QSAR with Few Compounds and Many Features

Douglas M. Hawkins,*,† Subhash C. Basak,‡ and Xiaofang Shi†

School of Statistics, 313 Ford Hall, 224 Church Street S. E., University of Minnesota, Minneapolis, Minnesota 55455, and Natural Resources Research Institute, 5013 Miller Trunk Highway, University of Minnesota, Duluth, Minnesota 55811

Received August 12, 2000

Fitting quantitative structure−activity relationships (QSAR) requires different statistical methodologies and, to some degree, philosophies depending on the "shape" of the data matrix. When few features are used and there are many compounds, it is a reasonable expectation that good feature subset selection may be made and that nonlinearities and nonadditivities can be detected and diagnosed. Where there are many features and few compounds, this is unrealistic. Methods such as ridge regression RR, PLS, and principal component regression PCR, which abjure feature selection and rely on linearity may provide good predictions and fair understanding. We report a development of ridge regression for the underdetermined case by using generalized cross-validation to choose the ridge constant and perform *F*-tests for additional information. Conventional regression diagnostics can be used in followup to identify nonlinearities and other departures from model. We illustrate the approach with QSAR models of four data sets using calculated molecular descriptors.

1. INTRODUCTION

During the past two decades an increasing number of quantitative structure−activity/property relationship (QSAR/QSPR) models have been using theoretical molecular descriptors for predicting biomedical, toxicological, and technological properties of chemicals.[1−7] QSARs/QSPRs are mathematical models that seek to predict complicated physicochemical/biological properties of chemicals from their simpler experimental or calculated properties.[8] The main problem with the use of experimental data as independent variables in QSARs is that they are not available for the majority of chemical structures, real or hypothetical.

Consequently, in recent years, there has been a shift of interest from the use of experimental data to the application of theoretical properties/parameters in the development of QSARs. In the area of drug design, one needs to evaluate many thousands (sometimes millions) of chemicals for their therapeutic effectiveness and toxicity before a chemical reaches from the chemist's desk to the pharmacist's desk. One can save a lot of resources if one can have reliable estimates of therapeutic and toxic potential of chemicals. Such estimated values can be used to differentiate promising compounds from the less useful ones. In the areas of the regulation of chemicals, the United States Environmental Protection Agency (USEPA) needs a substantial list of physical, environmental, and toxicological properties of chemicals to carry out a reasonable hazard assessment. The Toxic Substances Control Act (TSCA) Inventory, the list of chemicals in commerce in the United States, currently contains approximately 80,000 chemicals. Nearly 50% of the TSCA chemicals have no available experimental data; only 15% of these chemicals have experimental genotoxicity data. Therefore, hazard assessment of chemicals in U.S.A. in the foreseeable future has to be carried out from estimated properties calculated from their molecular structure. Modern combinatorial chemistry is creating libraries of chemicals consisting of millions, sometimes billions, of molecules. Most of these molecules have no available data except for the molecular structure. Scientists are using theoretical molecular descriptors to evaluate such combinatorial libraries and prioritize them for practical drug design and testing protocols.[9]

It is widely recognized that QSPR/QSAR equations, whether they are derived in a purely empirical fashion from an arbitrary set of molecular descriptors, or from a preselected set of descriptors, expected on theoretical grounds to have a connection with a particular property, can provide insight into the molecular and submolecular origin of physical and biological properties.[10]

A theoretical molecular descriptor associates a real number with the chemical structure based on certain aspects of chemical structure. The topological, geometrical, and quantum chemical indices, the three groups of theoretical parameters which have been most frequently used in QSAR/QSPR model development, are numerical descriptors of molecular architecture and quantify such aspects of molecular structure as size, shape, symmetry, complexity, branching, cyclicity, stereoelectronic character, etc. Individual molecular descriptors have been successfully used in the development of QSAR/QSPRs of congeneric and structurally homogeneous sets of molecules.

The formulation of useful QSAR/QSPR models for structurally and mechanistically diverse sets of chemicals requires the use of a set of diverse parameters, as opposed to a handful of indices sufficient for model building in the congeneric situation. A diversity of parameters enlarges the scope of our predictive power, bringing another problem at the same time: the problem of intercorrelation of independent

* Corresponding author phone: (218)720-4230; fax: (218)720-4328; e-mail: sbasak@nrri.umn.edu.
† School of Statistics.
‡ Natural Resources Research Institute.

10.1021/ci0001177 CCC: $20.00
Published on Web 01/12/2001

variables. Although different parameters are defined based on different theoretical backgrounds, they often are strongly correlated with each other. In such cases one needs to extract distinct and orthogonal or uncorrelated structural information from the collection of diverse predictors in order to develop useful QSAR/QSPR models. Principal components analysis, partial least squares, and ridge regression are methods that can be helpful in the development of QSAR models in such situations.

A quantitative structure−activity relationship, QSAR, is based on the reasonable premise that the biological activity of a compound is a consequence of its molecular structure and that, provided we can identify those aspects of molecular structure that are relevant to a particular biological activity, we can gain a better understanding of the mechanism by which the compound acts. This better understanding may also translate directly into designed compounds constructed to have molecular features that enhance the desired biological activity and not to have other structural features that are superfluous, inhibit the desirable activity, or lead to unwanted side effects. This understanding can be used to guide searches of structure databases to find compounds with an increased probability of being biologically active or environmentally friendly. The QSAR problem consists of trying to quantify the relationship between measurable or calculable properties of a molecule and the property/activity/toxicity of the molecule, while the field of chemometrics has emerged with a focus on analyzing observational data originating mostly from organic and analytical chemistry, food research, and environmental studies.

QSAR/QSPR data sets commonly introduce some concerns:

- There is often a high degree of collinearity among the predictor variables.
- Since the number of molecular characteristics that may be measured or calculated and used as predictors is limited only by the investigator's inventiveness, the data sets tend to be characterized by many predictor variables.
- Often this tendency is so pronounced as to lead to a data set that is "underdetermined," that is the number of such variables p exceeds (and often greatly) the number of observations n.

2. APPROACHES

QSAR models are calibrated using sets of compounds whose molecular structure and biological activity are known. Traditionally, multiple linear regression has been used. This is not because of any conviction that additive linear models truly hold exactly but as a convenient relatively standard way to capture general monotonic relationships. Also traditionally, regression subset methods were used, motivated by the "magic key" model that most of the measured features were irrelevant to the activity and that good models should flow from finding the minority of relevant features.

If the number of observations is substantially larger than the number of features, there are many other ways of fitting the models. For example, neural nets (and their statistical relatives Generalized Additive Models)[11] provide a way of escaping the straitjacket of linearity but not that of additivity. Recursive partitioning[12] escapes both the additivity and linearity assumptions as well as handling data sets with more than one mode of activity, but this is at the cost of substantial sample size.

Where the number of compounds is not subtantially larger than the number of features, the options for analysis narrow considerably, and it becomes difficult to escape the assumptions of additivity and linearity in one guise or another. Commonly used methodologies include ordinary least squares OLS regression, partial least squares/projection to latent spaces (PLS),[13−15] principal components regression (PCR),[16] and ridge regression (RR).[17] All these methods except OLS are also intended to work when the predictors are highly multicollinear and even when there is underdetermination.

3. ISSUES OF STATISTICAL METHODOLOGY

The output of a conventional OLS includes several important values:

1. An estimate of the coefficient of each predictor.
2. Standard errors quantifying the uncertainty in these estimates.
3. An estimate (with degrees of freedom) of the underlying residual variance.
4. Case diagnostics such as residuals, leverages, and influence for checking the constant-variance linear assumption and the impact of individual observations in the data set.

A byproduct of part (3) is a formal F test of additional information. To test whether adding a block of predictors improves the model, we use the residual sums of squares and degrees of freedom of the "full" model (using all predictors) and "restricted" model (omitting the additional predictors being tested) to carry out an F test for the additional information in the predictors. OLS however cannot be applied in an underdetermined setting. Any methodology to be used then should have ways of providing these useful outputs.

The methods mentioned (PLS, PCR, RR) are all in the general family of "linear smoothers" in which fitted values are computed from

$$\hat{Y} = \mathbf{A}Y$$

where Y is the vector of observed dependent values and $\hat{Y}$ is the vector of fitted values. The smoother matrix $\mathbf{A}$ differs for the different methods. The number of "fitting" degree of freedom for each estimator is given by trace ($\mathbf{A}$), the number of "error" degrees of freedom by n − trace($\mathbf{A}$), and a variance estimate is given by

$$||\hat{Y} - Y||^2/[n - \text{trace}(\mathbf{A})] \quad (1)$$

Because of this, the same approach as in OLS may in principle be used to test for additional information. We use the above formula to get the "full" model variance estimate and the "restricted" model variance estimate. From these we then get an F ratio as in conventional linear model fitting. This ratio is approximately F distributed and can be used to test the adequacy of the "restricted" model.

In methods that estimate β as a linear transform of Y, $\hat{\beta} = \mathbf{C}Y$, we can also get standard errors for the estimated coefficients, these having covariance matrix $\sigma^2\mathbf{C}\mathbf{C}^T$, where σ^2 is the residual variance of the (presumed linear constant-variance) regression model.

Jackknifing[18] provides another method of getting standard errors. In this approach, each observation in turn is omitted from the data set and the analysis repeated. Writing $\hat{\beta}$ for the estimate given by the full sample and $\hat{\beta}_i$ for the estimate found omitting case i defines the "pseudovalue" $t_i = n\hat{\beta} - (n - 1)\hat{\beta}_i$. Then the mean of the vectors t_i gives the "jackknife" estimate of the coefficient vector β, and their sample covariance matrix provides an estimate of the covariance matrix of the jackknife estimate. Jackknifing does not require linearity of $\hat{\beta}$ in Y, so it also addresses general estimators.

OLS does not work at all if $n < p$ and works poorly if the predictors are highly correlated in the sense that the standard errors of the coefficients become very large. The principal goal of PCR, PLS, and RR is to shrink the solution coefficient vector away from the OLS solution toward directions in the predictor-variable space of a large sample spread. PCR and PLS are seen to shrink more heavily away from the low spread directions than RR. A comparison among these methods was made by Frank and Friedman.[19] They observed that PLS, PCR, and RR tend to be quite similar in most situations, largely because they are applied to problems involving high collinearity in which variance tends to dominate the bias. This paper did however demonstrate that RR outperformed PLS and PCR. Butler and Denham[20] illustrate a further apparently undesirable property of PLS—that it does not necessarily shrink the fitted coefficient vector toward zero. Motivated largely by these papers, we will focus on RR for the remainder of this paper and compare its results on our illustrative data sets with those of PLS.

Frank and Friedman also considered subset regression and concluded that this approach was far inferior to the methods retaining all predictors. This had already been suggested by Copas;[21] it is not yet entirely clear whether there are identifiable circumstances in which subset selection may be a viable approach.

4. RR METHOD

Set up the conventional multiple linear regression model

$$Y = \mathbf{X}\beta + \epsilon$$

where Y is the $n \times 1$ vector of dependent variables, $\mathbf{X}$ is the $n \times p$ "design matrix" of predictor values, and β is the vector of regression coefficients. It is assumed that ϵ, the vector of true residuals, has $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2 I$, where I is an identity matrix of order n. The OLS estimator of β

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y$$

is undefined when the matrix $\mathbf{X}^T\mathbf{X}$ is singular. Simple ridge regression replaces this estimate by

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X} + kI)^{-1}\mathbf{X}^T Y$$

As a practical matter, it is conventional when using simple RR to first scale the predictors to correlation form so that the diagonal elements of $\mathbf{X}^T\mathbf{X}$ are all equal. Note that RR gives an estimated coefficient vector that is linear in Y, making it a linear smoother. The ridge parameter k controls the amount of "smoothing" done in ridge regression. If k is set very large, $\hat{\beta}$ will be close to zero,while if k is small, the coefficients $\hat{\beta}$ may well include numerically large values. Any k greater than zero is enough to make it possible to compute the ridge estimate, whatever the relativities of the number of cases and predictors or collinearity in the data. It is even possible to have predictors that do not vary in the data set; their fitted coefficients will be zero.

The key problem in ridge regression is choosing the value of k. There are several methods for addressing this problem other than Hoerl and Kennard's original proposal to do so graphically. All involve studying many different values of k and picking that which best meets some criterion. The standard criteria are as follows.

1. Cross-Validation (PRESS) Method. This method goes hand-in-glove with the jackknife. For each trial k, omit each case i in turn from the data set and compute the ridge regression estimate $\hat{\beta}_i$. Use this to predict the omitted case i and calculate the prediction error—the difference between the actual y observed and this hold-out prediction. The prediction sum of squares PRESS is the total of the squares of these prediction errors. The cross-validation criterion uses for k that value for which PRESS is a minimum.

2. Generalized Cross-Validation (GCV) Method. This method, due to Golub et al.,[22] is much faster than the PRESS criterion and also leads to a more stable solution. While it is motivated by the idea of validation with future data, rather than explicitly remove each case, it uses matrix results to simulate applying the RR model to the prediction of future similar cases. This leads to the objective of choosing k to minimize $V(k)$, defined by the following.

3. A final criterion, also motivated by ideas of applying the fitted model to future cases, is that suggested by Myers[23] (p 249). This selects k by minimizing

$$C_k = SSR_k/S^2 - n + 2\text{trace}(\mathbf{A})$$

$$\sigma^2 = SSR_k/(n - p)$$

where SSR_k is the sum of the squares of residuals from ridge regression, S^2 is the estimator of σ^2 from OLS, and $\mathbf{A} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + kI)^{-1}\mathbf{X}^T$. Some properties of this estimator are sketched by Shi and Wang.[24] Note that the C_k criterion requires that OLS be fitted first. It cannot therefore be applied to underdetermined problems.

For our work, we elected to select k using GCV. The model fits were then assessed using hold-one-back cross-validation as an independent criterion.

5. REAL DATA SETS

Here we discuss QSAR/QSPRs of four data sets. These all involve the use of a hierarchy of predictors to the prediction of measured activity. Two involve biological activity, and two are physicochemical properties.

The hierarchy of predictors is as follows: (1) topostructural (TSI) variables, (2) topochemical (TCI) variables, (3) geometrical (3D) properties, and (4) semiempirical quantum chemical (QC) variables. In view of the nature of these predictors, we are particularly interested in the potential value of adding each of the types of information to models already using the preceding types (comparing, for example, modeling using TSI, TCI, and 3D with modeling using TSI and TCI). The computer runs (and our discussion) use simplified codes

for these properties; a translation showing the equivalence between our codes and the properties is given in Tables 7−10. For a complete list of these parameters and their definitions and symbols see ref 25.

5.1. "Hall" Data Set. The first data set, which we will refer to as the Hall data set, comprises the LC_{50} of $n = 69$ benzene derivatives, measuring their toxicity to fish. It is taken from ref 26. The first group of predictors (TSI) has 35 predictors: idw, midw, w, id, hv, hd, ic_bar, max_ic, m1, m2, s0-s6, sc3, sc5, scy6, spc4-spc6, k0-k10, j. The second group (TCI) has 51 more predictors: i_orb, ic0-ic6, sic0-sic6, cic0-cic6, b0-b6, bc3, bc5, bcy6, bpc4-bpc6, v0-v6, vc3, vc5, vcy6, vpc4-vpc6, jx, jy, jb. The third group (3D) has 3 more predictors: edh, ed, vol; and the fourth group (QC) has 6 more predictors: homo, homo1, lumo, lumo1, hf, dm.

For this data set, the number of predictors 35 in the first group is less than the sample size of 69, so fitting by OLS is possible. But further in the hierarchy, the number of predictors, 86, 89, and 95, exceeds the number of cases and the models are underdetermined.

5.2. "TA" Data Set. The second data set, which we will abbreviate to the "TA" data set, comprises the mutagenicity of $n = 95$ aromatic and heteroaromatic amines. It is taken from the compilation of Debnath et al.[27] The TSI group of predictors has 40 variables: idw, midw, w, id, hv, hd, ic_bar, max_ic, m1, m2, s0-s6, sc3-sc6, scy3-scy6, spc4-spc6, k0-k10, j. The TCI group (TCI) adds 61 predictors: i_orb, ic0-ic6, sic0-sic6, cic0-cic6, b0-b6, bc3-bc6, bcy3-bcy6, bpc4-bpc6, v0-v6, vc3-vc6, vcy3-vcy6, vpc4-vpc6, jx, jy, jb; the 3D group 3 predictors: edh, ed, vol; and the QC group 6 predictors: homo, homo1, lumo, lumo1, hf, dm.

For this data set, the TSI model is fully determined, but the subsequent models are underdetermined.

5.3. "TSCABP" Data Set. This data set is of the boiling point of $n = 1037$ structurally diverse TSCA chemicals. The first group (TSI) has 40 predictors: idw, midw, w, id, hv, hd, ic_bar, max_ic, m1, m2, s0-s6, sc3-sc6, scy3-scy6, spc4-spc6, k0-k10, j; the second group (TCI) has 61 more predictors: i_orb, ic0-ic6, sic0-sic6, cic0-cic6, b0-b6, bc3-bc6, bcy3-bcy6, bpc4-bpc6, v0-v6, vc3-vc6, vcy3-vcy6, vpc4-vpc6, jx, jy, jb. For this data set we do not have 3D or QC measures. It is always the case that $n \geq p$, and so this data set is amenable to analysis using a wide array of modeling approaches.

5.4. "VP" Data Set. This final data set has 476 observations—the vapor pressures of a subset of 476 of the TSCA compounds. The first group (TSI) has 40 predictors: idw, midw, w, id, hv, hd, ic_bar, max_ic, m1, m2, s0-s6, sc3-sc6, scy3-scy6, spc4-spc6, k0-k10, j; the second group (TCI) has 61 more predictors: i_orb, ic0-ic6, sic0-sic6, cic0-cic6, b0-b6, bc3-bc6, bcy3-bcy6, bpc4-bpc6, v0-v6, vc3-vc6, vcy3-vcy6, vpc4-vpc6, jx, jy, jb. The third group (3D) has 3 more predictors: vol, edh, ed. Here too, all models of interest are fully determined.

6. RESULTS

6.1. Selecting *k*. In this section, we report on the results produced by ridge regression with the ridge parameter selected by GCV.

Table 1 shows some relevant summary statistics. The different rows correspond to the four data sets, starting modeling using the TSI predictors and then successively adding the TCI, the 3D, and finally the QC (so that the final line corresponds to using all four blocks of predictors). The columns show the unshrunk OLS PRESS and GCV and then the results for the ridge regression. These comprise the cross-validated R^2, the k value selected, the cross-validation PRESS, and the GCV criterion. Also shown for comparive purposes is the cross-validation PRESS produced by PLS. The unshrunk results are not available for the cases where there were fewer cases than predictors. In all cases, the variables were standardized prior to the start of analysis, so the PRESS values are in multiples of the standard deviation of the dependent variable.

Table 1. Picking the Ridge Constant *k*

	OLS		ridge				PLS
	PRESS	GCV	RSq	k	PRESS	GCV	PRESS
				Hall			
TSI	87.54	0.420	0.451	2.3E−05	40.00	0.435	43.83
+TCI			0.752	0.286	17.10	0.251	23.41
+3D			0.726	6.400	18.88	0.253	22.62
+QC			0.798	6.400	13.95	0.198	14.62
				TA			
TSI	46.37	0.676	0.412	10.326	30.77	0.330	34.66
+TCI			0.708	4.281	27.77	0.283	28.94
+3D			0.734	9.551	25.24	0.269	29.42
+QC			0.717	24.804	26.88	0.278	27.16
				TSCABP			
TSI	186.08	0.174	0.821	2.4E−04	185.95	0.173	187.03
TCI	30.93	0.023	0.970	1.9E−04	30.68	0.023	38.22
				VP			
TSI	350.19	0.539	0.456	135.418	258.76	0.509	265.43
+TCI	433.90	0.176	0.739	0.143	124.29	0.168	109.90
+3D	462.53	0.154	0.727	0.062	130.12	0.149	109.88

Some interesting features can be seen in this table. One is the effect of adding predictors. It is well-known that adding predictors in an OLS regression necessarily reduces the residual sum of squares—this is what makes formal testing for additional information relevant. This is not the case with the PRESS criterion. In the VP data set, in particular, the OLS PRESS increases as the number of predictors increases. This is due to instability in the fitted slopes, resulting in worse predictions of the hold-back case. When we go to the ridge analyses of the VP data set, this anomaly is less serious. The PRESS values are also far below the OLS PRESS values, a result of the way ridging stabilizes the fitted slopes.

The k values selected for a particular data set can vary substantially from one set of predictors to another. For example, the very large k value for VP using TSI is an indication that this set of features has little predictive value since this large k leads to heavy shrinkage of the slopes.

Comparing the PRESS values of the ridge regression with those of PLS, we see that in 11 out of the 13 models fitted, RR has a smaller PRESS than does PLS. In the table overall, the PRESS values of RR average some 7% lower than those of PLS. This observation is in line with the results of Frank and Friedman and provides support for our concentration on RR in the analysis of the data.

6.2. Comparison of Different Sets of Predictors. The next issue we address is the value of additional predictors. For this purpose, we fixed upon the k chosen by GCV for the full set of predictors and used this k value in regressions with the successive groups of predictors. Each such fit

Table 2. *F*-Test for Hall Data Set

	PRESS	full-sample ss	errdf	MS	*F*	*P*
TSI	35.790	26.447	59.30	0.446		
+TCI	19.814	10.561	52.14	0.203		
+3D	18.885	10.318	52.01	0.198		
+QC	13.955	7.592	50.45	0.150	11.61	0.0002

Table 3. *F*-Test for TA Data Set

	PRESS	full-sample ss	errdf	MS	*F*	*P*
TSI	31.005	25.069	83.75	0.299		
+TCI	27.761	19.064	78.05	0.244	4.31	0.0010
+3D	25.841	17.849	77.69	0.230		
+QC	26.883	17.181	75.68	0.227	1.46	0.2387

Table 4. *F*-Test for TSCABP Data Set

	PRESS	full-sample ss	errdf	MS	*F*	*P*
TSI	185.96	166.08	995.47	0.1668		
+TCI	30.68	19.44	935.18	0.0208	116.93	0.0000

Table 5. *F*-Test for VP Data Set

	PRESS	full-sample ss	errdf	MS	*F*	*P*
TSI	296.38	218.74	443.3	0.493		
+TCI	131.18	55.64	395.5	0.141		
+3D	130.13	48.71	426.3	0.124	24.30	0.0000

provided a residual sum of squares and an associated degrees of freedom as shown in eq 1. This then provided the equivalents of conventional "block" *F* tests for additional groups of predictors:

$$F = [(SS_{sub} - SS_{full})/(DF_{sub} - DF_{full})]/[SS_{full}/DF_{full}]$$

6.2.1. Hall. Table 2 shows the results for model fitting for the Hall data set. The successive columns are the PRESS value, the residual sum of squares, the approximate degrees of freedom of the residual sum of squares, and the residual mean square. Also shown are the relevant *F* ratios for formal testing.

In deciding where in the hierarchy of predictors to stop, we carry out block *F* tests from the bottom of the list. The test for addition of QC, the last block of predictors, to the set TSI + TCI + 3D gives an *F* ratio of 11.61 with degrees of freedom 1.56 and 50.45, *p*-value is 0.0002, a highly significant value indicating that within this hierarchical list we should use all four blocks of predictors. A less formal decision based on the sizes of the successive PRESS values would probably yield the same decision.

Note the near-identical degrees of freedom for the TSI + TCI and TSI + TCI + 3D models. These suggest that the addition of the 3D predictors to TSI + TCI has no real impact on the model fit.

Table 6 shows this full regression, listing the coefficients, standard errors, and *t* values of each of the predictors. The standard errors were found using the "linear transformation of *Y*" property of ridge regression, and the *t* value is the coefficient divided by its standard error. Some of the predictors have numerically large *t* values. Looking at those whose value is 3 or more, the largest positive is for k9 with a value of 6.61. The largest negatives are for dm (−4.65), edh (−3.80), ic3 (−3.01), and ic2 (−3.00). Where high values of k9 are predictive of higher activity, smaller values of dm, edh, ic3, and ic2 are associated with higher activity.

Table 6. Fitted Ridge Regression of Hall Data

pred	coeff	s.e.	*t*	pred	coef	s.e.	*t*
idw	0.003	0.008	0.43	midw	−0.087	0.086	−1.01
w	0.008	0.009	0.95	id	−0.105	0.085	−1.23
hv	1.207	0.643	1.88	hd	−0.113	0.109	−1.04
ic_bar	0.112	0.258	0.43	max_ic	0.006	0.145	0.04
m1	0.034	0.022	1.55	m2	0.026	0.022	1.20
s0	0.035	0.020	1.72	s1	0.008	0.035	0.24
s2	0.094	0.043	2.17	s3	0.032	0.054	0.60
s4	0.047	0.087	0.54	s5	−0.071	0.108	−0.66
s6	−0.091	0.106	−0.86	sc3	0.259	0.104	2.49
sc5	0.044	0.152	0.29	scy6	−2.648	1.223	−2.16
spc4	0.043	0.054	0.78	spc5	0.062	0.047	1.33
spc6	0.019	0.029	0.66	k0	0.021	0.022	0.95
k1	0.021	0.022	0.95	k2	0.049	0.026	1.89
k3	0.003	0.026	0.13	k4	−0.008	0.028	−0.29
k5	−0.008	0.030	−0.27	k6	0.001	0.021	0.06
k7	−0.051	0.028	−1.83	k8	−0.039	0.047	−0.85
k9	0.655	0.099	6.61	k10	0.000	0.000	
j	−0.361	0.294	−1.23	i_orb	0.177	0.138	1.28
ic0	−0.222	0.244	−0.91	ic1	−0.099	0.158	−0.63
ic2	−0.404	0.135	−3.00	ic3	−0.358	0.119	−3.01
ic4	−0.106	0.101	−1.05	ic5	−0.188	0.086	−2.18
ic6	−0.188	0.086	−2.18	sic0	0.052	0.502	0.10
sic1	0.071	0.342	0.21	sic2	−0.617	0.350	−1.76
sic3	−0.382	0.301	−1.27	sic4	0.224	0.239	0.94
sic5	0.019	0.172	0.11	sic6	0.019	0.172	0.11
cic0	−0.611	0.230	−2.66	cic1	−0.026	0.229	−0.11
cic2	0.229	0.164	1.40	cic3	0.076	0.106	0.71
cic4	−0.245	0.105	−2.34	cic5	−0.165	0.094	−1.76
cic6	−0.165	0.094	−1.76	b0	0.059	0.036	1.63
b1	−0.030	0.098	−0.30	b2	0.156	0.069	2.27
b3	0.019	0.106	0.18	b4	0.060	0.170	0.35
b5	0.086	0.284	0.30	b6	−0.334	0.417	−0.80
bc3	0.575	0.201	2.86	bc5	0.370	0.443	0.84
bcy6	−5.475	4.892	−1.12	bpc4	0.077	0.084	0.91
bpc5	0.112	0.082	1.36	bpc6	0.089	0.070	1.28
v0	0.078	0.117	0.67	v1	0.111	0.134	0.83
v2	0.252	0.130	1.94	v3	0.246	0.117	2.11
v4	−0.445	0.191	−2.33	v5	0.893	0.454	1.97
v6	−0.731	0.436	−1.68	vc3	0.443	0.188	2.36
vc5	0.865	0.587	1.47	vcy6	−5.475	4.892	−1.12
vpc4	0.238	0.105	2.26	vpc5	−0.059	0.086	−0.69
vpc6	0.059	0.079	0.75	jx	−0.449	0.616	−0.73
jy	0.036	0.258	0.14	jb	−0.024	0.298	−0.08
edh	−0.203	0.053	−3.80	ed	0.014	0.010	1.46
vol	−0.104	0.125	−0.83	homo	−0.080	0.043	−1.87
homo1	−0.050	0.042 0	−1.2	lumo	−0.075	0.029	−2.62
lumo1	−0.015	0.037	−0.42	hf	0.003	0.002	1.87
dm	−0.085	0.018	−4.65				

The status of these 5 predictors is clear; regardless of which other predictors from the set we wish to use, we should be sure to keep these 5. The status of coefficients with numerically small *t* values is less clear. While they are not useful in models using all 93 other predictors, it might be that if certain of these other predictors were removed from the model they might become valuable. This common situation arises when predictors share predictive power and we need some but not all of them to capture this power.

6.2.2. TA. The results for the TA data set are in Table 3. We use the full-model ridge *k* of model 24.804. Starting from the end, we see that adding QC to TSI + TCI + 3D does not give a significant improvement. Going up to the next level, adding 3D to TSI + TCI only consumes 0.36 degrees of freedom, suggesting that the addition does not alter the fit appreciably. Because of this degrees-of-freedom issue, we bypass the formal test of 3D and come up to the TCI predictors. Here the *F* ratio is 4.32 with 5.7 and 78.05 degrees of freedom for a highly significant *P* value of 0.001. We thus augment TSI with TCI but do not add 3D or QC.

Table 7. Symbols and Definitions of Topostructural Parameters

simplified symbol	common symbol	
idw	I_D^W	information index for the magnitudes of distances between all possible pairs of vertices of a graph
midw	$\overline{I}_D^W$	mean information index for the magnitude of distance
w	W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
id	I^D	degree complexity
hv	H^V	graph vertex complexity
hd	H^D	graph distance complexity
ic_bar	$\overline{IC}$	information content of the distance matrix partitioned by frequency of occurrences of distance h
m1	M_1	a Zagreb group parameter = sum of square of degree over all vertices
m2	M_2	a Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
s0-s6	${}^h\chi$	path connectivity index of order $h = 0-6$
sc3-sc6	${}^h\chi_C$	cluster connectivity index of order $h = 3-6$
spc4-spc6	${}^h\chi_{PC}$	path-cluster connectivity index of order $h = 4-6$
scy3-scy6	${}^h\chi_{Ch}$	chain connectivity index of order $h = 3-6$
k0-k10	P_h	number of paths of length $h = 0-10$
j	J	Balaban's J index based on distance

Table 8. Symbols and Definitions of Topochemical Parameters

simplified symbol	common symbol	
max_ic	O	order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
max_orb	O_{orb}	order of neighborhood when IC_r reaches its maximum value for the hydrogen-suppressed graph
i_orb	I_{ORB}	information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
ic0-ic6	IC_r	mean information content or complexity of a graph based on the rth ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
sic0-sic6	SIC_r	structural information content for rth ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
cic0-cic6	CIC_r	complementary information content for rth ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
b0-b6	${}^h\chi^b$	bond path connectivity index of order $h = 0-6$
bc3-bc6	${}^h\chi^b_C$	bond cluster connectivity index of order $h = 3-6$
bcy3-bcy6	${}^h\chi^b_{Ch}$	bond chain connectivity index of order $h = 3-6$
bpc4-bpc6	${}^h\chi^b_{PC}$	bond path-cluster connectivity index of order $h = 4-6$
v0-v6	${}^h\chi^v$	valence path connectivity index of order $h = 0-6$
vc3-vc6	${}^h\chi^v_C$	valence cluster connectivity index of order $h = 3-6$
vcy3-vcy6	${}^h\chi^v_{Ch}$	valence chain connectivity index of order $h = 3-6$
vpc4-vpc6	${}^h\chi^v_{PC}$	valence path-cluster connectivity index of order $h = 4-6$
jb	J^B	Balaban's J index based on bond types
jx	J^X	Balaban's J index based on relative electronegativities
jy	J^Y	Balaban's J index based on relative covalent radii

Table 9. Symbols for Geometric Parameters

simplified symbol	common symbol
vol	V_W
ed	${}^{3D}W$
edh	${}^{3D}W_H$

Table 10. Symbols for Quantum Chemical Parameters

simplified symbol	common symbol
homo	E_{HOMO}
homo1	E_{HOMO1}
lumo	E_{LUMO}
lumo1	E_{LUMO1}
hf	ΔHf
dm	μ

6.2.3. TSCABP. The results for the TSCABP data set are given in Table 4. The very small full-model ridge of 0.000194 suggests that the slopes are well determined and do not need to be shrunk appreciably. We can see that for the TSCABP data set, the first two groups (TSI + TCI) of predictors can explain the activity well. The comparison between the first two subsets gives $F = 116.995$, p-value is 0.0000. This F ratio confirms what the PRESS strongly implies—that using TCI in addition to TSI vastly improves the predictions.

6.2.4. VP. Finally, Table 5 shows the results for the VP data set. The F ratio for adding 3D to TSI + TCI is 24.30 with 2.3 and 393.2 degrees of freedom, whose P value is returned by software as zero. This means that adding 3D to TSI + TCI gives a highly significant improvement in model fit, leading to TSI + TCI + 3D as the suggested set of predictors.

7. CHECKING THE MODEL FIT

Finally, we return to the opening point that the linear regression model is fitted, not because we believe it is exactly correct, but because it is a relatively easy way to capture smoothly increasing or smoothly decreasing relationships between the predictors and dependent variable. An important part of any regression modeling (including ridge regression) is graphical or formal checks on the model assumptions of linearity and constant variance. We will not go into these in depth; the checks discussed in detail in texts such as Cook and Weisberg[28] can be used in this context as in any other regression study.

One particularly simple but powerful diagnostic is the plot of residuals against fitted values. This plot detects some forms of nonlinearity (which need to be fixed by changing the model) and nonconstant variance (which is fixed by weighting). It can also identify outlying cases.

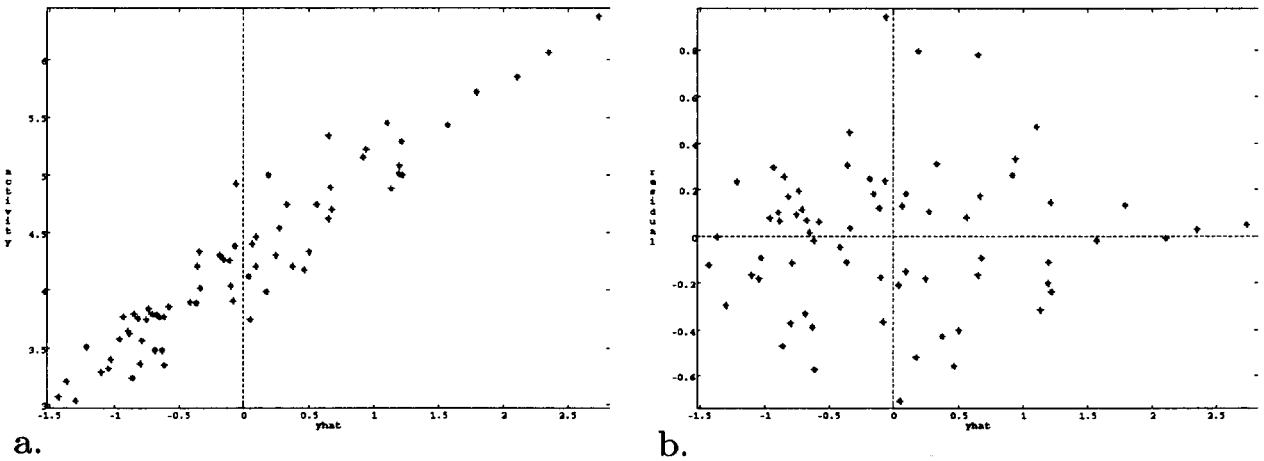

Figure 1. Diagnostics for Hall data: (a) activity vs predicted value for Hall data set and (b) residual vs predicted value for Hall data set.

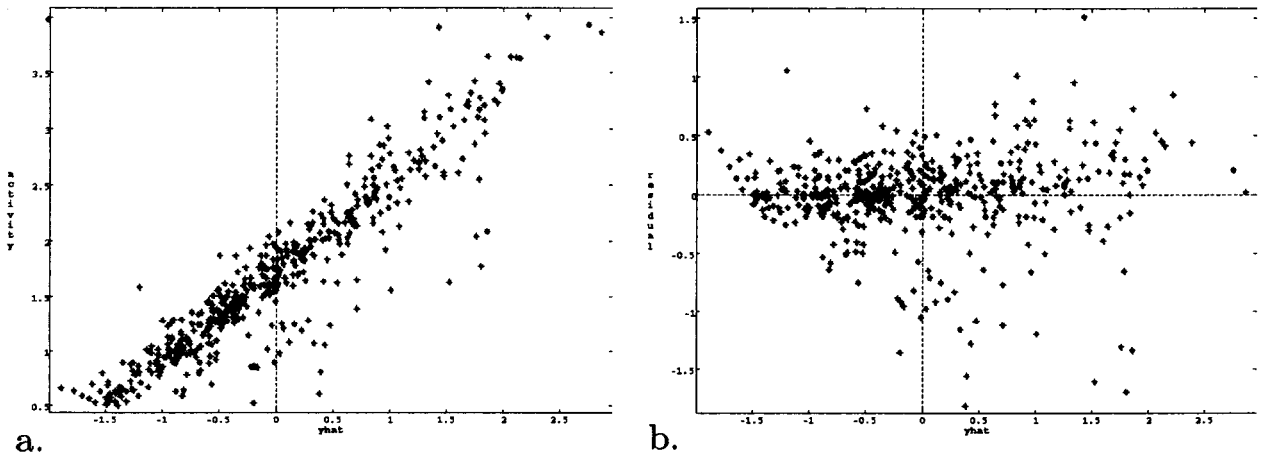

Figure 2. Diagnostics for VP data: (a) activity vs predicted value for VP data set and (b) residual vs predicted value for VP data set.

To illustrate the first point, a generalization of the linear regression model is the model of a nonlinear link function of a linear combination as is done in for example the "generalized linear model" approach of McCullagh and Nelder.[29] As shown in ref 30, if the correct model is

$$E(Y) = g(X'\beta)$$

where g is an arbitrary monotonic function, then under some assumptions on the X, fitting the linear regression of Y on X provides a consistent estimate of β, and looking at the plot of the Y_i against the fitted values $x_i'\hat{\beta}$allows one to diagnose the form of the nonlinear function g. Thus, using standard followup to a linear regression fit will lead to a correct analysis even though it uses a wrong starting point.

Figure 1 shows the plot of observed versus predicted activity and that of fitted residuals versus predicted activity for the Hall data set as modeled using all predictors. The residual plot shows no cause for great concern; there are three cases whose residuals seem high but not so high as to make them clear outliers. There is no evidence of curvature. There is no apparent widening or narrowing of the plot such as would indicate nonconstant variance. The plot of actual versus predicted activity tells the same story, adding the information that visually the predicted values seem to capture the actual values well. It is interesting that we seem to be able to do quite well on a data set with many more predictors than cases.

Figure 2 shows the corresponding plots of the VP data set using all predictors. These plots show a feature that is algebraically impossible in an OLS fit: the residuals do not add to zero. The plot of actual versus predicted value suggests that the data fall on two lines; a main line which follows the regression and a second lower line of compounds whose activity is systematically below that of the main group. This feature is interesting and invites further study—perhaps by separating the compounds into these two groups and studying their differences.

8. CONCLUSION

Despite the appearance of a number of new technologies for studying QSA/QSP relationships, the traditional approach of multiple linear regression is still useful. Its extension into ridge regression allows the analysis of data sets where the number of features exceeds the number of cases and provides seamless handling of exact or approximate multicollinearities and predictors that do not vary in the data set. Choosing the ridge parameter can be done using prediction sum of squares, by generalized cross-validation or (when the data set is of full rank) by a Mallow-type C_k criterion. As GCV is always applicable while C_k is not, we recommend GCV. This also gives standard errors of the regression coefficients, as does using the jackknife as part of a PRESS analysis.

As with conventional regression modeling, it is important to follow up the ridge fit with a study of residuals. Simple

plots can identify certain types of nonlinearity and nonconstant variance. These provide much protection against the common fear of using the linear regression model inappropriately.

ACKNOWLEDGMENT

This is contribution number 277 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported, in part, by grant F49620-98-1-0015 from the United States Air Force.

REFERENCES AND NOTES

(1) Gute, B. D.; Basak, S. C. Predicting acute toxicity of benzene derivatives using theoretical molecular descriptors: a hierarchical QSAR approach. *SAR QSAR Environ. Res*. **1997**, *7*, 117−131.
(2) Basak, S. C.; Grunwald, G. D. Predicting mutagenicity of chemicals using topological and quantum chemical parameters: A similarity based study. *Chemosphere* **1995**, *31*, 2529−2546.
(3) Basak, S. C.; Grunwald, G. D. Predicting genotoxicity of chemicals using nonempirical parameters. In *Proceeding of XVI International Cancer Congress*; 1995a; pp 413−416.
(4) Basak, S. C. Information theoretic indices of neighborhood complexity and their applications. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon & Breach Science Publishers: The Netherlands, 1999; pp 563−593.
(5) Basak, S. C.; Grunwald, G. D.; Gute, B. D.; Balasubramanian, K.; Opitz, D. Use of statistical and neural net approaches in predicting toxicity of chemicals. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 885−890.
(6) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Use of topostructural topochemical, and geometric parameters in the prediction of vapor pressure: A hierarchical approach. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 651−655.
(7) Basak, S. C.; Gute, B. D.; Ghatak, S. Prediction of complement inhibitory activity of benzamidines using topological and geometrical parameters. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 255−260.
(8) Hansch, C.; Leo, A. *Exploring QSAR: Fundamental and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.
(9) Grassy, G.; Calas, B.; Yasri, A.; Lahana, R.; Woo, J.; Iyer, S.; Kaczorek, M.; Floc'h R.; Buelow, R. Computer-assisted rational design of Immunosuppressive compounds. *Nature Biotechnol.* **1998**, *16*, 748−752.
(10) Katritzky, A. R.; Petrukhin, R.; Tatham, D.; Basak, S.; Benfenati, E.; Karelson M.; Maran, U. The Interpretation of Quantitative Structure−Property and −Activity Relationships. *J. Chem. Inf. Comput. Sci*. **2000**, in preparation.
(11) Hastie, T. J.; Tibshirani, R. J. *Generalized Additive Models*; Chapman & Hall: 1990.
(12) Hawkins, D. M.; Young, S. S.; Rusinko, A. Analysis of a large structure−activity data set using recursive partitioning. *Quantitative Structure−Activity Relationships 16*, 296−302.
(13) Wold, H. Soft modeling by latent variables: the nonlinear iterative partial least squares approach. In *Perspectives in Probability and Statistics, Papers in Honor of M. S. Bartlett*; Gani, J., Ed.; Academic Press: London, 1975.
(14) Höskuldsson, A. PLS regression methods. *J. Chemometrics* **1988**, *2*, 211−228.
(15) Höskuldsson, A. A combined theory for PCA and PLS. *J. Chemometrics* **1995**, *9*, 91−123.
(16) Massy, W. F. Principal components regression in exploratory statistical research. *J. Am. Statistical Assoc.* **1965**, *60*, 234−246.
(17) Hoerl, A. E.; Kennard, R. W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **1970**, *8*, 27−51.
(18) Efron, B. *The Jackknife, the Bootstrap and Other Resampling Plans*; Society for Industrial and Applied Mathematics: 1982.
(19) Frank, I. E.; Friedman, J. H. A statistical view of some chemometrics regression tools. *Technometrics* **1993**, *35*, 2, 109−135.
(20) Butler, N. A.; Denham, M. C. The peculiar shrinkage properties of partial least squares regression. *J. Royal Statistical Soc., Ser. B* **2000**, *62*, Part 3, 585−593.
(21) Copas, J. B. Regression, Prediction and Shrinkage (with Discussion). *J. Royal Statistical Soc., Ser. B* **1983**, *45*, 311−354.
(22) Golub, G. H.; Heath, M.; Wahba, G. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **1979**, *21*, 2, 215−223.
(23) Myers, R. H. *Classical and modern regression with applications*; Duxbury Press: Boston, 1986.
(24) Shi, L.; Wang, X. Local influence in ridge regression. *Computational Statistics Data Analysis* **1999**, *31*, 341−353.
(25) Basak, S.; Balaban, A.; Grunwald, G.; Gute, B. Topological indices: Their nature and mutual relatedness. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 891−898.
(26) Hall, L. H.; Kier, L. B.; Phipps, G. Structure−activity relationships on the toxicities of benzene derivatives: 1. An additivity model. *Environ. Toxicol. Chem*. **1984**, *3*, 355−365.
(27) Debnath, A. K.; Debnath, G.; Shusterman, A. J.; Hansch, C. A QSAR investigation of the role of hydrophobocity in regulating mutagenicity in the Ames test: 1. Mutagenicity of aromatic and heteroaromatic amines in *Salmonella typhimurium* TA98 and TA100. *Environ. Mol. Mutagen*. **1992**, *19*, 37−52.
(28) Cook, R. D.; Weisberg, S. *Applied Regression Including Computing and Graphics*; Wiley: 1999.
(29) McCullagh, P.; Nelder, J. A. *Generalized Linear Models*; Chapman and Hall: London, 1983.
(30) Li, K.-C.; Duan, N. Regression analysis under link violation. *Annals Statistics* **1989**, *17*, 1009−1052.

CI0001177