedigree information. This can be poor if pedigrees are (1) incorrect or (2) missing.

The numerator relationship matrix can be calculated with the rules presented by Bourdon:

1. The numerator relationship between X and Y is the average of the numerator relationship of between X and the parents of Y (S and D).

$$r_{xy} = \frac{1}{2}(r_{xs} + r_{xd}) \tag{1}$$

2. The diagonal elements (numerator relationship between an individual and itself) is 1 plus the inbreeding coefficient

(half the relationship of parents).

$$r_{xx} = 1 + F_x = 1 + \frac{1}{2} r_{sd} \tag{2}$$

2.2 TA matrix

With advances in molecular biology it became possible to obtain genotypes (e.g. SNP) for each animal. Nejati-Javaremi 1997 [17] describes the allelic identity at a locus:

$$TA_l = 2\frac{\sum_{i=1}^{2}\sum_{j=1}^{2} I_{ij}}{4} = \frac{\sum_{i=1}^{2}\sum_{j=1}^{2} I_{ij}}{2} \tag{3}$$

Where I_{ij} is the identity of the ith allele of the first individual with jth allele of the second (2 alleles by 2 animals = 2x2 = 4). I_{ij} is 1 if they are identical, otherwise 0. But this is only for 1 locus, therefore we must consider all loci. The total allelic identity (relationship) is averaged over all loci (L):

$$TA_{xy} = \frac{\sum_{l=1}^{L} TA_l}{L} \tag{4}$$

Consider the pedigree (Figure 1) and genotype matrix (**M**) (0,1,2 coding):

Figure 1: Pedigree from Negati-Javaremi et al. 1997

Individual 1: $A_2\ B_1\ C_1\ D_1\ E_2$ / $A_2\ B_1\ C_2\ D_1\ E_2$

Individual 2: $A_1\ B_1\ C_2\ D_1\ E_1$ / $A_2\ B_2\ C_2\ D_2\ E_1$

mated to produce the following offspring:

Individual 3: $A_2\ B_1\ C_1\ D_1\ E_2$ / $A_1\ B_1\ C_2\ D_2\ E_1$

Individual 4: $A_2\ B_1\ C_2\ D_1\ E_2$ / $A_2\ B_1\ C_2\ D_1\ E_1$

Individual 5: $A_2\ B_1\ C_2\ D_1\ E_2$ / $A_1\ B_2\ C_2\ D_1\ E_1$

Individual 6: $A_2\ B_1\ C_2\ D_1\ E_2$ / $A_2\ B_1\ C_2\ D_1\ E_1$

$$\mathbf{M}_{(n\times m)} = \begin{bmatrix} 0 & 2 & 1 & 2 & 0 \\ 1 & 1 & 0 & 1 & 2 \\ 1 & 2 & 1 & 1 & 1 \\ 0 & 2 & 0 & 2 & 1 \\ 1 & 1 & 0 & 2 & 1 \\ 0 & 2 & 0 & 2 & 1 \end{bmatrix} \tag{5}$$

$$\mathbf{A}_{(n\times n)} = \begin{bmatrix} 1 & 0 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0 & 1 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 1 \end{bmatrix} \tag{6}$$

Utilizing the previous equations we can obtain a total allelic relationship matrix (**TA**) using the genotypes from above.

$$\mathbf{TA}_{(n\times n)} = \begin{bmatrix} 1.8 & 0.8 & 1.2 & 1.6 & 1.2 & 1.6 \\ 0.8 & 1.4 & 1.0 & 1.2 & 1.2 & 1.2 \\ 1.2 & 1.0 & 1.2 & 1.2 & 1.0 & 1.2 \\ 1.6 & 1.2 & 1.2 & 1.8 & 1.4 & 1.8 \\ 1.2 & 1.2 & 1.0 & 1.4 & 1.4 & 1.4 \\ 1.6 & 1.2 & 1.2 & 1.8 & 1.4 & 1.8 \end{bmatrix} \tag{7}$$

Lets do an example of how to get the relationship between animal 1 and 2 using the 5 SNP genotypes. The numerator for TA is:

$$\frac{(0+1+0+1)}{2} + \frac{(1+0+1+0)}{2} + \frac{(0+0+1+1)}{2} + \frac{(1+0+1+0)}{2} + \frac{(0+0+0+0)}{2} = 4 \tag{8}$$

And the total allelic relationship (TA) is:

$$TA_{12} = \frac{4}{5} = 0.8 \tag{9}$$

2.3 G Matrix

VanRaden [19] is most commonly cited for the **G** matrix. Matrix **M** is the marker information matrix of order n (number of individuals) by m (number of markers). If elements in **M** are set to -1, 0, and 1, the diagonals count the number of homozygous loci and off-diagonals count the number of alleles shared [19].

Let one row of $\boldsymbol{P}_{n\times m} = 2(p_i - 0.5)$, then $\mathbf{Z}_{(n\times m)}$ = $\mathbf{M}_{(n\times m)} - \mathbf{P}_{(n\times m)}$. So each row is simply filled with 2 × allele frequencies minus 0.5 and repeated across all rows. In this example **M** is coded -1, 0 , 1.

$$\mathbf{G}_{(n\times n)} = \frac{\mathbf{Z}_{(n\times m)}\mathbf{Z'}_{(m\times n)}}{2\sum_{i=1}^{m} p_i(1-p_i)} \tag{10}$$

Where p_i = the allele frequency of that loci. The denominator is used to scale **G** to be analogous to the **A** matrix. There are many ways to scale this matrix, it will not be discussed here. Keep in mind that the allele frequencies should be from the base population, however in practice the current allele frequencies of the

entire population are used instead.

You can also keep **M** 0, 1, and 2 coding and define **P** as $2(p_i)$. Please see Mrode's book [16] for more information.

Another alternative from VanRaden [19] is:

$$\mathbf{G}_{(n\times n)} = \mathbf{Z}_{(n\times m)}\mathbf{D}_{(m\times m)}\mathbf{Z'}_{(m\times n)} \tag{11}$$

with $\mathbf{D}_{(m\times m)}$ being a diagonal with elements

$$D_{ii} = \frac{1}{m(p_i(1-p_i))} \tag{12}$$

where i is the loci from 1, 2, ... m. However **D** can also be defined as a function of the allele substitution effect (α), such as Tiezzi 2015 [18].

$$D_{ii} = w_i = 2p_i(1-p_i)\hat{\alpha}_i^2 \tag{13}$$

2.4 H Matrix

The **H** matrix combines the **G** matrix and **A** matrix. The **H** matrix was derived independent from both Legarra [10] and Christensen and Lund [3]. Legarra [11] describes both derivations briefly.

First we must separate genotyped vs non-genotyped animals. Let n_1 be the number of non-genotyped individuals and n_2 be the number of genotypes individuals. In the **A** matrix and define **A** as:

$$\mathbf{A}_{(n\times n)} = \begin{bmatrix} \mathbf{A}_{11(n_1\times n_1)} & \mathbf{A}_{12(n_1\times n_2)} \\ \mathbf{A}_{21(n_2\times n_1)} & \mathbf{A}_{22(n_2\times n_2)} \end{bmatrix} \tag{14}$$

Where $\mathbf{A}_{11}$ is the numerator relationship matrix of non-genotypes animals, $\mathbf{A}_{12}$ and $\mathbf{A}_{21}$ is the numerator relationship between genotyped and non-genotyped, and $\mathbf{A}_{22}$ is the numerator relationship matrix between genotyped animals. We can define n_1 as the number of non-genoytped animals and n_2 as the number of genotyped animals.

The inverse of **A** can be defined as:

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix} \tag{15}$$

Note the superscript numbers indicating that they are subsections of the *inverse* matrix.

Misztal [14] showed that a simple version of **H** exists.

$$\mathbf{H}_{(n\times n)} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{G} \end{bmatrix} \tag{16}$$

However, this should not be used and instead use the later equations. Aguilar [1] explains that the covariance of the joint distribution of $\mathbf{u}_1$ and $\mathbf{u}_2$, the breeding values of non-genotyped and genotyped individuals respectively, is **H**.

And from Legarra [11] we get **H**:

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{bmatrix} \tag{17}$$

$$\mathbf{H} = \begin{bmatrix} \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{bmatrix} \tag{18}$$

Aguilar [1] figured out a simple equation for $\mathbf{H}^{-1}$

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix} \tag{19}$$

Finally from Lourenco [12] it is clear to see how the matrices and inverses can be weighted.

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \tau(\alpha\mathbf{G} + \beta\mathbf{A}_{22})^{-1} - \omega\mathbf{A}_{22}^{-1} \end{bmatrix} \tag{20}$$

The MME are accomplished by simply plugging in $\boldsymbol{H^{-1}}$ for $\boldsymbol{A^{-1}}$.

$$\begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'Z} \\ \boldsymbol{Z'X} & \boldsymbol{Z'Z} + \boldsymbol{H^{-1}}\lambda \end{bmatrix} \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{u} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'y} \\ \boldsymbol{Z'y} \end{bmatrix} \tag{21}$$

An example is provided in the Legarra's original derivation paper [10] of what happens if some are genotyped and how it changes relationships of other related individuals. Figure 2 shows the pedigree used in the example. The **A** matrix presented in his paper is displayed in Figure 3.

Figure 2: The pedigree from Legarra et al. 2009

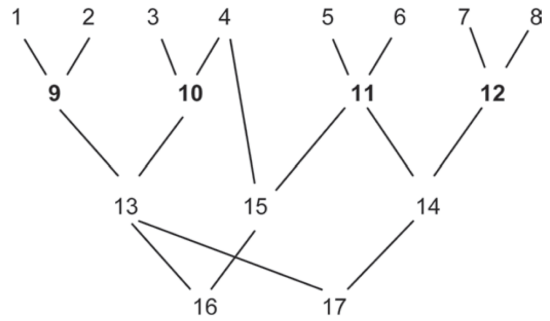

Figure 1. Example pedigree. Genotyped animals are in bold.

The numerator relationship was then calculated for this pedigree in Figure 3.
Utilizing the genomic relationship of animals 9 - 12, **G** was substituted in and changed the relationships "downstream" (Figure 4). This is obviously incorrect because for animals 9 and 10 to be related 0.7, their parents would have to be related (not 0 as in **A**), yet 1 and 2 (parents of 9) are not related (9 has a diagonal of 1). Legarra [10] calls this projecting forward and backward in the pedigree.
Theoretically, this should increase the accuracy for not only genotyped, but non-genotpyed animals. However in practice this may be limited (see Christensen [4])

2.5 Single-step under selection

Vitezica et al. (2011) [20] investigated the effect of selection using single-step GBLUP. The idea being that the **G** matrix does not go back to a base population, the base is genotyped animals. If the best animals are genotyped then there is a problem because there is a difference in average EBV between non-genotyped and genotyped.

Fernando et al. (2014) added a column to the **X** matrix and did this adjustment as a fixed effect. Wether it be fixed or random is beyond the scope of this introduction.

Let μ be a zero or non-zero constant to adjust **G** because of non-random selection of genotyped individuals. And let α be the difference in means between $\boldsymbol{A_{22}}$ and **G**.

$$p(\mu) \sim N(0, \alpha\sigma_u^2) \tag{22}$$

$$p(\boldsymbol{u_2}|\mu) \sim N(\mathbf{1}\mu, \boldsymbol{G}\sigma_u^2) \tag{23}$$

$$p(\boldsymbol{u_2}|\alpha) \sim N(\mathbf{0}, (\boldsymbol{G} + \mathbf{11'}\alpha)\sigma_u^2) \tag{24}$$

$$p(\boldsymbol{u}) \sim N(\mathbf{0}, \boldsymbol{H}^{\dagger}\sigma_u^2) \tag{25}$$

$\boldsymbol{H}^{\dagger}$ is equivalent to **H** with **G** substituted for $\boldsymbol{G} + \mathbf{11'}\alpha$.

$$\begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'Z} \\ \boldsymbol{Z'X} & \boldsymbol{Z'Z} + \boldsymbol{H}^{\dagger -1}\lambda \end{bmatrix} \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{u} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'y} \\ \boldsymbol{Z'y} \end{bmatrix} \tag{26}$$

where:

$$\boldsymbol{H}^{\dagger -1}_{(n \times n)} = \begin{bmatrix} \boldsymbol{A^{11}} & \boldsymbol{A^{12}} \\ \boldsymbol{A^{21}} & \boldsymbol{A^{22}} + (\boldsymbol{G} + \mathbf{11'}\alpha)^{-1} - \boldsymbol{A_{22}^{-1}} \end{bmatrix} \tag{27}$$

$$\alpha = \frac{1}{n^2}\left(\sum_i \sum_j \boldsymbol{A_{22(i,j)}} - \sum_i \sum_j \boldsymbol{G_{(i,j)}}\right) \tag{28}$$

The μ value of genetic effects of genotyped animals ($\boldsymbol{u_2}$) can be written $\mu = \frac{1}{n}\mathbf{1'u_2}$ and is therefore a random variabled due to $\boldsymbol{u_2}$ obviously being random in animal models. This value

Figure 3: The numerator relationship matrix from Legarra et al. 2009

Table 1. Numerator relationship matrix **A** for the pedigree in Figure 1[1]

1.00								0.50				0.25			0.13	0.13
	1.00							0.50				0.25			0.13	0.13
		1.00							0.50			0.25			0.13	0.13
			1.00						0.50			0.25		0.50	0.38	0.13
				1.00						0.50			0.25	0.25	0.13	0.13
					1.00					0.50			0.25	0.25	0.13	0.13
						1.00					0.50		0.25			0.13
							1.00				0.50		0.25			0.13
0.50	0.50							**1.00**				0.50			0.25	0.25
		0.50	0.50						**1.00**			0.50		0.25	0.38	0.25
				0.50	0.50					**1.00**			0.50	0.50	0.25	0.25
						0.50	0.50				**1.00**		0.50			0.25
0.25	0.25	0.25	0.25					0.50	0.50			1.00		0.13	0.56	0.50
				0.25	0.25	0.25	0.25			0.50	0.50		1.00	0.25	0.13	0.50
			0.50	0.25	0.25			0.25	0.50			0.13	0.25	1.00	0.56	0.19
0.13	0.13	0.13	0.38	0.13	0.13			0.25	0.38	0.25		0.56	0.13	0.56	1.06	0.34
0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.25	0.25	0.25	0.25	0.50	0.50	0.19	0.34	1.00

[1]Cells with 0 are empty to show the pattern. Coefficients for genotyped animals are in bold. Matrix $\mathbf{A}_g$ is obtained by setting the out-of-diagonal coefficients of genotyped animals to 0.7.

Figure 4: The relationship matrix after substituting **G** into the matrix from Legarra et al. 2009

Table 2. Modified relationship matrix $\mathbf{A}_p$ including genomic information for genotyped animals and their progeny for the pedigree in Figure 1[1]

1.00								0.50				0.25			0.13	0.13
	1.00							0.50				0.25			0.13	0.13
		1.00							0.50			0.25			0.13	0.13
			1.00						0.50			0.25		0.50	0.38	0.13
				1.00						0.50			0.25	0.25	0.13	0.13
					1.00					0.50			0.25	0.25	0.13	0.13
						1.00					0.50		0.25			0.13
							1.00				0.50		0.25			0.13
0.50	0.50							**1.00**	**0.70**	**0.70**	**0.70**	0.85	0.70	0.35	0.60	0.78
		0.50	0.50					**0.70**	**1.00**	**0.70**	**0.70**	0.85	0.70	0.60	0.73	0.78
				0.50	0.50			**0.70**	**0.70**	**1.00**	**0.70**	0.70	0.85	0.50	0.60	0.78
						0.50	0.50	**0.70**	**0.70**	**0.70**	**1.00**	0.70	0.85	0.35	0.53	0.78
0.25	0.25	0.25	0.25					0.85	0.85	0.70	0.70	1.35	0.70	0.48	0.91	1.03
				0.25	0.25	0.25	0.25	0.70	0.70	0.85	0.85	0.70	1.35	0.43	0.56	1.03
			0.50	0.25	0.25			0.35	0.60	0.50	0.35	0.48	0.43	1.00	0.74	0.45
0.13	0.13	0.13	0.38	0.13	0.13			0.60	0.73	0.60	0.53	0.91	0.56	0.74	1.33	0.74
0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.78	0.78	0.78	0.78	1.03	1.03	0.45	0.74	1.53

[1]Cells with 0 are empty to show the pattern. Coefficients for genotyped animals are in bold.

Figure 5: The correct **H** matrix from Legarra et al. 2009

Table 3. Modified relationship matrix **H** including genomic information for genotyped animals and all relatives for the pedigree in Figure 1[1]

1.00		0.18	0.18	0.18	0.18	0.18	0.18	0.50	0.35	0.35	0.35	0.43	0.35	0.26	0.34	0.39
	1.00	0.18	0.18	0.18	0.18	0.18	0.18	0.50	0.35	0.35	0.35	0.43	0.35	0.26	0.34	0.39
0.18	0.18	1.00		0.18	0.18	0.18	0.18	0.35	0.50	0.35	0.35	0.43	0.35	0.18	0.30	0.39
0.18	0.18		1.00	0.18	0.18	0.18	0.18	0.35	0.50	0.35	0.35	0.43	0.35	0.68	0.55	0.39
0.18	0.18	0.18	0.18	1.00		0.18	0.18	0.35	0.35	0.50	0.35	0.35	0.43	0.34	0.34	0.39
0.18	0.18	0.18	0.18		1.00	0.18	0.18	0.35	0.35	0.50	0.35	0.35	0.43	0.34	0.34	0.39
0.18	0.18	0.18	0.18	0.18	0.18	1.00		0.35	0.35	0.35	0.50	0.35	0.43	0.26	0.31	0.39
0.18	0.18	0.18	0.18	0.18	0.18		1.00	0.35	0.35	0.35	0.50	0.35	0.43	0.26	0.31	0.39
0.50	0.50	0.35	0.35	0.35	0.35	0.35	0.35	**1.00**	**0.70**	**0.70**	**0.70**	0.85	0.70	0.53	0.69	0.78
0.35	0.35	0.50	0.50	0.35	0.35	0.35	0.35	**0.70**	**1.00**	**0.70**	**0.70**	0.85	0.70	0.60	0.73	0.78
0.35	0.35	0.35	0.35	0.50	0.50	0.35	0.35	**0.70**	**0.70**	**1.00**	**0.70**	0.70	0.85	0.68	0.69	0.78
0.35	0.35	0.35	0.35	0.35	0.35	0.50	0.50	**0.70**	**0.70**	**0.70**	**1.00**	0.70	0.85	0.53	0.61	0.78
0.43	0.43	0.43	0.43	0.35	0.35	0.35	0.35	0.85	0.85	0.70	0.70	1.35	0.70	0.56	0.96	1.03
0.35	0.35	0.35	0.35	0.43	0.43	0.43	0.43	0.70	0.70	0.85	0.85	0.70	1.35	0.60	0.65	1.03
0.26	0.26	0.18	0.68	0.34	0.34	0.26	0.26	0.53	0.60	0.68	0.53	0.56	0.60	1.18	0.87	0.58
0.34	0.34	0.30	0.55	0.34	0.34	0.31	0.31	0.69	0.73	0.69	0.61	0.96	0.65	0.87	1.41	0.80
0.39	0.39	0.39	0.39	0.39	0.39	0.39	0.39	0.78	0.78	0.78	0.78	1.03	1.03	0.58	0.80	1.53

[1]Cells with 0 are empty to show the pattern. Coefficients for genotyped animals are in bold.

can be estimated with both pedigree and single-step, μ_p and μ_s respectively. The equations are:

$$\mu_s \sim N\left(0, \frac{1}{n^2}\mathbf{1}'(\boldsymbol{G}+\mathbf{1}\mathbf{1}'\alpha)\mathbf{1}\sigma_u^2\right) \tag{29}$$

$$var(\mu_p) = \sigma_u^2\left(\frac{1}{n^2}\sum_i\sum_j \boldsymbol{A_{22(i,j)}}\right) \tag{30}$$

$$var(\mu_s) = \sigma_u^2\left(\alpha + \frac{1}{n^2}\sum_i\sum_j \boldsymbol{G(i,j)}\right) \tag{31}$$

The equivalent model from the appendix using the genetic group model from Quaas (1988):

$$\boldsymbol{Q}_{(n\times 1)} = \begin{bmatrix} (\boldsymbol{A_{12}A_{22}^{-1}}\mathbf{1})_{(n_1\times 1)} \\ \mathbf{1}_{(n_2\times 1)} \end{bmatrix} \tag{32}$$

$$\begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'Z} & \mathbf{0} \\ \boldsymbol{Z'X} & \boldsymbol{Z'Z}+\boldsymbol{H}^{\dagger -1}\lambda & -\boldsymbol{H^{-1}Q}\lambda \\ \mathbf{0} & -\boldsymbol{Q'H^{-1}}\lambda & -\boldsymbol{Q'H^{-1}Q}\lambda+\alpha^{-1}\lambda \end{bmatrix}\begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{u} \\ \mu \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'y} \\ \boldsymbol{Z'y} \\ 0 \end{bmatrix} \tag{33}$$

This can also be written in an alternative form.

$$\begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'Z} \\ \boldsymbol{Z'X} & \boldsymbol{Z'Z}+\boldsymbol{H}^{*-1}\lambda \end{bmatrix}\begin{bmatrix} \boldsymbol{b} \\ \theta \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'y} \\ \boldsymbol{Z'y} \end{bmatrix} \tag{34}$$

where $\mathbf{Z}$ is expanded with 0's. The updated $\mathbf{H}$ matrix, $\boldsymbol{H^*}$ is equal to:

$$\boldsymbol{H}^{*-1}_{(n\times n)} = \begin{bmatrix} \boldsymbol{H^{11}} & \boldsymbol{H^{12}} & \mathbf{0} \\ \boldsymbol{H^{21}} & \boldsymbol{H^{22}} & -\boldsymbol{G^{-1}}\mathbf{1} \\ \mathbf{0} & -\mathbf{1}'\boldsymbol{G^{-1}} & \mathbf{1}'\boldsymbol{G^{-1}}\mathbf{1}+\alpha^{-1} \end{bmatrix} \tag{35}$$

The alternative form is:

$$\boldsymbol{H}^{*-1}_{(n\times n)} = \begin{bmatrix} \boldsymbol{A^{11}} & \boldsymbol{A^{12}} & \mathbf{0} \\ \boldsymbol{A^{21}} & \boldsymbol{A^{22}}+\boldsymbol{G^{-1}}-\boldsymbol{A_{22}^{-1}} & -\boldsymbol{G^{-1}}\mathbf{1} \\ \mathbf{0} & -\mathbf{1}'\boldsymbol{G^{-1}} & \mathbf{1}'\boldsymbol{G^{-1}}\mathbf{1}+\alpha^{-1} \end{bmatrix} \tag{36}$$

There are still alternatives to this in the paper, but they will not be covered here. Please refer to the paper for more details.

2.6 A and G compatibility

The genomic relationship matrix needs to be compatible with the numerator relationship matrix ($\mathbf{A}$) matrix. VanRaden [19] says that dividing **MM'** by $\sum_{i=1}^{m} 2pq$ will standardize $\mathbf{G}$ to $\mathbf{A}$. There are many ways to standardize $\mathbf{G}$. Keep in mind this can be trait specific. Please see Chen [2], Vitezica [20], and Lourenco [12] for more details.

2.7 Single-step derivation

Two separate derivations of the $\mathbf{H}$ matrix exist in the literature. They were essentially derived at the same time, but Legarra's was published first in JDS (see Legarra [10]). Christensen and Lund (2010) published their derivation in GSE (see Christensen and Lund [3]). Both of these were described more briefely in Legarra's review paper of single-step (see Legarra [11]).

3 Imputation

As mentioned earlier, genotyping is expensive so generally rely on lower density SNP panels for less important animals. Then we can impute from the lower density panel to the higher density panel. This can be implemented by first genotyping heavily with the higher density panel and then using impuation software such as Beagle (Website) or AlphaImpute (see Hickey's Website).

Yijian Huang (at SPG currently) has a nice paper on imputation in swine populations [9]. They examined genotyping strategies to maximize rate of genetic gain (or imputation accuracy) and mimimizing cost.

4 Quality Control

Quality control is an important aspect of the **G** and therefore **H** matrix. This is extremely important for genomic analyses and should not looked over. The following is a list of quality control measures (with default) in the BLUPF90 programs (explained later)

- Minor Allele Frequency (MAF) (0.05)

  Monomorphic SNP (1 to ignore)
- Call rate of SNPs and Individuals (0.90 for both)
- Parent-progeny conflicts (SNPs and animals)

The last one (conflicts) is set with an option for the *percentage* of all SNP that are in conflict, default is 1%. The following is an explaination of a parent-progeny conflict.

4.1 Example of parent-progeny conflict

Say sire S is a parent to individual X according to the pedigree. If S has genotype A_1A_1 and X has genotype A_2A_2, this is obviously not possible. Sire S would have to give one A_1 allele to his offspring X. In this case you would have to reject that sire S is the truly the sire of animal X. In the following section, the default is to remove or flag animals that have 1% of genotypes that are in error. This can become a problem with imputed genotypes to the closest 0, 1, or 2 and should be cautioned. Trios (sire, dam, and offspring) with all high density genotypes would be ideal, but rarely seen so far in swine.

5 BLUPF90 Programs

The University of Georgia (UGA) has created a suite of programs for mixed model equations in animal breeding. They have now implemented ssGBLUP with the help of Ignacio Aguilar. You can visit the UGA Website. The following programs are available:

- BLUPF90
- REMLF90
- AIREMLF90
- GIBBS1F90
- GIBBS2F90
- GIBBS3F90
- POSTGIBBS1F90
- THRGIBBS1F90
- RENUMF90 - recodes pedigree, data, and parameter file
- INBLUPF90
- SEEKPARENTF90

Genomic programs include:

- PREGSF90
- POSTGSF90

Although PREGSF90 is built into the application programs.

They are are very easy to start using with a little help. Documentation is very poor, but there is a Yahoo group that is *usually* helpful, but like any mailing list, sometimes people have a short temper.

Figure 6: The Table of accuracies comparing ssGBLUP to Bayes methods in Legarra et al. 2014

Table 1
Accuracy of Single Step versus other methods in some species.

Authors	Single step	Multiple step	Pedigree BLUP	Species, trait
Aguilar et al. (2010)	0.70	0.70	0.60	Dairy cattle, final score
Baloche et al. (2014)	0.47	0.43	0.32	Milk yield, dairy sheep
C.Y. Chen et al. (2011)[a]	0.36		0.20	Breast meat, chicken
C. Chen et al. (2011)	0.37	0.09	0.28	Leg Score, chicken
Christensen et al. (2012)[a]	0.35	0.35	0.18	Daily gain, pigs
Aguilar et al. (2011)	0.39		0.26	Conception rate at first parity

[a] Predictive abilities: $r(y, \hat{u})$.

6 Comparisons to Bayesian methodology

Comparisons between single-step and Bayesian methodology has been reviewed by Legarra [10]. The following is a Table out of his paper (Table 6).

In general it can be stated that without a large effect QTL, Bayesian methods and ssGBLUP are equivalent methods in terms of accuracy and bias. Slight variations can depend on species, amount of data, and genetic architecture of the trait in a specific population.

Currently, a newer Bayesian single step has been proposed by Fernando et al. (2014) [6].

7 Problems with ssGBLUP

The most commonly cited argument against using this methodology is when **G** gets large (n > p) it will become computationally infeasible to invert the **G** matrix used in the MME equations. However, this has been addressed by Misztal in his paper [15] using recursion to get the inverse of **G** directly. Then using solving methods, such as PCG, you can solve the MME equations. With advances being made in computers (see HP's new Machine, Website) we will see how much this impacts evaluations in the future, but it *could* become a problem.

The other very common problem with single-step is convergence. When the dairy industry tried single-step they struggled a lot (personal communication, Chris Maltecca). Personally, I have had a huge problem with convergence in AIREML and basic EMREML. Models are highly sensitive and rarely converge. The consensus so far has been to use the variance component estimates from the pedigree evaluations instead. Another problem relates to the so called "missing heritability", which is when the heritability decreases pedigree and genomic evaluations (GBLUP or ssGBLUP).

8 Implementation in the Industry

The review by Legarra [11] discusses implementation thus far (up to 2014) along with my own personal communication with swine companies.

So far, probably the largest implementor of single-step GBLUP is the swine industry. This includes, but not limited to: PIC, SPG, TOPIGS, and soon The Maschhoffs. At least two are using the UGA programs, it is unknown about the other companies.
In dairy cattle, national evaluations are still using multiple step procedures, but it seems as if there are many in the dairy industry trying single-step.
It is unknown what is being implmented in poultry, but Anna Wolc (at ISU working for Hy-Line) has published a paper comparing methods (see Wolc [21]) so they are at least looking into them.
According to Legarra's 2014 paper, there were no studies with real beef data. But one from Lourenco et al. (2013) using simulated data. Much is still ran from breed associations.
France is using single-step with corrections to make **G** and **A** compatible for dairy sheep[11]. France is testing ssGBLUP for the diary goat [11].

References

[1] I. Aguilar, I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. Single step, a general approach for genomic selection. *J. Dairy. Sci.*, 93:743–752, 2010.

[2] C. Y. Chen, I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. Effect of different genomic relationship matrices on accuracy and scale. *J. Anim. Sci*, 89:2673–2679, 2011.

[3] O. F. Christensen and M. S. Lund. Genomic prediction when some animals are not genotyped. *Gen. Sel. Evol.*, 42(2), 2010.

[4] O. F. Christensen, P. Madsen, B. Nielsen, T. Ostersen, and G. Su. Single-step methods for genomic evaluation in pigs. *Animal*, 6(10), 2012.

[5] G. de los Campos, J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. L. Calus. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, 193:327–345, 2013.

[6] R. L. Fernando, Jack C. M. Dekkers, and D. J. Garrick. A class of bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. *Gen. Sel. Evol.*, 46(50), 2014.

[7] D. J. Garrick, J. F. Taylor, and R. L. Fernando. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Gen. Sel. Evol.*, 41(55), 2009.

[8] C. R. Henderson. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics*, 32(1):69–83, 1976.

[9] Y. Huang, J. M. Hickey, M. A. Cleveland, and C. Maltecca. Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. *G. Sel. Evol.*, 44(25), 2012.

[10] A. Legarra, I. Aguilar, and I. Misztal. A relationship matrix including full pedigree and genomic information. *J. Dairy. Sci.*, 92:4656–4663, 2009.

[11] A. Legarra, O. F. Christensen, I. Aguilar, and I. Misztal. Single step, a general approach for genomic selection. *Livest. Sci.*, 166:54–65, 2014.

[12] D. A. L. Lourenco, I. Misztal, S. Tsuruta, T. J. Lawlor, S. Forni, and J. I. Weller. Are evaluations on young genotyped animlas benefiting from the past generations. *J. Dairy. Sci.*, 97:3930–3942, 2014.

[13] T. H. E. Meuwissen, B. J. Hayes, and M. E. Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157:1819–1829, 2001.

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

[15] I. Misztal, A. Legarra, and I. Aguilar. Using recursion to compute the inverse of the genomic relationship matrix. *J. Dairy. Sci.*, 97:3943–3952, 2014.

[16] R. A. Mrode. *Linear models for the prediction of animal breeding values*. CABI, Wallingford, Oxfordshire OX10 8DE, UK, 2014.

[17] A. Nejati-Javaremi, C. Smith, and J. P. Gibson. Effect of total allelic relationship on accuracy of evaluation and response to selection. *J. Anim. Sci.*, 75:1738–1745, 1997.

[18] F. Tiezzi and C. Maltecca. Accounting for trait architecture in genomic predictions of us holstein cattle using a weighted realized relationship matrix. *Gen. Sel. Evol.*, 47(24), 2015.

[19] P. M. VanRaden. Efficient methods to compute genomic predictions. *J. Dairy. Sci.*, 91:4414–4423, 2008.

[20] Z. G. Vitezica, I. Aguilar, I. Misztal, and A. Legarra. Bias in genomic predictions

for populaitons under selection. *Genet. Res. Camb.*, 93:357–366, 2011.

[21] A. Wolc, C. Stricker, J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, R. Preisinger, D. Habier, R. Fernando, D. J. Garrick, S. J. Lamont, and J. C. M. Dekkers. Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model. *Gen. Sel. Evol.*, 43(5), 2011.