Genet. Res., Camb. (2010), **92**, *pp.* 209–225. © Cambridge University Press 2010
doi:10.1017/S0016672310000157

Radial basis function regression methods for predicting quantitative traits using SNP markers

NANYE LONG[1]*, DANIEL GIANOLA[1,2], GUILHERME J. M. ROSA[2], KENT A. WEIGEL[2], ANDREAS KRANIS[3] AND OSCAR GONZÁLEZ-RECIO[4]
[1] *Department of Animal Sciences, University of Wisconsin, Madison, WI 53706, USA*
[2] *Department of Dairy Science, University of Wisconsin, Madison, WI 53706, USA*
[3] *Aviagen Ltd, Newbridge, Midlothian EH28 8SZ, UK*
[4] *Instituto Nacional de Investigacion y Tecnologia Agraria y Alimentaria, 28040 Madrid, Spain*

(*Received 25 May 2009 and in revised form 6 March 2010*)

Summary

A challenge when predicting total genetic values for complex quantitative traits is that an unknown number of quantitative trait loci may affect phenotypes via cryptic interactions. If markers are available, assuming that their effects on phenotypes are additive may lead to poor predictive ability. Non-parametric radial basis function (RBF) regression, which does not assume a particular form of the genotype–phenotype relationship, was investigated here by simulation and analysis of body weight and food conversion rate data in broilers. The simulation included a toy example in which an arbitrary non-linear genotype–phenotype relationship was assumed, and five different scenarios representing different broad sense heritability levels (0·1, 0·25, 0·5, 0·75 and 0·9) were created. In addition, a whole genome simulation was carried out, in which three different gene action modes (pure additive, additive + dominance and pure epistasis) were considered. In all analyses, a training set was used to fit the model and a testing set was used to evaluate predictive performance. The latter was measured by correlation and predictive mean-squared error (PMSE) on the testing data. For comparison, a linear additive model known as Bayes A was used as benchmark. Two RBF models with single nucleotide polymorphism (SNP)-specific (RBF I) and common (RBF II) weights were examined. Results indicated that, in the presence of complex genotype–phenotype relationships (i.e. non-linearity and non-additivity), RBF outperformed Bayes A in predicting total genetic values using SNP markers. Extension of Bayes A to include all additive, dominance and epistatic effects could improve its prediction accuracy. RBF I was generally better than RBF II, and was able to identify relevant SNPs in the toy example.

1. INTRODUCTION

A challenge when predicting future phenotypes (or total genetic values) for complex quantitative traits is that phenotypic variation often reflects the concerted action of a large number of genes (quantitative trait loci (QTLs)) with small allelic effects (Risch, 2000). QTLs acting together may affect the trait in a cryptic manner, inducing non-additivity in relationships between phenotypes and QTLs (Gianola & de los Campos, 2008; Mackay, 2008; Yamamoto *et al.*, 2008). Since QTL genotypes are not usually observable, molecular markers such as single nucleotide polymorphisms (SNPs) are used as proxies.

Standard linear parametric models that assume linear relationships between phenotypes and markers may have poor predictive ability in the presence of interaction. To take into account non-additive gene action (e.g. epistasis), interaction effects can be added to some additive linear model. However, in a whole-genome analysis, this usually produces an over-saturated model, due to the large number of markers (p) and a small sample size (N), with $N \ll p$. One solution is using random-effects models with a Bayesian treatment. For example, Xu & Jia (2007) used an

* Corresponding author. Nanye Long, Department of Animal Sciences, University of Wisconsin, Madison, WI 53706, USA. e-mail: nlong@wisc.edu

empirical Bayes method to estimate main effects of 127 markers and interactions between all marker pairs simultaneously. If more markers and higher order interactions are considered, a large sample is needed for these parametric linear models to produce stable estimates (e.g. of additive and epistatic effects).

For phenotypic prediction under cryptic forms of epistasis, model-free non-parametric approaches have been proposed by Gianola *et al.* (2006), Gianola & van Kaam (2008), Gianola & de los Campos (2008) and de los Campos *et al.* (2009*a*). In non-parametric regression, no assumption is made regarding the form of the genotype–phenotype relationship. Rather, this relationship is described by a smoothing function and driven primarily by the data. For example, non-parametric methods based on reproducing kernel Hilbert spaces (RKHS) regression require finding a suitable kernel, a semi-positive definite matrix that is used instead of the standard incidence matrix $\boldsymbol{X}$ in regression. Recently, Bennewitz *et al.* (2009) applied non-parametric additive models to evaluate breeding value (BV) prediction with simulated data, and found that their accuracy was moderate to high, and in some cases slightly higher than BLUP.

Another non-parametric approach is radial basis function (RBF) regression (Powell, 1987). RBF regression uses a linear combination of a set of RBFs, each of them centred at a training data point. Because of its capability of approximating complex non-linear functions, RBF regression has been used in signal processing Hu & Hwang, 2001). The genotype–phenotype relationship is expected to be complex as well, which makes this method appealing.

In this study, RBF regression methods for predicting genetic values of quantitative traits were developed and examined. Two types of RBF models were considered: one had a common weight parameter for all SNPs, and the other one had SNP-specific parameters and was thus more general. For comparison, a method known as Bayes A (Meuwissen *et al.*, 2001) was used as a benchmark. First, a toy example was developed, in which an arbitrary non-linear relationship between eight SNPs and the phenotype was created. Then, two chicken datasets representing body weight (BW) and food conversion rate (FCR, the ratio of BW gain to the amount of feed fed) were used in a real data analysis. Finally, a large-scale whole genome simulation was carried out to compare RBF and Bayes A under three different gene action modes: pure additive, additive+dominance and pure epistasis.

The paper is organized as follows. Section 2 is methodological and provides a description of RBF regression methods and related techniques. Two RBF models are proposed and a Bayesian implementation is formulated. Section 3 describes briefly the Bayes A model. Criteria used for checking model predictive ability and a principal component analysis (PCA) for SNP dimension reduction are described in sections 4 and 5, respectively. Simulation and data analysis are presented in sections 6 and 7, followed by a discussion of implications, limitations and scope for RBF regression in section 8. Concluding remarks are given in section 9.

2. RBF REGRESSION MODELS

RBF regression is an approach to approximation problems in high-dimensional spaces (Powell, 1987; Broomhead & Lowe, 1988; Poggio & Girosi, 1990). The expectation function (given $\boldsymbol{x}$) in a RBF regression model takes the form

$$F(\boldsymbol{x})=\sum_{j=1}^{m} k(\|\boldsymbol{x}-\boldsymbol{x}_j\|)\,\alpha_j, \qquad (1)$$

where $\{k(\|\boldsymbol{x}-\boldsymbol{x}_j\|)|j=1,\ 2,\ \ldots,\ m\}$ is a set of m RBFs, which are fixed and non-linear on $\boldsymbol{x}$, and $\|\cdot\|$ denotes the Euclidean norm. Each $\boldsymbol{x}_j\in\mathbb{R}^p$ is a training point (e.g. a vector of SNP covariates on individual j) taken as centre of the RBFs, and p is the dimension of the input, e.g., the number of SNPs. A regression coefficient α_j is associated with each basis function. The functions $k(\cdot)$ are called 'radial' because their value depends only on the distance (r) from the origin ($\boldsymbol{x}$) to the centre point $\boldsymbol{x}_j$, i.e., $r=\|\boldsymbol{x}-\boldsymbol{x}_j\|$. Examples are the multiquadric $k(r)=\sqrt{c^2+r^2}$ for some $c\geqslant 0$, the thin-plate-spline $k(r)=r^2\log r$, and the Gaussian $k(r)=\exp(-\theta r^2)$ for some $\theta\geqslant 0$.

(i) *k-medoids clustering algorithm for finding centres*

Given N observations or training data points, one possibility is using each training data point $\boldsymbol{x}_j$ as a centre in (1), so $N=m$. This can be computationally expensive for a large number of observations, since computation grows polynomially with N (Haykin, 1999). Besides, this may produce over-fitting unless regularization is imposed (Ripley, 1996; Haykin, 1999). Therefore, a number (m) of representative centres smaller than N is often used, to reduce model complexity. Moody & Darken (1989) used k-means clustering for selecting centres, which chooses $m<N$ centres or centroids in the input space such that the sum of squared distances from each training point to its nearest centroid is minimum. In k-means clustering, a centroid's coordinates are the mean of coordinates of the points in the cluster. This is not suitable for discrete coordinates, such as SNPs, because SNP covariates are coded using the number of allele copies, and thus cannot be a fractional number as it would be if means were taken. To overcome this restriction, a variant called the k-medoids algorithm (Kaufman & Rousseeuw, 1990) chooses an existing data point, or medoid, for each cluster. The objective

https://doi.org/10.1017/S0016672310000157 Published online by Cambridge University Press

function to be minimized is the sum of distances from each data point to its closest medoid.

(ii) *Bayesian Gaussian RBF models*

In this study, phenotypes were approximated with a RBF regression model incorporating SNP information. A Bayesian analysis was adopted to infer all parameters in model (1) simultaneously. The RBF chosen was Gaussian, expressed as

$$\begin{aligned} k_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \exp\left[-\sum_{k=1}^{p} \theta_k \left(x_i^{[k]} - x_j^{[k]}\right)^2\right] \\ &= \exp\left[-(\boldsymbol{x}_i - \boldsymbol{x}_j)' \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j)\right]. \end{aligned} \qquad (2)$$

Here $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)'$ is a vector of unknown and non-negative weights and p is the number of SNPs. Genotypes at each SNP were coded as $x = 0$, 1 and 2 for A_2A_2, A_1A_2 and A_1A_1, respectively, and $\boldsymbol{x}_i$ contained the p codes for genotypes of individual i. The weights $\boldsymbol{\theta}$ can be viewed as inverse variances for each input SNP, and the 'covariance' matrix Σ, which was common to all centres, was $\mathrm{diag}(\frac{1}{\theta_1}, \ldots, \frac{1}{\theta_p})$. The larger the value of a particular θ_k, the more relevant the associated input (SNP) is to the outcome. Conversely, zero or a very small value of θ_k implicitly excludes that SNP from the basis functions. When p is large, as in the case of genome-wide association studies involving tens of thousands of SNPs, only a small portion is expected to be relevant. While it may be necessary to assume different weights for each SNP *a priori*, p distinct θ values may lead to an exceedingly greedy specification when p is large, as discussed later.

If an equal weight (variance) is assumed for all SNPs, the RBF reduces to

$$k_{\theta}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left[-\theta \sum_{k=1}^{p} \left(x_i^{[k]} - x_j^{[k]}\right)^2\right], \quad \theta \geqslant 0. \qquad (3)$$

In this case, only a single θ needs to be estimated. RBFs that take the forms of (2) and (3) will be denoted as RBF I and RBF II, respectively, in what follows.

Given N observations with p SNPs each, a linear model with m ($m \leqslant N$) RBFs is

$$y_i = \boldsymbol{w}_i' \boldsymbol{\beta} + \sum_{j=1}^{m} k_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j)\alpha_j + e_i, \quad i = 1, 2, \ldots, N. \qquad (4)$$

Here,

$\boldsymbol{\beta}$: nuisance parameter vector
$\boldsymbol{w}_i$: incidence vector linking $\boldsymbol{\beta}$ to observation i
$\boldsymbol{x}_i$: $p \times 1$ input vector for observation i
$\boldsymbol{\alpha} = \{\alpha_j | j = 1, 2, \ldots, m\}$: $m \times 1$ vector of non-parametric regression coefficients
$k_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j)$: RBF based on $\boldsymbol{x}_i$, $\boldsymbol{x}_j$
$\boldsymbol{\theta} = \{\theta_k | k = 1, 2, \ldots, p\}$: $p \times 1$ vector of weights

and $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_e^2)$ is a vector of residuals, where σ_e^2 is the residual variance. It is important to note that this model is linear in $\boldsymbol{\alpha}$ but not in $\boldsymbol{\theta}$. In matrix notation,

$$\mathbf{y} = \boldsymbol{W\beta} + \boldsymbol{K_\theta}\,\boldsymbol{\alpha} + \boldsymbol{e}, \qquad (5)$$

where $\boldsymbol{y} = \{y_i | i = 1, 2, \ldots, N\}$ is a vector of phenotypic values, $\boldsymbol{W} = \{\boldsymbol{w}_i' | i = 1, 2, \ldots, N\}$, and the radial basis matrix $\boldsymbol{K_\theta}$ of order $N \times m$ is

$$\boldsymbol{K_\theta} = \begin{bmatrix} \exp\left[-\sum_{k=1}^{p} \theta_k \left(x_1^{[k]} - x_1^{[k]}\right)^2\right] & \cdots & \exp\left[-\sum_{k=1}^{p} \theta_k \left(x_1^{[k]} - x_m^{[k]}\right)^2\right] \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ \exp\left[-\sum_{k=1}^{p} \theta_k \left(x_N^{[k]} - x_1^{[k]}\right)^2\right] & \cdots & \exp\left[-\sum_{k=1}^{p} \theta_k \left(x_N^{[k]} - x_m^{[k]}\right)^2\right] \end{bmatrix}. \qquad (6)$$

The above formulation is for RBF I; RBF II is obtained by replacing the weight vector $\boldsymbol{\theta}$ by a scalar variable θ, which is common to all input SNPs.

The conditional likelihood for the Bayesian hierarchical model is

$$p(\boldsymbol{y} | \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}, \sigma_e^2) = \prod_{i=1}^{N} N \times \left(y_i;\ \boldsymbol{w}_i'\boldsymbol{\beta} + \sum_{j=1}^{m} k_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j)\alpha_j, \sigma_e^2\right). \qquad (7)$$

Park & Casella (2008) suggested assigning a double exponential prior to each of the regression coefficients, α_i, in (1), due to its equivalence to regularization with an l_1 norm penalty as in the (least absolute shrinkage and selection operator LASSO) method (Tibshirani, 1996). Specifically, $\boldsymbol{\alpha}$ was assigned a product of double exponential distributions with density

$$p(\boldsymbol{\alpha} | \lambda, \sigma_e^2) = \prod_{j=1}^{m} \frac{\lambda}{2\sqrt{\sigma_e^2}} \exp\left(-\frac{\lambda |\alpha_j|}{\sqrt{\sigma_e^2}}\right), \qquad (8)$$

where λ is a positive parameter. Since θ_k ($k = 1, 2, \ldots, p$), is non-negative, this parameter was assigned an exponential prior with parameter ρ, i.e.

$$p(\boldsymbol{\theta} | \rho) = \prod_{k=1}^{p} \mathrm{Expon}(\rho) = \prod_{k=1}^{p} \rho \exp(-\rho\theta_k). \qquad (9)$$

https://doi.org/10.1017/S0016672310000157 Published online by Cambridge University Press

Priors for all parameters and fully conditional distributions are provided in Appendix A.

3. BENCHMARK MODEL: BAYES A

Bayes A poses the linear model

$$\mathbf{y} = \boldsymbol{W}\boldsymbol{\beta} + \boldsymbol{X}\boldsymbol{g} + \boldsymbol{e}, \tag{10}$$

where $\boldsymbol{W}$ and $\boldsymbol{\beta}$ are as in (5); matrix $\boldsymbol{X}$ links additive SNP effects to the phenotypes, with element X_{ij} being the genotype code (as for the RBF models) of individual i at SNP locus j ($j = 1, 2, \ldots, p$); $\boldsymbol{g}$ is a vector of additive SNP effects. Bayes A assumes *a priori* that $g_j \sim N(0, \sigma^2_{g_j})$ $(j = 1, 2, \ldots, p)$ are independent, and that the variance of marker effect ($\sigma^2_{g_j}$) differs for every locus j. Further, the prior distribution assigned to $\sigma^2_{g_j}$ in Bayes A is a scaled inverted chi-square distribution, $\chi^{-2}(\nu, S)$, for all SNPs, where S and ν are the scale and degrees of freedom, respectively. In our Bayes A implementation, instead of using a scaled inverted chi-square distribution, a uniform (0, u) distribution was assigned to each $\sigma^2_{g_j}$. The value of u was determined using information on phenotypic variance, number of SNP loci and allele frequency at each locus (see Appendix C). It was found that the two priors (scaled inverted chi-square and uniform) resulted in virtually the same predictive performance on a toy example and the two chicken data sets (see Appendix D). For $\boldsymbol{\beta}$, a flat prior was used, and an Inverse Gamma (0·001, 0·001) distribution was assigned to the residual variance σ^2_e.

4. PREDICTIVE ABILITY

Predictive performance on a test data set (i.e. a sample not used for model training) was used to compare models. Given M posterior samples of the parameter vector obtained with the training data set, M predicted values of the response variable were obtained for each test individual. Given the mth ($m = 1, 2, \ldots, M$) posterior draw of the parameter vector ($\boldsymbol{\beta}_m$, $\boldsymbol{\alpha}_m$, $\boldsymbol{\theta}_m$, etc.), the two RBF models generate a predicted value for test individual i as

$$\hat{y}_{im} = \boldsymbol{w}_i{}'\hat{\boldsymbol{\beta}}_m + \hat{\boldsymbol{k}}^{(\text{test}\rightarrow\text{train})'}_{\boldsymbol{\theta}_{m^i}}\hat{\boldsymbol{\alpha}}_m, \tag{11}$$

where $\boldsymbol{w}_i{}'$ is a nuisance covariate vector pertaining to individual i, and $\hat{\boldsymbol{k}}^{(\text{test}\rightarrow\text{train})'}_{\boldsymbol{\theta}_{m^i}}$ is the ith row of the rectangular radial basis matrix linking test and training observations, $\hat{\boldsymbol{K}}^{(\text{test}\rightarrow\text{train})}_{\boldsymbol{\theta}_m}$. For Bayes A, the predicted response value was

$$\hat{y}_{im} = \boldsymbol{w}_i{}'\hat{\boldsymbol{\beta}}_m + \boldsymbol{x}_i{}'\hat{\boldsymbol{g}}_m. \tag{12}$$

Here, $\boldsymbol{w}_i{}'$ and $\hat{\boldsymbol{\beta}}_m$ are as in (11) and $\boldsymbol{x}_i{}'$ is the genotype coding of individual i. After (11) or (12) was computed for each test individual i, its final predicted response, $\hat{y}_i$, was the average of M predicted values, which were used for computing the following criteria.

Correlations. Two types of correlations were of interest. (1) Correlation between expected phenotype $E(y|\boldsymbol{x}, \boldsymbol{w})$ using true values of the parameters, and its predicted value $\hat{y}$. This correlation is denoted as corr($E(y)$, $\hat{y}$) hereinafter. In the absence of nuisance covariates, the expected phenotype can be regarded as true genomic value, and corr($E(y)$, $\hat{y}$) represents accuracy of marker-assisted prediction of genomic value. (2) Correlation between observed and predicted phenotype, denoted by corr(y, $\hat{y}$). When $E(y)$ is known, corr($E(y)$, $\hat{y}$) is a better measure than corr(y, $\hat{y}$), for the observed phenotype (y) is a noisy version of the expected phenotype $E(y)$. Since only in a simulation study one knows $E(y)$, the measure corr($E(y)$, $\hat{y}$) was used only in simulation settings, and corr(y, $\hat{y}$) used only in the real data analysis.

Predictive mean squared errors (PMSE). Two different PMSE measures were used, depending on whether y (observed phenotype) or $E(y)$ (expected phenotype) was the target. The two PMSE criteria were: $\text{PMSE1} = n^{-1}_{\text{test}}\sum^{n_{\text{test}}}_{i=1}(E(y_i) - \hat{y}_i)^2$ and $\text{PMSE2} = n^{-1}_{\text{test}}\sum^{n_{\text{test}}}_{i=1}(y_i - \hat{y}_i)^2$, where n_{test} is the number of observations in the testing set. PMSE1 was used for the simulation study and PMSE2 was used for the real data analysis.

5. PCA FOR REDUCING SNP DIMENSIONALITY

In the BW and FCR datasets, the number (thousands) of SNPs made RBF I implementation computationally intensive, because weights (θ's) need to be sampled from a non-standard distribution for every SNP and $\boldsymbol{K}_{\boldsymbol{\theta}}$ has to be reconstructed accordingly. For Bayes A, computational burden was also heavy but less so because $\boldsymbol{X}$ is fixed in the course of iteration. For RBF II, this was not an issue, because only one θ needs to be estimated, regardless of the number of SNPs. To alleviate computations, SNP dimensionality was reduced prior to applying the three methods (RBF I, RBF II and Bayes A). This was done by means of a PCA, where principal components of the original SNP incidence matrix $\boldsymbol{X}$ were found and used to form a number of 'mega-SNPs', which was much smaller than the number of SNPs. PCA as a data reduction technique has been widely used in population genetics (e.g. McVean, 2009) and genetic association studies (e.g. Price *et al.*, 2006). For example, principal components can be used to represent ancestry differences among individuals in a case-control study, and adjusting for them is necessary prior to association tests. Procedures of PCA-based modelling and prediction in the current study are given below.

https://doi.org/10.1017/S0016672310000157 Published online by Cambridge University Press

(i) *PCA-based modelling of RBF I, RBF II and Bayes A*

The method of PCA (Manly, 2005; Varmuza & Filzmoser, 2009) used is described briefly here. Let $\boldsymbol{X}$ be the $N \times p$ SNP incidence matrix, where each column $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ represents genotypes of an SNP locus over N individuals. Performing PCA on matrix $\boldsymbol{X}$ gives

$$\boldsymbol{T} = \boldsymbol{X}\boldsymbol{P}. \tag{13}$$

Here, $\boldsymbol{P}$ $(p \times p)$ is a loading matrix containing in its columns all eigenvectors (i.e., principal components) of the covariance matrix of $\boldsymbol{X}$. $\boldsymbol{T}$ is the score matrix $(N \times p)$ whose columns are uncorrelated score vectors. Each score vector in $\boldsymbol{T}$ can be regarded as a regressor for a given mega-SNP. Note that for each eigenvector in $\boldsymbol{P}$, its associated eigenvalue corresponds to the variance of the associated score vector in $\boldsymbol{T}$. Usually columns of $\boldsymbol{P}$ are ordered in a way that the eigenvalues decrease. Since the primary aim of PCA is dimension reduction, the number of mega-SNPs used in PCA regressions is as small as possible, provided that a large proportion of the total variance of $\boldsymbol{X}$ can be explained. The score matrix containing the first a mega-SNPs is

$$\boldsymbol{T}_a = \boldsymbol{X}\boldsymbol{P}_a, \tag{14}$$

where $\boldsymbol{P}_a$ consists of the first a columns in $\boldsymbol{P}$. Therefore, the dimension of SNP inputs is reduced from p to a, where a (number of mega-SNPs) can be much less than p. Subsequent modelling (RBF I, RBF II and Bayes A) was based on this mega-SNP incidence matrix $\boldsymbol{T}_a$, rather than on the original SNP matrix $\boldsymbol{X}$.

For RBF models, the k-medoids algorithm used to reduce the number of basis functions from N to m was applied to $\boldsymbol{T}_a$ instead of $\boldsymbol{X}$; each entry in the radial basis matrix $\boldsymbol{K}_{\boldsymbol{\theta}}$ in (6) was computed using mega-SNP distance between two individuals. In Bayes A, $\boldsymbol{X}$ in (10) was replaced with $\boldsymbol{T}_a$, and the vector of SNP effects $\boldsymbol{g}$ was replaced with a vector of mega-SNP effects, denoted by $\boldsymbol{m}$. This yields a principal component regression, but cast in a Bayesian framework.

(ii) *Mega-SNP grouping for RBF I*

As noted, assigning SNP-specific weights as in RBF I is appealing but greedy. An alternative consists of grouping SNPs first, and then assigning group-specific weights. In the context of mega-SNPs, a natural grouping criterion is the variances accounted for by them. To do this, eigenvalues corresponding to each of the mega-SNPs were ordered and cut-points were chosen to divide them into a few groups, with the range of eigenvalues in each group being roughly equal. Then, a common θ-parameter was assigned to mega-SNPs within a group. Thus, the number of weights to be inferred reduced to the number of groups, which was arbitrarily set to four for both BW and FCR data.

(iii) *Prediction using mega-SNPs*

RBF models. Observations from the test data were first projected onto the space of principal components obtained from the training data to form a score matrix for the test data. That is,

$$\boldsymbol{T}_{\text{test}} = \boldsymbol{X}_{\text{test}}\boldsymbol{P}_a, \tag{15}$$

where $\boldsymbol{P}_a$ is the loading matrix defined previously, computed using training data; $\boldsymbol{X}_{\text{test}}$ is the SNP incidence matrix for the test data. Based on the training and testing score matrices ($\boldsymbol{T}_a$ and $\boldsymbol{T}_{\text{test}}$), a test-to-training radial basis matrix (e.g. (11)) can be built for prediction using RBF models.

Bayes A. For Bayes A-based prediction, a modification was needed. Bayes A becomes (for simplicity nuisance covariates are omitted here)

$$\boldsymbol{y} = \boldsymbol{T}_a\boldsymbol{m} + \boldsymbol{e}, \tag{16}$$

where $\boldsymbol{m}$ is a vector of mega-SNP effects defined earlier. Note that $\boldsymbol{T}_a = \boldsymbol{X}\boldsymbol{P}_a$ as in (14), so that

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{P}_a\boldsymbol{m} + \boldsymbol{e}, \tag{17}$$

$$= \boldsymbol{X}\boldsymbol{g} + \boldsymbol{e}, \tag{18}$$

where $\boldsymbol{g} = \boldsymbol{P}_a\boldsymbol{m}$. Hence, $\hat{\boldsymbol{g}} = \boldsymbol{P}_a\hat{\boldsymbol{m}}$. This means that SNP effects ($\boldsymbol{g}$) can be estimated by transforming estimates of mega-SNP effects ($\boldsymbol{m}$) using the loading matrix ($\boldsymbol{P}_a$). Therefore, predictions on testing observations can be made using the original SNP genotypes directly.

6. Data analysis

(i) *Outline*

The Bayesian analysis was performed using the R package R2WinBUGS (Sturtz *et al.*, 2005), which provides functions to call WinBUGS from R. The study consisted of three parts. First, a toy example with eight SNPs was used to illustrate the behaviour of RBF I and RBF II, as compared to Bayes A, under a hypothetical non-linear genotype–phenotype relationship. Subsequently, the three models were applied to two chicken datasets (provided by Aviagen Ltd) for BW and FCR, respectively. In the chicken data analysis, predictors in each model were not the original thousands of SNPs, but a smaller number of mega-SNPs derived from PCA. This was so because handling thousands of SNPs in our current implementation of RBF I with WinBUGS was too computationally intensive, so some reduction in input

https://doi.org/10.1017/S0016672310000157 Published online by Cambridge University Press

dimension was required. Although RBF II and Bayes A did not face this issue, mega-SNPs were used as predictors, to keep consistency between methods. A third part of the data analysis was a simulation-based whole genome marker-assisted prediction of genetic value. As with the chicken data, dense markers (about 2000) were used for prediction. In the simulation, total genetic value was determined by QTL in different manners, and model comparison was made correspondingly. In this part, the two models being compared were RBF II and Bayes A.

In the toy example and chicken data analysis, the number of basis functions was fixed at about 1/6 of training sample size; this was unlikely to produce overfitting according to our experience. The effect of the number of basis functions on RBF's predictive performance was investigated further in the whole genome simulation study, where RBF II was fitted using different numbers of basis functions.

(ii) *A toy example*

The simulated training data set had 300 unrelated individuals. Each observation consisted of two nuisance covariates (w_1 and w_2), eight SNP covariates (x_1, x_2, … , x_8) and a phenotypic value. Only three SNPs had effects on the phenotype. The phenotype for individual i (y_i) was generated as

$$y_i = w_{i1} + 2w_{i2} + e^{x_{i1}} \sin(x_{i2} - 0{\cdot}5)\, x_{i3}^2 + e_i,\quad i = 1, 2, \ldots, 300. \tag{19}$$

Here, regression coefficients on w_{i1} and w_{i2} are 1 and 2, respectively; x_{i1}, x_{i2} and x_{i3} are the genotypic codes for the three relevant SNPs among all eight SNPs, with a non-linear effect on phenotype, and e_i is a residual, distributed as $N(0, \sigma_e^2)$. The covariates were drawn by independent sampling from $w_{i1} \sim$ uniform$(-1, 1)$, $w_{i2} \sim$ uniform$(0, 2)$, and x_{ip} was sampled independently with equal probability (1/3) from {0, 1, 2}, for $p = 1, 2, \ldots, 8$. Hence, the eight SNPs were in linkage equilibrium, and five SNPs were completely uninformative about the genetic signal in the phenotype.

The residual variance was specified such that the ratio of genetic variance (caused by signal SNP variations) to residual variance, $\sigma_G^2 : \sigma_e^2$, would take the following five different values: 1 : 9 (scenario 1); 1 : 3 (scenario 2); 1 : 1 (scenario 3); 3 : 1 (scenario 4) and 9 : 1 (scenario 5). These scenarios correspond to broad sense heritabilities of about 0·10, 0·25, 0·50, 0·75 and 0·90, respectively. To assess σ_G^2, the variance of the term involving x_1, x_2, x_3 in (19), which is a non-linear function over the space of (x_1, x_2, x_3), was calculated. Parameter σ_G^2 was estimated empirically, by generating a large number (100 000) of realizations of e^{xi1} sin$(x_{i2} - 0{\cdot}5)$ x_{i3}^2, and their variance was taken as an indication of genetic variance in the population. Subsequently, σ_e^2 was adjusted accordingly. A testing data set with sample size $n_{\text{test}} = 500$ was simulated for each of the five scenarios under the same data generating distribution as for the training set.

Predictors in all three models were the eight SNPs (x_1, x_2, … , x_8) and the two nuisance covariates (w_1, w_2). RBF I assigned a specific θ to each of the eight SNPs, whereas RBF II assigned a common θ to them. The number of basis functions chosen by the k-medoids clustering algorithm was 50. The Bayes A was fitted in three different ways: (1) containing only additive effects of the eight SNPs (Bayes A^{A}); (2) containing additive and dominance effects of each SNP locus (Bayes A^{AD}); and (3) containing additive and dominance effects at each SNP locus, and their pair-wise epistatic interactions (Bayes A^{ADE}). The latter two were an enrichment to the first one by introducing non-additive marker effects. Bayes A^{A} was as in (10); Bayes A^{AD} and Bayes A^{ADE} were obtained by expanding the columns of the $\boldsymbol{X}$ matrix to represent all additive, dominance and epistasis terms needed. For example, in the case of two loci, the genetic value assumed by Bayes A^{ADE} can be represented as (Cordell, 2002)

$$g = a_1x_1 + d_1z_1 + a_2x_2 + d_2z_2 + i_{aa}x_1x_2 + i_{ad}x_1z_2 + i_{da}z_1x_2 + i_{dd}z_1z_2. \tag{20}$$

Here, x_i and z_i are dummy variables coding additive and dominant effects at locus i, respectively. One can set $x_i = 1$, $z_i = -0{\cdot}5$ for one homozygote, $x_i = -1$, $z_i = -0{\cdot}5$ for the other homozygote and $x_i = 0$, $z_i = 0{\cdot}5$ for a heterozygote. Coefficients a and d represent additive and dominance effects, respectively; i_{aa}, i_{ad}, i_{da} and i_{dd} correspond to epistatic effects. Following assumptions of Bayes A, all coefficients were assigned normal priors with heterogeneous variances.

For the toy example data, the length of the Markov chain was 20 000, and the first half was discarded as burn-in. Chains were then thinned at a rate of 5, such that 2000 samples were saved for inference.

(iii) *BW and FCR data analysis*

BW control is a key factor in rearing meat-type poultry and efficiency of food conversion is economically important in poultry production (Bell & Weaver, 2002). Heritability is 0·3 for BW and 0·2 for FCR, as used in the genetic evaluation program in Aviagen Ltd. Both datasets contained information on whole genome SNPs and these were used to predict the two traits.

BW data consisted of mean 42-day BWs of the progeny of each of 200 sires. Birds were raised in a low-hygiene environment, representing a commercial production setting. Phenotypic values of each sire were adjusted progeny means. Specifically, for each progeny of a sire, its phenotype was adjusted with estimates of fixed effects, maternal environmental random effects and of dam's BV, using best linear unbiased prediction. That is, adjusted phenotype = phenotype − fixed effect − maternal random effect − 1/2 dam's BV. Then, a sire's phenotype was the average of its offspring's adjusted phenotypes. To make the analysis more reliable, sires with less than 20 progeny records were removed, leaving 192 sires. These 192 sires were further partitioned into a training set (143 sires) and testing set (49 sires), such that sires in the testing set were sons of those in the training set. Phenotypic means (standard deviations) of the training and testing sets were 4·43 (6·00) and 5·57 (6·45), respectively. SNPs with monomorphic genotypes or minor allele frequency (MAF) less than 0·05 were removed, leaving 6947 SNPs for each sire.

FCR data were adjusted progeny means of FCR records of 394 sires from a commercial broiler line. The adjustment procedure was as for the BW records. The FCR data have been used by González-Recio *et al.* (2009), where predictive ability of Bayes A was compared against that of RKHS. Training data contained 333 sires and the test data contained 61 sires that were sons of sires in the training set. SNPs with monomorphic genotypes or MAF less than 0·05 were removed, leaving 3481 SNPs. Means (standard deviations) of sire phenotypes in the training and testing sets were 1·23 (0·10) and 1·21 (0·08), respectively.

For both datasets, PCA was applied to the SNP incidence matrix (143 × 6947 for BW and 333 × 3481 for FCR) of the training set, and mega-SNPs were formed after extracting the minimum number of principal components that could explain about 90% of the total variance. The RBF models and Bayes A (the one that assumes additivity only) fitted to the training data were based on mega-SNPs, and a mega-SNP grouping strategy was adopted for RBF I, as described earlier. Since phenotypic values were pre-adjusted for other effects, only an intercept was included in each model along with SNP genotypes. For RBF I and RBF II, the number of basis functions was set to 25 for BW data (training sample size = 143), and 50 for FCR data (training sample size = 333). Residuals in each model were assumed to be independently distributed normal random variables, with zero mean and different variances depending on the number of progeny of each sire. That is, for sire i with N_i progeny that were used to compute its mean, its residual variance was σ_e^2/N_i. As before, the length of Markov chain was 20 000, with the first half as burn-in. Chains were thinned at a rate of 5, giving 2000 samples for inference.

(iv) *Whole genome simulation*

The simulation basically followed that of Meuwissen *et al.* (2001) and Solberg *et al.* (2008). The simulated genome contained 10 chromosomes, each of 1 Morgan length. Along each chromosome, there were in total 302 loci (202 SNP markers and 100 putative QTLs) equi-distantly spaced. Markers (M) and QTL (Q) were positioned as

$$M_1 - M_2 - Q_1 - M_3 - M_4 - Q_2 - \cdots - M_{199} - M_{200}$$
$$Q_{100} - M_{201} - M_{202}.$$

The population evolved during 1000 generations of random mating and random selection, with a population size of 100 (50 males and 50 females) in each generation. Mutation rates at QTL and SNP markers were $2{\cdot}5 \times 10^{-5}$ and $2{\cdot}5 \times 10^{-3}$, respectively. After 1000 generations, the population size was increased to 1000 at generation $t = 1001$ by mating each sire with 20 dams, with one offspring per mating pair. In $t = 1002$, 1000 offspring were born from random mating of individuals in $t = 1001$. This resulted in 57 segregating QTL (5·7% of the total number of QTLs) and 2004 segregating SNP markers (99% of the total number of markers) in generations $t = 1001$ and $t = 1002$.

The total genetic values resulting from the QTL were generated for individuals in generations $t = 1001$ and $t = 1002$. Three hypothetical gene action scenarios were considered.

(1) Purely additive: each QTL locus had an additive effect only, without dominance or epistasis.
(2) Additive + dominance: each QTL had an additive as well as a dominance effect, and there was no epistasis between QTLs.
(3) Pure epistasis: there was no additive or dominance effects at any of the individual QTLs. Epistasis existed only between pairs of QTLs. The forms of epistasis included additive × dominance ($a \times d$), dominance × additive ($d \times a$) and dominance × dominance ($d \times d$). Additive and $a \times a$ epistasis effects were excluded, to prevent the additive variance from dominating the total genetic variance.

Given the genetic values and the genetic variance (estimated empirically from the genetic values generated), the error variance was chosen to keep heritability (in a broad sense) at 0·37 for each scenario. In scenario 1, the total genetic variance was completely additive. In scenario 2, 38% of the genetic variance was additive and the remaining 62% was due to dominance. In scenario 3, 42% of the variance was additive, 36% was dominance and 22% was epistatic. Note that there was additive genetic variance even though gene action was completely non-additive. Details of this simulation as well as variance component estimation are in Appendix B.

https://doi.org/10.1017/S0016672310000157 Published online by Cambridge University Press

Table 1. *Correlation and PMSE1 on test data in the toy example. Five scenarios (1–5) of broad sense heritability are 0·1, 0·25, 0·5, 0·75 and 0·9, respectively. corr ($E(y)$, $\hat{y}$): correlation between expected and predicted phenotype; PMSE1 $= n_{test}^{-1}\sum_{i=1}^{n_{test}}(E(y_i)-\hat{y}_i)^2$, where $n_{test}=500$ is sample size in the testing set. Bayes A^A, Bayes A^{AD} and Bayes A^{ADE} represent a Bayes A model that contains additive marker effects only, additive + dominance effects and additive + dominance + pair-wise epistatic effects, respectively.*

Scenario	RBF I	RBF II	Bayes A^A	Bayes A^{AD}	Bayes A^{ADE}
corr($E(y)$, $\hat{y}$)					
1	0·72	0·59	0·50	0·56	0·52
2	0·86	0·56	0·53	0·51	0·70
3	0·95	0·74	0·54	0·55	0·80
4	0·99	0·79	0·55	0·57	0·85
5	0·99	0·83	0·56	0·58	0·87
PMSE1					
1	23·50	32·50	43·14	49·06	87·27
2	11·32	30·26	33·91	37·60	36·63
3	4·54	19·84	31·53	32·24	20·22
4	0·91	16·24	30·42	30·43	14·04
5	0·35	13·84	29·94	29·58	11·73

The training set was formed by randomly picking 300 individuals from generation $t=1001$. The testing set was formed by randomly picking 500 individuals from $t=1002$. The training (testing) set was repeatedly sampled from $t=1001$ ($t=1002$) 50 times and, for each replicate, RBF II and Bayes A (assuming additivity) were used to predict genetic values of individuals in the testing set using SNP markers. Specifically, RBF II was fitted with a varying number of basis functions (50, 100, 150, 200, 250 and 300) to allow for an examination of its effect on model's prediction accuracy. Results of the 50 replications were averaged for a final evaluation of each method. This procedure was performed for each of the three gene action scenarios described earlier. The Markov chain in the Bayesian implementation of the two models was run for 100 000 iterations, with the first half as burn-in. Thinning rate was 10, yielding 5000 samples for inference.

7. RESULTS

(i) *Toy example*

Table 1 shows correlation (corr($E(y)$, $\hat{y}$)) and PMSE1 on the test data (sample size = 500) for the five competing models (RBF I, RBF II, Bayes A^A, Bayes A^{AD} and Bayes A^{ADE}). RBF I presented the best predictive ability in the five heritability scenarios. Even at low heritability (scenarios 1 and 2 with heritability of 0·1 and 0·25, respectively), RBF I attained large prediction correlations (0·72 and 0·86). RBF II was worse than RBF I, indicating that assigning SNP-specific weights improved ability to infer non-linear signals and enhanced prediction accuracy. Bayes A^A and Bayes A^{AD} were similar in correlation and PMSE1, indicating that introducing dominance effects in addition to additive effects was not helpful. Prediction correlation was improved by Bayes A^{ADE}, especially for moderate to high heritability ($\geqslant 0·25$). There, correlations of Bayes A^{ADE} increased by about 40–50% over those of Bayes A^A and Bayes A^{AD}, and were about 5–25% higher than RBF II. Concerning PMSE, RBF I was clearly superior and RBF II was better than any of the Bayes A versions except when heritability was 0.75 and 0.9.

Overall, Bayes A^A and Bayes A^{AD} seemed inadequate for dealing with the datasets used in the toy example, which were generated in a non-linear and interactive manner. Bayes A^{ADE} further accommodated interaction effects and improved Bayes A's predictive ability considerably, but it was still inferior to RBF I.

Absolute values of estimates of SNP effects $\boldsymbol{g}$ in Bayes A^A can be regarded as signal intensities of each SNP. A larger absolute value implies a greater relevance to the phenotype. Similarly, values of $\boldsymbol{\theta}$ in RBF I can be used as relevance measure also. Figure 1 shows 'signals' of the eight SNPs, as inferred by RBF I and Bayes A^A. RBF I was able to pick the three truly relevant SNPs (1–3) and to flag the other five as irrelevant, especially in scenarios 3–5, where the signal: noise ratio was large. Even in the two noisiest scenarios (1 and 2), SNPs 1–3 still stood out. For Bayes A^A, ambiguity existed due to large absolute values of estimates of some irrelevant SNPs. For example, SNP 7 appeared to be 'relevant' across scenarios 2 through 5. This indicated that Bayes A was not able to infer SNP effects accurately when the underlying relationship between genotype and phenotype was not additive. This is consistent with its poorer predictive ability.

(ii) *BW and FCR data*

In the BW analysis, about 90% of the total variance of the SNP training incidence matrix (143 × 6947) was

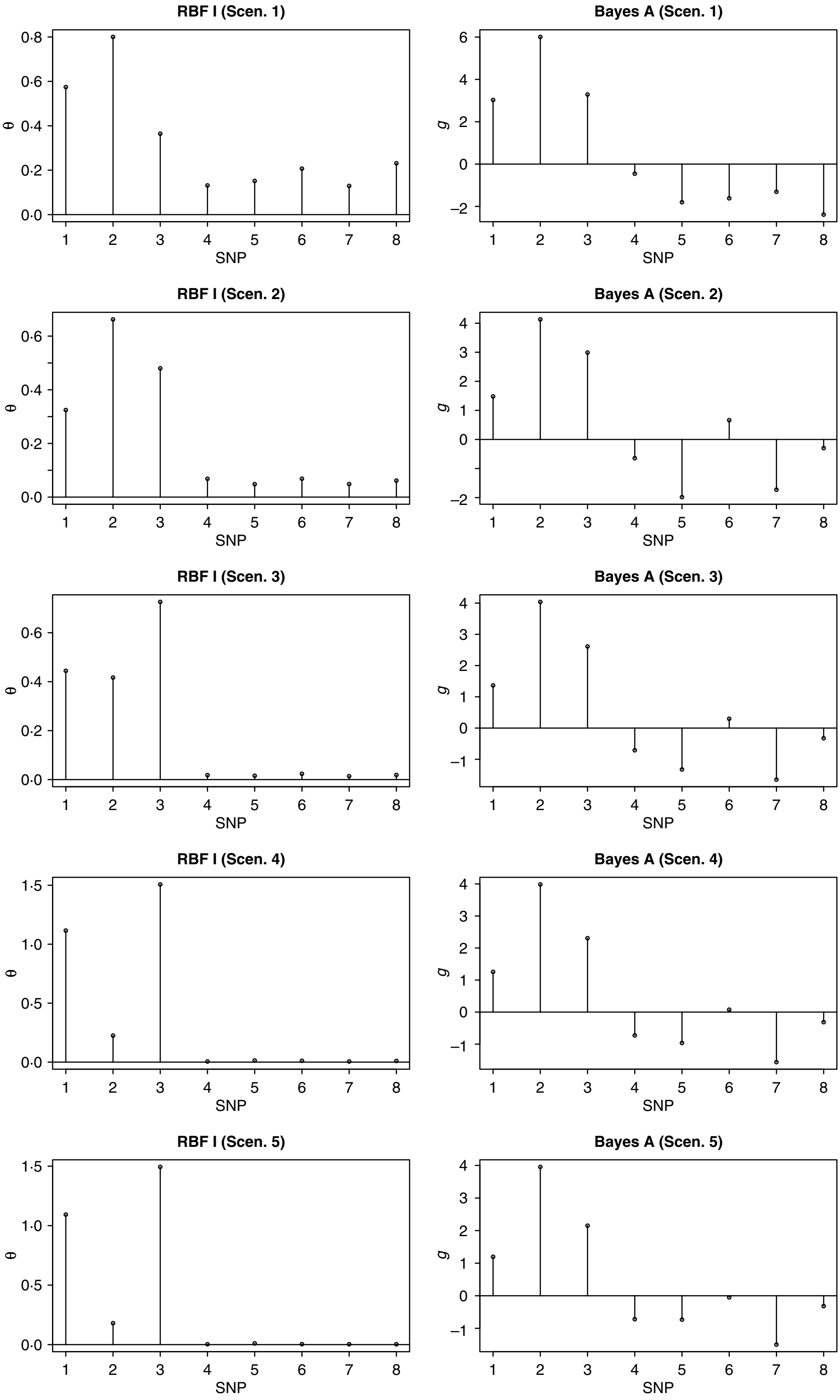


Fig. 1. Posterior means of $\boldsymbol{\theta}$ in RBF I and of SNP effects $\boldsymbol{g}$ in Bayes A in the toy example. The five scenarios correspond to different ratios between the genetic and residual variances (1–1 : 9; 2–1 : 3; 3–1 : 1; 4–3 : 1 and 5–9 : 1).

explained by 105 principal components, resulting in 105 mega-SNPs. The range of their eigenvalues was 8·2–102·6. Cut-points for a roughly 'equal-interval' division of mega-SNPs (for RBF I) were 32, 56 and 80, resulting in four groups and, therefore, 4 θ's. Using the 105 mega-SNPs, all three models were fitted

Table 2. *Correlation and PMSE2 in the testing sets of BW and FCR data. corr(y, ŷ): correlation between observed and predicted phenotype. PMSE2* $=n_{test}^{-1}\sum_{i=1}^{n_{test}}(y_i-\hat{y}_i)^2$*, where* n_{test} *is the sample size in the testing set.*

	BW		FCR	
	corr(y, $\hat{y}$)	PMSE2	corr(y, $\hat{y}$)	PMSE2
RBF I	0·49	32·49	0·23	0·0059
RBF II	0·45	35·49	0·18	0·0063
Bayes A	0·35	35·86	0·18	0·0067

to the training data (143 sires). Predictive performance on the testing data (49 sires) is summarized in Table 2. The highest predictive correlation (corr(y, $\hat{y}$)) was achieved by RBF I (0·49), followed by RBF II (0·45). Bayes A produced the lowest correlation (0·35). PMSE2 was smallest with RBF I (32·5) and larger in RBF II (35·5) and Bayes A (35·9). Overall, RBF I and RBF II were better than Bayes A.

In the FCR analysis, 220 mega-SNPs (explaining 90% of the total variance of the training set 333 × 3481$\boldsymbol{X}$ matrix) were formed; the range of the associated eigenvalues was 1·7–35·4. Cut-points for a roughly 'equal-interval' division (for RBF I) were 10, 20 and 30, yielding four groups and therefore 4 θ's, as in the BW data. Prediction results on the testing data (61 sires) are given in Table 2. RBF I had the best performance (corr(y, $\hat{y}$) = 0·23, PMSE2 = 0·0059). RBF II and Bayes A were the same in correlation (0·18) but RBF II had a slightly smaller value for PMSE2 (0·0063) than Bayes A (0·0067).

(iii) *Whole genome simulation*

As in the toy example, corr($E(y)$, $\hat{y}$) and PMSE1 were used, and the former can be interpreted as accuracy of marker-assisted prediction of total genetic value. Both metrics were summarized from all 50 replications, and density plots are presented in Figs 2 and 3 for prediction accuracy and PMSE1, respectively. The corresponding summaries are given in Table 3.

In the scenario of pure additivity, Bayes A had an average accuracy of 0·44, whereas RBF II with 300 basis functions (i.e. using each observation as the centre of a basis function) achieved an average accuracy of 0·43. Increasing the number of basis functions in RBF II gradually from 50 to 300 increased accuracy from 0·28 to 0·43. The improvement was relatively large from 50 to 100 basis functions. Beyond that, the change was small. The PMSE1 of Bayes A was 0·82; RBF II attained a PMSE1 from 0·86 with 50 basis functions moving down to 0·75 with 300 basis functions. Using an appropriate number of basis functions, RBF II can be as accurate as Bayes A or even better than Bayes A in PMSE.

In additive + dominance scenario, there was no major difference between Bayes A and RBF II in accuracy. Both were able to make predictions with an average accuracy of 0·13–0·14. Increasing the number of basis functions increased RBF II's accuracy somehow, but not much. PMSE1 was larger in Bayes A than in RBF II: PMSE 1 was 7·25 for Bayes A, whereas it ranged from 5·36 (50 basis functions) to 5·61 (300 basis functions) with RBF II. The overall performance of RBF II was therefore better than that of Bayes A.

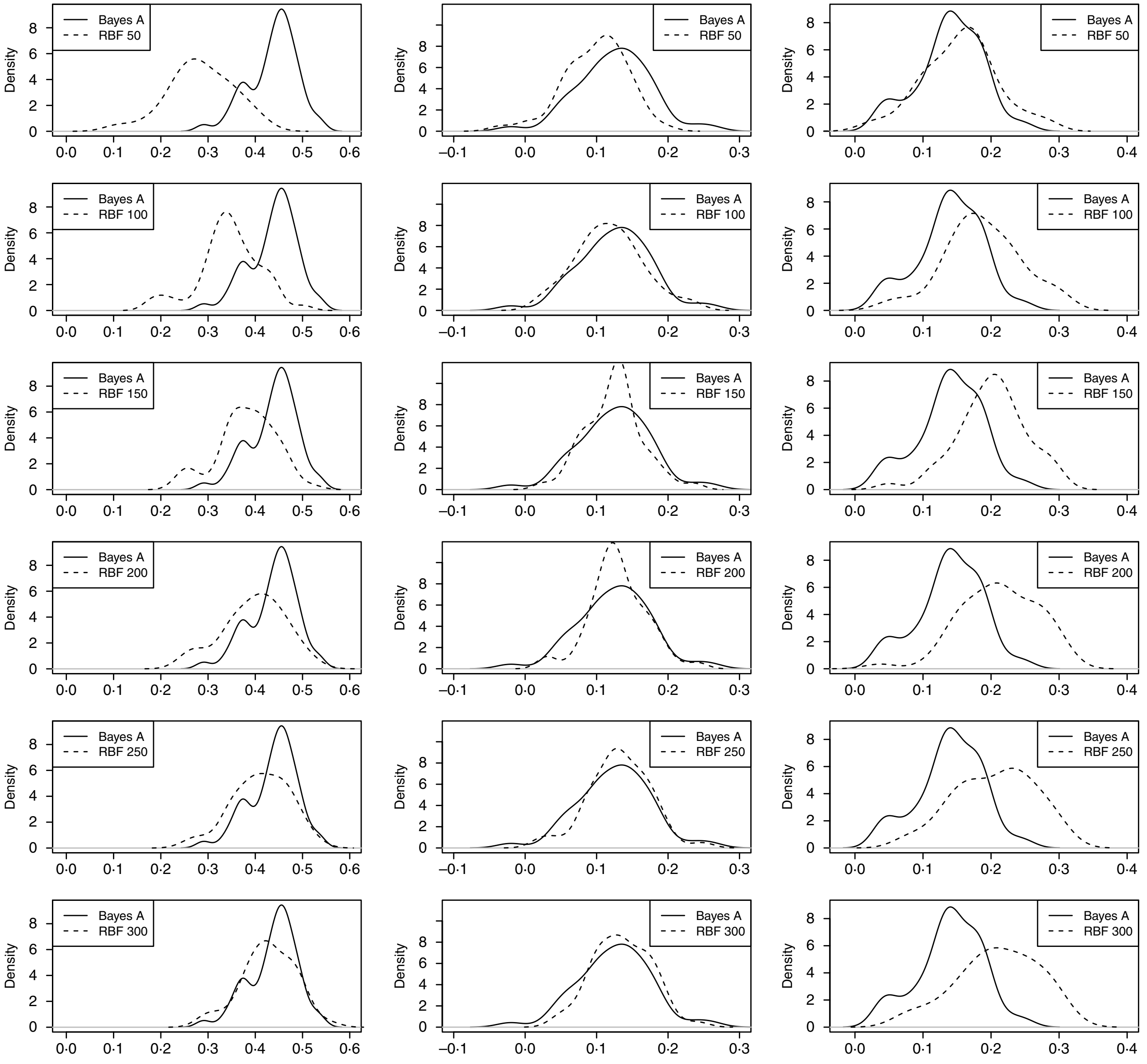
In the scenario of pure epistasis, the accuracy of Bayes A was 0·14. For RBF II with 50 basis functions, the accuracy was 0·15, and increased to 0·19 and 0·20 with 100 and 150 basis functions, respectively. Then, accuracy stabilized at 0·21 with 200, 250 and 300 basis functions. The PMSE1 was clearly smaller for RBF II (about 1) than for Bayes A (3·23). In short, RBF II was superior to Bayes A in predictive ability.

8. DISCUSSION

The Bayes A method of Meuwissen *et al.* (2001) is a linear model that assumes additive marker effects with each marker assigned a different variance, *a priori*. This produces heterogeneous shrinkage of marker effects, which may be more realistic than assuming equal variances and homogenous shrinkage for all markers, e.g., a BLUP model (Meuwissen *et al.*, 2001; de los Campos *et al.*, 2009*b*). However, an additive model may not work well in the presence of non-linear phenotype–genotype relationships (e.g. dominance or epistasis), in which case a substantial portion of genetic variance might be non-additive. In theory, one can include all possible forms of gene actions (additive, dominance and epistasis) in a Bayes A model but this becomes infeasible with hundreds of thousands of markers as model dimensionality is intractable. On the other hand, as pointed out by Gianola *et al.* (2006), non-parametric regression on marker genotypes does not make strong assumptions about the form of the marker–phenotype relationships and is expected to capture all possible forms of interaction. In the toy example where an arbitrary non-linear function was used to model genotype–phenotype relationship, RBFs attained a reasonably high accuracy whereas Bayes A was useful only when it contained additive, dominant and epistatic effects. In the analysis of two chicken datasets (offspring–parent settings), RBFs also outperformed a Bayes A that assumed additivity. Notably, model dimension of RBF II is independent of the number of markers, an important advantage over Bayes A.

In order to assess the value of RBF in various scenarios of gene action including pure additivity, a high

https://doi.org/10.1017/S0016672310000157 Published online by Cambridge University Press



(a) Pure additive (b) Additive+dominance (c) Pure epistasis

Fig. 2. Density of prediction accuracy of Bayes A and RBF II in the whole genome simulation (based on 50 replications), for each of the three scenarios (pure additive, additive+dominance and pure epistasis). Prediction accuracy is corr($E(y)$, $\hat{y}$). RBF50–RBF300 indicate RBF II with 50–300 basis functions, respectively.

level of dominance and pure epistasis, a whole genome simulation study was carried out. The training data simulated contained 2000 SNPs and 300 individuals so as to make it a 'large p small n' problem. Genetic value was generated by QTL but was predicted using SNP markers, relying on linkage disequilibrium (LD) between QTL and markers. Due to computational reasons discussed below, RBF I could not be considered so the comparison was between RBF II and Bayes A. It was found that RBF II, which makes no assumption of additivity of marker effects, was competitive for genetic value prediction relative to Bayes A (which assumed additivity) even when genetic values were purely additive. When dominance plus additive effects were considered (leading to 38% of the total genetic variance being additive and 62% being due to dominance), RBF II showed some advantage over Bayes A, especially in PMSE. Furthermore, when epistatic effects were simulated (42% additive variance, 36% dominance variance and 22% epistasis variance), RBF II was clearly better than Bayes A.

Increasing the number of SNPs increases computing time for RBF I because SNP-specific weights need to be estimated. When sampling from their conditional posterior distribution, every time a single weight is updated, the entire radial basis matrix has to be recomputed, making the process more time-consuming than RBF II. For example, with WinBUGS, RBF I with 50 basis functions and eight

https://doi.org/10.1017/S0016672310000157 Published online by Cambridge University Press

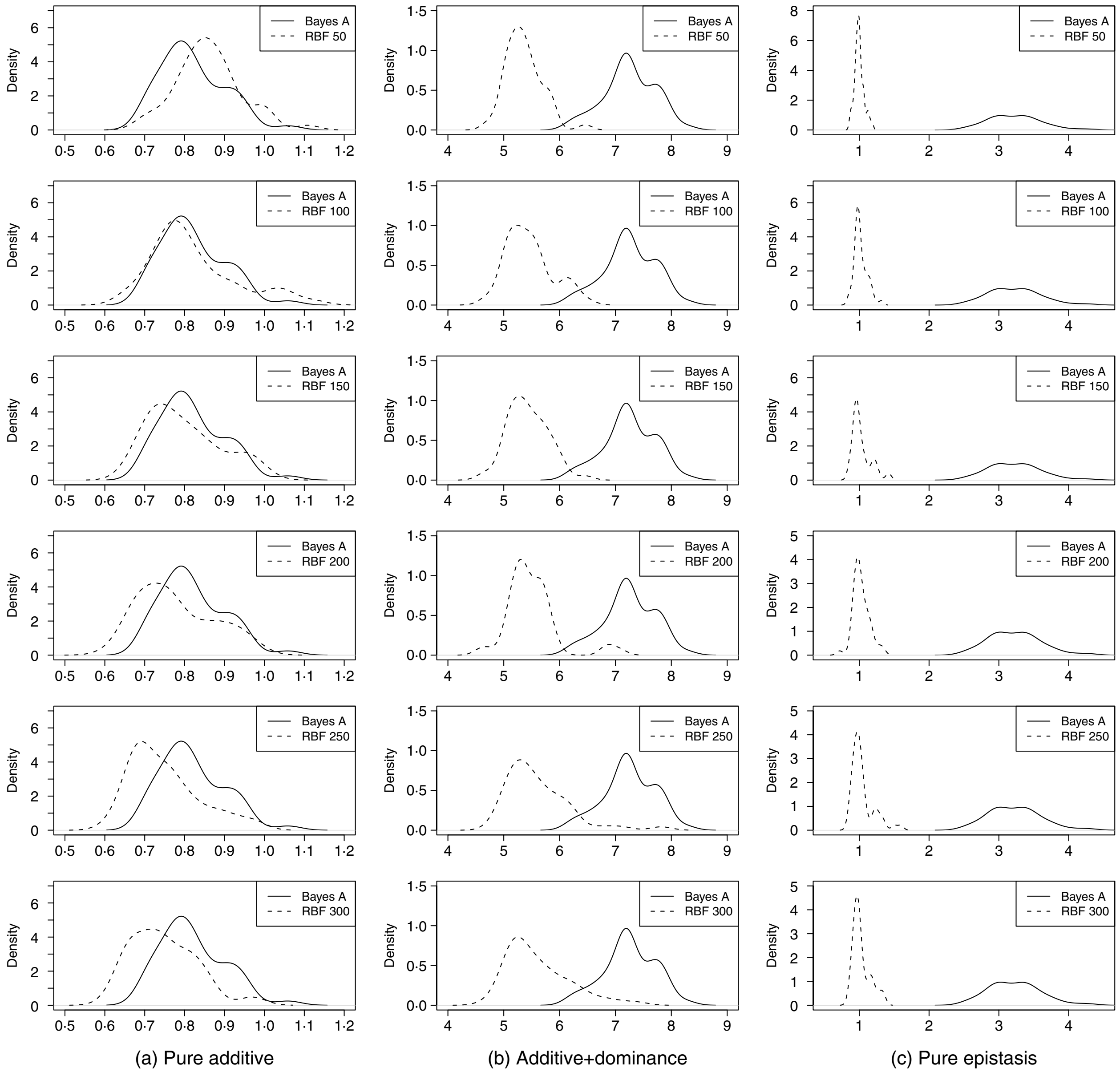


Fig. 3. Density of PMSE1 of Bayes A and RBF II in the whole genome simulation (based on 50 replications), for each of the three scenarios (pure additive, additive + dominance and pure epistasis). PMSE1 $=n_{\text{test}}^{-1}\sum_{i=1}^{n_{\text{test}}}(E(y_i)-\hat{y}_i)^2$, where $n_{\text{test}}=500$ is sample size in the testing set. RBF50–RBF300 indicate RBF II with 50–300 basis functions, respectively.

SNPs required 11 h for running 20 000 iterations, while RBF II spent only 4·9 h. It must be noted that this software is not tailored for large-scale calculation, so computational requirements may be lowered considerably if problem-specific software were developed. The WinBUGS environment limits the scope of RBF I unless some data reduction is done with respect to SNP dimension. For instance, in our analysis of BW and FCR, instead of using SNP-specific weights, group-specific weights were used, so that the effective number of weights to be estimated was reduced from the number of SNPs to the number of groups. Grouping can be done on the basis of chromosome, LD, or some other measure. LD was not useful for grouping here. Pairwise LD (Fig. 4) among SNPs was very weak for both BW and FCR data, as estimated by the r^2 measure (squared correlation between two SNP loci). This means that LD-based grouping would produce almost as many groups as there are SNPs. Instead, PCA-based grouping was adopted, which consisted of two steps: forming mega-SNPs and then grouping mega-SNPs based on their eigenvalues.

To compare mega-SNPs with real SNPs, Bayes A and RBF II were also fitted to BW and FCR data using real SNPs. Results indicated that the two yielded similar prediction performance. For example, Bayes A with real SNPs had correlations of 0·36 (BW) and 0·19 (FCR), very close to those obtained from using mega-SNPs: 0·35 (BW) and 0·18 (FCR).

https://doi.org/10.1017/S0016672310000157 Published online by Cambridge University Press

Table 3. *Prediction accuracy and PMSE1 for the three scenarios (pure additive, additive + dominance and pure epistasis) in the whole genome simulation. Prediction accuracy is* $corr(E(y), \hat{y})$. *PMSE1* $= n_{test}^{-1}\sum_{i=1}^{n_{test}} (E(y_i) - \hat{y}_i)^2$, *where* $n_{test} = 500$ *is the sample size in the testing set. RBF50–RBF300 indicate RBF II with 50–300 basis functions, respectively. Results were averages of 50 replications, with standard deviations given in parentheses.*

	Pure additive		Additive + dominance		Pure epistasis	
	Accuracy	PMSE1	Accuracy	PMSE1	Accuracy	PMSE1
Bayes A	0·44 (0·05)	0·82 (0·08)	0·13 (0·05)	7·25 (0·46)	0·14 (0·05)	3·23 (0·37)
RBF50	0·28 (0·07)	0·86 (0·08)	0·10 (0·04)	5·36 (0·32)	0·15 (0·06)	1·00 (0·07)
RBF100	0·34 (0·06)	0·82 (0·11)	0·12 (0·05)	5·46 (0·42)	0·19 (0·06)	1·02 (0·09)
RBF150	0·38 (0·06)	0·80 (0·09)	0·12 (0·04)	5·43 (0·36)	0·20 (0·05)	1·03 (0·13)
RBF200	0·40 (0·07)	0·78 (0·09)	0·13 (0·04)	5·50 (0·46)	0·21 (0·06)	1·03 (0·11)
RBF250	0·41 (0·06)	0·75 (0·09)	0·13 (0·04)	5·60 (0·57)	0·21 (0·06)	1·04 (0·16)
RBF300	0·43 (0·06)	0·75 (0·08)	0·14 (0·04)	5·61 (0·57)	0·21 (0·06)	1·03 (0·12)

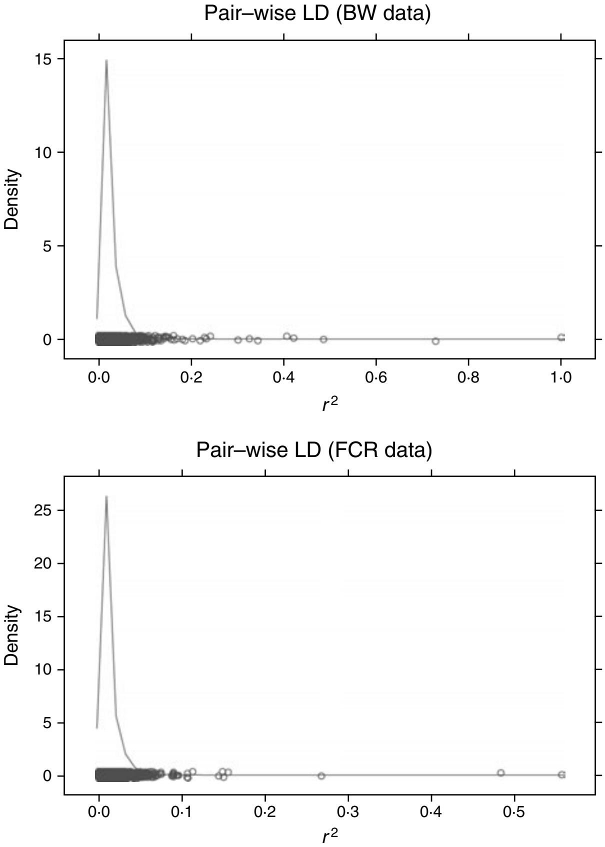


Fig. 4. Density plots of r^2 (squared correlation between two SNP loci) of pairwise SNP LD in BW and FCR data. Exhaustive pairwise computations were not feasible (24 126 931 pairs from 6947 SNPs in BW; 6 056 940 pairs from 3481 SNPs in FCR), so repeated sampling of all SNPs was used. Each repetition consisted of randomly sampling 50 SNPs, and pairwise r^2 was calculated. Values of r^2 from all repetitions were collected to approximate the LD level among the SNPs, and used for plotting. Here, 10 repetitions were performed for each dataset.

Likewise, RBF II with real SNPs had correlations of 0·41 (BW) and 0·17 (FCR), close to those from using mega-SNPs: 0·45 (BW) and 0·18 (FCR).

RBF I tended to have better predictive ability than RBF II. Also, by assigning SNP-specific weights, RBF I successfully flagged the three relevant SNPs in the toy example (Fig. 1). Nonetheless, RBF II is a promising, simpler, non-parametric method for making predictions. It achieved reasonably high accuracy in the toy example across a wide range of heritabilities. Although slightly worse, RBF II had similar prediction accuracy to RBF I in the real data analysis. More importantly, RBF II outperformed Bayes A when the underlying data generating system was far from additive.

As suggested by the whole genome simulation study, increasing the number of basis functions in RBF can improve prediction accuracy in general. However, the improvement gets smaller as more basis functions are included. At some point, accuracy reaches a plateau. Hence, one may not need as many basis functions as there are observations in order to get satisfactory predictions, provided that a smaller number of basis functions are properly chosen.

When all training data points are to be used as centres for the RBFs, some smoothness (regularization) must be imposed on the regression coefficients $\boldsymbol{\alpha}$ in the model $\mathbf{y} = 1\mu + \boldsymbol{K\alpha} + \boldsymbol{e}$, in order to avoid over-fitting. In a Bayesian treatment, regularization is done via a prior over the regression coefficients. In the present study, $\boldsymbol{\alpha}$ was assigned a double exponential prior which puts more mass near 0 and in the tails, relative to a normal prior (Tibshirani, 1996). With a normal prior $N(0, \boldsymbol{K}^{-1}\sigma_\alpha^2)$ on $\boldsymbol{\alpha}$, the model becomes a RKHS regression model (e.g. Wahba, 2002) from a Bayesian point of view (Gianola & van Kaam, 2008), in which a scaled inverted chi-square distribution is usually assigned as a prior to σ_α^2. RKHS requires the kernel matrix (equivalently radial basis matrix in RBF) $\boldsymbol{K}$ to be semi-positive definite. RBF regression does not require so. González-Recio *et al.* (2008, 2009) presented applications of RKHS to mortality and FCR in broilers, and their results suggested that RKHS regression using SNPs can produce more reliable predictions of genomic value than standard parametric additive model currently in use.

In addition to RBF and RKHS, another non-parametric method for modelling genotypic data is kernel regression, introduced by Gianola *et al.* (2006). When interest is in additive effects, kernel regression can be built in an additive manner, in which the total genetic value is the sum of non-parametric function values at each locus. In this way, additive and dominance effects at each of the loci can be captured, but not epistatic interactions (Gianola & van Kaam, 2008). Bennewitz *et al.* (2009) compared additive kernel regression with a BLUP model, and found that with differential amount of smoothing for the markers, the kernel regression was better than BLUP in predicting genomic BVs.

Since additive effects are central in selection improvement, it is sensible to ask whether or not it is useful to accommodate non-additivity in the model. Further, as shown by Hill *et al.* (2008), and corroborated here, a large portion of the genetic variance turns out to be additive. Our research indicates that non-parametric methods have similar or even better predictive ability than additive parametric models. Further, they are more general in situations in which prediction of performance is a focal point, such as in personalized medicine or genome-assisted management programmes.

9. CONCLUSION

Non-parametric RBF regressions were investigated by simulation and analysis of two broiler datasets. In the presence of a complex genotype–phenotype relationship (i.e. non-linearity and non-additivity), the RBF models outperformed a linear additive model, Bayes A, in predicting total genetic values of quantitative traits using SNP markers. The RBF methods had similar or even better predictive ability when gene action was purely additive. A RBF model with SNP-specific weights (RBF I) was generally better than one with a common weight for every SNP (RBF II). There is a potential for RBF I to discover relevant markers from a large pool of genetic markers scattered across the entire genome. When dealing with a massive number of markers, computational demand in RBF I is intensive, which remains an issue to be addressed.

APPENDIX A. BAYESIAN HIERARCHICAL RBF MODEL

The priors for all parameters in the Bayesian hierarchical model (see (5) and (7)) are as follows

$$p(\boldsymbol{\beta}) \propto \text{constant}, \qquad \text{(A.1)}$$

$$p(\boldsymbol{\alpha}|\lambda, \sigma_e^2) = \prod_{j=1}^{m} \frac{\lambda}{2\sqrt{\sigma_e^2}} \exp\left(-\frac{\lambda|\alpha_j|}{\sqrt{\sigma_e^2}}\right), \qquad \text{(A.2)}$$

$$p(\lambda|\gamma_1, \delta_1) = \text{Gamma}\,(\gamma_1, \delta_1) \propto (\lambda)^{\gamma_1 - 1} \exp\,(-\delta_1 \lambda), \qquad \text{(A.3)}$$

$$p(\boldsymbol{\theta}|\rho) = \prod_{k=1}^{p} \text{Expon}\,(\rho) = \prod_{k=1}^{p} \rho \exp\,(-\rho\theta_k), \qquad \text{(A.4)}$$

$$p(\rho|\gamma_2, \delta_2) = \text{Gamma}\,(\gamma_2, \delta_2) \propto \rho^{\gamma_2 - 1} \exp\,(-\delta_2 \rho), \qquad \text{(A.5)}$$

$$p(\sigma_e^2|a, \nu) = \text{Inverse Gamma}\,(a, \nu) \propto (\sigma_e^2)^{-a-1} \exp\left(-\frac{\nu}{\sigma_e^2}\right). \qquad \text{(A.6)}$$

The residual variance σ_e^2 was assigned a vague Inverse Gamma($a = 0{\cdot}001$, $\nu = 0{\cdot}001$) prior, throughout; the Gamma distribution parameters for λ and ρ were tuned to ensure convergence in all analyses. The joint posterior density of all parameters is proportional to the product of (7)–(A.6). The fully conditional distributions can be shown to be:

$$p(\boldsymbol{\beta}|\text{else}) = N\left[(\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'(\mathbf{y} - \boldsymbol{K_\theta\alpha}), (\boldsymbol{W}'\boldsymbol{W})^{-1}\sigma_e^2\right], \qquad (27)$$

$$p(\sigma_e^2|\text{else}) \propto (\sigma_e^2)^{-N-a-1} \times \exp\left\{-\frac{1}{2\sigma_e^2}[(\mathbf{y} - \boldsymbol{W\beta} - \boldsymbol{K_\theta\alpha})'(\mathbf{y} - \boldsymbol{W\beta} - \boldsymbol{K_\theta\alpha}) + 2\nu]\right\} \times \exp\left(-\frac{\lambda}{\sqrt{\sigma_e^2}}\sum_{j=1}^{m}|\alpha_j|\right), \qquad (28)$$

$$p(\boldsymbol{\alpha}|\text{else}) \propto \exp\left[-\frac{1}{2\sigma_e^2}(\mathbf{y} - \boldsymbol{W\beta} - \boldsymbol{K_\theta\alpha})'(\mathbf{y} - \boldsymbol{W\beta} - \boldsymbol{K_\theta\alpha})\right] \times \prod_{j=1}^{m} \exp\left(-\frac{\lambda|\alpha_j|}{\sqrt{\sigma_e^2}}\right), \qquad (29)$$

$$p(\lambda|\text{else}) \propto \lambda^{m+\gamma_1-1} \exp\left[-\left(\sum_{j=1}^{m}\frac{|\alpha_j|}{\sqrt{\sigma_e^2}} + \delta_1\right)\lambda\right] = \text{Gamma}\left(m + \gamma_1, \sum_{j=1}^{m}\frac{|\alpha_j|}{\sqrt{\sigma_e^2}} + \delta_1\right), \qquad (30)$$

$$p(\boldsymbol{\theta}|\text{else}) \propto \exp\left[-\frac{1}{2\sigma_e^2}(\mathbf{y} - \boldsymbol{W\beta} - \boldsymbol{K_\theta\alpha})'(\mathbf{y} - \boldsymbol{W\beta} - \boldsymbol{K_\theta\alpha}) - \rho\sum_{k=1}^{p}\theta_k\right], \qquad (31)$$

https://doi.org/10.1017/S0016672310000157 Published online by Cambridge University Press

$$p(\rho|\text{else}) = \rho^{p+\gamma_2-1} \exp\left(-\rho\left(\delta_2 + \sum_{k=1}^{p} \theta_k\right)\right), \tag{32}$$

$$= \text{Gamma}\left(p+\gamma_2, \delta_2 + \sum_{k=1}^{p} \theta_k\right). \tag{33}$$

Hence, not all fully conditionals are recognizable distributions. For parameters with recognized distributions ($\boldsymbol{\beta}$, λ and ρ), Gibbs sampling (i.e. using the fully conditionals as proposal distributions) can be used. For the other parameters, Metropolis–Hastings sampling can be employed.

APPENDIX B. GENERATING GENETIC VALUES AND EVALUATING ADDITIVE GENETIC VARIANCE IN THE WHOLE GENOME SIMULATION

Pure additive. The additive effect (a) was equal to the allele substitution effect, such that for genotypes A_1A_1, A_1A_2 and A_2A_2, their genotypic values were $2a$, a and 0, respectively. The value of a at each QTL locus was sampled from a normal distribution with mean 0 and variance 0·1.

Additive + dominance. The additive effects were as in the pure additive scenario. To simulate a marked dominance effect, its value (d) assigned to each QTL locus was obtained by multiplying the additive effect (a) at that locus by arandom number sampled from uniform (−5, 5) distribution. For a QTL locus with additive effect a and dominance effect d, its genotypic values were $2a$ for A_1A_1, $a+d$ for A_1A_2, and 0 for A_2A_2. If genotypic frequencies at that locus are p^2, $2pq$ and q^2 for A_1A_1, A_1A_2 and A_2A_2, respectively, the total genetic variance can be decomposed into additive (σ_A^2) and dominance variance (σ_D^2) as follows (Falconer & Mackay, 1996):

$$\sigma_A^2 = 2pq[a+(q-p)d]^2,$$
$$\sigma_D^2 = (2pqd)^2.$$

For simplicity, independence between QTLs was assumed and, as a result, the total additive (dominance) variance was summed over all loci.

Pure epistasis. Only segregating QTLs were involved in epistatic interactions. Among the 57 segregating QTLs, 56 were randomly chosen to form 28 QTL pairs. Each pair was assigned an $a \times d$ interaction effect i_{ad}, a $d \times a$ effect i_{da} and a $d \times d$ effect $i_{d \times d}$. The absolute value of each of i_{ad}, i_{da} and i_{dd} was the same, and set to mean of $|a_1|$, $|d_1|$, $|a_2|$ and $|d_2|$, where a_1 and d_1 were additive and dominance effects at the first locus, and a_2 and d_2 were those at the second locus. The sign of each of the three epistatic effects was sampled from {1, −1} with probability 0·5.

Given a pair of QTLs (i = 1, 2), its epistatic value was given by $i_{ad}x_1z_1 + i_{da}z_1x_2 + i_{dd}z_1z_2$, where x_i and z_i were additive and dominance codes at locus i, respectively. For A_1A_1 genotype, $x_i = 1$, $z_i = -0{\cdot}5$; for A_1A_2, $x_i = 0$, $z_i = 0{\cdot}5$; and for A_2A_2, $x_i = -1$, $z_i = -0{\cdot}5$ (Cordell, 2002).

The total genetic value was the sum of epistatic values produced by each of the 28 QTL pairs. The procedure of estimating additive and epistatic variance components followed Cockerham (1954), assuming independence between two loci of each QTL pair and between QTL pairs.

APPENDIX C. DETERMINE *u* IN THE UNIFORM PRIOR DISTRIBUTION (0, *u*) FOR VARIANCES OF MARKER EFFECTS IN BAYES A

The construction of the prior distribution was based on some assumptions and approximations. The purpose was getting a rough value of the marker variance $\sigma_{g_j}^2$ ($j = 1, 2, \ldots, p$), which was the variance of the prior distribution for marker effect $g_j : g_j \sim N(0, \sigma_{g_j}^2)$. The assumptions were:

1. Allele frequencies at all loci come from the same distribution. Similarly, marker effects at all loci come from the same distribution $g_j \sim N(0, \sigma_g^2)$.
2. Consider only additive genetic variance at each locus j, which is $2p_jq_jg_j^2$ (p_j, q_j are allele frequencies and g_j is additive marker effect at locus j).
3. The total genetic value (G) of an individual is assumed to be obtained by adding genetic values at all marker loci, with marker loci treated as independent (no LD). Then, the variance of the total genetic value (var(G)) is the sum of variances at individual loci.
4. Approximate var(G) by the variance of observed phenotypic values (var(y)). This leads to an overstatement of genetic variance and, thus, of the variance of each marker. As a prior, this allows a wide range of values for marker variance.

Based on assumptions 1 and 3, the expectation of the variance contributed by each of the p loci is the same. Hence,

$$E(H_jg_j^2) = \frac{\text{var}(y)}{p},$$

$H_j = 2p_jq_j$ represents heterozygosity.

Because H_j and g_j^2 are independent,

$$E(H_j)\,E(g_j^2) = E(H_j)\,\sigma_g^2 = \frac{\text{var}(y)}{p}, \quad \text{since } g_j \sim N(0, \sigma_g^2).$$

$E(H_j)$ was estimated by averaging H_j across all p loci (due to assumption 1) and var(y) was calculated from the empirical distribution of phenotypic values. Thus, σ_g^2 was available. Finally, the upper bound u in the uniform (0, u) distribution was set to be twice as

https://doi.org/10.1017/S0016672310000157 Published online by Cambridge University Press

large as the estimated value of σ_g^2, to make the mean of this uniform distribution equal to the estimated σ_g^2.

This procedure was also applied to the principal component regression, as used for the two broiler datasets. In that case, the variance of the prior distribution of mega-SNP effects $\boldsymbol{m}$ (16) was the target. The expectation of the variance caused by each of the a mega-SNPs ($\boldsymbol{t}_j$, $j=1, \ldots, a$) was approximated as

$$E[\mathrm{var}(\boldsymbol{t}_j m_j)] = E[\mathrm{var}(\boldsymbol{t}_j) m_j^2]$$
$$= E[\mathrm{var}(\boldsymbol{t}_j)]E(m_j^2) = E[\mathrm{var}(\boldsymbol{t}_j)]\sigma_m^2 = \frac{\mathrm{var}(y)}{a},$$

and $E[\mathrm{var}(\boldsymbol{t}_j)]$ was estimated by averaging $\mathrm{var}(\boldsymbol{t}_j)$ across all a mega-SNPs. Thus, σ_m^2 was available.

APPENDIX D. COMPARE BAYES A WITH UNIFORM PRIOR AND SCALED INVERTED CHI-SQUARE PRIOR ON THE VARIANCES OF MARKER EFFECTS

The uniform prior was described in the preceding section. For the scaled inverted chi-square distribution, its degrees of freedom (ν) was fixed at 4·2, and the scale S was chosen so as to make the mean of the distribution equal the estimated value of σ_g^2, as for the uniform prior. Bayes A (additive effects only) with each of the two priors was fitted to the five simulated data used in the toy example, as described in section 6(ii), and to the two chicken datasets (BW and FCR) using real SNPs. Predictive correlations (corr($E(y)$, $\hat{y}$)) obtained with each of the two priors were similar.

	Scenarios in toy example					Chicken data	
	1	2	3	4	5	BW	FCR
Uniform	0·50	0·53	0·54	0·55	0·56	0·36	0·19
Scaled inverted chi-square	0·50	0·52	0·54	0·55	0·56	0·35	0·19

Support by the Wisconsin Agriculture Experiment Station, and by grants NRICGP/USDA 2003-35205-12833, NSF DEB-0089742 and NSF DMS-044371 is acknowledged. Two anonymous reviewers and the editor are thanked for useful comments.

References

Bell, D. D. & Weaver, W. D. (ed.) (2002). *Commercial Chicken Meat and Egg Production*. New York, NY: Springer.

Bennewitz, J., Solberg, T. & Meuwissen, T. (2009). Genomic breeding value estimation using nonparametric additive regression models. *Genetics Selection Evolution* **41**, 20.

Broomhead, D. & Lowe, D. (1988). Multivariable functional interpolation and adaptive networks. *Complex Systems* **2**, 321–355.

Cockerham, C. C. (1954). An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics* **39**, 859–882.

Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics* **11**, 2463–2468.

de los Campos, G., Gianola, D. & Rosa, G. J. M. (2009*a*). Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *Journal of Animal Science* **87**, 1883–1887.

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K. & Cotes, J. (2009*b*). Predicting quantitative traits with regression models for dense molecular markers and pedigrees. *Genetics* **182**, 375–385.

Falconer, D. S. and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*. 4th edn. Harlow, Essex, UK: Longmans Green.

Gianola, D. & de los Campos, G. (2008). Inferring genetic values for quantitative traits non-parametrically. *Genetical Research* **90**, 525–540.

Gianola, D., Fernando, R. & Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* **173**, 1761–1776.

Gianola, D. & van Kaam, J. (2008). Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* **178**, 2289–2303.

González-Recio, O., Gianola, D., Long, N., Weigel, K. A., Rosa, G. J. M. & Avendano, S. (2008). Nonparametric methods for incorporating genomic information into genetic evaluations: An application to mortality in broilers. *Genetics* **178**, 2305–2313.

González-Recio, O., Gianola, D., Rosa, G., Weigel, K. & Kranis, A. (2009). Genome-assisted prediction of a quantitative trait measured in parents and progeny: application to food conversion rate in chickens. *Genetics Selection Evolution* **41**, 3.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*. 2nd edn. Upper Saddle River, NJ: Prentice-Hall.

Hill, W. G., Goddard, M. E. & Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics* **4**, e1000008.

Hu, Y. H. & Hwang, J.-N. (ed.) (2001). *Handbook of Neural Network Signal Processing*. Boca Raton, FL: CRC Press.

Kaufman, L. & Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York, NY: Wiley.

Mackay, T. F. C. (2009). The genetic architecture of complex behaviors: lessons from *Drosophila*. *Genetica* **136**, 295–302.

Manly, B. F. J. (2005). *Multivariate Statistical Methods: A Primer*. 3rd edn. Boca Raton, FL: Chapman and Hall/CRC.

McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genetics* **5**, e1000686.

Meuwissen, T. H., Hayes, B. J. & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829.

Moody, J. & Darken, C. J. (1989). Fast learning in networks of locally-tuned processing units. *Neural Computation* **1**, 281–294.

Park, T. & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association* **103**, 681–686.

https://doi.org/10.1017/S0016672310000157 Published online by Cambridge University Press

Poggio, T. & Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science* **247**, 978–982.

Powell, M. J. D. (1987). Radial basis functions for multivariable interpolation: a review. In *Algorithms for Approximation*, pp. 143–167. New York, NY: Clarendon Press.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–909.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. New York, NY: Cambridge University Press.

Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856.

Solberg, T. R., Sonesson, A. K., Woolliams, J. A. & Meuwissen, T. H. E. (2008). Genomic selection using different marker types and densities. *Journal of Animal Science* **86**, 2447–2454.

Sturtz, S., Ligges, U. & Gelman, A. (2005). R2WinBUGS: a package for running WinBUGS from R. *Journal of Statistical Software* **12**, 1–16.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B* **58**, 267–288.

Varmuza, K. & Filzmoser, P. (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics*. Boca Raton, FL: CRC Press.

Wahba, G. (2002). Soft and hard classification by reproducing kernel Hilbert space methods. *Proceedings of the National Academy of Sciences of the USA* **99**, 16524–16530.

Xu, S. & Jia, Z. (2007). Genome-wide analysis of epistatic effects for quantitative traits in barley. *Genetics* **175**, 1955–1963.

Yamamoto, A., Zwarts, L., Callaerts, P., Norga, K., Mackay, T. F. C. & Anholt, R. R. H. (2008). Neurogenetic networks for startle-induced locomotion in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the USA* **105**, 12393–12398.

https://doi.org/10.1017/S0016672310000157 Published online by Cambridge University Press