Theor Appl Genet (2011) 123:1065–1074
DOI 10.1007/s00122-011-1648-y

ORIGINAL PAPER

Application of support vector regression to genome-assisted prediction of quantitative traits

Nanye Long · Daniel Gianola · Guilherme J. M. Rosa · Kent A. Weigel

Received: 14 March 2011 / Accepted: 22 June 2011 / Published online: 8 July 2011

Abstract A byproduct of genome-wide association studies is the possibility of carrying out genome-enabled prediction of disease risk or of quantitative traits. This study is concerned with predicting two quantitative traits, milk yield in dairy cattle and grain yield in wheat, using dense molecular markers as predictors. Two support vector regression (SVR) models, ε-SVR and least-squares SVR, were explored and compared to a widely applied linear regression model, the Bayesian Lasso, the latter assuming additive marker effects. Predictive performance was measured using predictive correlation and mean squared error of prediction. Depending on the kernel function chosen, SVR can model either linear or nonlinear relationships between phenotypes and marker genotypes. For milk yield, where phenotypes were estimated breeding values of bulls (a linear combination of the data), SVR with a Gaussian radial basis function (RBF) kernel had a slightly better performance than with a linear kernel, and was similar to the Bayesian Lasso. For the wheat data, where phenotype was raw grain yield, the RBF kernel provided clear advantages over the linear kernel, e.g., a 17.5% increase in correlation when using the ε-SVR. SVR with a RBF kernel also compared favorably to the Bayesian Lasso in this case. It is concluded that a nonlinear RBF kernel may be an optimal choice for SVR, especially when phenotypes to be predicted have a nonlinear dependency on genotypes, as it might have been the case in the wheat data.

Communicated by C. Schön.

N. Long (✉) · D. Gianola · G. J. M. Rosa
Department of Animal Sciences, University of Wisconsin,
1675 Observatory Dr., Animal Science Bldg, Madison,
WI 53706, USA
e-mail: nlong@wisc.edu

D. Gianola · K. A. Weigel
Department of Dairy Science, University of Wisconsin,
Madison, WI 53706, USA

D. Gianola · G. J. M. Rosa
Department of Biostatistics and Medical Informatics,
University of Wisconsin, Madison, WI 53706, USA

Introduction

The enormous amount of data stemming from high-throughput genotyping assays has prompted genome-wide association studies (GWAS) for many traits in many species. It has been increasingly recognized that, in addition to detection of causal variants (e.g., single nucleotide polymorphisms or SNPs), another utility of GWAS is to carry out genome-assisted prediction of genetic merit of individuals for the trait in question. In essence, this is equivalent to genomic selection as proposed by Meuwissen et al. (2001) for animal and plant breeding. Accurate prediction of genetic merit plays a crucial role in genetic improvement of livestock and plants, and also in personalized medicine where preventive and therapeutic decisions can be made for patients according to their genetic profiles. For a recent review of genome-enabled prediction methods for human disease susceptibility, see de los Campos et al. (2010).

Results from GWAS (e.g., Visscher 2008; Wei et al. 2009; Yang et al. 2010) have confirmed that one key to successful disease risk assessment or prediction of genetic values is the use of a large ensemble of markers, as opposed to setting stringent significance thresholds and capitalizing on those validated large-effect loci only. Additionally, albeit their relatively large effects, the small number of validated variants usually explain a very small fraction of phenotypic variation.

Prediction of genetic values can be carried out using parametric or nonparametric approaches. Parametric Bayesian regression models, such as BayesA (Meuwissen et al. 2001), are the most commonly used, and have the appeal that can be related to a body of theory of quantitative genetics. However, such methods are not flexible enough to incorporate complex gene action (e.g., dominance and epistasis) due to a rapid increase in model dimension. On the other hand, nonparametric methods, such as reproducing kernel Hilbert spaces (RKHS) regression and radial basis function (RBF) regression (e.g., Gianola et al. 2006; Gianola and van Kaam 2008; Long et al. 2010), provide an alternative to making predictions without imposing a specific functional structure between phenotypes and genotypes. The core component of such methods is the choice of a suitable kernel matrix capable of capturing unknown genetic complexity, and the resulting model is a regression of phenotypes on a number of kernel function evaluations.

This paper focuses on regression modeling for quantitative traits using support vector machines (Vapnik 1995), or SVM. SVM is regarded as a state-of-the-art machine learning algorithm for classification and regression problems in various fields. However, there have been few reported applications of SVM to genome-wide prediction of quantitative traits; two are Maenhout et al. (2007) and Moser et al. (2009) in plant and animal breeding, respectively. Generally speaking, the advantage of SVM lies in its use of nonlinear kernel functions to explore nonlinearities, which is in common with the aforementioned RKHS and RBF regressions.

Two sets of data were used to conduct empirical evaluations of SVM regressions. The first came from a Holstein population, where the trait of interest was milk yield (MY). The second was a set of wheat lines, with the trait being grain yield. We built SVM regression models using a large number of genetic markers as inputs for predicting phenotypes. Prediction results from SVM regressions were compared with those of a benchmark model, the Bayesian Lasso. Bayesian Lasso (Park and Casella 2008) is a well-established parametric regression method for genomic selection; especially, it has the nice feature of being fairly robust with respect to prior distributions of a regularization parameter controlling shrinkage of coefficients (de los Campos et al. 2009).

This paper is organized as follows. The “Data” section describes the two data sets (MY and wheat). Section “Support vector regression” describes two SVM regression models used in this study and details about parameter tuning for them. Section “Bayesian Lasso” briefly reviews the Bayesian Lasso method. After presenting the “Results” section we discuss our findings and related issues in “Discussion” and conclude our work in “Conclusions”.

Materials and methods

Data

MY data was from a sample of 4,703 Holstein sires. Phenotypes were sires’ predicted transmitting ability (PTA, half of the breeding value) for milk yield, obtained from conventional progeny testing. A panel of 32,518 whole-genome SNP markers was available for all sires, and they were used as model predictors. Within these sires, 3,305 (born between years 1952 and 1998) were assigned to a training set, and the remaining 1,398 sires (born between 1999 and 2002) were assigned to a test set. All models were fitted using data on training sires and genomic predictions of PTAs were made for testing sires. Genotypes and phenotypes were provided by the USDA Bovine Functional Genomics Lab and Animal Improvement Programs Lab, respectively, as in an earlier study (Weigel et al. 2009). Figure 1a depicts empirical distributions of the phenotypic and marker allele frequency data from all 4,703 sires. The minor allele frequencies of these 32,518 SNPs had an approximately uniform distribution with a slight excess of frequencies toward 0.5. This is probably due to SNP prescreening so as to increase informativeness of the final set of SNPs.

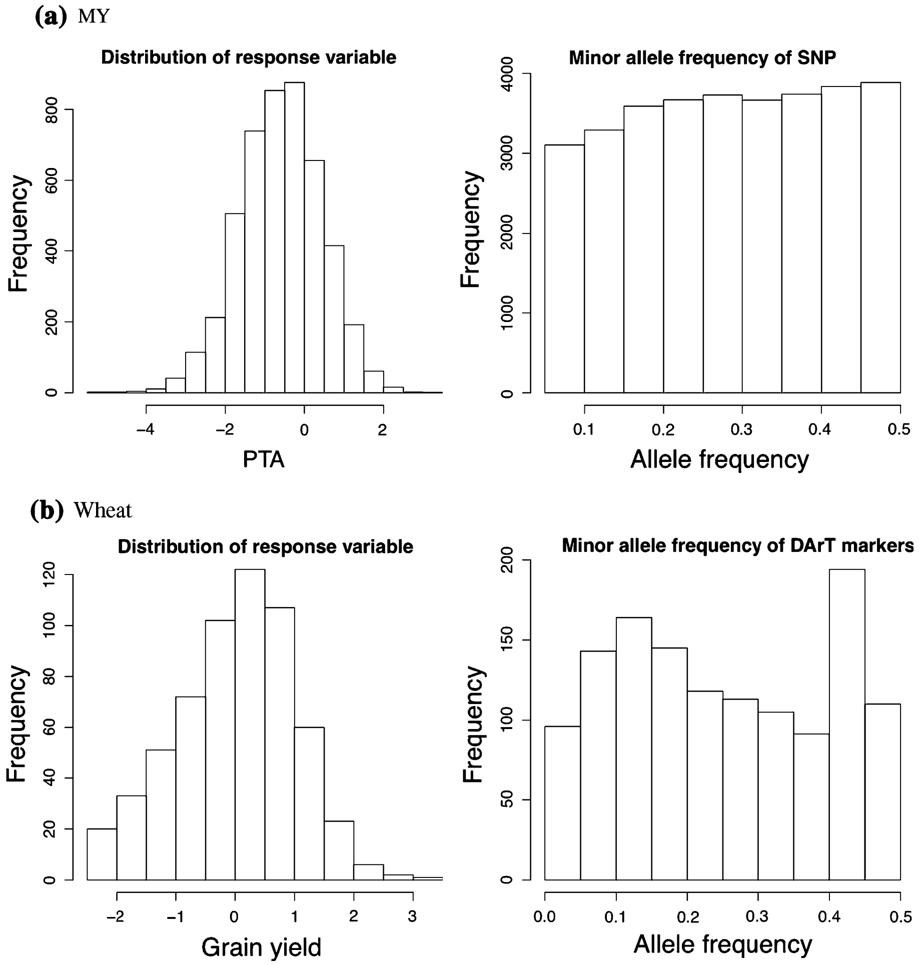
The wheat data contained 599 wheat lines, each genotyped with 1279 DArT markers (Diversity Array Technology). DArT markers may take on one of two values, denoting presence or absence of an allele. The data was from several international trials conducted at the International Maize and Wheat Improvement Center, Mexico. Edited data can be downloaded from R package BLR (http://cran.r-project.org/web/packages/BLR/index.html) and more information can be found in Crossa et al. (2007) and de los Campos et al. (2009). The phenotypic trait considered here was the average grain yield for each line collected in one the four macroenvironments chosen in these trials.

Distributions of phenotypes and genotypes are shown in Fig. 1b. The minor allele frequencies of DArT appeared to have a bimodal distribution. Unlike the MY data, where division into training-test sets was naturally based on birth years of sires, the wheat data was partitioned randomly into a training set (480 lines) and a test set (119 lines). This was repeated 50 times by sampling lines at random.

Support vector regression

The SVM developed by Vapnik (1995) is grounded in statistical learning theory, where the goal is to achieve good generalization performance (i.e., low prediction error of testing data), given a finite amount of training data. Useful reviews on SVM are Burges (1998), Cristianini and

Fig. 1 Distributions of the response variable and minor allele frequency of markers in MY and wheat data sets


(a) MY

(b) Wheat

Shawe-Taylor (2000), Smola and Schölkopf (2004), Bishop (2006) and Cherkassky and Mulier (2007) and references therein. Support vector regression (SVR) is an important application of the SVM methodology. In this study, two SVR models were investigated. One is the classic ε-SVR (Vapnik 1995) and the other is a least-squares (LS) SVR (Suykens et al. 2002). Both SVRs minimize a regularized loss function. However, they differ in the loss function chosen, as described below.

General formulation

Consider learning a mapping $f(\mathbf{x})$: $R^p \rightarrow R$, given a set of training data

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_i, y_i), \ldots, (\mathbf{x}_n, y_n),\ \mathbf{x}_i \in R^p, y_i \in R.$$

Here, $\mathbf{x}_i$ is a p-dimensional input vector, such as a vector of genotypic codes of p SNP markers; y_i is a real-valued response variable, e.g., phenotype. Specifically, one assumes that $f(\mathbf{x})$ is a linear function of the form $f(\mathbf{x}) = \mathbf{w}'\mathbf{x} + b$, with $\mathbf{w}$ being a vector of unknown weights (i.e., regression coefficients) and b being the bias. According to the SVM theory, learning $f(\mathbf{x})$ is by minimizing the following regularized loss function:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} L(e_i). \qquad (1)$$

In (1), $\|\mathbf{w}\|^2 = \mathbf{w}'\mathbf{w}$, represents model complexity; $e_i = y_i - f(\mathbf{x}_i)$ is the error associated with the ith ($i = 1,\ldots,n$) training data point; $L(\cdot)$ denotes the loss function, and C is a positive regularization parameter controlling the trade-off between model complexity and training error.

ε-SVR

In ε-SVR, the so-called ε-insensitive loss function (Vapnik 1995) is used for L, giving

$$L_\varepsilon(e) = \begin{cases} 0 \text{ if } |e| < \varepsilon \\ |e| - \varepsilon \text{ otherwise} \end{cases}$$

Hence, the loss function is zero ("insensitive") for any absolute error smaller than a predefined value ε. For an

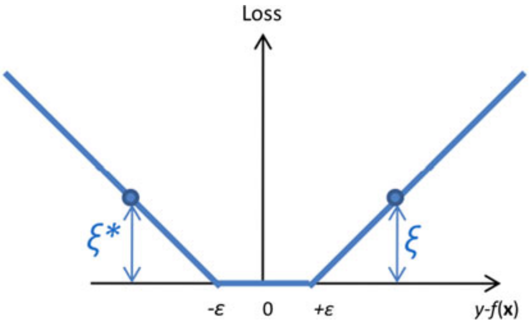


Fig. 2 Vapnik's ε-insensitive loss function for regression. The loss function is zero if the absolute error is less than ε, a predefined value. For an error with an absolute value larger than ε, the loss is equal to the extra error beyond ε. The loss is denoted by ξ or ξ^*, depending on the sign of the error

error larger than ε, the loss is equal to the difference between the absolute error and ε. Figure 2 illustrates the ε-insensitive loss function. It is more convenient to express the ε-insensitive loss function in terms of slack variables (ξ, ξ^*). By definition, one has $\xi > 0$ and $\xi^* = 0$ for data points with positive errors, and $\xi^* > 0$ and $\xi = 0$ for points with negative errors. Taken together, the optimization problem can now be formulated as

$$\min_{\mathbf{w},b,\xi,\xi^*}\left(\frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{n}\left(\xi_i+\xi_i^*\right)\right), \tag{2}$$

subject to

$$\xi_i \geq 0,\ \xi_i^* \geq 0,$$
$$y_i \leq f(\mathbf{x}_i)+\varepsilon+\xi_i,$$
$$y_i \geq f(\mathbf{x}_i)-\varepsilon-\xi_i^*,\ \text{for } i=1,\ldots n.$$

Both C and ε affect model complexity. A larger C or a smaller ε results in a more complicated model but smaller training errors.

LS-SVR

LS-SVR uses a common squared loss function, leading to the following optimization problem:

$$\min_{\mathbf{w},b,e}\left(\frac{1}{2}\|\mathbf{w}\|^2+\frac{C}{2}\sum_{i=1}^{n}e_i^2\right). \tag{3}$$

As in ε-SVR, here, C is a positive regularization parameter. In fact, (3) is essentially a ridge regression problem.

Solutions

The standard method for solving the SVR optimization problem is to construct a Lagrange function from (2) or (3) so as to cast the original problem into a dual space of Lagrange multipliers and find solutions therein (see, e.g., Nocedal and Wright 1999). The final solutions of the two SVRs take the same form, that is, $f(\mathbf{x})=\sum_{i=1}^{n}\alpha_i\mathbf{x}'_i\mathbf{x}+b$, where α_i's and b can be obtained by solving either a quadratic programming problem (ε-SVR) or a set of linear equations (LS-SVR) in the dual space. It is useful to note that the solution is a linear combination of inner products $\mathbf{x}'_i\mathbf{x}$ and, thus, linear in the input data $\mathbf{x}$. An important property of the ε-SVR solution is sparseness, meaning that a fraction of α_i's are equal to zero and thereby vanishing in the final model $f(\mathbf{x})$. Only data points with $\alpha_i > 0$ are relevant and are termed "support vectors". However, sparsity does not hold for LS-SVR.

The use of kernel functions for nonlinear SVR modeling

A feature of SVR is that its formulation (in the dual space) and solution depends only on the inner products (e.g., $\mathbf{x}'_i\mathbf{x}_j$). A kernel function returns the value of the inner product between two vectors in some transformed or feature space (which can be infinite dimensional) and, therefore, can be used in place of the inner products in the SVR solution. Generally, a kernel can be expressed as $k(\mathbf{x},\ \mathbf{z})=\varphi(\mathbf{x})'\varphi(\mathbf{z})$, where $\mathbf{x}$ and $\mathbf{z}$ are two vectors in the original space, whereas $\varphi(\mathbf{x})$ and $\varphi(\mathbf{z})$ are the corresponding vectors in the feature space. Since a kernel is a function of two vectors in the original space, one does not need to know explicitly the mapping $\varphi(\mathbf{x})$, which is a great convenience for formulating a nonlinear SVR.

A linear kernel performs an identity mapping, that is, $k(\mathbf{x},\ \mathbf{z})=\mathbf{x}'\mathbf{z}$, as used in the SVR described earlier. On the other hand, a nonlinear kernel is often used to potentially increase SVR's predictive power. A commonly used nonlinear kernel is the Gaussian RBF function (RBF kernel for short); its general form is $k(\mathbf{x},\ \mathbf{z})=$ exp $(-\|\mathbf{x}-\mathbf{z}\|^2/\sigma^2)$, which is indexed by a bandwidth parameter σ. Given the choice of kernel function, solving the SVR problem follows the same Lagrange method mentioned before, and the resulting nonlinear SVR model is a linear combination of a set of kernel functions: $f(\mathbf{x})=\sum_{i=1}^{n}\alpha_i k(\mathbf{x},\mathbf{x}_i)+b$.

Selection of tuning parameters

As noted, to solve the optimization problem of ε-SVR, i.e., (2), one needs to preset two tuning parameters: C and ε. In LS-SVR, there is only one regularization parameter C in (3). For both SVRs, when a Gaussian RBF kernel is used, an additional tuning parameter, the bandwidth σ, must also be determined. Parameter tuning for each of the two SVRs in this study is detailed below.

LS-SVR

Tuning C or σ for LS-SVR was done by grid search over a range of values, and each value was evaluated by cross-validation (CV) on the training set (using correlation as a criterion). A fivefold CV was used, and the value giving rise to the best CV performance was chosen. The grid values of C for LS-SVR ranged from 0.5 to 70 for both the MY and wheat data. This grid was obtained by locally refining an initial broader grid, after several trials. For σ, the candidate set of values was chosen based on the distribution of Euclidean distances between training samples. Linking σ to sample Euclidean distance was proposed by Coen et al. (2006), and the basic rationale is as follows. Note that a (scaled) Gaussian RBF kernel function has form

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|/\sqrt{p}}{\sigma}\right)^2,$$

where $\mathbf{x}_i$ ($\mathbf{x}_j$) is the vector of genotypes of individual i (j). The scaling via $\sqrt{p}$ (dimension of $\mathbf{x}_i$) is applied because the value of the Euclidean norm $\|\mathbf{x}_i - \mathbf{x}_j\|$ grows with p. The idea is that the range of $\|\mathbf{x}_i - \mathbf{x}_j\|/\sqrt{p}$ (the scaled Euclidean distance) can be used as a reasonable guide for the range of σ values; σ being too large or too small relative to the distribution of $\|\mathbf{x}_i - \mathbf{x}_j\|/\sqrt{p}$ tends to make the kernel function approach its extreme values 1 or 0, thereby affecting prediction accuracy adversely.

Figure 3 presents distributions of the scaled Euclidean distances of all pairs of training samples in the MY and wheat data. Figure 4 illustrates (using MY data) that a σ value deviating far away from the range of the sample Euclidean distances (about 0.7 to 1) leads to a degraded CV performance; smaller values of σ seem to have a stronger adverse effect than larger values. Guided by the distributions shown in Fig. 3, the grid values chosen for σ ranged from 0.1 to 4.5 for MY, and from 0.1 to 1 for wheat; these ranges covered the optimal values according to our tuning results.

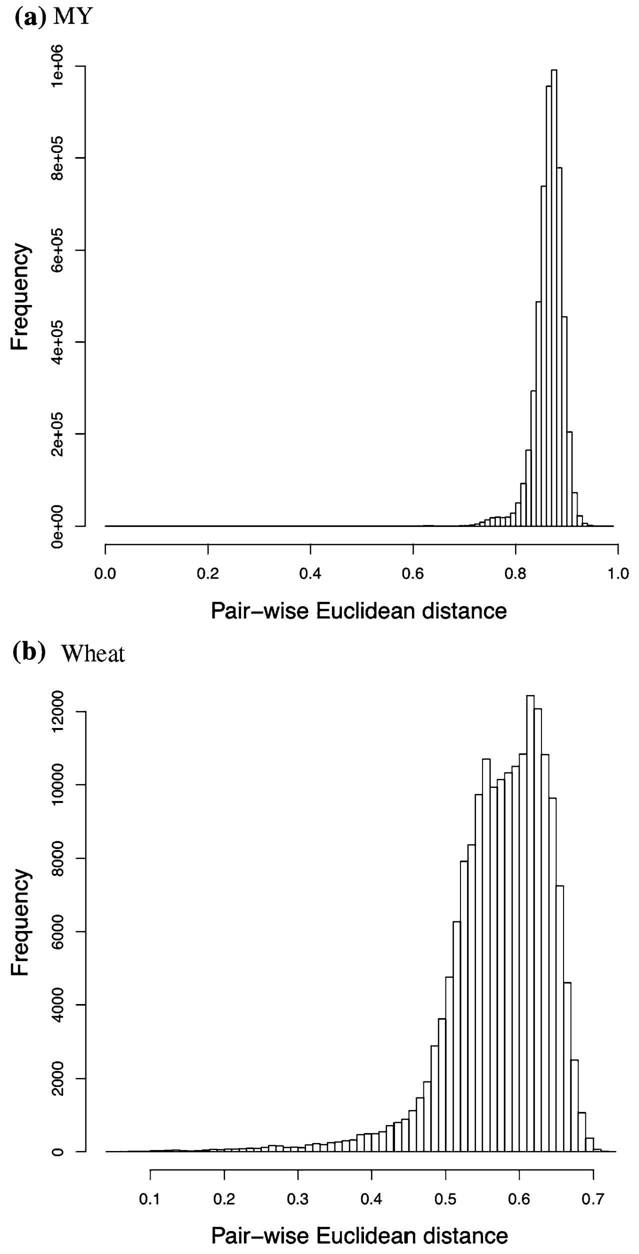


Fig. 3 Histograms of Euclidean distances between training sample points in the MY and wheat data. Distances ($\|\mathbf{x}_i - \mathbf{x}_j\|/\sqrt{p}$) were computed for all pairs of training data points

ε-SVR

In ε-SVR with a linear kernel, C and ε were selected via a fivefold CV on the training set, by grid search. The grid values of C ranged from 10^{-5} to 1 for MY data, and from 10^{-3} to 1 for wheat data; the grid values of ε ranged from 10^{-4} to 1 for MY data, and from 10^{-5} to 1 for wheat data.

For ε-SVR with a RBF kernel, the additional tuning of the bandwidth parameter σ implies a grid search over a 3-dimensional parameter space (C, ε, σ). For a large data set, such as the MY data (with 3,305 records and 32,518 predictors), computational burden in this case is enormous. Alternatively, Cherkassky and Ma (2004b) proposed a practical and analytical method to choose C and ε directly from the training data, assuming that a Gaussian RBF kernel is used. Specifically, the value of C is chosen as

$$C = \max\left(|\bar{y} + 3\sigma_y|,\ |\bar{y} - 3\sigma_y|\right)$$

where $\bar{y}$ and σ_y are the mean and standard deviation of the response values, respectively. This can effectively handle outliers in the training data. For ε, its value is given by

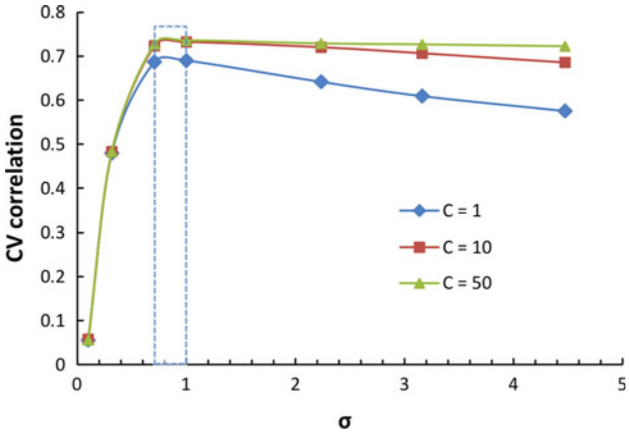


Fig. 4 Influence of bandwidth parameter (σ) in the Gaussian RBF kernel on cross-validation correlation of LS-SVR, for different values of C (the regularization parameter). MY data was used. The range of sample Euclidean distances is indicated in the *box*

$$\varepsilon = 3\sigma_e\sqrt{\frac{\ln n}{n}}$$

where σ_e is the residual standard deviation, and n is training sample size. It is assumed that the residual standard deviation is known or can be estimated from data. In our case, however, σ_y was used instead of σ_e, as the latter was unknown. Although one can first perform some prior analysis (e.g., via the Bayesian Lasso) to obtain an estimate of σ_e, it may lead to an unfair comparison with other methods that do not utilize data driven information. After C and ε were determined analytically, the bandwidth σ in the RBF kernel was chosen by grid search in conjunction with a fivefold CV evaluation. The grid of values searched for σ was the same as those in LS-SVR.

Implementation

For each of ε-SVR and LS-SVR, two types of kernels were considered: a linear kernel and a Gaussian RBF kernel. The toolbox LIBSVM (Chang and Lin 2001) was used to implement ε-SVR, and another toolbox, LS-SVMlab (v1.7) (Pelckmans et al. 2007), was used for LS-SVR.

Bayesian Lasso

The Bayesian Lasso was used to estimate all marker effects simultaneously, and was employed as a benchmark. This method was first proposed by Park and Casella (2008) and has been applied to QTL mapping (e.g., Yi and Xu 2008) and genomic selection (e.g., de los Campos et al. 2009) recently.

Consider the linear regression model $\mathbf{y} = \mathbf{1}_n\mu + \mathbf{X}\beta + \mathbf{e}$, where $\mathbf{y}$ is an $n \times 1$ vector of phenotypes; μ is a common effect to all individuals; $\mathbf{X}$ is an $n \times p$ incidence matrix of marker genotypes (coded as 0, 1 or 2 for bi-allelic SNP); β is a $p \times 1$ vector of unknown coefficients, i.e., regressions on markers; and $\mathbf{e}$ is a vector of independent and identically distributed residuals distributed as $N(\mathbf{0}, \mathbf{I}\sigma_e^2)$, where σ_e^2 is the residual variance.

Bayesian Lasso assigns the same double exponential prior distribution to each element of β, β_j (j = 1,…, p). This is equivalent to the following two steps (Park and Casella 2008):

$$p(\beta_j|\tau_j) \sim N(0, \tau_j^2),\ j = 1, \ldots, p$$
$$p(\tau_j) \sim \text{Exponential}(\lambda),\ j = 1, \ldots, p$$

After integrating out the τ_j's, the marginal distribution for each β_j (given λ) can be shown to be double exponential, with density

$$p(\beta_j|\lambda) \sim \frac{\sqrt{2\lambda}}{2}\exp\left(-\sqrt{2\lambda}|\beta_j|\right).$$

Compared to a Normal prior, the double exponential produces stronger shrinkage of coefficients that are close to zero and less shrinkage of those with large absolute values (de los Campos et al. 2009). Usually, λ is assigned a conjugate Gamma prior with its hyperparameters (shape α and rate β) chosen by the user. Specifically, β controls the amount of shrinkage: a small value produces a strong shrinkage, reflecting the prior belief that most marker effects are nil. Further, the fully conditional posterior distribution of λ is Gamma$\left(\alpha + p,\ \beta + \sum_{j=1}^{p}\tau_j^2\right)$. Hence, a small value for β also corresponds to a vague prior.

In our Bayesian Lasso implementation, the prior distributions used were as follows. (1) μ was assigned a flat prior. (2) The prior for the error variance, σ_e^2, was a scaled inverted Chi-square distribution with degrees-of-freedom $v_e = 4.2$, and scale $S_e^2 = 1$. 3) For λ, a vague Gamma ($\alpha = 1$, $\beta = 0.001$) distribution was chosen. All posterior distributions were sampled via Gibbs sampling. A single chain was run for 80,000 iterations, with the first 50,000 discarded as burn-in. The remaining iterations were thinned at a rate of 30. The posterior mean (after burn-in and thinning) of each parameter was used as its point estimate. The Bayesian Lasso was coded in Fortran 90.

Results

Table 1 summarizes results obtained with the MY and wheat data sets. Two measures of predictive performance were used, predictive correlation and predictive mean squared error (PMSE). Given observed response values in the test set ($\mathbf{y}$) and their predicted values ($\hat{\mathbf{y}}$), the predictive correlation was the Pearson's correlation between $\mathbf{y}$ and $\hat{\mathbf{y}}$,

Table 1 Predictive correlations and predictive mean squared errors (PMSE) on the testing set by different methods: ε-SVR, LS-SVR and Bayesian Lasso

Criterion	ε-SVR		LS-SVR		Bayesian Lasso
	Linear	RBF	Linear	RBF	
MY data					
Correlation	0.669	0.692	0.689	0.700	0.700
PMSE	0.600	0.572	0.544	0.532	0.500
Wheat data					
Correlation	0.497 (0.054)	0.584 (0.050)	0.517 (0.056)	0.584 (0.056)	0.515 (0.056)
PMSE	0.799 (0.086)	0.686 (0.071)	0.765 (0.083)	0.688 (0.079)	0.768 (0.078)

"Linear" or "RBF" represents kernel functions used in the SVRs. In MY data, the training-test partition was fixed according to birth years of bulls, whereas in the Wheat data, the training-test partition was random and repeated 50 times with standard errors given in parentheses. Bayesian Lasso results for MY data are from Vázquez et al. (2010)

and PMSE was $(\mathbf{y}-\hat{\mathbf{y}})'(\mathbf{y}-\hat{\mathbf{y}})/n_{test}$, where n_{test} is the size of test sample.

For MY data, the following was observed.

1. LS-SVR was slightly better than ε-SVR when either a linear or a RBF kernel was used. The largest difference between the two SVR's was found with respect to PMSE when both used a linear kernel, in which case LS-SVR reduced PMSE by 9% relative to ε-SVR.
2. For either LS-SVR or ε-SVR, the RBF kernel was slightly better than the linear kernel. The largest difference between the two kernels was found for PMSE when ε-SVR was used, where the RBF kernel reduced PMSE by 4.7% relative to the linear kernel.
3. Overall, the best SVR performance was attained when using LS-SVR with a RBF kernel; in this case, the predictive correlation (0.7) of SVR was equal to that of the Bayesian Lasso, while the PMSE (0.532) of SVR was a little larger than that of Bayesian Lasso (0.50).

For the wheat data, results are expected to be less variable than those from the MY data, as they were averaged over 50 training–testing replicates. Clear differences were found between kernels for each of the two SVR's. The improvement of the RBF kernel over the linear kernel was 17.5% in correlation and 14.1% in PMSE in the case of ε-SVR. For LS-SVR, the improvement was 13% in correlation and 10% in PMSE. Prediction results of Bayesian Lasso were closest to those of LS-SVR with a linear kernel. Moreover, when using a Gaussian RBF kernel, the two SVR's had similar predictive ability, and were better than the Bayesian Lasso in this respect.

While our primary interest was in assessing predictive ability of various procedures, we also checked the extent of agreement of the predicted values yielded by them. This was done for the MY data only because the wheat data analysis involved 50 training-test sets, and it was not practical to display results replicate by replicate. To do this, $\hat{\mathbf{y}}$ values predicted by all methods were displayed in pair-wise scatter plots as shown in Fig. 5. Overall, there was a strong consistency between any two of the five methods. Predictions from different methods were highly correlated and very close to each other. The most similar predictions were those between the Bayesian Lasso and the linear LS-SVR. On the other hand, some scatter plots deviated somewhat from the 45° line, such as the ε-SVR Linear and the LS-SVR RBF pair.

Discussion

The SVR approach enables modeling of nonlinear relationships between phenotypes and SNPs via the use of a nonlinear kernel, such as the Gaussian RBF kernel. Compared to a linear kernel (which reduces SVR to a linear regression model), the RBF kernel generally had a better predictive ability in the two data sets analyzed here. In the MY data, the response variable (PTA) is expected to reflect the sum of additive genetic effects of all loci affecting the trait. Therefore, a linear model should presumably be a sensible choice. However, for both ε-SVR and LS-SVR, a slight improvement in prediction accuracy was achieved by using a RBF kernel relative to using a linear kernel. On the other hand, the response variable in the wheat data was grain yield phenotype and, as such, may have a nonlinear dependency on marker genotypes due to, e.g., dominance or epistasis (Maccaferri et al. 2008). In this case, RBF had a significant advantage over the linear kernel in predictive correlation and PMSE (e.g., 17.5% increase in correlation for ε-SVR and 13% for LS-SVR). Hence, the relative advantages of different methods are clearly data-dependent. This was also reflected in the comparison of SVR vs. the benchmark Bayesian Lasso, a purely additive model. Bayesian Lasso seemed to be the best choice for MY data, when considering predictive correlation and PMSE, but it

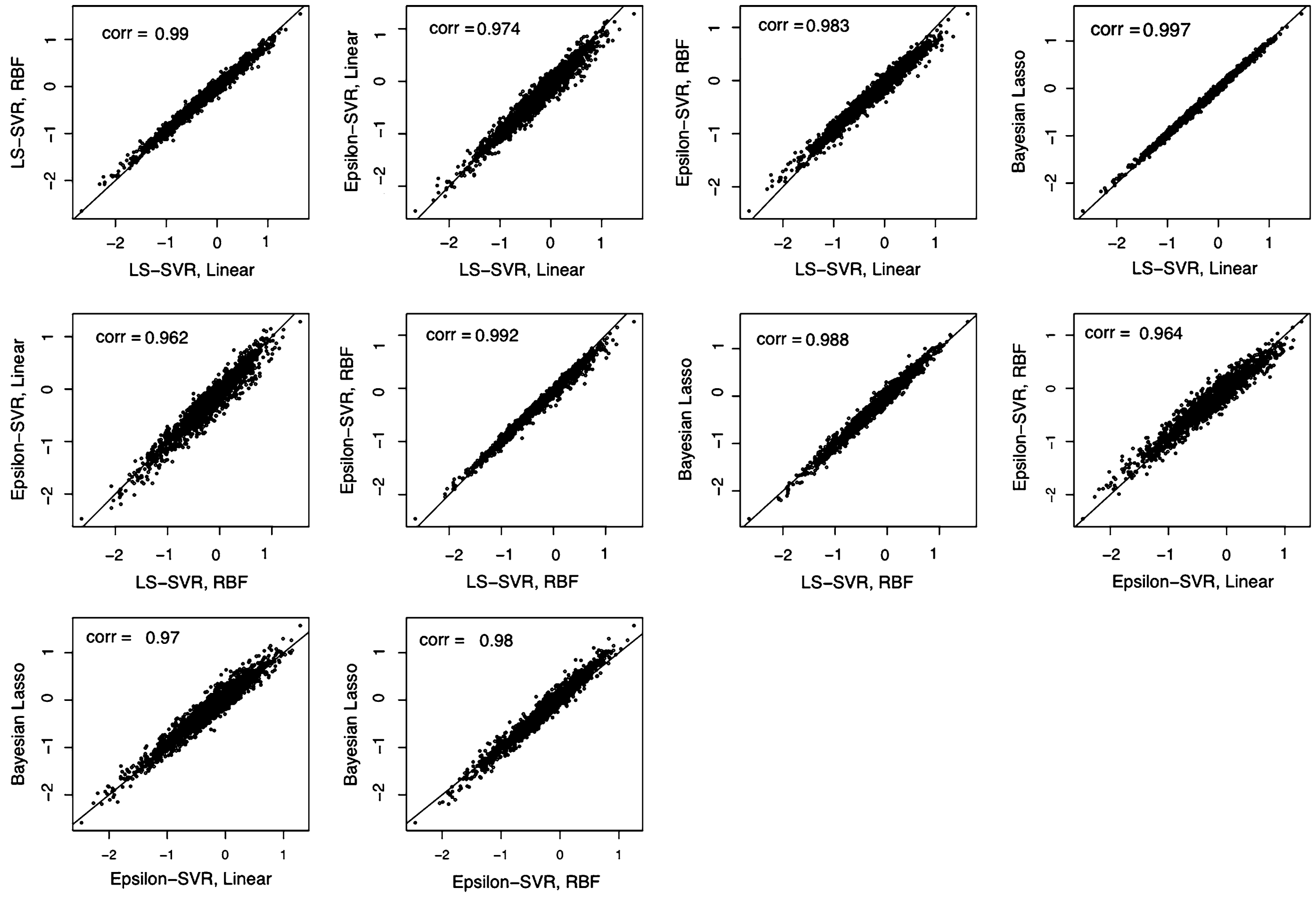


Fig. 5 Predicted response values in MY data by different methods. Each *plot* gives predictions from two methods, with the correlation and a 45° line indicated

was inferior to the nonlinear SVR when predicting wheat yield.

As pointed out by Ben-Hur et al. (2008), in many bioinformatics applications the linear kernel is competitive, and its performance is not worse than that of more flexible kernels, especially in a "small *n* large *p*" setting. Our empirical results indicated that the choice of kernel depends on the phenotype-marker relationship, in agreement with Crossa et al. (2010). While the nonlinear kernel should be better in handling complex relationships (e.g., epistasis), it may overfit the data if the response variable is governed by a simple (e.g., additive) structure and if only a finite sample is available for model training. Moreover, non-linear SVR models suffer from lack of obvious interpretability. For example, although they are potentially capable of handling interactions among markers, one cannot identify main or interactive effects. While this is an impediment to the understanding of genetic mechanisms, such models are nonetheless useful from the perspective of prediction.

The main difference between the two SVR models studied lies in the loss function adopted in the regularized optimization problem. The ε-SVR uses an ε-insensitive loss function, which ignores "small" errors and assigns "large" errors an absolute-value loss; the LS-SVR uses a common squared loss. From the point of view of prediction, the appropriateness of a loss function depends on the actual noise structure of the data. For example, the ε-insensitive loss typically outperforms other loss functions when the noise has a bimodal distribution (Cherkassky and Ma 2004a). Our MY and wheat data analysis indicated that LS-SVR seemed to be a better choice than ε-SVR, especially when a linear kernel was in use, although the difference was small. Moreover, LS-SVR has a practical advantage, because it requires one less tuning parameter than ε-SVR, as mentioned before.

The linear LS-SVR and the Bayesian Lasso use the same squared loss function but different penalty functions on estimated marker effects. LS-SVR imposes a L_2 penalty while the Bayesian Lasso imposes a L_1 penalty. Another important difference is that the Bayesian Lasso produces heterogeneous shrinkage on marker effects, a feature that has been strongly advocated for handling genome-wide markers for complex trait prediction (e.g., Meuwissen et al.

2001). In contrast, LS-SVR assumes that all markers are equally important. This is a consequence of solving a dual problem, where a weight is assigned to each sample rather than to each predictor variable. Nonetheless, this had little influence on prediction performance as found here. The two linear models had very similar predictive correlations and PMSE values for both MY and wheat data (Table 1).

A SVM characterized by a kernel representation is a way of handling complexity. A nice fact is that kernel design can be decoupled from the learning algorithm, such that context-specific kernels can be constructed for the task at hand. Besides the commonly used kernels (e.g., Gaussian RBF, polynomial), a large number of complicated kernels for specific types of data exists, such as those for text, protein/DNA sequences. Details on kernels can be found in Shawe-Taylor and Cristianini (2004).

It is possible to find parameter-free kernels for a specific context (Watkins 2000; Maenhout et al. 2007; González-Recio et al. 2009). This offers great computational advantage because tuning of kernel parameters is avoided. For example, genetic similarity measured by the complement of modified Rogers' distance (or MRD) (Wright 1978; Goodman and Stuber 1983) has been used for constructing kernels that do not contain parameters (e.g., Maenhout et al. 2007). They evaluated ε-SVR in predicting phenotypic performance of untested hybrids on a real maize data set. One of the kernels employed in the study was based on simple sequence repeat markers. They used MRD to compute dissimilarity (d_{kl}) between hybrids k and l, and then took its complement (s_{kl}) as element (k, l) of the kernel:

$$s_{kl} = 1 - d_{kl} \text{ where } d_{kl} = \frac{1}{\sqrt{2s}} \sqrt{\sum_{i=1}^{s} \sum_{j=1}^{n_i} \left(p_{ij}^{k} - p_{ij}^{l}\right)^2}$$

Above, s is the number of loci, n_i is the number of allele for locus i and p_{ij}^{k}, p_{ij}^{l} represent allele frequency of the jth allele at locus i for hybrids k and l, respectively. It was found that s_{kl} produced predictions that were similar to those obtained with a Gaussian RBF kernel. This suggests that context-specific kernels may be good alternatives to commonly used kernels, as long as they produce positive definite matrices and are constructed in a biologically meaningful way.

Conclusions

Two support vector regression models (ε-SVR and LS-SVR) were investigated using milk yield and wheat grain yield data in terms of their ability to predict quantitative traits using a large number of genetic markers. For each method, a linear kernel and a Gaussian RBF kernel were compared. In general, the RBF kernel had better predictive performance than the linear kernel, and its superiority was clearer in a situation in which phenotypes may be affected by non-additive marker effects.

Acknowledgments This work was supported by the Wisconsin Agriculture Experiment Station, Aviagen Ltd., and by grants NRICGP/USDA 2003-35205-12833, NSF DEB-0089742 and NSF DMS-044371. We thank the editor and reviewers for their insightful comments.

References

Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Ratsch G (2008) Support vector machines and kernels for computational biology. PLoS Comput Biol 4(10):e1000173

Bishop CM (2006) Pattern recognition and machine learning. Springer, New York

Burges CJ (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc 2:121–167

Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Cherkassky V, Ma Y (2004a) Comparison of loss functions for linear regression. In: Proceedings of the International Joint Conference on Neural Network

Cherkassky V, Ma Y (2004b) Practical selection of SVM parameters and noise estimation for SVM regression. Neural Netw 17(1):113–126

Cherkassky VS, Mulier F (2007) Learning from data: concepts, theory, and methods, 2nd edn. Wiley, Hoboken

Coen T, Saeys W, Ramon H, Baerdemaeker JD (2006) Optimizing the tuning parameters of least squares support vector machines regression for NIR spectra. J Chemometr 20:184–192

Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, New York

Crossa J, Burgueño J, Dreisigacker S, Vargas M, Herrera-Foessel SA, Lillemo M, Singh RP, Trethowan R, Warburton M, Franco J, Reynolds M, Crouch JH, Ortiz R (2007) Association analysis of historical bread wheat germplasm using additive genetic covariance of relatives and population structure. Genetics 177(3):1889–1913

Crossa J, de los Campos G, Perez P, Gianola D, Burgueno J, Araus JL, Makumbi D, Singh R, Dreisigacker S, Yan J, Arief V, Banziger M, Braun H-J (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724

de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel KA, Cotes J (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigrees. Genetics 182(1):375–385

de los Campos G, Gianola D, Allison DB (2010) Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat Rev Genet 11:880–886

Gianola D, van Kaam J (2008) Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178(4):2289–2303

Gianola D, Fernando R, Stella A (2006) Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173(3):1761–1776

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. Genet Sel Evol 41(1):3

Goodman M, Stuber C (1983) Races of maize: VI. Isozyme variation among races of maize in Bolivia. Maydica 28:169–187

Long N, Gianola D, Rosa GJM, Weigel KA, Kranis A, González-Recio O (2010) Radial basis function regression methods for predicting quantitative traits using SNP markers. Genet Res 92(3):209–225

Maccaferri M, Sanguineti MC, Corneti S, Ortega JLA, Salem MB, Bort J, DeAmbrogio E, del Moral LFG, Demontis A, El-Ahmed A, Maalouf F, Machlab H, Martos V, Moragues M, Motawaj J, Nachit M, Nserallah N, Ouabbou H, Royo C, Slama A, Tuberosa R (2008) Quantitative trait loci for grain yield and adaptation of durum wheat (*Triticum durum* Desf.) across a wide range of water availability. Genetics 178(1):489–511

Maenhout S, Baets BD, Haesaert G, Bockstaele EV (2007) Support vector machine regression for the prediction of maize hybrid performance. Theor Appl Genet 115:1003–1013

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819–1829

Moser G, Tier B, Crump RE, Khatkar MS, Raadsma HW (2009) A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet Sel Evol 41(1):56

Nocedal J, Wright SJ (1999) Numerical optimization. Springer, New York

Park T, Casella G (2008) The Bayesian lasso. J Am Stat Assoc 103:681–686

Pelckmans K, Suykens JAK, Gestel TV, Brabanter JD, Lukas L, Hamers B, Moor BD, Vandewalle J (2007) LS-SVMlab: a MATLAB/C toolbox for least squares support vector machines. Software available at http://www.esat.kuleuven.be/sista/lssvmlab/

Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge University Press, New York

Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. Stat Comput 14(3):199–222

Suykens J, Gestel TV, Brabanter JD, Moor BD, Vandewalle J (2002) Leaset squares support vector machines. World Scientific, Singapore

Vapnik V (1995) The nature of statistical learning theory, 2nd edn. Springer, New York

Vázquez AI, Rosa GJM, Weigel KA, de los Campos G, Gianola D, Allison DB (2010) Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. J Dairy Sci 93:5942–5949

Visscher PM (2008) Sizing up human height variation. Nat Genet 40(5):489–490

Watkins C (2000) Dynamic alignment kernels. In: Smola AJ, Bartlett PL, Schölkopf B, Schuurmans D (eds) Advances in large margin classifiers. MIT Press, Cambridge

Wei Z, Wang K, Qu H-Q, Zhang H, Bradfield J, Kim C, Frackleton E, Hou C, Glessner JT, Chiavacci R, Stanley C, Monos D, Grant SFA, Polychronakos C, Hakonarson H (2009) From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. PLoS Genet 5(10):e1000678

Weigel KA, de los Campos G, González-Recio O, Naya H, Wu XL, Long N, Rosa GJM, Gianola D (2009) Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. J Dairy Sci 92(10):5248–5257

Wright S (1978) Variability within and among natural populations. In: Evolution and the genetics of populations

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42(7):565–569

Yi N, Xu S (2008) Bayesian LASSO for quantitative trait loci mapping. Genetics 179(2):1045–1055