Selective Shrinkage of Genomic Effects using Synthetic Dependencies in Neighboring Chromosome Regions

D. Wittenburg[1] **and N. Reinsch**[1]
[1]Leibniz Institute for Farm Animal Biology, Dummerstorf, Germany

ABSTRACT: As the number of model parameters increases with still growing number of SNPs, multicollinearity between covariates can affect the results of whole genome prediction. Selecting appropriate SNPs may counteract this phenomenon. Additionally, dependencies between single SNPs or chromosome regions can directly be incorporated in prediction methods. In this study, relationships between regions were modelled synthetically via a base-function approach; the genetic effect at some locus is a linear combination of base-function effects referring to the underlying and preceding regions. This B-spline approach was combined with a stochastic variable selection method to identify regions with non-zero impact. Application to milk performance data of 1,295 Holstein cows showed little reduced estimates of genetic variance components with decreased standard error and drastically reduced computing time compared to the analysis including all SNPs. Due to the synthetic structure of dependencies, this approach is applicable to different species.

Keywords: SNP selection; model complexity; B-spline

Introduction

In genome-based phenotype prediction, tens or hundreds of thousands covariates (e.g. single SNPs or haplotypes) are considered to have potential impact on a trait of interest. Because of dependencies among the predictor variables, which are not necessarily linear, multicollinearities are likely to occur. Multicollinearity may cause higher standard errors of the affected coefficients, and wrong inferences on these variables may be drawn (Farrar and Glauber (1967)).

In whole genome prediction, similarities between individuals (rows in some design matrix $\boldsymbol{X}$) rather than relationships between covariates (columns in $\boldsymbol{X}$) were often taken into account (e.g. Gianola et al. (2006); Piepho (2009)). Dependencies between SNPs were exploited in haplotype-based approaches (e.g. Meuwissen et al. (2002); Calus et al. (2009); Hickey et al. (2012)). Alternatively, in single-SNP approaches, each SNP effect can be corrected for the impact of its preceding neighbor in an antedependence model (Yang and Tempelman (2012)). Similarly, the natural order of SNPs may be utilized via the fused LASSO, where an additional penalty term is obtained from the successive differences of coefficients (Tibshirani and Saunders (2005)). Jointly considering similarities between relatives and correlations between single SNPs due to linkage disequilibrium is possible but requires parameters, which are not generally available (Gianola et al. (2009)). In whole genome association analyses, empirical correlations among SNP genotypes were employed in ranking of SNPs (Zuber et al. (2012)).

In this study, aiming at genomic prediction, effects of chromosome segments were estimated while considering dependencies between covariates in a synthetic manner, i.e. detached from biological interpretations. For this purpose two approaches were combined: The simplest way to decrease the effect of multicollinearity was to reduce the number of SNPs in the statistical model. Furthermore, the relationship between SNPs was considered by incorporating effects of the preceding chromosome regions.

Material and Methods

Statistical method. Dependencies between covariates are incorporated via B-spline interpolation. B-splines are piecewise polynomial functions of degree d; they are often used in non-parametric modelling (de Boor (2001)). Given a set of K knots on the x level, the outcome y at some x value is interpolated as linear combination of B-spline functions, $y(x) = \sum_{k=1}^{K} b_k B_k(x)$. For a simple definition of knots in genomic analyses, a set of SNPs is selected as representatives of certain chromosome regions via "pairwise tagging" method (de Bakker et al. (2005)). These tagSNPs exceed a user-defined value of correlation with SNPs in their neighborhood. The set of tagSNPs was completed by few equidistantly chosen extra knots to the left and right on each chromosome to properly build the B-spline functions. As SNPs exhibit a natural order based on either physical or genetic distances, SNP effects $\boldsymbol{g} = (g_1, \ldots, g_p)'$ are interpolated as $\boldsymbol{g} = \boldsymbol{B}\boldsymbol{b}$, with a $(p \times K)$-matrix $\boldsymbol{B}$ containing the known values of B-spline functions at positions $1, \ldots, p$ with reference to the knots $1, \ldots, K$, and the coefficients $\boldsymbol{b} = (b_1, \ldots, b_K)'$ are the base-function effects. The number of chromosome regions with impact on a current locus is specified by the degree of B-splines. As an example, with $d = 2$, the SNP effect at some locus $j \in \{1, \ldots, p\}$ is a linear combination of three base functions representing the effects of the underlying chromosome segment and the two previous segments, see Figure 1. Note that the shape of B-spline functions is non-uniform due to non-equidistant knots.

The statistical model $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{a} + \boldsymbol{X}\boldsymbol{B}\boldsymbol{b} + \boldsymbol{e}$ is applied to a milk performance trait $\boldsymbol{y} = (y_1, \ldots, y_n)'$, where $\boldsymbol{a} = (a_1, \ldots, a_q)'$ denotes fixed effects with the corresponding $(n \times q)$-design matrix $\boldsymbol{W}$, and the $(n \times p)$-matrix $\boldsymbol{X}$ contains a coding of observed SNP genotypes.

To discover chromosome regions with substantial effect on a trait, the important base functions have to be identified. For this purpose, the B-spline approach is combined with a stochastic variable selection method which enables selective shrinkage of effects (SVS, Ishwaran and Rao (2005)). For genomic data analysis, this approach has been extended to include different kinds of genetic effects (Wittenburg and Reinsch (2011)). As a consequence, the model includes two design matrices $\boldsymbol{X}_a$ and $\boldsymbol{X}_d$ with properly chosen coefficients of SNP effects (e.g. using the NOIA model of Álvarez-Castro & Carlborg (2007)) when

additive and dominance effects, respectively, are explored; each design matrix is right-multiplied by $\boldsymbol{B}$. The combined approach is here called SVS-B. To reach segment-wise shrinkage, a hyper-variance parameter, which also allows for the classification of zero or non-zero contribution, is defined for each coefficient of base function. The prior distribution of the hyper-variance is defined as a continuous bimodal distribution; its density is characterized by a spike near zero and a slab at the right tail. Due to the spike, the posterior expectation of a potential zero effect b_k of k-th base function ($k=1,...,K$) is shrunk towards zero, otherwise b_k is enlarged. After estimating the base-function effects via SVS, in which the concatenated design matrix [$\boldsymbol{XB}$] is processed, the genetic effects are obtained as $\hat{\boldsymbol{g}} = \boldsymbol{B}\hat{\boldsymbol{b}}$.

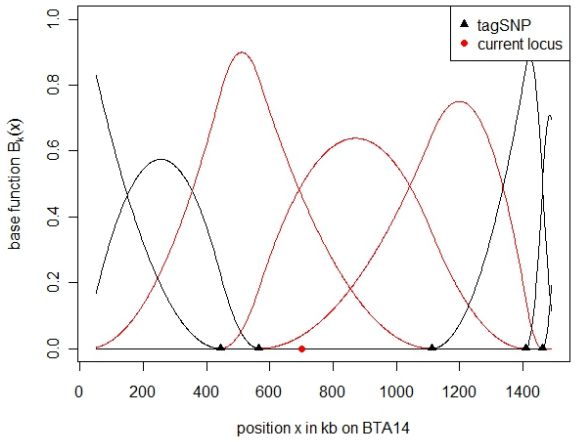

Figure 1. Base functions on beginning of BTA14. The effect at the current locus is obtained as linear combination of three base functions (red curves).

To verify features of the proposed approach, SVS-B is compared with the "classical" SVS method, which directly estimates genetic effects of all SNPs (SVS-A) and on tagSNPs only (SVS-T).

Dairy cattle data. Genotypes and phenotypes of $n = 1{,}295$ Holstein-Friesian cows, which were kept on 18 dairy farms in north-east Germany, were analyzed. Milk samples were collected between 21st and 120th day of their first lactation during the milk performance test from May to November 2009. Only first-lactation cows were selected to avoid variation due to parity and effects of selection due to culling of cows with low milk yield. Cows were descendants of 192 sires, but 22 cows had unknown sire. SNP genotypes were determined using the Illumina BovineSNP50 BeadChip. After several quality checks, $p = 37{,}180$ SNPs were retained. Physical distances between SNPs rely on Btau4.2. The genetic impact on three exemplary milk traits was studied (fat%, protein% and lactose%), but only results on fat% are presented here. The traits were standardized a priori.

In the design matrix $\boldsymbol{X}$, the SNP alleles were coded according to major allele frequencies in the population. Sample frequencies might be misleading, because the sample data were unbalanced in terms of half-sib families (e.g. a sire had on average 6.6 daughters, ranging from 1 to 106). To avoid bias due to abundance of minor alleles in large families, only half-sib families with more than 20 but less than 50 daughters were chosen for counting alleles resembling population frequency. Coding of SNPs is important for SVS-B. The correlation between SNPs – at least in the narrow neighborhood – was expected to be positive, implying that the more frequent alleles were inherited together. Then the use of B-splines leads to plausible inferences of SNP effects from effects of chromosome segments.

Algorithm. The SVS approach was implemented as a Gibbs sampling algorithm in Fortran90. Three chains with 100,000 iterations each were realized (40,000 samples were discarded as burn-in phase; thinning interval of two to reduce possible autocorrelation between samples).

Results

Knot selection was based on the complete data. Claiming a minimum correlation of 0.15 between SNPs in a neighborhood, in total $K = 14{,}687$ tagSNPs were determined. Arbitrarily, quadratic B-splines were used.

As an example, applying the statistical model including additive and dominance effects of SNPs to fat% yielded the estimated variance components and broad-sense heritability listed in Table 1. For all traits, highest genetic components and smallest residual variance component were estimated with SVS-A followed by SVS-B. Minimum standard error (SE) was obtained with SVS-B followed by SVS-T (for fat% and lactose%) or SVS-A (for protein%).

Table 1. Estimates of variance components and heritability for fat% (SE in brackets)*.

Method	σ_a^2	σ_d^2	σ_e^2	H^2
SVS-A	0.220 (0.026)	0.130 (0.043)	0.551 (0.052)	0.388 (0.054)
SVS-B	0.204 (0.025)	0.086 (0.021)	0.623 (0.038)	0.318 (0.034)
SVS-T	0.180 (0.033)	0.069 (0.034)	0.643 (0.047)	0.279 (0.048)

*σ_a^2 variance of breeding values, σ_d^2 variance of dominance deviations, σ_e^2 residual variance, H^2 broad sense heritability. Estimate and SE were obtained as posterior mean and standard deviation, respectively, from Gibbs samples. SVS combined with A: all SNPs, B: B-spline, T: tagSNPs.

Figure 2(a) shows the estimated additive (black pin) and dominance (red pin) genetic effects on fat%. In total 15 (12) significant additive (dominance) base-functions effects were identified with SVS-B. The tagSNP at *DGAT1* locus had greatest additive genetic impact on fat%. Figure 2(b) extracts the additive genetic effects around the *DGAT1* locus. It shows how the effects of base functions are spread over different SNPs.

Serving as a measure of accuracy of prediction, the correlation between estimated total genetic values and observed phenotypes based on a ten-fold cross validation was calculated. Only little differences were observed between the methods (e.g. fat, SVS-A: $\rho = 0.302$, SVS-B: $\rho = 0.287$, SVS-T: $\rho = 0.301$). The deviance information criterion was slightly in favor of SVS-A. Though SVS-B

was competitive concerning predictive accuracy, it required only a third of computing time of SVS-A (in average 11.8 h vs. 31.1 h per chain of Gibbs Sampling on a 2.93 GHz multi-user system). The calculation of variance of breeding values and dominance deviations in each Gibbs sampling iteration was most time-consuming.

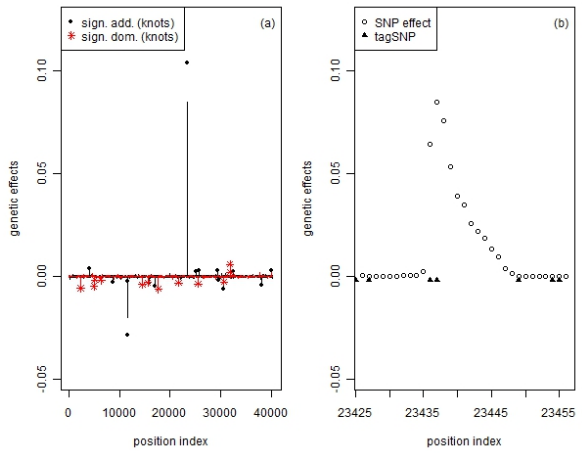

Figure 2. (a) Additive (black) and dominance (red) genetic effects of SNPs on fat%. Significant base-function effects are marked by a black circle (additive) and red star (dominance) at the corresponding tagSNP position. **(b) Window around *DGAT1* locus for additive genetic SNP effects on fat%.**

Discussion

Though the marker density increases, the number of model parameters can be kept small. The combined SVS and B-spline approach allowed for selective shrinkage of effects of chromosome segments; segments were represented by tagSNPs. Shrinkage is selective because a zero or non-zero impact of the underlying region was determined. For different chromosome regions, different degrees of shrinkage were desired, which increased complexity of the model. SVS-B deals with synthetic dependencies between regions: although parameters are embedded, they have no practical interpretation (De Boor (2001)). Tuning the degree of B-spline functions, however, specifies the number of preceding chromosome segments with impact on the current region. Defining tagSNPs as knots led to non-equidistant partitions of the genome, which is biologically reasonable. Alternative ways of assigning representatives to regions might be deduced. Pure model reduction (SVS-T) vs. combined approach (SVS-B) led to slightly higher accuracy of genetic value prediction, but variance components were estimated less precisely. Thus, with nearly the same model complexity, SVS-B benefits from relationships between chromosome segments, whose effects are used for genomic prediction. Due to reduced consequences of multicollinearity, it may be particularly valuable when further inferences are drawn based on SE of estimates (e.g. confidence intervals of genomic effects).

Some general questions concerning SVS still exist. The choice of hyper-parameters was verified via simulation study (Wittenburg and Reinsch (2011)). This setting does not necessarily fit to real data, where $p \gg n$. As an example, raw data analyses led to unrealistically small estimates of the residual variance component (but still small SE). The prior choice of hyper-parameters better suits to standardized traits. Rather than standardization, a cross-validation approach might help searching for an appropriate prior parameter setting.

For the selected milk traits, variation of dominance deviations was found to contribute to total genetic variation by roughly 30%. Omitting dominance effects led to slightly higher DIC values and similar predictive accuracy.

Conclusion

The combined SVS and B-spline approach enables selective shrinkage of effects of chromosome segments, which are sufficient for genomic prediction while accounting for dependencies between segments. Though initial SNP selection as representatives for chromosome regions requires prior knowledge, the incorporation of dependencies between segments is synthetic and allows for an application to different species.

Literature Cited

Álvarez-Castro, J. M. and Carlborg, Ö. (2007). Genetics 176:1151-1167.

Calus, M. P. L., Meuwissen, T. H. E., Windig, J. J. et al. (2009). Gen. Sel. Evol. 41:11.

De Bakker, P. I. W., Yelensky, R., Pe'er, I. et al. (2005). Nat. Genet. 37:1217-1223.

De Boor, C. (2001). A practical Guide to Splines. Springer.

Farrar, D. E. and Glauber, R. R. (1967). Rev. Econ. Stat. 49:92-107.

Gianola, D., de los Campos, G., Hill, W. G. et al. (2009). Genetics 183:347-363.

Gianola, D., Fernando, R. L. and Stella, A. (2006). Genetics 173: 1761-1776.

Hickey, J. M., Kinghorn, B. P., Tier, B. et al. (2012). J. Anim. Breed. Gen. 130:259-269.

Ishwaran, H. and Rao, J. S. (2005). Ann. Stat. 33:730-773.

Meuwissen, T. H. E., Karlsen, A., Lien, S. et al. (2002). Genetics 161:373-379.

Piepho, H. P. (2009). Crop Sci. 49:1164-1176.

Tibshirani, R. and Saunders, M. (2005). J. R. Stat. Soc. B 67:91-108.

Wittenburg, D. and Reinsch, N. (2011). In: Proc. 62nd EAAP Meeting, Stavanger, Norway. p. 116.

Yang, W. and Tempelman, R. J. (2012). Genetics 190:1491-150.

Zuber, V., Silva, A. P. D. and Strimmer, K. (2012). BMC Bioinformatics 13:284.