Journal of Computer-Aided Molecular Design **18:** 537–547, 2004.
DOI 10.1007/s10822-004-4063-5

QSAR modeling based on the bias/variance compromise: A harmonious and parsimonious approach

John H. Kalivas*, Joel B. Forrester & Heather A. Seipel
Department of Chemistry, Idaho State University, Pocatello, ID 83209, USA

Received 19 April 2004; accepted in revised form 29 July 2004

Key words: calibration, harmonious, multiple linear regression, parsimonious, partial least squares, principal component regression, ridge regression

Summary

Modeling quantitative structure–activity relationships (QSAR) is considered with an emphasis on prediction. An abundance of methods are available to develop such models. Using a harmonious approach that balances the bias and variance of predictions, the best calibration models are identified relative to the bias and variance criteria used. Criteria utilized to determine the adequacy of models are the root mean square error of calibration (RMSEC) and validation (RMSEV), respective R^2 values, and the norm of the regression vector. QSAR data from the literature are used to demonstrate concepts. For these data sets and criteria used, it is suggested that models obtained by ridge regression (RR) are more harmonious and parsimonious than models obtained by partial least squares (PLS) and principal component regression (PCR) when the data is mean-centered. The most harmonious RR models have the best bias/variance tradeoff, reflected by the smallest RMSEC, RMSEV, and regression vector norms and the largest calibration and validation R^2 values. The most parsimonious RR models have the smallest effective rank.

Introduction

Quantitative structure–activity relationships (QSAR) is concerned with understanding how molecular structural variation affects biological activity for a set of compounds [1, 2]. For example, QSAR is used to explain why a certain drug produces a particular effect and eventually predict the effect of newly synthesized compounds. Prediction is accomplished by using a mathematical model that describes a relationship between the prediction property of interest and molecular descriptors for a series of compounds based on physicochemical, quantum mechanical, and conformational variables. The multivariate calibration model commonly used is

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}, \tag{1}$$

where $\mathbf{y}$ symbolizes the $m \times 1$ vector of quantitative information for the property of interest for m calibration samples, e.g., inhibition, toxicity, etc., $\mathbf{X}$ denotes the $m \times p$ matrix of respective values for p molecular descriptors (variables), e.g., number of bonds, molecular polarizability, etc., $\mathbf{b}$ signifies the $p \times 1$ vector of unknown regression coefficients, and $\mathbf{e}$ represents the $m \times 1$ vector of normally distributed errors with mean zero and covariance matrix $\sigma^2\mathbf{I}$.

A common approach to solving Equation 1 for an estimate of the regression vector is the least squares (LS) solution

*To whom correspondence should be addressed. Fax: +1-208-282-4373; E-mail: kalijohn@isu.edu

$$\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \quad (2)$$

which is often termed the multiple linear regression (MLR) solution. This estimate is restricted to situations where $m \geq p$ and matrix $\mathbf{X}$ is well-conditioned [3]. With the rapid expansion of molecular descriptors that can be calculated where the number of descriptors exceeds the number of compounds, it has become routine to use a selection algorithm to choose molecular descriptors in order to use Equation 2.

Alternatively, one of the popular multivariate methods of principal component regression (PCR), partial least squares (PLS), or ridge regression (RR) can be implemented, thereby avoiding the variable selection needed with MLR. However, these methods necessitate determining meta parameters in order to obtain good values for $\hat{\mathbf{b}}$ from the calculation

$$\hat{\mathbf{b}} = \mathbf{X}^{+}\mathbf{y}, \quad (3)$$

where the $+$ superscript designates a generalized inverse of $\mathbf{X}$. For PLS and PCR, the meta parameters for obtaining $\mathbf{X}^{+}$ are the number of basis vectors (factors or latent vectors) and RR requires a ridge value.

Depending on meta parameter values, the levels of prediction bias and variance relative to the LS model will vary where bias refers to the level of prediction accuracy from the model (how close the predicted values are to known or actual values) and variance relates to the level of uncertainty for the predicted values. As more factors are used with PCR and PLS, the degree of bias decreases while variance increases. Similarly, as the ridge parameter decreases and converges to zero, the level of bias also decreases while variance increases. This is known as the bias/variance tradeoff for *biased* modeling methods such as PCR, PLS, RR, and other less used biased methods such as continuum regression, cyclic subspace regression (CSR), continuum CSR, generalized ridge regression (GRR), ridge PCR, ridge PLS, generalized ridge PCR, generalized ridge PLS, significance regression, reduced rank regression (RRR), softly shrunk RRR, principal covariates regression, etc. [4–17].

Regardless of the modeling method used, the final model is typically determined from a plot of an error criterion against respective model meta parameter values. Error criteria regularly used include the root mean square error of cross-validation (RMSECV), calibration (RMSEC), and validation (RMSEV). Unfortunately, these criteria are indicators of prediction bias and do not include direct information on prediction variance expected from the model (see the *Prediction Variance* section in the Background section). Thus, overfitting QSAR data with PCR, PLS, RR, and other modeling methods is a concern that is easily overlooked.

Recent work using spectroscopic data demonstrated that prediction variance regression diagnostics are also important and should be used in conjunction with prediction bias measures [18–21, J. Forrester and J.H. Kalivas, submitted]. As with the spectroscopic data sets, when a variance indicator such as the Euclidean norm (2-norm) of the regression vector $\|\hat{\mathbf{b}}\|_2$ is plotted against a bias measure such as the RMSEC, optimal QSAR models should be discerned relative to respective scalings of the molecular descriptors. The goal is to simultaneously minimize both prediction bias and variance. This requires selecting a harmonious model from the plotted curve where an acceptable bias/variance tradeoff is obtained. Such a plot can be considered a Pareto plot [19, 22–26] and will also be termed the harmonious plot in this paper. This emphasis on including the bias/variance tradeoff for a harmonious model is supported by the H-principle [27, 28]. Thus, the goal is to not only determine a model with an acceptable prediction bias level as evaluated by an error criterion (RMSECV, etc.), but also a model with suitable prediction variance is needed.

In a harmonious plot, optimization criteria are used as axis and values for the criteria corresponding to solutions for the optimization problem are plotted. In the case of minimization, an edge should develop with solutions on the edge bending toward the origin forming an L-shaped curve. Solutions on the edge forming the curve are deemed superior and Pareto optimal because no solutions can be identified where all criteria decrease relative to criteria values for another solution. The area containing solutions to the right of the curve is termed inferior because improved solutions can be found where all criteria decrease relative to the inferior solutions. Because of the L-shape, the harmonious curve has also been termed the L-curve [29, 30].

When $\|\hat{\mathbf{b}}\|_2$ was plotted against RMSEC or RMSEV for PLS, PCR, RR, and GRR on the same harmonious plot in recent studies using spectroscopic data, it was found that RR and GRR are Pareto optimal to PLS and PLS is Pareto optimal to PCR [19, 21]. It was also observed that GRR and RR essentially formed the same models [19]. Other modeling methods have also been plotted with spectroscopic data and RR (GRR) has consistently been found to be Pareto optimal based on $\|\hat{\mathbf{b}}\|_2$ and RMSEC or RMSEV [19]. Work provided in references [31] and [32] has shown that RR is Pareto optimal to PCR.

The use of PLS or PCR with Equation 1 is common to QSAR modeling, but to the authors' knowledge, RR is not normally used. Thus, an important aspect of this paper is to demonstrate the Pareto optimality of RR relative to PLS and PCR with QSAR data. When inter-comparing PLS, PCR, and RR models, selection of the proper meta parameters is critical. Typically, model selection is based only on a prediction error measure such as the RMSECV and hence, improper selection of respective model meta parameters often occurs, resulting in wrong conclusions, e.g., PLS supplies better bias error values than RR when in fact, the wrong ridge value was used. If the correct ridge value is used, then RR may indeed be providing lower bias values than PLS. By using the harmonious plot of a variance indicator against a bias criterion with all the models on the same plot, there can be no confusion as to how the different modeling methods are performing. With the harmonious plot, inter- and intra-model comparisons are possible. This paper also demonstrates this with QSAR data.

Another desire in modeling is to identify the more parsimonious model. This is the model that usually contains the least number of variables. With respect to MLR, the model with the least number of molecular descriptors is considered the more parsimonious model. For RR, PLS, PCR, and others, it is difficult to ascertain which is the more parsimonious model. Recent work has shown that a measure of effective rank (ER) can be used to perform inter- and intra-model comparisons for parsimony [H. Seipel and J.H. Kalivas, in press]. Such a comparison has not been performed with QSAR data and this paper provides an ER evaluation.

Background

Prediction variance

Derivations in references [33] and [34] show that the variance of a predicted value for a compound, $V(\hat{y}_{\text{unk}})$, can be estimated by

$$V(\hat{y}_{\text{unk}}) = \left(s_{\mathbf{e}}^2 + s_{\mathbf{y}}^2 + s_{\mathbf{X}}^2\|\hat{\mathbf{b}}\|_2^2\right)\left(\frac{1}{m} + h_{\text{unk}}\right) + s_{\mathbf{x}_{\text{unk}}}^2\|\hat{\mathbf{b}}\|_2^2 + s_{e_{\text{unk}}}^2, \quad (4)$$

where $s_{\mathbf{e}}^2$, $s_{\mathbf{y}}^2$, and $s_{\mathbf{X}}^2$ represent estimated variances in $\mathbf{e}$, $\mathbf{y}$, and $\mathbf{X}$, respectively, $s_{e_{\text{unk}}}^2$ denotes the estimated residual variance for the unknown compound's (compound with unknown activity) molecular descriptors $\mathbf{x}_{\text{unk}}$ with variance $s_{\mathbf{x}_{\text{unk}}}^2$ ($y_{\text{unk}} = \mathbf{x}_{\text{unk}}^{\text{t}}\mathbf{b} + e_{\text{unk}}$) and $h_{\text{unk}} = \mathbf{x}_{\text{unk}}^{\text{t}}(\mathbf{X}^{\text{t}}\mathbf{X})^{+}\mathbf{x}_{\text{unk}}$ symbolizes the leverage value for the unknown compound. Equation 4 reveals that the variance of prediction is directly proportional to h_{unk} and $\|\hat{\mathbf{b}}\|_2^2$, i.e., an increase in either term causes an increase in the prediction variance.

An equation similar to Equation 4 from the literature is

$$V(\hat{y}_{\text{unk}}) = \sum_{j=1}^{w}\hat{b}_j^2 V(x_{\text{unk},j}) + \kappa\sum_{i=1}^{m}h_{\text{unk},i}^2 V(y_i), \quad (5)$$

where κ is a scalar that takes into account the errors in $\mathbf{X}$ and $h_{\text{unk},i}$ is obtained from $\mathbf{h}_{\text{unk}}^{\text{t}} = \mathbf{x}_{\text{unk}}^{\text{t}}\mathbf{X}^{+}$ [35]. The first term in Equation 5 reflects errors in measured values for the compound and the second term relates to error in the $\mathbf{y}$ values for the calibration samples. Assuming $V(y_i)$ to be constant ($s^2 \approx \sum(\hat{y}_i - y_i)^2/dof$ where dof denotes the degrees of freedom), $V(x_{\text{unk},j})$ to be constant $\left(s_{\mathbf{x}_{\text{unk}}}^2\right)$ and noting that $\sum_{i=1}^{m}h_{\text{unk},i}^2 = h_{\text{unk}}$ in Equation 4, then Equation 5 simplifies to

$$V(\hat{y}_{\text{unk}}) \approx s^2 h_{\text{unk}} + s_{\mathbf{x}_{\text{unk}}}^2\|\hat{\mathbf{b}}\|_2^2. \quad (6)$$

As with Equation 4, Equation 6 shows that prediction variance is directly proportional to h_{unk} and $\|\hat{\mathbf{b}}\|_2^2$.

If $s_{\mathbf{e}}^2 = s_{e_{\text{unk}}}^2$ and $s_{\mathbf{X}}^2 = s_{\mathbf{x}_{\text{unk}}}^2$ are assumed and when $s_{\mathbf{X}}^2$ dominates the other variance terms, then Equation 4 reduces to

$$V(\hat{y}_{\text{unk}}) \approx s_{\mathbf{X}}^2\|\hat{\mathbf{b}}\|_2^2\left(1 + \frac{1}{m} + h_{\text{unk}}\right). \quad (7)$$

Further noting that $1 \gg \frac{1}{m}$ and when h_{unk} is small (the compound is near the mean of the calibration and hence, $h_{\text{unk}} \ll 1$), then Equation 7 becomes

$$V(\hat{y}_{\text{unk}}) \approx s_{\mathbf{x}}^{\mathbf{2}} \|\hat{\mathbf{b}}\|_2^2. \tag{8}$$

Similar assumptions with Equation 6 result in Equation 8 which is the usual equation when error in $\mathbf{X}$ dominates. If $s_{\mathbf{e}}^2$ has a dominating role, Equation 4 reduces to

$$V(\hat{y}_{\text{unk}}) \approx s_{\mathbf{e}}^2 \left(1 + \frac{1}{m} + h_{\text{unk}}\right), \tag{9}$$

which is the common classical equation for prediction variance. Equations 4–9 show that the variance of prediction is affected by the magnitudes of h_{unk} and $\|\hat{\mathbf{b}}\|_2^2$ with respect to variance estimates $s_{\mathbf{e}}^2$, $s_{\mathbf{X}}^2$, and $s_{\mathbf{y}}^2$. For example, differences in prediction variance from compound to compound reside in a compound's relative position in the calibration space for the particular model. Thus, the closer the compound is to the mean of the calibration, the smaller h_{unk} becomes, thereby reducing the prediction variance. Similarly, a sizeable $\|\hat{\mathbf{b}}\|_2^2$ indicates that an inflated prediction variance is probable.

The preliminary studies performed in this paper make use of $\|\hat{\mathbf{b}}\|_2$ as a variance inflation factor to describe the potential variance that predicted values may incur. While $\|\hat{\mathbf{b}}\|_2^2$ is the term in the variance equations, $\|\hat{\mathbf{b}}\|_2$ is also reflective of the potential variance inflation. Actual variances will depend on the magnitude of errors in $\mathbf{X}$ and $\mathbf{y}$.

It should be kept in mind that because $\|\hat{\mathbf{b}}\|_2$ is only an indicator of variance, it is probably best used in an absolute sense for intra-model studies and not inter-model comparisons. For example, when the optimal λ value for RR or the optimal number of basis vectors for PCR, etc. are sought. If inter-modeling comparisons between RR, PLS, PCR, etc. are to be made, a more complete variance equation should be used.

Tikhonov regularization

Estimating $\mathbf{b}$ in Equation 1 by MLR, PLS, or PCR is regularly described as a minimization problem in the 2-norm of

$$\min\left(\|\mathbf{Xb} - \mathbf{y}\|_2^2\right) \tag{10}$$

[36, 37]. In essence, Equation 4 is a minimization of the accuracy error (bias). However, this can result in an overfitted model when biased methods are used. To circumvent this, Tikhonov regularization can be used with

$$\min\left(\|\mathbf{Xb} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{Lb}\|_2^2\right) \tag{11}$$

where $\mathbf{L}$ represents a matrix of values usually approximating a derivative operator such as the first or second in order to generate a smoothed regression vector and λ denotes a scalar that must be optimized which assists in controlling the shape of the regression vector with $\mathbf{L}$ [38, 39].

Minimization of expression (11) with $\mathbf{L} = \mathbf{I}$ is called Tikhonov regularization in standard form and is RR with ridge parameter λ. Note that while expression (11) represents RR when $\mathbf{L} = \mathbf{I}$, the standard approach in the chemistry literature for determining λ is not to use expression (11), but instead to use a ridge trace plot [40] or only a bias measure such as RMSEC, RMSECV, or RMSEV. When $\mathbf{L}$ is not the identity matrix, then expression (11) is said to represent Tikhonov regularization in general form. With QSAR data, it is not expected that $\mathbf{L} \neq \mathbf{I}$ would be beneficial because adjacent molecular descriptors (columns) in $\mathbf{X}$ are not related in a smoothed way as with spectroscopic data.

With expression (11), bias is minimized in the first term and variance is simultaneously minimized with the second term. Thus, a harmonious model is achieved balancing bias and variance. As previously noted, this significantly reduces the chance of obtaining an over- or underfitted model. The proper model can be identified with expression (11) by using a harmonious plot to form the L-shaped curve.

Application of harmonious plots to determine the number of respective factors for PCR and PLS and the ridge value for RR has been performed with spectroscopic data. Criteria evaluated consisted of variance measures $\|\hat{\mathbf{b}}\|_2$ and A-optimality ($\text{trace}(\mathbf{X}^T\mathbf{X})^+$) and bias diagnostics RMSEC, RMSECV, R^2, and pseudo-degrees of freedom [18–21, J. Forrester and J.H. Kalivas, submitted]. Recent work investigated using leverages as a variance indicator [J. Forrester and J.H. Kalivas, submitted]. In all cases, the characteristic L-shaped curves were obtained and the same PCR, PLS, and RR models were identified as optimally harmoni-

ous. By using A-optimality for a variance indicator, the harmonious L-shaped curve is sharper in some cases, thereby providing an easier determination of the optimal model.

As an example, points composing a PLS harmonious curve would correspond to PLS models formed with increasing number of PLS factors. Low rank, underfitted models, define the portion of the curve parallel to the *x*-axis (bias axis) with low variance and high bias values. High rank, overfitted models, delineate that part of the harmonious curve positioned parallel to the *y*-axis (variance axis) corresponding to high variance and low bias models. Ideally, the best model should occur at the corner of the plotted L-shaped curve reflecting the more harmonious model providing a good balance between minimization of variance and bias. The final model selected from this corner region is ultimately up to the user.

Cross-validation

One of the most well-known methods for model selection (variable and/or meta parameter selection) is cross-validation. The most common forms are leave-one-out cross-validation (LOOCV) and *v*-fold cross-validation to calculate the RMSECV values. The method of *v*-fold is analogous to LOOCV except that the data set is split into *v* disjunct groups of approximately the same size and the modeling algorithm is run *v* times with $v-1$ groups for the calibration set and one group for the validation set. As with LOOCV, each group is left out once. When $v = m$, *v*-fold is LOOCV.

Essentially, these two cross-validation approaches tend to identify too many variables and/or meta parameter values that result in overfitted models, as noted by large error values for a validation set and corresponding small R^2 values. This deficiency can be overcome by using leave-multiple-out cross-validation (LMOCV) [41, 42]. The approach has also been called Monte Carlo cross-validation [43] and repeated learning-testing method [44]. A repeated *v*-fold cross-validation has also been described [42]. In LMOCV, the data set is split into a validation set with *d* samples (compounds for QSAR) and a calibration set with the remaining $m-d$ compounds, i.e., *d* compounds are left out. Respective RMSEC and RMSEV values are averaged over a large number of different splits. A restriction to the process is that validation set size *d* needs to be much greater than the calibration set size $m-d$. In ref. [45], theoretical reasons are provided justifying the use of $d = m - m/(\ln(m)-1)$. It is not uncommon for *d* to range from 0.6 to 0.8 *m*.

In this paper, the split size recommended in ref. [45] is used. The split is repeated 300 times and the average value for the variance indicator $\|\hat{\mathbf{b}}\|_2$ is plotted against average bias RMSEC, RMSEV, and respective R^2 values to form harmonious curves.

Experimental

Materials

A Micron Millennia Pentium III 1 GHz computer with MatLab 6.5 (The Math Works, Inc., Natick, MA) was used to perform all data analyses. Programs were written by the authors and implemented MatLab routines for RR, PLS, and PCR.

Data sets

All data sets were mean centered prior to computations. As noted in the *Cross-validation* section, the data sets were split according to the ratio in ref. [45] and 300 splits were used. All $\|\hat{\mathbf{b}}\|_2$, RMSEC, RMSEV, and respective R^2 values reported are the mean of the 300 splits.

Carbonic anhydrase (CA) inhibitors

It is thought that some CA enzymes contribute to the production of eye humor which through excess secretion in the eye can cause permanent eye damage and diseases such as macular edema and open-angle glaucoma. Drugs inhibiting the activity of CA are needed. This data set consists of 142 compounds assayed for inhibition of three CA isozymes: CA I, CA II, and CA IV [46]. The log of respective inhibition values (K_i) were modeled using 63 molecular descriptors. Descriptor selection was also performed for artificial neural networks (ANNs) in ref. [46]. It was found that eight molecular descriptors were key for CA I and CA IV and nine were optimal for CA II. Results in this paper are presented for CA I and CA IV using PLS, PCR, and RR with the full descriptor set and descriptors identified as

best for the ANNs. Similar conclusions to those presented in this paper can be made for CA II. It should be noted that the variables deemed best for the ANNs may not be best for PLS, PCR, and RR. However, the focus of this paper is not on how to select the best molecular descriptors but how to compare modeling methods and select the best model for a given data set of compounds and descriptors.

Dihydrofolate reductase inhibitors

Inhibition of dihydrofolate reductase (DHFR) is important in combating diseases originating from pathogens *Toxoplasma gondii* (*tg*) and *Pneumocystis carinii* (*pc*) for patients without healthy immune systems. This data set consists of 320, 334, and 340 compounds assayed for inhibition of, respectively, *tg*DHFR, *pc*DHFR and the mammalian standard *rl*DHFR [47]. The log of the corresponding 50% inhibition concentration (IC_{50}) values were modeled using 83, 84, and 84 molecular descriptors for *tg*DHFR, *pc*DHFR, and *rl*DHFR, respectively. Descriptor selection was also performed for ANNs in ref. [47]. It was found that a total of 10 molecular descriptors were key for all three DHHR's. Results in this paper are presented for *tg*DHFR and *pc*DHFR using PLS, PCR, and RR with the full set of descriptors and selected descriptors deemed best for ANNs. Similar conclusions to those presented in this paper can be made for *rl*DHFR. As with the CA data set, it should be noted that the variables considered best for the ANNs may not be best for PLS, PCR, and RR. Again, the focus of this paper is not on how to select the best molecular descriptors but how to compare modeling methods and select the best model for a given data set of compounds and descriptors.

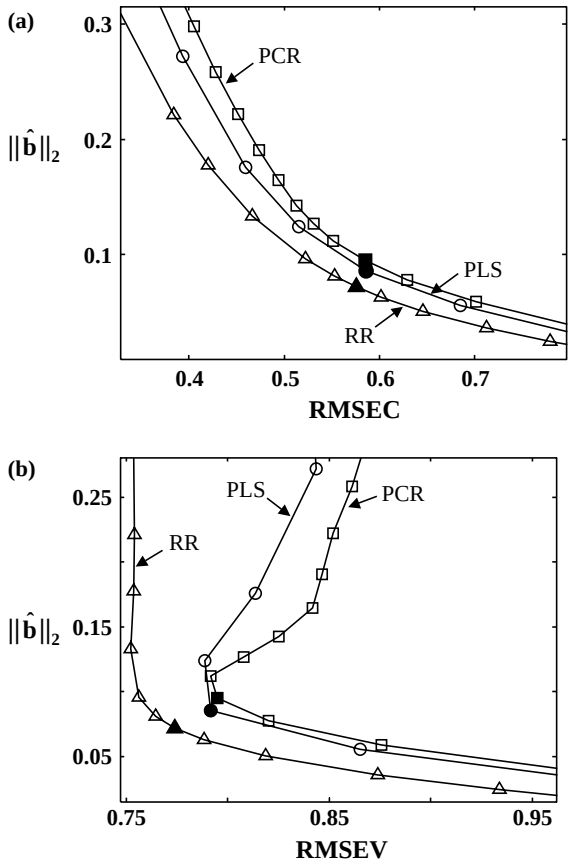

Figure 1. CA IV harmonious plots using 63 descriptors. (a) $\|\hat{\mathbf{b}}\|_2$ against RMSEC for PCR (□), PLS (○), and RR (△). Filled symbols denote optimal models with a ridge value of 750 and 6 and 8 PLS and PCR factors, respectively. Ridge values vary from 45 in the upper left corner to 7050 in the lower right corner. PLS and PCR models varying from 9 and 16 factors in the upper left corner, respectively, to 5 and 6 factors in the lower right corner, respectively. (b) $\|\hat{\mathbf{b}}\|_2$ against RMSEV for PCR (□), PLS (○), and RR (△). Filled symbols denote optimal models as in (a). Ridge values and PLS models vary as in (a). PCR models varying from 15 factors in the upper center to 6 factors in the lower right corner.

Results and discussion

CA IV and CA I

Plotted in Figure 1a is the harmonious RMSEC plot using all 63 molecular descriptors for predicting CA IV. At the 6 or 7 factor PLS models, the curve begins to increase in variance with less change in the RMSEC bias measures. This occurs for PCR in the 7 to 10 factor range and for RR between ridge parameter values 800 and 400. Figure 1b shows the corresponding RMSEV values and clearly discerns the optimal models for PLS and PCR, namely, the six factor PLS and eight factor PCR models. Figure 1b reveals that increasing the number of factors by one results in little improvement for the RMSEV values while the variance indicator increases. The exact value for the best RR model is still not determinable from Figure 1b due to the continuous nature of the ridge value compared to the discrete meta parameter values for PLS and PCR. However, the plot in Figure 1b shows that a narrow range for the best

ridge parameter value is obtainable. A range is expected because, as just stated, the ridge parameter is continuous to infinite decimal places. From Figure 1, it is decided that the RR model with a ridge value of 750 provides a good balance of bias and variance.

Note that just a plot of the RMSEV values without the regression vector norm would mark the seven and nine factor PLS and PCR models, respectively, as optimal because the RMSEV values minimize at these points. The optimal RR model would be difficult to determine from a plot of only RMSEV values. Additionally, an RMSEV plot does not always have a minimum present (see J. Forrester and J.H. Kalivas (submitted) for examples). Including a variance indicator provides the needed additional information to better determine harmonious models. Figure 1 and corresponding criteria values listed in Table 1 suggest that RR performs the best and should be considered the more harmonious model. As noted in the *Prediction and variance* and *Tikhonov regularization* sections, determination of the actual most harmonious model requires a more complete variance expression for the predicted values.

Shown in Figure 2 are harmonious plots for CA IV based on the eight descriptors deemed best in ref. [46]. Again, the better models are easily discerned from the plots and RR appears to provide the more harmonious model (see Table 1 for meta parameter values). The MLR model is obtained when all respective PLS or PCR factors are used or setting the ridge value to zero. From the plots in Figure 2 and respective values listed in Table 1, it is observed that the MLR model is not appropriate. The RMSEC and corresponding R^2 value for MLR are the smallest and largest, respectively, but the corresponding RMSEV and corresponding R^2 validation values are the largest and smallest, respectively. Thus, the MLR model is clearly overfitting the data.

Using Table 1, the full descriptor CA IV models can be compared to the models based on eight descriptors. While the calibration fit is better for the eight descriptor models, the regression vector norms are substantially larger, indicating that prediction variances may be larger for the eight descriptor data set. As the *Prediction variance* section shows, actual prediction variances will depend on actual values for the error terms.

Similar harmonious plots to those obtained for CA IV were observed for the CA I prediction property. From the results listed in Table 2, it is seen that the full descriptor situations provide slightly better bias and variance values compared to respective eight descriptor models. The MLR model with eight descriptors provides the best model fit values, but the regression vector norm is substantially larger, effectively assuring inflated variances. Again, RR appears to represent the more harmonious model for the full set and subset of descriptors.

With respect to parsimony, it would normally be stated that the eight descriptor models are more parsimonious than the full descriptor models. However, what is important is not the number of descriptors, but the amount of calibration space being used to form the corresponding models, i.e., the degrees of freedom involved in forming a model. The degrees of freedom are important in many regression diagnostic calculations. For example, the RMSEC is typically defined as

$$\mathrm{RMSEC} = \sqrt{\frac{\sum(\hat{y}_i - y)^2}{dof}} \tag{12}$$

Table 1. Model values for CA IV.

Model[a]	No. of descriptors	$\|\hat{\mathbf{b}}\|_2$	RMSEC	RMSEV	R^2_{cal}	R^2_{val}	ER
RR (750)	63	0.0721	0.575	0.774	0.870	0.751	0.891
PLS (6)	63	0.0855	0.585	0.791	0.855	0.747	1.057
PCR (8)	63	0.0948	0.585	0.795	0.853	0.747	1.170
RR (6)	8	0.460	0.536	0.671	0.880	0.816	1.036
PLS (4)	8	0.463	0.546	0.676	0.875	0.813	1.043
PCR (5)	8	0.464	0.551	0.677	0.879	0.814	1.045
MLR	8	3.553	0.504	0.700	0.894	0.802	8

[a]Parentheses contain ridge values for RR and the number of factors for PLS and PCR.

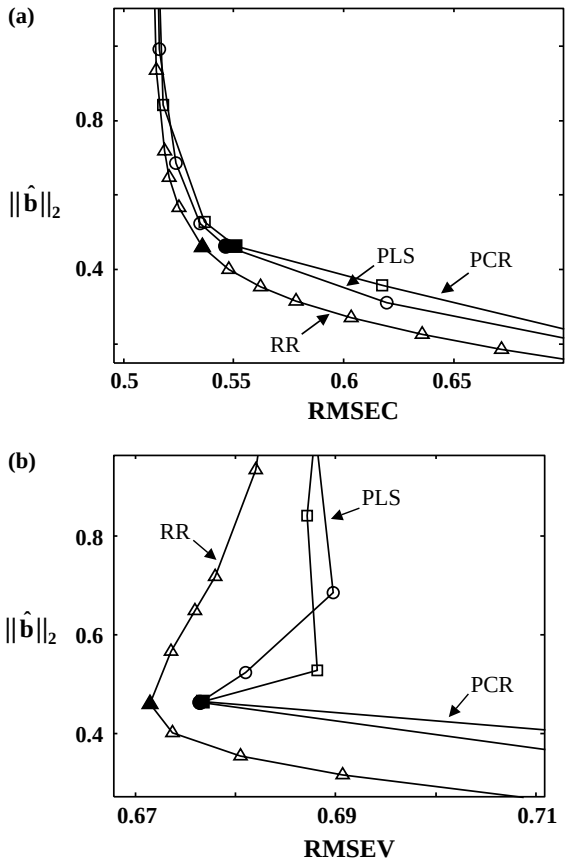

Figure 2. CA IV harmonious plots using 8 descriptors. (a) $\|\hat{\mathbf{b}}\|_2$ against RMSEC for PCR (□), PLS (○), and RR (△). Filled symbols denote optimal models with a ridge value of 6 and 4 and 5 PLS and PCR factors, respectively. Ridge values vary from 0.2 in the upper left corner to 126 in the lower right corner. PLS and PCR models vary from 7 each in the upper left corner to 3 and 4 corresponding factors in the lower right corner, respectively. (b) $\|\hat{\mathbf{b}}\|_2$ against RMSEV for PCR (□), PLS (○), and RR (△). Filled symbols denote optimal models as in (a). Ridge values vary from 0.2 in the upper left corner to 30 in the lower right corner. PLS and PCR models vary from 6 and 7 factors in the upper center, respectively, to 4 and 5 factors in the lower left corner, respectively.

Unfortunately, there is disagreement on how to estimate the *dof* term. For MLR it is generally agreed that the *dof* is set to m minus the number of variables (descriptors). However, when the number of variables is greater than the number of compounds, this is no longer possible. When PCR or PLS is used in this situation the *dof* term is sometimes set to m minus the number of respective PCR or PLS factors. However, setting the *dof* to this is incorrect as the factors used by PCR and PLS are not the same, i.e., the basis vectors are not equivalent [4, 16, 48]. When the number of variables is less than the number of compounds, as with MLR, and PCR or PLS is used, it is uncertain whether the number of variables or factors should be used in calculating *dof*. Recent work compared three newer definitions for *dof* that can be used for PCR, PLS, RR, and MLR [H. Seipel and J.H. Kalivas, in press]. Two were found to be the more practical and of these, the ER is evaluated in this paper as a parsimonious measure where $dof = m - \mathrm{ER}$. The ER measure is defined as

$$\mathrm{ER} = \mathrm{rank}(\mathbf{X}) \frac{\|\hat{\mathbf{b}}\|_2^2}{\|\hat{\mathbf{b}}_{\mathrm{LS}}\|_2^2} \tag{13}$$

where $\mathrm{rank}(\mathbf{X}) \leq \min(m, p)$ denotes the mathematical rank of $\mathbf{X}$ and $\hat{\mathbf{b}}_{\mathrm{LS}}$ represents the regression vector for the LS solution. Equation 13 calculates the size of the effective calibration space being used relative to LS and the fraction is then scaled by the rank of $\mathbf{X}$ in order for ER to be in the range $0 \leq \mathrm{ER} \leq \mathrm{rank}(\mathbf{X})$. It should be noted that the RMSEC values tabulated in the tables are based on dividing by m, not *dof*, in order to maintain impartiality from favoring one definition over another.

The ER values listed in Table 1 for CA IV show that the ERs are nearly the same, with the smallest being RR when all 63 descriptors are used. Thus, besides forming the more harmonious model, RR also forms the parsimonious model. This is also true for the eight descriptor data set. While the full descriptor RR model is apparently the most parsimonious model for CA IV, the RR, PLS, and PCR models for CA I based on eight descriptors are more parsimonious than any of the full descriptor models. The method of RR maintains the overall smallest ER with eight descriptors. However, it should be kept in mind that harmonious and parsimonious differences between RR, PCR, and PLS are small.

It is interesting to note that the ERs for the CA IV eight descriptor models are similar in magnitude to the full descriptor models, yet for CA I, the ERs for the eight descriptor models are much smaller than the full descriptor models. This is because for the optimal models shown, the degree of shrinkages in the magnitudes of the regression vectors relative to the LS models in the maximum

Table 2. Model values for CA I.

Model[a]	No. of descriptors	$\|\hat{\mathbf{b}}\|_2$	RMSEC	RMSEV	R^2_{cal}	R^2_{val}	ER
RR (670)	63	0.0667	0.487	0.695	0.868	0.717	0.903
PLS (6)	63	0.0767	0.497	0.703	0.851	0.720	1.038
PCR (7)	63	0.0742	0.519	0.705	0.835	0.717	1.005
RR (22)	8	0.209	0.582	0.699	0.792	0.720	0.172
PLS (3)	8	0.227	0.586	0.705	0.787	0.717	0.187
PCR (4)	8	0.267	0.581	0.711	0.790	0.712	0.220
MLR	8	9.722	0.453	0.657	0.878	0.762	8

[a]Parentheses contain ridge values for RR and the number of factors for PLS and PCR.

dimensional calibration space are approximately 17% (the square root of the right term (ratio) in Equation 12) for the CA IV and CA I full descriptor models and the degrees of shrinkage for the CA IV and CA I eight descriptor models are, respectively, about 36% and 15%. When the degrees of shrinkage are scaled by 34 and 8 for the full and eight descriptor data sets (the ranks of **X**), the tabulated ERs are obtained. Thus, even though there is less shrinkage for the CA IV eight descriptor models, the size of the full dimensional calibration space is smaller, resulting in the similar ER. Similarly, because the degrees of shrinkage are nearly the same for the CA I full and eight descriptor models, when the values are scaled by the dimensional size of the respective calibration spaces, the ERs are smaller for the eight descriptor models.

Besides the bias/variance tradeoff that must be considered in selecting a model, a harmonious/parsimonious compromise appears to also be pertinent. That is, with a subset of selected molecular descriptors from a greater set, the ER, RMSEC, and RMSEV values can be smaller with larger respective R^2 values, yet $\|\hat{\mathbf{b}}\|_2$ can be greater [49, 50], implying larger uncertainties in predicted values. In wavelength selection studies [20, 21], even though $\|\hat{\mathbf{b}}\|_2$ increased with a subset of wavelengths, predicted concentrations sometimes maintained slightly smaller variances estimated by Monte Carlo simulations compared to variances for the complete set of wavelengths. Therefore as noted previously, $\|\hat{\mathbf{b}}\|_2$ is more a measure of potential variance and not useful for discerning between models based on different numbers of molecular descriptors, especially when differences between $\|\hat{\mathbf{b}}\|_2$ are small. For a true harmony/parsimony comparison between models with different number of molecular descriptors and between different modeling methods, more complete variance expressions should apparently be used.

As noted in the Experimental section, the eight descriptors were selected as being best for a neural network. No attempt was made in this study to use a descriptor selection algorithm to select optimal descriptors for PLS, PCR, RR, or MLR.

tgDHFR and pcDHFR

With *tg*DHFR and *pc*DHFR, the RMSEC harmonious plots for RR, PLS, and PCR using the full set of 84 and 83 descriptors, respectively, are similar to those displayed in Figure 1a. However, the RMSEV and regression vector norm values only increased with each additional factor or decreased in the ridge value. These observations are clear indicators that the data set is overfitted with the full descriptor sets and RR, PLS, or PCR cannot model the data.

Tabulated in Tables 3 and 4 are the results when only 10 descriptors are used for *tg*DHFR and *pc*DHFR, respectively. Again, it is observed that RR is estimated to be more harmonious and parsimonious than PLS and PCR. The MLR models do provide better values for the bias measures, but the variance measure is substantially larger. Similar to Figure 2, the MLR models are significantly above the corner of the harmonious plot, i.e., the MLR models are located where little to no change occurs for bias measures while the variance measure $\|\hat{\mathbf{b}}\|_2$ substantially increases. Thus, MLR is not appropriate and RR, PLS, and PCR provide better results with the selected descriptors.

Table 3. Model values for *tg*DHFR using 10 descriptors.

Model[a]	$\|\hat{\mathbf{b}}\|_2$	RMSEC	RMSEV	R^2_{cal}	R^2_{val}	ER
RR (11)	0.597	0.808	0.919	0.677	0.580	0.689
PLS (5)	0.607	0.822	0.929	0.657	0.578	0.701
PCR (6)	0.646	0.821	0.927	0.657	0.581	0.746
MLR	8.664	0.699	0.902	0.766	0.634	10

[a]Parentheses contain ridge values for RR and the number of factors for PLS and PCR.

Table 4. Model values for *pc*DHFR using 10 descriptors.

Model[a]	$\|\hat{\mathbf{b}}\|_2$	RMSEC	RMSEV	R^2_{cal}	R^2_{val}	ER
RR (17)	0.421	0.867	0.979	0.605	0.478	6.19×10^{-3}
PLS (5)	0.500	0.865	0.996	0.603	0.478	7.36×10^{-3}
PCR (6)	0.500	0.868	0.993	0.600	0.478	7.36×10^{-3}
MLR	679	0.798	1.020	0.603	0.495	10

[a]Parentheses contain ridge values for RR and the number of factors for PLS and PCR.

Conclusions

This paper has shown that as with spectroscopic data, RR provides QSAR models that generate more robust predictions in terms of bias values compared to PLS and PCR when the data is mean-centered. This is also in agreement with simulated data presented in ref. [48]. It should be kept in mind that when the proper meta parameters are chosen, differences between the three methods are small. This agrees with other comparison studies showing that in the optimal model region for RR, PLS, and PCR, the methods are nearly equivalent [51]. Results from this investigation have also shown that a variance indicator is just as important in assessing models as a bias measure. When both are plotted, the proper meta parameters can easily be selected to allow fair inter-model comparisons.

Besides bias and variance measures to discern the harmonious model, the ER measure should also be used to assess parsimony. The parsimony tends to improve with selection of molecular descriptor subsets; however, the variance indicator $\|\hat{\mathbf{b}}\|_2$ always degrades, suggesting a harmony/parsimony compromise. Further studies are needed with the more complete variance expressions to discern this.

For this study, slightly better results were obtained when all the molecular descriptors were used for the CA I data set compared to a subset of variables deemed optimal for an ANN. Small improvements were generated for CA IV in proceeding from the full set of descriptors to using only eight descriptors, provided that MLR is not used. For the *tg*DHFR and *pc*DHFR data sets, only with selection of key molecular descriptors could acceptable results be obtained for the validation set, even though the calibration results were satisfactory with the full set of descriptors. Thus, depending on the data set, variable selection may or may not be useful.

Needed is a modeling process that will select variables and simultaneously formulate the best model with the variables. This may be possible with a generalization of expression (11) written as

$$\min\left(\|\mathbf{Xb}-\mathbf{y}\|_a^a+\lambda\|\mathbf{Lb}\|_b^b\right) \tag{14}$$

where a and b represent the same or different norms in the range $1 \le a, b \le \infty$ [38, 52]. Using a minimum cut-off of 1 for the 1-norm assures all norms to be convex. Expression (14) with $a = b = 1$ for the 1-norm was applied in ref. [53] and provided stability with respect to accounting for possible outliers. Using $a = 2$ and $b = 1$ produced a 'sparse spike train' solution for reflection seismograms, i.e., a solution with the least number of non-zero coefficients [54]. While the regression vector 1-norm is not directly related to prediction variance, it is a variance indicator in that it does capture the size of the model. Full or selected variables QSAR

modeling requires discontinuities in $\hat{\mathbf{b}}$. Such solutions are possible with expression (14) using $a = 2$ and $b = 1$. This type of investigation is on-going in our laboratory and similar work with a 1-norm has been performed with support vector machines [55].

Additional variance and bias indicators as well as the ER could be used simultaneously to generate a sharper corner in order to better characterize the harmony/parsimony tradeoff. Visual inspection is possible with a third measure. Beyond this it requires other methods to represent multidimensional plots, such as color [18].

Acknowledgement

The authors are grateful to Peter Jurs and Brian Mattioni for providing the data sets.

References

1. van de Waterbeemd, H. (Ed.) Chemometric Methods in Molecular Design, VCH, New York, 1995.
2. Vandeginste, B.G.M., Massart, D.L., Buydens, L.M.C., De Jong, S., Lewi, P.J. and Smeyers-Verbeke, J., Handbook of Chemometrics and Qualimetrics: Part B, Chapter 37, Elsevier, Amsterdam, The Netherlands, 1998.
3. Kalivas, J.H. and Lang, P.M., Mathematical Analysis of Spectral Orthogonality, Marcel Dekker, New York, 1994.
4. Kalivas, J.H., J. Chemom., 13 (1999) 111.
5. Kalivas, J.H., Chemom. Intell. Lab. Syst., 45 (1999) 215.
6. Goldstein, M. and Smith, A.F.M., J. Royal Stat. Soc. B, 36 (1974) 284.
7. Gunst, R.F. and Mason, R.L., J. Am. Stat. Assoc., 72 (1977) 616.
8. Hansen, P.C., Computing, 40 (1988) 185.
9. Lowerre, J.M., Technometrics, 16 (1974) 461.
10. Bingham, C. and Larntz, K., J. Am. Stat. Assoc., 72 (1977) 97.
11. Hocking, R.R., Speed, F.M. and Lynn, M.J., Technometrics, 18 (1976) 425.
12. de Jong, S., Wise, B.M. and Ricker, N.L., J. Chemom., 15 (2001) 85.
13. de Jong, S. and Kiers, H.A.L., Chemom. Intell. Lab. Syst., 14 (1992) 155.
14. Aldrin, M., Am. Stat., 54 (2000) 29.
15. Holcomb, T.R., Hjakmarsson, H., Morari, M. and Tyler, M.L., J. Chemom., 11 (1997) 282.
16. Kalivas, J.H., Anal. Chim. Acta, 428 (2001) 31.
17. Xu, Q.S., Liang, Y.Z. and Shen, H.L., J. Chemom., 15 (2001) 135.
18. Green, R.L. and Kalivas, J.H., Chemom. Intell. Lab. Syst., 60 (2002) 173.
19. Kalivas, J.H. and Green, R.L., Appl. Spectrosc., 55 (2001) 1645.
20. Kalivas, J.H., Anal. Chim. Acta, 505 (2004) 9.
21. Anderson, K.J. and Kalivas, J.H., Appl. Spectrosc., 57 (2003) 309.
22. Cohon, J.L. Multiobjective Programming and Planning, Academic Press, New York, 1978.
23. Censor, Y., Appl. Math. Optim., 4 (1977) 41.
24. Da Cunha, N.O. and Polak, E., J. Math. Anal. Appl., 19 (1967) 103.
25. Zadeh, L.A., IEEE Trans. Autom. Contr., AC-8 (1963) 1.
26. Smilde, A.K., Knevelman, A. and Coenegracht, P.M.J., J. Chromatogr., 369 (1968) 1.
27. Höskuldsson, A., Chemom. Intell. Lab. Syst., 14 (1992) 139.
28. Höskuldsson, A., Chemom. Intell. Lab. Syst., 32 (1996) 37.
29. Hansen, P.C., SIAM Rev., 34 (1990) 503.
30. Hansen, P.C., In Johnston, P. (Ed.), Computational Inverse Problems in Electrocardiology, WIT Press, South Hampton, 2001.
31. C.L. Lawson and R.J. Hanson, Solving Least Squares Problems, Prentice-Hall, Englewood Cliffs, NJ, 1974.
32. Hansen, P.C., SIAM J. Sci. Stat. Comput., 11 (1990) 503.
33. Faber, K. and Kowalski, B.R., J. Chemom., 11 (1997) 181.
34. Faber, K. and Kowalski, B.R., Chemom. Intell. Lab. Syst., 34 (1996) 283.
35. Lorber, A. and Kowalski, B.R., J. Chemom., 2 (1988) 93.
36. Næs, T., Isaksson, T., Fern, T. and Davies T. A User Friendly Guide to Multivariate Calibration and Classification, NIR Publications, Chichester, UK, 2002.
37. Weisberg, S. Applied Linear Regression, Wiley, New York, 1985.
38. Hansen, P.C. Rank-deficient and Discrete Ill-posed Problems: Numerical Aspects of Linear Inversion, SIAM, Philadelphia, PA, 1998.
39. Tikhonov, A.N., Soviet Math. Dokl., 4 (1963) 1035.
40. Hoerl, A.E. and Kennard, R.W., Technometrics, 12 (1970) 55.
41. Baumann, K., Trends Anal. Chem., 22 (2003) 395.
42. Baumann, K., von Korff, M. and Albert, H., J. Chemom., 16 (2002) 351.
43. Xu, Q.S. and Liang, Y.Z., J. Chemom., 56 (2001) 1.
44. Burman, P., Biometrika, 76 (1989) 503.
45. Shao, J., J. Am. Stat. Assoc., 88 (1993) 486.
46. Mattioni, B.E. and Jurs, P.C., J. Chem. Inf. Comput. Sci., 42 (2002) 94.
47. Mattioni, B.E. and Jurs, P.C., J. Mol. Graph. Model., 21 (2003) 391.
48. Frank, I.E. and Friedman, J.H., Technometrics, 35 (1993) 109.
49. Lorber, A. and Kowalski, B.R., J. Chemom., 2 (1988) 67.
50. Mark, H. Principles and Practice of Spectroscopic Calibration, Wiley, New York, 1991.
51. Geladi, P., In Andrews, D.L. and Davies, A.M.C. (Eds.), Frontiers in Analytical Spectroscopy, The Royal Society of Chemistry, London, 1995.
52. Dax, A., SIAM J. Optim., 2 (1992) 602.
53. Taylor, H.L., Banks, S.C. and McCoy, J.F., Geophysics, 44 (1979) 39.
54. Santosa, F. and Symes, W., SIAM J. Sci. Stat. Comput., 7 (1986) 1307.
55. Song, M., Breneman, C.M., Bi, J., Sukumar, N., Bennett, K.P., Cramer, C. and Tugcu, N., J. Chem. Inf. Comput. Sci., 42 (2002) 1347.