Kernel-Based BLUP with Genomic Data

U. Ober[*], M. Erbe[*], M. Schlather[†] and H. Simianer[*]

Introduction

Best linear unbiased prediction (BLUP, Henderson 1973) of breeding values based on phenotypic measurements and additive genetic relationships between individuals is a well established methodology in animal breeding (Robinson, 1991). With the availability of high-throughput genotyping facilities, genotypes for massive numbers of single nucleotide polymorphisms (SNPs) are available. Meuwissen et al. (2001) suggested to include SNP information in a statistical model of prediction. In geostatistics, the prediction of a spatial variable in any point of the considered space based on a given limited set of measurements is of interest. A standard approach in this case is "kriging" whose application is naturally limited to two, three or four dimensions.

We suggest the application of a high-dimensional kriging-extension to the genomic prediction problem, where one dimension reflects a genotype realization at one SNP. Our model for genomic data combines a polygenic effect and a general function of SNP genotypes. We present two kriging variants using the Matérn covariance function to reflect the functional dependency of the covariances from the distance of genotypes. Finally, we compare their predictive performance to a common genomic BLUP as a reference method in a simulation study.

Material and methods

Kriging. "Kriging" is used whenever spatial prediction of a so-called regionalized variable has to be performed based on a few, isolated measurements of the quantity. To this end, it is assumed that the regionalized variable is a realization of a random function with a certain covariance structure which is described by a parameterized covariance function. The "kriging" approach consists of two steps: (i) estimation of the unknown parameters and hidden variables (in particular by ML) and (ii) best linear unbiased prediction (BLUP) of the regionalized variables under the auxiliary assumption that the quantities estimated in step (i) are the true ones. Many variants of this unique kriging principle have been published (Cressie, 1993).

The model for polygenic and genomic data. We assume to have q animals with pedigree information, n of them having genotypic and phenotypic measurements ($n \ll q$). We use $y = W\beta + Zu + g(x) + e$ as a model for the given data, where y is a n-vector of phenotypes, β is a f-vector of nuisance location parameters and x_i is a $(p \times 1)$-vector of dummy SNP instance variates (genotype) observed on animal i. Let $\{g(x_i), x_i \in \mathbb{R}^p\}$ be a Gaussian zero-mean random field with covariance structure given by the Matérn covariance function

$$\mathrm{Cov}(g(x_i), g(x_j)) = K(x_i, x_j) = \sigma_K^2 \cdot \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\|x_i - x_j\|/h\right)^{\nu} \mathfrak{K}_\nu \left(\sqrt{2\nu}\|x_i - x_j\|/h\right).$$

[*]Georg-August-University Göttingen, Animal Breeding and Genetics Group, 37075 Göttingen, Germany
[†]Georg-August-University Göttingen, Institute for Mathematical Stochastics, 37077 Göttingen, Germany

Here, $\|\cdot\|$ is the Euclidean norm, $\nu > 0$ is a smoothness parameter, h is a scale parameter, σ_K^2 is the variance parameter and $\mathfrak{K}_\nu(\cdot)$ is the modified Bessel function of the second kind of order ν. Let $K = K(x_i, x_j)_{1\leq i,j\leq n}$ and let $g(x) = (g(x_1), \ldots, g(x_n))^T$. Furthermore, let $u \sim \mathcal{N}(0, \sigma_u^2 A)$ be a $(q \times 1)$-vector of additive genetic effects of q individuals, where σ_u^2 is the additive genetic variance due to unmarked polygenes and A is the additive relationship matrix. Let $e \sim \mathcal{N}(0, \sigma_e^2 \mathbf{I})$ be a n-vector of environmental residual effects. We assume that u, e and $g(x)$ are independent. W is a $(n \times f)$- and Z is a $(n \times q)$-incidence matrix.

Two kriging approaches and one reference model. We consider two kriging approaches for predicting the total genetic breeding value (BV) $z_0^T u + g(x_0)$ of a certain genotyped animal indexed by 0. The two approaches differ in the sets of quantities that are estimated and subsequently used for predictions. The reference model is a common genomic BLUP approach.

Model 1: Universal Kriging. We use $y \sim \mathcal{N}(W\beta, \sigma_u^2 ZAZ^T + K + \sigma_e^2 \mathbf{I})$ and estimate the parameters $\sigma_u, \sigma_e, \nu, h$ and σ_K by maximizing $\log(f_y)$ of the corresponding density function f_y. Then we predict $g(x_0)$ via $\hat{g}(x_0) = a_g^T y$ where a_g is chosen such that the prediction is a best linear unbiased one, i.e. we apply the BLUP principle and minimize $\mathbb{E}(\hat{g}(x_0) - g(x_0))^2$ with $\hat{g}(x_0) = a_g^T y$ under the condition $a_g^T W = 0$. Note that the condition assures that $\hat{g}(x_0)$ is unbiased. This approach is called "universal kriging". Let $K_0 = (K(x_1, x_0), \ldots, K(x_n, x_0))^T$. Minimizing the Lagrange functional $\mathbb{E}(\hat{g}(x_0) - g(x_0))^2 + 2a_g^T W\lambda$ with respect to a_g and λ yields the linear system

$$\begin{bmatrix} W & \sigma_u^2 ZAZ^T + K + \sigma_e^2 \mathbf{I} \\ 0 & W^T \end{bmatrix} \cdot \begin{bmatrix} \lambda \\ a_g \end{bmatrix} = \begin{bmatrix} K_0 \\ 0 \end{bmatrix}.$$

Analogously, $z_0^T u$ can be predicted by universal kriging.

Model 2: Simple Kriging. We consider the joint density function $f_{y,u,g}$ of y, u and g in order to estimate the parameters $\sigma_u, \sigma_e, \nu, h$ and σ_K and the hidden variables u and $g(x)$. Maximizing $J = \log(f_{y,u,g})$ with respect to β, u and $g(x)$ is equivalent to solving

$$\begin{bmatrix} W^T W & W^T Z & W^T \\ Z^T W & Z^T Z + \frac{\sigma_e^2}{\sigma_u^2} A^{-1} & Z^T \\ W & Z & \mathbf{I} + \sigma_e^2 K^{-1} \end{bmatrix} \cdot \begin{bmatrix} \hat{\beta} \\ \hat{u} \\ \hat{g}(x) \end{bmatrix} = \begin{bmatrix} W^T y \\ Z^T y \\ y \end{bmatrix}.$$

By re-employing the resulting estimators for β, u and $g(x)$ into J, one can calculate J for fixed $\sigma_u, \sigma_e, \nu, h$ and σ_K. Thus, J can be maximized with respect to these parameters by repeating this procedure iteratively. After convergence, we have estimators for $\sigma_u, \sigma_e, \nu, h, \sigma_K, u$ and $g(x)$. Now, we perform a BLUP of $g(x_0)$ via $\hat{g}(x_0) = a_g^T g(x)$ by minimizing $\mathbb{E}(\hat{g}(x_0) - g(x_0))^2$. This approach is called "simple kriging". Note that $\hat{g}(x_0)$ is always unbiased. The solution is $\hat{g}(x_0) = K_0^T K^{-1} g(x)$.

Model 3: Genomic BLUP. Based on the model $y = W\beta + Zu + Xg + e$, we perform a genomic BLUP and obtain the MME

$$\begin{bmatrix} W^T W & W^T Z & W^T X \\ Z^T W & Z^T Z + \frac{\sigma_e^2}{\sigma_u^2} A^{-1} & Z^T X \\ X^T W & X^T Z & X^T X + \frac{\sigma_e^2}{\sigma_g^2} G^{-1} \end{bmatrix} \cdot \begin{bmatrix} \hat{\beta} \\ \hat{u} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} W^T y \\ Z^T y \\ X^T y \end{bmatrix}.$$

In contrast to the initial model, it is $g \sim \mathcal{N}(0, \sigma_g^2 G)$ with G being a genomic relationship matrix calculated using the approach of VanRaden (2008). The matrix X is a known incidence matrix.

Simulation. We use the software "LDSO" (Ytournel, 2008). The base generation (1000 unrelated animals) is followed by 1001 generations (1000 animals each, random mating, sex ratio $1:1$ up to generation 1021). Then, the population size is gradually reduced, kept constant at 100 in generations 1011–1016 and is gradually increased to 500 in generation 1021, followed by 10 generations, each composed of 50 males and 450 females (random mating). The initial genomic data consists of 30'000 evenly spaced biallelic SNPs in linkage equilibrium on three chromosomes (1M length each, initial allele frequency of 0.5). In generation 1021, 100 SNPs are randomly chosen to act as QTL (MAF > 0.05), whose allele substitution effects are gamma distributed $\Gamma(0.42, 5.4)$. We randomly choose 1000 SNPs per chromosome (MAF > 0.05) to make up the final map. In generation 1027, we choose $\sigma^2_{poly} = 0.5\sigma^2_{QTL}$ and $\sigma^2_e = 3.5\sigma^2_{QTL}$ for the environmental variance to get a heritability of $h^2 = 0.3$. The polygenic effects in generation 1027 are normally distributed $\mathcal{N}(0, \sigma^2_{poly})$. From generation 1028 on, the polygenic effect u of an offspring is $0.5(u_{dam}+u_{sire})+m$, where m is a normally distributed Mendelian sampling term. BVs are obtained as the sum of polygenic effects and additive QTL-effects. Phenotypic values are the sum of the BVs and the environmental effects $e \sim \mathcal{N}(0, \sigma^2_e)$. Finally, one dataset consists of 5500 animals (generations 1021–1031), the last 5000 of them having pedigree information and the last 2000 of them being geno- and phenotyped.

Statistical analyses. All approaches are implemented in R. For each method, we predict the BVs of the last generation. Parameters and hidden variables are estimated with the information on 1500 animals (generations 1028 – 1030). The ML estimation of the parameters and hidden variables for the kriging-approaches is done using the R-package "RandomFields" Version 2.0.23 (cf. Schlather, 2001) and its function "fitvario". For the estimation of the variance components for the genomic BLUP we use "ASReml" (Gilmour et al., 2002). For each method, we compute the correlation between the predicted BVs and the true BVs in the estimation set (1500 animals) and in the prediction set (500 animals). The ultimate goal of animal breeding is to identify the genotypically best animals as parents of the next generation. Therefore, we calculate the average true BV of the 50 animals with the highest predicted BVs. We perform 100 independent simulations and summarize the results by calculating averages.

Results and discussion

The results are shown in Table 1. As can be seen, all models yield very similar correlations between predicted and true simulated BVs, in both the estimation and the validation set. This also holds for the average true BV of the best (according to their estimated BV) 10 percent of animals in the validation set.

The conceptual equivalence of kriging with BLUP has already been described by Wahba (1990). The idea of kriging in a space spanned by genomic data was first mentioned by Piepho (2009). The results demonstrate that our suggested kriging variants with the parameterized Matérn covariance function are able to compete with the common genomic BLUP. This also highlights the flexibility of the basic kriging principle, which is shown to work well even in an extension from 3 or 4 to 3000 dimensions.

Table 1: Average correlations between predicted and true BVs and average true BVs of the predicted best 10 percent of animals.

Model / Approach	cor(pred.BV, true.BV)[a]		av. true BV[b]
	validation set	estimation set	best 50
1 Universal Kriging	0.5929 ± 0.0048	0.7061 ± 0.0039	33.51 ± 2.33
2 Simple Kriging	0.5841 ± 0.0049	0.7182 ± 0.0028	33.52 ± 2.34
3 Genomic BLUP	0.5954 ± 0.0048	0.7280 ± 0.0025	33.62 ± 2.31

[a] average correlation between true BVs and predicted BVs with corresponding standard error
[b] average true BV of the 50 animals with the highest predicted BVs (averaged over all datasets) with standard error

The application of kriging in the suggested high-dimensional setup has additional consequences: distances between genotypes range within values that are clearly distinct from 0. In the simulated situation with 3000 genotypes coded as 0, 1, or 2, the average Euclidean distance in a typical replicate of the simulated data is 47.142 ± 0.001 with range $[25.120, 57.758]$. In the low-dimensional geostatistical modeling, distances are often close to zero.
The general approach of basing the prediction on a covariance function offers a number of possibilities for more differentiated modeling. In spatial statistics, the use of the Euclidean distance is a natural choice. In a genomic context, other distance metrics may be more adequate. With dense marker maps it is found that the genome is structured in haplotype blocks within which the loci are in high linkage disequilibrium. Hence, it might be adequate to account for this non-independence in the definition of the scale. A further option is to implement a feature selection which could give a higher weight to SNPs in genomic regions which are found to be relevant for the physiological pathways underlying the studied trait complex.

Acknowledgments

This research was funded by the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr "Synbreed – Synergistic plant and animal breeding".

References

Cressie, N. (1993). *Statistics for spatial data*. John Wiley & Sons, New York, Chichester.

Gilmour, A., Gogel, B., Cullis, B. *et al.* (2002). *ASReml user guide release 1.0*. VSN International Ltd., Hemel Hempstead, UK.

Henderson, C. (1973). *J. Anim. Sci.*, 1973:10–41.

Meuwissen, T., Hayes, B., and Goddard, M. (2001). *Genetics*, 157:1819–1829.

Piepho, H. (2009). *Crop Sci.*, 49:1165–1176.

Robinson, G. (1991). *Statist. Sci.*, 6:15–51.

Schlather, M. (2001). *R News*, 1(2):18–20.

VanRaden, P. (2008). *J. Dairy Sci.*, 91:4414–4423.

Wahba, G. (1990). *Spline models for observational data*. Society for Industrial Mathematics.

Ytournel, F. (2008). *Linkage disequilibrium and QTL fine mapping in a selected population*. PhD thesis, Station de Génétique Quantitative et Appliquée, INRA.