Deflated preconditioned conjugate gradient method for solving single-step single nucleotide polymorphism BLUP

J. Vandenplas[1], H. Eding[2], M.P.L. Calus[1] & C.Vuik[3]

[1] Animal Breeding and Genomics, Wageningen University & Research, P.O. Box 338, 6700 AH Wageningen, The Netherlands
jeremie.vandenplas@wur.nl (Corresponding Author)
[2] CRV BV, Wassenaarweg 20, 6843 NW Arnhem, the Netherlands
[3] DIAM, TU Delft, Mekelweg 4, 2628 CD Delft, the Netherlands

Summary

A deflated preconditioned conjugate gradient (DPCG) method was implemented for solving iteratively the system of hybrid single-step single nucleotide polymorphism Best Linear Unbiased Prediction (ssSNPBLUP). Using a small dataset, we showed the presence of large unfavourable eigenvalues associated with a poorly-conditioned preconditioned coefficient matrix of hybrid ssSNPBLUP. These large unfavourable eigenvalues were deflated with the DPCG method, which improved the conditioning and reduced the number of iterations by up to a factor 6, in comparison to the traditionally used PCG method.

Keywords: single-step, genomic, convergence, deflation, preconditioned conjugate gradient

Introduction

Single-step genomic evaluations simultaneously combine phenotypic and pedigree information of genotyped and non-genotyped animals with genomic information. Currently, national datasets may contain up to 1 million genotypes composed of several thousands of single nucleotide polymorphisms (SNP). In addition to single-step genomic Best Linear Unbiased Prediction (ssGBLUP) that fits breeding values as a random effect (Legarra et al., 2014), different equivalent models that explicitly fit SNP covariates as random effects (single-step SNPBLUP; ssSNPBLUP), were proposed to deal with such large datasets (Legarra and Ducrocq, 2012; Fernando et al., 2016; Mäntysaari and Strandén, 2016; Taskinen et al., 2017).

An iterative method that is usually implemented in animal breeding for solving large systems of linear equations, is the preconditioned conjugate gradient (PCG) method (Strandén and Lidauer, 1999). It seems therefore a logical choice for solving systems of ssSNPBLUP. However, in comparison to ssGBLUP, PCG convergence issues were reported for different types of ssSNPBLUP (Manzanilla Pech, 2017; Taskinen et al., 2017). Therefore, the primary aim of this study was to compare properties of the coefficient matrices of ssSNPBLUP and ssGBLUP, and to relate this to observed convergence patterns. The second aim was to implement a deflated PCG method (DPCG) for solving ssSNPBLUP.

Material and methods

Hybrid ssSNPBLUP

In the following, the subscripts *g* and *n* refer to genotyped and non-genotyped animals, respectively. In this study, we investigated a hybrid ssSNPBLUP model that fits SNP effects (**g**) and a residual polygenic effect ($\mathbf{a}_g$) for genotyped animals, and an additive genetic effect ($\mathbf{u}_n$) for the non-genotyped animals (Legarra and Ducrocq, 2012; Fernando et al., 2016; Mäntysaari and Strandén, 2016). The univariate mixed model for hybrid ssSNPBLUP is:

$$\mathbf{y} = \mathbf{Xb} + \begin{bmatrix} \mathbf{W}_n & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_g & \mathbf{W}_g\mathbf{Z} \end{bmatrix} \begin{bmatrix} \mathbf{u}_n \\ \mathbf{a}_g \\ \mathbf{g} \end{bmatrix} + \mathbf{e} \qquad (1)$$

where **y** is the vector of records, **b** is the vector of fixed effects, and **e** is the vector of residuals. The matrices **X**, $\mathbf{W}_n$, and $\mathbf{W}_g$ are incidence matrices relating records to the corresponding effects. The matrix **Z** is the matrix of SNP genotypes centered by their observed means. Without loss of generality, we assumed diagonal (co)variance structures for the residual and SNP effects.

The system of equations of the hybrid ssSNPBLUP model (1) is as follows:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}_n'\mathbf{W}_n & \mathbf{X}_g'\mathbf{W}_g & \mathbf{X}_g'\mathbf{W}_g\ \mathbf{Z} \\ \mathbf{W}_n'\mathbf{X}_n & \mathbf{W}_n'\mathbf{W}_n\ + \mathbf{A}^{nn}\gamma & \mathbf{A}^{ng}\gamma & \mathbf{A}^{ng}\mathbf{Z}\gamma \\ \mathbf{W}_g'\mathbf{X}_g & \mathbf{A}^{gn}\gamma & \mathbf{W}_g'\mathbf{W}_g\ + \mathbf{A}^{gg}\frac{\gamma}{w} + \left(1 - \frac{1}{w}\right)\mathbf{Q}\gamma & \mathbf{W}_g'\mathbf{W}_g\ \mathbf{Z} + \mathbf{QZ}\gamma \\ \mathbf{Z}'\mathbf{W}_g'\mathbf{X}_g & \mathbf{Z}\mathbf{A}^{gn}\gamma & \mathbf{Z}'\mathbf{W}_g'\mathbf{W}_g\ + \mathbf{Z}'\mathbf{Q}\gamma & \begin{matrix}\mathbf{Z}'\mathbf{W}_g'\mathbf{W}_g\ \mathbf{Z}' + \\ \mathbf{Z}'\mathbf{QZ}'\gamma + \mathbf{I}\frac{m\gamma}{1-w}\end{matrix} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}}_n \\ \hat{\mathbf{a}}_g \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}_n'\mathbf{y} \\ \mathbf{W}_g'\mathbf{y} \\ \mathbf{Z}'\mathbf{W}_g'\mathbf{y} \end{bmatrix}$$

where γ is the ratio of the residual variance (σ_e^2) and the additive genetic variance (σ_u^2); w is the proportion of residual polygenic effect (0.05 in this study); m is twice the sum of the products between the allele frequency and its counter-part for each SNP; **I** is an identity matrix; $\mathbf{A}^{nn}$, $\mathbf{A}^{gn}$, $\mathbf{A}^{ng}$, and $\mathbf{A}^{gg}$ are submatrices of the inverse of the pedigree relationship matrix $\mathbf{A}^{-1}$; and the matrix **Q** is equal to $\mathbf{A}^{gn}(\mathbf{A}^{nn})^{-1}\mathbf{A}^{ng}$.

PCG and effective spectral condition number

In animal breeding, systems of linear equations, $\mathbf{Cx} = \mathbf{b}$, are usually solved iteratively using the PCG method (Strandén and Lidauer, 1999; see Table 1 for algorithm) as:

$$\mathbf{M}^{-1}\mathbf{Cx} = \mathbf{M}^{-1}\mathbf{b} \qquad (2)$$

where $\mathbf{M}^{-1}$ is the preconditioner, **C** is the coefficient matrix, **x** is the vector of solutions, and **b** is the right-hand-side vector.

The effective spectral condition number of a positive semi-definite matrix (e.g., **C**), denoted $\kappa(\mathbf{C})$, is defined as the ratio of its largest and smallest positive eigenvalues (Nabben and Vuik, 2006). Because convergence of CG methods highly depends on the effective condition number of the coefficient matrix, the preconditioner $\mathbf{M}^{-1}$ aims to improve the condition number from $\kappa(\mathbf{C})$ to $\kappa(\mathbf{M}^{-1}\mathbf{C})$. The smaller $\kappa(\mathbf{M}^{-1}\mathbf{C})$ is, the more well-conditioned $\mathbf{M}^{-1}\mathbf{C}$ is, which is expected to result in faster convergence of the PCG method.

Deflated PCG

The DPCG method is a two-level PCG method to iteratively solve ill-conditioned linear systems. The DPCG method treats the unfavourable eigenvalues of the set of eigenvalues, called spectrum, of $\mathbf{M}^{-1}\mathbf{C}$ that may remain after preconditioning **C** (Nicolaides, 1987; Vuik et al., 1999). In DPCG, the preconditioned system (2) is transformed to the following system:

$$\mathbf{M}^{-1}\mathbf{PCx} = \mathbf{M}^{-1}\mathbf{Pb}$$

where $\mathbf{P} = \mathbf{I} - \mathbf{CZ}_d(\mathbf{Z}_d'\mathbf{CZ}_d)^{-1}\mathbf{Z}_d'$ is a second-level preconditioner, called deflation matrix, and $\mathbf{Z}_d$ is called deflation-subspace matrix.

The deflation-subspace matrix $\mathbf{Z}_d$ contains k columns, called 'deflation vectors', that should approximate the same space as the span of the unfavourable eigenvectors. A proposed approach to set up deflation vectors is the subdomain deflation approach (Frank and Vuik, 2001). This approach divides the computational domain $\mathbb{R}^n$ (with $\mathbf{x} \in \mathbb{R}^n$) into k subdomains, with each i^{th} ($i = 1, ..., k$) subdomain corresponding to the i^{th} deflation vector. An entry of the deflation vector $\mathbf{z}_i$ is equal to 1 if the corresponding entry is included in the i^{th} subdomain; otherwise the entry of $\mathbf{z}_i$ is equal to 0. Therefore, each row of $\mathbf{Z}_d$ contains only one non-zero element. Advantages of the subdomain deflation approach are that $\mathbf{Z}_d$ is sparse, that additional computations for DPCG that involve the deflation matrix (in comparison to PCG) can be implemented efficiently, and that it gives good results if k is large enough (Frank and Vuik, 2001). An algorithm for DPCG is given in Table 1.

Data and model

Datasets and variance components were extracted from the Dutch single-step genomic evaluation for ovum pick-up (OPU) and embryo transfer of dairy cattle (Cornelissen et al., 2017). After extraction, the data file included information for 61,592 OPU sessions from 4,109 animals, and the pedigree included 37,021 animals. Genotypes of 6,169 animals without phenotype were available. Due to computational limits, genotypes included 9,994 segregating SNP randomly sampled from (imputed) 50K SNP genotypes. The univariate mixed model included random effects (additive genetic, permanent environmental, and residual), fixed covariables (heterosis and recombination), and fixed cross-classified effects (herd-year, year-month, parity, age (in months), technician, assistant, interval, gestation, session, and protocol).

Statistical analyses

The hybrid ssSNPBLUP was compared against ssGBLUP. Both systems of hybrid ssSNPBLUP and ssGBLUP were solved using the PCG method. Additionally, the system of hybrid ssSNPBLUP was also solved using the DPCG method with different subdomain decompositions. For setting up the deflation-subspace matrix $\mathbf{Z}_d$, we divided the hybrid ssSNPBLUP domain as follows: (1) all cross-classified fixed effects were included in the same subdomain, to avoid singular $\mathbf{Z}_d'\mathbf{C}\mathbf{Z}_d$; (2) covariables and random effects were included in their own subdomain; and (3) each set of m consecutive SNPs were included in a same subdomain. Sets of m = 1, 5, 50, and 200 SNPs were used. For both PCG and DPCG methods, a Jacobi preconditioner, defined as $\mathbf{M} = diag(\mathbf{C})$, was used (Strandén and Lidauer, 1999). All systems were iterated until squared relative residual norm reached criteria of 10^{-12}.

The spectrum of $\mathbf{M}^{-1}\mathbf{C}$ of ssGBLUP, of $\mathbf{M}^{-1}\mathbf{C}$ of hybrid ssSNPBLUP, and of $\mathbf{M}^{-1}\mathbf{PC}$ of hybrid ssSNPBLUP with 1 SNP per subdomain, were computed using Intel(R) Math Kernel Library (MKL) 11.3 subroutines. To avoid expensive computation of eigenvalues, only the smallest and largest positive eigenvalues of $\mathbf{M}^{-1}\mathbf{PC}$ of hybrid ssSNPBLUP with 5, 50, and 200 SNPs per subdomain were estimated using the Lanczos algorithm based on information obtained from the PCG method (Paige and Saunders, 1975). The different $\kappa(\mathbf{M}^{-1}\mathbf{C})$ and $\kappa(\mathbf{M}^{-1}\mathbf{PC})$ were computed for comparing the different systems.

Results and discussion

The number of equations was equal to 41,949 for ssGBLUP and to 51,943 for hybrid ssSNPBLUP. Figure 1 shows the spectrum of $\mathbf{M}^{-1}\mathbf{C}$ of ssGBLUP and of hybrid ssSNPBLUP, and the spectrum of $\mathbf{M}^{-1}\mathbf{PC}$ of hybrid ssSNPBLUP with 1 SNP per subdomain. All eigenvalues lower than 10^{-10} were set to 10^{-10}, and were considered as zero eigenvalues. Similar patterns for the different spectra were observed. The smallest positive eigenvalues of the different $\mathbf{M}^{-1}\mathbf{C}$ and $\mathbf{M}^{-1}\mathbf{PC}$ were equal to $1.1*10^{-4}$, whatever the model or the definition of subdomains. The largest eigenvalue of $\mathbf{M}^{-1}\mathbf{C}$ was equal to 11.9 for ssGBLUP, and to 181.0 for hybrid ssSNPBLUP (Table 2). The increase of the largest eigenvalues can therefore be attributed to the additional SNP equations fitted in hybrid ssSNPBLUP. When deflation was applied, the largest eigenvalue of $\mathbf{M}^{-1}\mathbf{PC}$ varied from 6.0 with 1 or 5 SNPs per subdomain to 99.4 with 200 SNPs per subdomain. Figure 1 also shows that the largest eigenvalues of $\mathbf{M}^{-1}\mathbf{C}$ of hybrid ssSNPBLUP disappeared from the spectrum of $\mathbf{M}^{-1}\mathbf{PC}$. These results show that the unfavourable largest eigenvalues were removed from the spectrum of $\mathbf{M}^{-1}\mathbf{C}$ of hybrid ssSNPBLUP, without affecting the rest of the spectrum. The deflation vectors spanned thereby approximately the same space as the span of the eigenvectors corresponding to the largest eigenvalues of $\mathbf{M}^{-1}\mathbf{C}$. Consequences of deflation were that the condition number of hybrid ssSNPBLUP decreased from $1.7*10^{6}$ to between $5.4*10^{4}$ with 1 SNP per subdomain and $9.3*10^{5}$ with 200 SNPs per subdomain. Smaller condition numbers means faster convergence for a CG solver.

All iterative solvers, PCG and DPCG, converged to the same solutions for all linear systems of ssGBLUP and hybrid ssSNPBLUP. When the PCG method was used, the number of iterations to reach convergence for hybrid ssSNPBLUP was more than 5 times higher than the number of iterations for ssGBLUP (Table 2; Figure 2). However, when the DPCG method with 1 or 5 SNPs per subdomain was used, the number of iterations decreased by a factor 6, and was slightly less than the one for ssGBLUP. Fifty and 200 SNPs per subdomain also led to a decrease of the number of iterations by a factor 1.25 and 1.70, respectively. Decreases of number of iterations with DPCG were in agreement with the condition numbers (Table 2). Figure 2 illustrates convergence by iteration for the PCG and DPCG methods. A flat pattern is observed for the PCG method applied on hybrid ssSNPBLUP. The DPCG method allowed to reduce this flat pattern to finally observe a pattern similar to the one of ssGBLUP.

The reduction of the number of iterations by the DPCG method can be performed at a relatively low additional computational cost, compared to the decrease of number of iterations, because subdomain deflation allows the use of sparse matrices (e.g., $\mathbf{Z}_{d}$) and parallelisation. The DPCG method could be also implemented for other ssSNPBLUP models for which similar convergence issues were observed. Future studies will investigate the DPCG method on large datasets and applied in multivariate models.

Conclusion

We implemented a DPCG method for solving iteratively a system of linear equations of hybrid ssSNPBLUP. We showed that the DPCG method treated the largest unfavourable eigenvalues of the preconditioned coefficient matrix, and reduced the number of iterations by up to a factor 6, in comparison to the PCG method.

Table 1. Algorithm for preconditioned conjugate gradient (PCG) and deflated PCG (DPCG) methods[1].

```
Select an initial guess $\mathbf{x}_0$; $\mathbf{r}_{init} = \mathbf{b} - \mathbf{C}\mathbf{x}_0$ ; $\mathbf{r}_0 = \mathbf{\Psi}\mathbf{r}_{init}$ ; $\mathbf{p}_{-1} = \mathbf{0}$ ; $\tau_{-1} = 1$
for $j$ = 0,..., until convergence
    $\mathbf{y}_j = \mathbf{M}^{-1}\mathbf{r}_j$
    $\tau_j = \mathbf{r}_j{}'\,\mathbf{y}_j$
    $\beta_j = \tau_j / \tau_{j-1}$
    $\tau_{j-1} = \tau_j$
    $\mathbf{p}_j = \mathbf{y}_j + \boldsymbol{\beta}_j\,\mathbf{p}_{j-1}$
    $\mathbf{w}_j = \mathbf{\Psi}\mathbf{C}\,\mathbf{p}_j$
    $\boldsymbol{\alpha}_j = \mathbf{r}_j{}'\,\mathbf{y}_j / \mathbf{p}_j{}'\,\mathbf{w}_j$
    $\mathbf{x}_{j+1} = \mathbf{x}_j + \boldsymbol{\alpha}_j\,\mathbf{p}_j$
    $\mathbf{r}_{j+1} = \mathbf{r}_j - \boldsymbol{\alpha}_j\,\mathbf{w}_j$
end
$\mathbf{x}_{final} = \boldsymbol{\upsilon}$
```

[1] For PCG: $\mathbf{\Psi} = \mathbf{I}$, $\boldsymbol{\upsilon} = \mathbf{x}_{j+1}$; For DPCG: $\mathbf{\Psi} = \mathbf{P}$, $\boldsymbol{\upsilon} = \mathbf{Z}_d\,(\mathbf{Z}_d{}'\mathbf{C}\mathbf{Z}_d)^{-1}\mathbf{Z}_d{}'\mathbf{b} + \mathbf{P}'\mathbf{x}_{j+1}$.

Table 2. Largest eigenvalues and effective condition numbers (κ) of preconditioned (deflated) coefficient matrices, and number of iterations to reach convergence.

Model	Method[1]	Largest eigenvalue	*κ*	*Number of iterations*
ssGBLUP	PCG	11.9	$1.1*10^5$	273
ssSNPBLUP	PCG	181.0	$1.7*10^6$	1497
	DPCG (200)	99.4	$9.3*10^5$	1195
	DPCG (50)	40.5	$3.8*10^5$	880
	DPCG (5)	6.0	$5.4*10^4$	233
	DPCG (1)	6.0	$5.4*10^4$	240

[1] *Number of SNPs per subdomain are within brackets.*

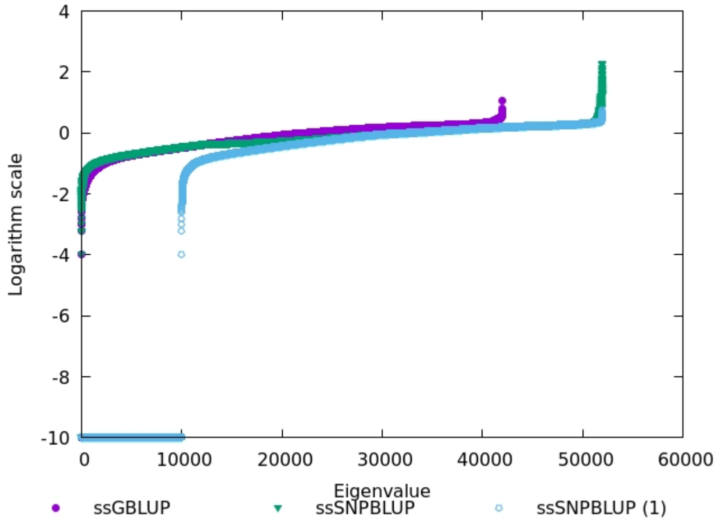

Figure 1. Eigenvalues of the preconditioned coefficient matrices of ssGBLUP and of ssSNPBLUP, and of the preconditioned deflated coefficient matrix of ssSNPBLUP with 1 SNP per subdomain.

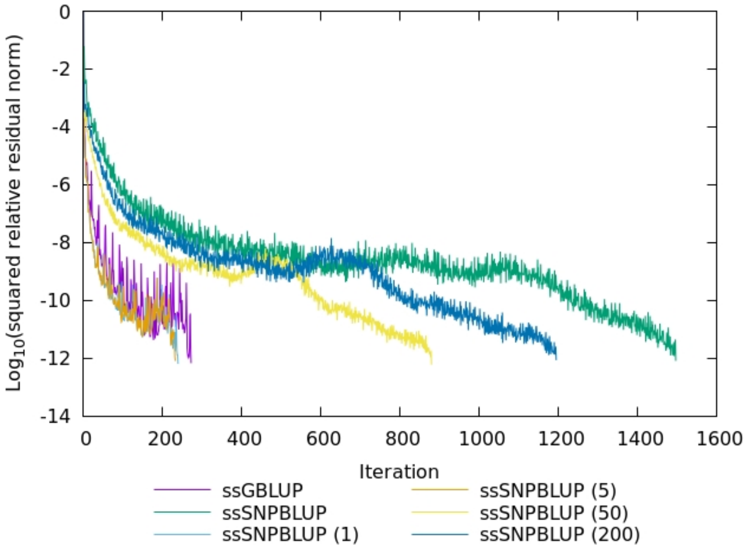

Figure 2. Square relative residual norms for ssGBLUP and ssSNPBLUP using the PCG method and for ssSNPBLUP using the DPCG method. Number of SNPs per subdomain are within brackets.

Acknowledgments

The Dutch Ministry of Economic Affairs, Agriculture, and Innovation (Public–private partnership “Breed4Food” Code BO-22.04-011-001-ASG-LR-3) and CRV BV (Arnhem, the Netherlands) are acknowledged for financial support and for providing the data.

References

Cornelissen, M.A.M.C., E. Mullaart, C. Van der Linde, and H.A. Mulder. 2017. Estimating variance components and breeding values for number of oocytes and number of embryos in dairy cattle using a single-step genomic evaluation. J. Dairy Sci. 100:4698–4705.

Fernando, R.L., H. Cheng, B.L. Golden, and D.J. Garrick. 2016. Computational strategies for alternative single-step Bayesian regression models with large numbers of genotyped and non-genotyped animals. Genet. Sel. Evol. 48:96.

Frank, J., and C. Vuik. 2001. On the Construction of Deflation-Based Preconditioners. SIAM J. Sci. Comput. 23:442–462.

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

Legarra, A., and V. Ducrocq. 2012. Computational strategies for national integration of phenotypic, genomic, and pedigree data in a single-step best linear unbiased prediction. J. Dairy Sci. 95:4629–4645.

Mäntysaari, E.A., and I. Strandén. 2016. Single-step genomic evaluation with many more genotyped animals. Page in Book of abstracts of the 67th annual meeting of the EAAP: 29 August-2 September 2016; Belfast.

Manzanilla Pech, C.I.V.M. 2017. Genetic improvement of feed intake and methane emissions of cattle. phd Thesis. Wageningen University, Wageningen.

Nabben, R., and C. Vuik. 2006. A comparison of deflation and the balancing preconditioner. SIAM J. Sci. Comput. 27:1742–1759.

Nicolaides, R.A. 1987. Deflation of conjugate gradients with applications to boundary value problems. SIAM J. Numer. Anal. 24:355–365.

Paige, C., and M. Saunders. 1975. Solution of Sparse Indefinite Systems of Linear Equations. SIAM J. Numer. Anal. 12:617–629.

Strandén, I., and M. Lidauer. 1999. Solving large mixed linear models using preconditioned conjugate gradient iteration. J. Dairy Sci. 82:2779–2787.

Taskinen, M., E.A. Mäntysaari, and I. Strandén. 2017. Single-step SNP-BLUP with on-the-fly imputed genotypes and residual polygenic effects. Genet. Sel. Evol. 49:36.

Vuik, C., A. Segal, and J.A. Meijerink. 1999. An efficient preconditioned CG method for the solution of a class of layered problems with extreme contrasts in the coefficients. J. Comput. Phys. 152:385–403.