Prediction of Properties from Simulations: A Re-examination with Modern Statistical Methods

R. A. Mansson,[†] J. G. Frey,[‡] J. W. Essex,[‡] and A. H. Welsh*[,†,§]

School of Mathematics and School of Chemistry, University of Southampton, Highfield, Southampton SO17 1BJ, U.K., and Centre for Mathematics and its Applications, The Australian National University, Canberra ACT 0200, Australia

Received February 15, 2005

We discuss models fit to data collected by Duffy and Jorgensen to predict solvation free energies and partition equilibria of drugs, organic molecules, aromatic heterocycles, and other molecules. These data were originally examined using linear regression, but here more recently developed statistical models are applied. The data set is complicated due to the presence of discrepant observations and also curvature in the response. In some cases it is possible to discard a small number of the observations to get good fit to the data, but, in others, discarding an increasing proportion of the observations does not improve the fit. Our general preference is to use robust parameter estimation which downweights to reduce the influence of discrepant observations on the fitted models. Models are selected for four responses using linear or more complicated representations of the explanatory variables, such as cubic polynomials, B-splines, or smoothers via generalized additive models (GAMs). Variables are chosen using the traditional approach of formal tests to assess contribution to the fit of a model, and resampling methods including bootstrap are also considered to assess the prediction error for given models. Results of our analysis indicate that GAMs are an improvement on linear models for describing the data and making predictions. In general robust regression models and GAMs have the smallest conditional expected loss of prediction over the four responses. In addition, robust regression models offer the advantage of identifying molecules that perform poorly in the fit. In general, models were identified that yielded an improvement of approximately 50% in the conditional expected loss of prediction compared with the original parametrization of Duffy and Jorgensen. It was also found that the use of cross-validation to compare models was unreliable, and bootstrapping is preferred.

1. BACKGROUND

Duffy and Jorgensen[1] obtained averaged descriptors for a large number of organic solutes via Monte Carlo (MC) statistical mechanics simulations. They then used multiple linear regression to model the relationship between gas to liquid free energies of solvation in hexadecane, octanol, and water and the octanol/water partition coefficients and these descriptors.

Duffy and Jorgensen's results represent successful modeling within the framework of multiple linear regression methods. However, more recent additions to the statistical toolkit offer the possibility of further refinements which may deepen our understanding and improve prediction. The purpose of this paper is to explore the application of these additional methods to Duffy and Jorgensen's data.

We find that we can improve on the results achieved by multiple linear regression methods by using methods which allow us to incorporate curvature in the effects of the descriptors (B-splines, Generalized Additive Models) and to downweight the effect of molecules whose behavior differs from the bulk of the other molecules (robust methods).

* Corresponding author phone national: (02) 6125-9773; phone international +61-2-6125-9773; fax: (02) 6125-5549; e-mail: Alan.Welsh@anu.edu.au.

[†] School of Mathematics, University of Southampton.

[‡] School of Chemistry, University of Southampton.

[§] The Australian National University.

We describe the methods we use and our modeling strategy in section 2. The results of our analysis are presented in section 3 with general observations on model building in section 4. Finally, we present our conclusions in section 5.

2. STATISTICAL MODELING

For the Duffy and Jorgensen data we are working in a situation where there is no prior model assumed for the data, and the aim of the investigation is to derive an empirical approximation that is a useful description of the relationship between the response (logP values) and the descriptors. We consider various different classes of models for the data, following a general strategy for our analysis as follows:

(1) Fit linear regression models to the data and investigate diagnostic plots to see whether the model assumptions are valid.

(2) Where the assumption of linearity fails, extend the linear models using B-splines (piecewise cubic polynomials) to capture curvature. Compare these with the even more flexible smoothing splines used in Generalized Additive Models.

(3) When the assumption of normality fails due to the discrepant observations in the tails of the residual distribution, use robust parameter estimation to deal with these observations.

10.1021/ci050056i CCC: $30.25 © 2005 American Chemical Society
Published on Web 10/07/2005

(4) Compare the different models using the bootstrap to compute estimates of the conditional expected loss of prediction for the different models.

This approach has been used separately for each of the four responses in the set of data.

2.1. Linear Models. The simplest model uses linear terms to describe the relationship between the response, *Y*, and the *p* descriptors $X_1,\cdots, X_p$. This statistical model for *n* observations is expressed in the form

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i \quad i = 1,\cdots, n \tag{1}$$

where ϵ_i are random errors which we assume are independent and normally distributed with zero mean and constant variance σ^2. These assumptions are investigated as part of the model fitting process by diagnostics which are discussed later. The Wilkinson-Rogers' notation can be used to express the model in a straightforward manner, and for the linear regression model above we have

$$Y \sim X_1 + \cdots + X_p \tag{2}$$

In our analysis, we start with a model with all variables included and aim to reduce the complexity of the model by selecting a subset of the variables that provides a good approximation to the underlying relationship. Initially we use an automated procedure involving the Akaike Information Criterion (AIC), but this method is conservative so after running the automated method we use formal tests to simplify our model. Formally backward elimination is used, and at each step we consider excluding each remaining descriptor individually. The least significant descriptor is removed from the model, and the process continues until no further variables can be excluded, based on an F-test using the residual sum of squares for the two models under comparison. The final comparison and selection of models is based on a bootstrap criterion discussed in section 2.5 below.

If there is curvature in the regression model because the linear approximation is poor, we can extend the model to include piecewise cubic polynomials to represent the curvature. We consider single knot B-splines that smoothly join two cubic polynomials. It is sensible to describe the linear model with B-splines using the Wilkinson-Rogers' notation, and for *p* descriptors we have

$$Y \sim bs(X_1) + \cdots + bs(X_p) \tag{3}$$

The model assumptions are the same and are checked in the same way as for models with linear terms.

2.2. Robust Parameter Estimation. The curvature can sometimes be caused by discrepant observations (outliers) in the data. Robust fitting methods try to get a good fit to the majority of the data by minimizing the impact of a small number of discrepant data points. One approach to robust regression given by Rousseeuw and Leroy[2] is least trimmed squares (LTS) where discrepant observations are excluded if they affect the fit too strongly. LTS regression has high breakdown, but this property does not apply when some of the explanatory variables are discrete with a restricted range, e.g. number of nitro or acid groups. In this case we used robust parameter estimation to downweight rather than exclude the discrepant observations. We use M-estimators with the Tukey bisquare weight function to fit our robust regression models. M-estimators minimize a criterion function of the residuals, i.e., we minimize the sum

$$\sum_{i=1}^{n} \rho(Y_i - \sum_{j=1}^{p} \beta_j X_{ij}) \tag{4}$$

There are a number of choices for the function ρ, but as mentioned above we use the Tukey bisquare function. We minimize this function by solving numerically the system of equations

$$\sum_{i=1}^{n} \psi(Y_i - \sum_{j=1}^{p} \beta_j X_{ij}) \mathbf{X}_i = 0 \tag{5}$$

using iteratively reweighted least squares, where $\psi = \dot{\rho}$ is the influence function—the derivative of the criterion function ρ. Linear terms in this model can be replaced by B-splines to make the model more flexible in handling curvature. The effect of these models is to reduce the influence of discrepant observations on the overall fit. Variables are also eliminated one at a time using backward elimination and a robust test based on the change in the criterion function between the two models with the scale estimate held constant. There is a high breakdown version of the estimator, where the initial value for the M-estimator is chosen as an S-estimator. This estimator is implemented in the **rlm** function of the MASS (Modern Applied Statistics with S) library. Hampel et al.[3] and Staudte and Sheather[4] are useful references for these methods.

2.3. Generalized Additive Models. These models[5,6] relate the response to smooth, nonlinear functions of the descriptors. For a set of *n* observations and *p* descriptors a GAM can be written as

$$Y_i = \beta_0 + \sum_{j=1}^{p} f_j(X_{ij}) + \epsilon_i \quad i = 1,\cdots, n \tag{6}$$

where $f_j(X_{ij})$ is a smooth function of the *j*th descriptor. These models can be specified in the Wilkinson-Rogers' notation as

$$Y \sim s(X_1) + \cdots + s(X_p) \tag{7}$$

The nonlinearity for each descriptor is assessed using effect plots and estimates of the degrees of freedom associated with each descriptor. When the degrees of freedom is close to one it is possible to use a linear term to describe the relationship. We use the **mgcv** implementation of GAMs in the R system,[7] which is discussed by Wood.[8]

We again start with a full model with all descriptors included, and, at each stage, the least significant term is removed from the model until no reduction in the number of variables is possible. When using smoothing splines within the GAM framework the software automatically selects the smoothness for each of the explanatory variables. When the degrees of freedom is close to one we refit the model with the smoothing spline replaced by a linear term. The GAM plots show the relationship between the response and the variable in question. In a linear model these would all be linear; in a GAM we can see how far from linear the

relationship actually is. Diagnostic plots are also used to investigate the model assumptions.

2.4. Model Diagnostics. When a model is chosen we investigate the appropriateness of (1) the form of the expectation function, (2) additivity of the error, (3) constant variance, (4) normality, and (5) independence. These assumptions are usually interrelated and so can be considered simultaneously. We use the model residuals, $\hat{\epsilon}_i$, which are the difference between the observed response Y_i and the fitted values $\hat{Y}_i$ calculated by the model to produce graphical displays to investigate these assumptions. In the case of robust regression we need to incorporate the final model weights into the diagnostics to deal with the discrepant points.

We plot the residuals against fitted values, $\hat{Y}_i$, to investigate the expectation function, additivity of errors, and constant variance. These assumptions are plausible if the points form an even, horizontal band with no particular patterns; a problematic plot is one where there is a pattern, for example curvature or increasing variability in the residuals as the fitted values increase. To check the normal assumptions for the residuals we plot the ordered residuals against the expected normal quantiles. Normality is plausible if there is a straight line in the plot; points below the line on the left and above the line on the right are discrepant points. Dependence can be investigated if we know the order in which the data were collected.

2.5. Model Evaluation. There are often a number of models with similar performances to be compared, and we can use the estimated conditional expected loss of prediction to discriminate between these models. Cross-validation is a popular technique that is used extensively for variable selection, but recent work has suggested that it has shortcomings, for example see ref 9. We prefer using bootstrapping to estimate prediction error for competing models as discussed by Shao[10] and Wisnowski et al.[11] in the context of model selection.

The steps in cross-validation are as follows:

(1) We need to choose the proportion of the data to be omitted in each calculation. For example we can leave a single observation out or approximately 10% of the observations when performing 10-fold cross-validation.

(2) Fit the model to the reduced set of data and use this model to predict the observations that have been held out. Calculate the estimated conditional expected loss of prediction by comparing the observations and their predictions.

(3) Run the last step excluding each subset of data, i.e., n times for leave-one-out and 10 times for the 10-fold cross-validation. Take an average of the estimated conditional expected loss of prediction over all the subsets of data.

The bootstrap procedure[12] follows these general steps:

(1) We compute the residuals ϵ_i for the full model (i.e., the model with all the variables) and then adjust them based on the models and fitting methods we are considering. In the nonrobust case the raw residuals are adjusted using the formula

$$r_i = \frac{\epsilon_i - \bar{\epsilon}}{\sqrt{1 - p/n}} \tag{8}$$

where $\bar{\epsilon}$ is the mean of the raw residuals, p is the number of parameters in the model, and n is the number of observations. In the case of robust regression models there are problems with the convergence of the fitting procedure for some of the bootstrap samples, and the residuals are leverage corrected discussed by Davison and Hinkley[12] (p 312). The formula is

$$r_i = \frac{\epsilon_i}{\sqrt{1 - dh_i}},$$
$$\text{where } d = \frac{2\sum(\epsilon_i/s)\psi(\epsilon_i/s)}{\sum\dot{\psi}(\epsilon_i/s)} - \frac{n\psi^2(\epsilon_i/s)}{\{\sum\dot{\psi}(\epsilon_i/s)\}^2} \tag{9}$$

and h_i is the leverage of the ith observation.

(2) We then generate B bootstrap samples (we take $B = 200$ samples in most cases unless otherwise stated) of n bootstrap residuals by sampling from the adjusted residuals independently with replacement.

(3) For every bootstrap sample the n bootstrap residuals are added to the fitted values from the model fitted to the original data to generate bootstrap sample observations. These bootstrap observations corresponding to each model j and each bootstrap sample b are denoted y^*_{bij}, $b = 1,\cdots, B$, $j = 1,\cdots, n$.

(4) Each model j is fitted to each bootstrap sample b to produce parameter estimates, $\hat{\beta}^*_{bj}$, and we compute the conditional expected loss of prediction for each model as

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{ij})^2 + \frac{1}{nB}\sum_{b=1}^{B}\sum_{i=1}^{n}\{(y_i - \hat{y}^*_{bij})^2 - (y^*_{bij} - \hat{y}_{bij}{}^*)^2\} \tag{10}$$

where y_{ij} is the fitted value for the ith observation from model j, and $\hat{y}^*_{bij}$ is the fitted value for the ith observation in bootstrap sample b from model j.

When using robust regression models there are problems with convergence in some bootstrap samples, and these results are excluded when calculating prediction errors for the models. However these correspond to a small proportion of the models over all bootstrap samples that we generate. We estimate the three quantities in the prediction error robustly by taking 10% trimmed versions of each sum.

Some researchers select the model which minimizes the cross-validation or the bootstrap estimate of the conditional expected loss of prediction. We prefer the more stable procedure of selecting the smallest model (in terms of degrees of freedom) for which the estimated conditional expected loss of prediction is close to the minimum achieved in the models under consideration.

All the analysis reported in this paper was carried out in R which is available for free download for all major platforms from www.r-project.org.

3. DATA ANALYSIS

Exploring the correlation structure of the features, we found some relationships that were fairly strong, without being strong enough to compromise the model, so these variables were left in the model. However, in all of our analyses, there are four explanatory variables associated with solvent-accessible surface area, see Table 1 of Duffy and Jorgensen.[1] Since SASA = FOSA + FISA + ARSA, the

Table 1: Models for the Response log *P*(*Hexadecane*/*Gas*) Based on a Variety of Methods, Identified by the Code in the First Column[a]

		cross-validation		bootstrap		
call	model formula	l-o-o	10-fold	av	10% tr	df
lm	ESXC+ESXL+SASA+FOSA+ARSA+DIPL+HBDN+HBAC+INME+ INST+amine+nitro acid	0.163	0.154	0.134	0.135	13
lm	ESXC+SASA+FOSA	0.170	0.158	0.184	0.185	4
lm	SASA+ARSA+DIPL	0.206	0.186	0.240	0.240	4
lm	HBDN+SASA+DIPL	0.388	0.386	0.562	0.562	4
lm	DIPL+ARSA+ESXL	0.332	0.297	0.455	0.456	4
lm	bs(ESXL,4)+bs(SASA,4)+bs(FOSA,4)+bs(HBAC,4)+bs(INME,4)+bs(INST,4)	0.272	1.888	0.121	0.121	25
lm	bs(ESXL,4)+bs(SASA,4)+bs(FOSA,4)+bs(HBAC,4)	0.227	0.210	0.129	0.129	17
lm	bs(ESXL,4)+bs(SASA,4)+bs(HBAC,3)+bs(FOSA,3)	0.229	0.224	0.133	0.133	15
gam	s(SASA,6)+s(FOSA,6)+s(ARSA,6)+s(DIPL,6)+s(HBDN,5)+ESXL	0.145	0.134	0.136	0.135	8.15
gam	s(SASA,6)+s(ARSA,6)+s(DIPL,6)+ESXL+HBDN+FOSA	0.153	0.134	0.131	0.130	8.15
gam	s(ARSA,6)+ESXL+HBDN+FOSA+SASA+DIPL	0.164	0.154	0.138	0.138	8.77
rlm[b]	ESXL+SASA+FOSA+ARSA+DIPL+HBDN+HBAC+INME+INST+amine			0.116	0.117	11
rlm[b]	ESXL+SASA+FOSA+ARSA+DIPL+HBDN+HBAC+INME+INST			0.114	0.116	10
rlm[b]	ESXL+SASA+FOSA+DIPL+HBDN+HBAC+INME+INST			0.116	0.117	9
rlm[b]	ESXL+SASA+FOSA+DIPL+HBDN+HBAC+INST			0.116	0.118	8
rlm[b]	ESXL+SASA+FOSA+DIPL+HBDN+HBAC			0.121	0.123	7
rlm[b]	ESXL+SASA+FOSA+DIPL+HBDN			0.125	0.127	6
rlm	SASA+FOSA+DIPL+HBDN			0.139	0.141	5
rlm	SASA+FOSA+DIPL			0.159	0.161	4
rlm[b]	bs(ESXC,4)+bs(ESXL,4)+bs(SASA,4)+bs(FOSA,4)+bs(ARSA,3)+ bs(DIPL,4)+bs(HBAC,4)+amine			0.112	0.113	29
rlm[b]	bs(ESXC,4)+bs(ESXL,4)+bs(SASA,4)+bs(FOSA,4)+bs(DIPL,4)+bs(HBAC,4)			0.105	0.105	25
rlm[b]	bs(ESXC,4)+bs(ESXL,4)+bs(SASA,4)+bs(FOSA,4)+bs(DIPL,4)			0.107	0.107	21
rlm	bs(ESXL,4)+bs(SASA,4)+bs(FOSA,4)+bs(DIPL,4)			0.125	0.126	17
rlm	bs(SASA,4)+bs(FOSA,4)+bs(DIPL,4)			0.173	0.174	13

[a] lm − linear model, gam − generalized additive model, rlm − robustly fitted linear model. [b] Denotes good models.

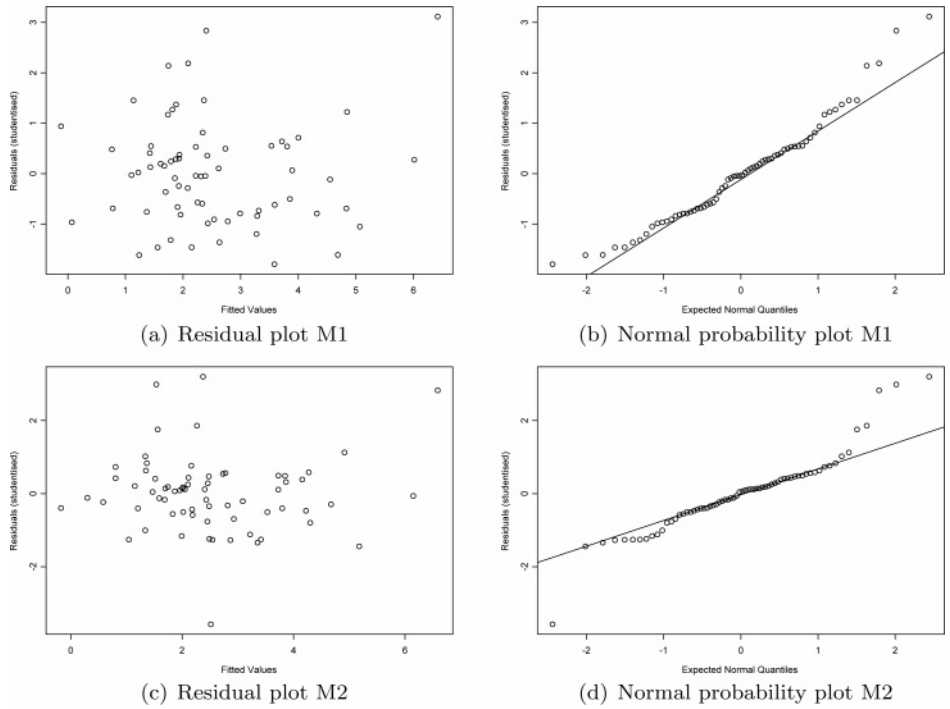
four variables are not all identifiable (at least one of the variables is redundant). The components are therefore confounded so their separate effects cannot be investigated, and the parameters in models including all the components are not identified. The standard approach to this problem is to impose identifiabilty constraints on the parameters: the simplest of these are equivalent to leaving one of the components out of the model. As identifiability constraints are not estimable, the choice is completely arbitrary. i.e., it makes no difference which component is omitted from the model. We arbitrarily selected to remove FISA and because it does not appear in any model reported by Duffy and Jorgensen.[1]

3.1. Hexadecane/Gas. Duffy and Jorgensen[1] modeled experimental free energies of solvation in hexadecane for 68 molecules. They obtained a three variable model with SASA, ARSA, and DIPL for which the conditional expected loss of prediction is estimated to be 0.240. The diagnostics (shown in Figure 1) for this model suggest that there are discrepant observations (the points off the line in the normal probability plot) and curvature (the points do not form an even horizontal band in the residual plot) so that an improved model might be sought. When backward elimination was used from a model with most of the explanatory variables (excluding FISA), we obtained a different subset of contributing variables for the model. The three contributing variables found by backward elimination are ESXC, SASA, and FOSA (M2) with an estimated conditional expected loss of prediction of 0.185. Diagnostics for this model (also shown in Figure 1) show reduced curvature but also suggest the presence of discrepant data points.

We also considered B-splines for the explanatory variables and GAMs with smoothing splines to further reduce curvature and improve the fit of the model to the data. Both of these extensions to the current model did not further improve the fit (evaluated by the diagnostics) and increased the complexity of the model.

Robust methods were used to identify the discrepant observations, and we considered two approaches to dealing with these molecules. Both methods calculate weights for the observations to reflect their influence on the parameter estimates. One method used Tukey biweight M-estimators, and nonzero weights were allocated to all of the observations. Alternatively, least trimmed squares (LTS) regression gave weights of either zero or one, which correspond to fitting the model to a subset of the original data set. LTS regression was used with a full model to identify molecules to exclude, and then the model was refitted with this subset of the data. Zero weights were suggested for eight of the molecules which is a large proportion of the data so the other robust regression approach was used.

A robust regression model with 11 explanatory variables was fitted to the data using the Tukey biweight M-estimator. Figure 2 shows these weights and identifies the down-weighted molecules as 2,2,2-TFE, tetrahydrofuran, dimethyl disulfide, anthracene, cyclohexane, ethyl fluoride, dimethyl sulfide, and fluorobenzene. Variables considered not to be contributing significantly to the fit were removed one at a time using the robust test procedure. The model simplified to three variables SASA, ESXC, and FOSA. Some of the variables were then modeled using smoothing splines or cubic polynomials. Residual and normal probability plots were created for the residuals from the robust fit. Figure 3 shows a plot of residuals versus fitted values for the four variable model with SASA, FOSA, DIPL, and HBDN. Here the majority of the observations are described well by the model, but the influence of some of the molecules on the parameter estimates (via the residuals) is reduced. This



(a) Residual plot M1 (b) Normal probability plot M1

(c) Residual plot M2 (d) Normal probability plot M2

Figure 1. Residual plots for regression models with three variables fitted to the *logP*(*hexadecane*/*gas*) data. M1 is SASA + ARSA + DIPL and M2 is SASA + FOSA + ESXC.

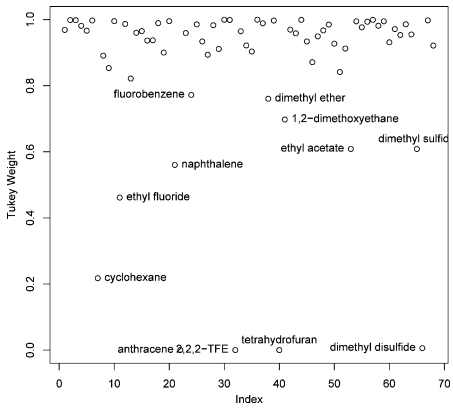


Figure 2. Plot of the Tukey weights for the robust fit of the *full* model to the data. Compounds with weights less than one are identified in the plot.

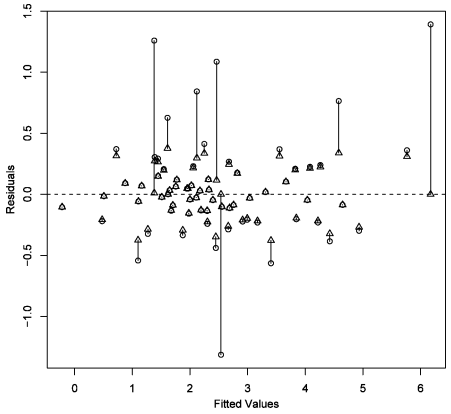


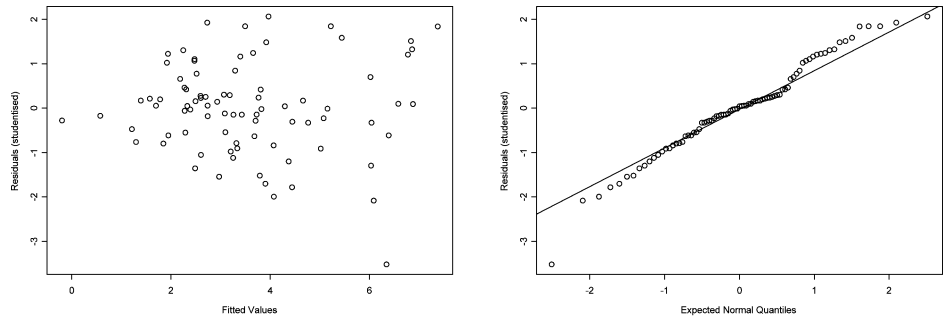
Figure 3. Residual plot for a four variable robust regression model. For the cases where the observations have their importance reduced, the weighted and unweighted residuals are shown joined by a solid line.

approach is preferable to LTS regression approach because we do not discard a large proportion of the data.

We also used cross-validation and bootstrapping to compare models based on their predictive powers. Table 1 provides a summary of a collection of models fitted to the log *P*(*hexadecane*/*gas*) response. Leave-one-out and 10-fold cross-validation were applied, and these suggested that the GAMs using smoothing splines with up to 6 degrees of freedom are the best way to represent the data. One noteworthy fact was the poor performance of some of the B-spline models in the situation where the data was divided into 10 groups for cross-validation. We computed bootstrap estimates of the expected conditional loss of prediction based on 200 bootstrap samples for all the models including those using robust parameter estimation. The average and a 10% trimmed version are reported in Table 1, which indicates that the GAM models are good, but the robust approach where we allow a few of the observations to be fitted poorly provides models with the smallest expected conditional loss of prediction. There is not a substantial improvement if we use B-splines within the robust regression model.

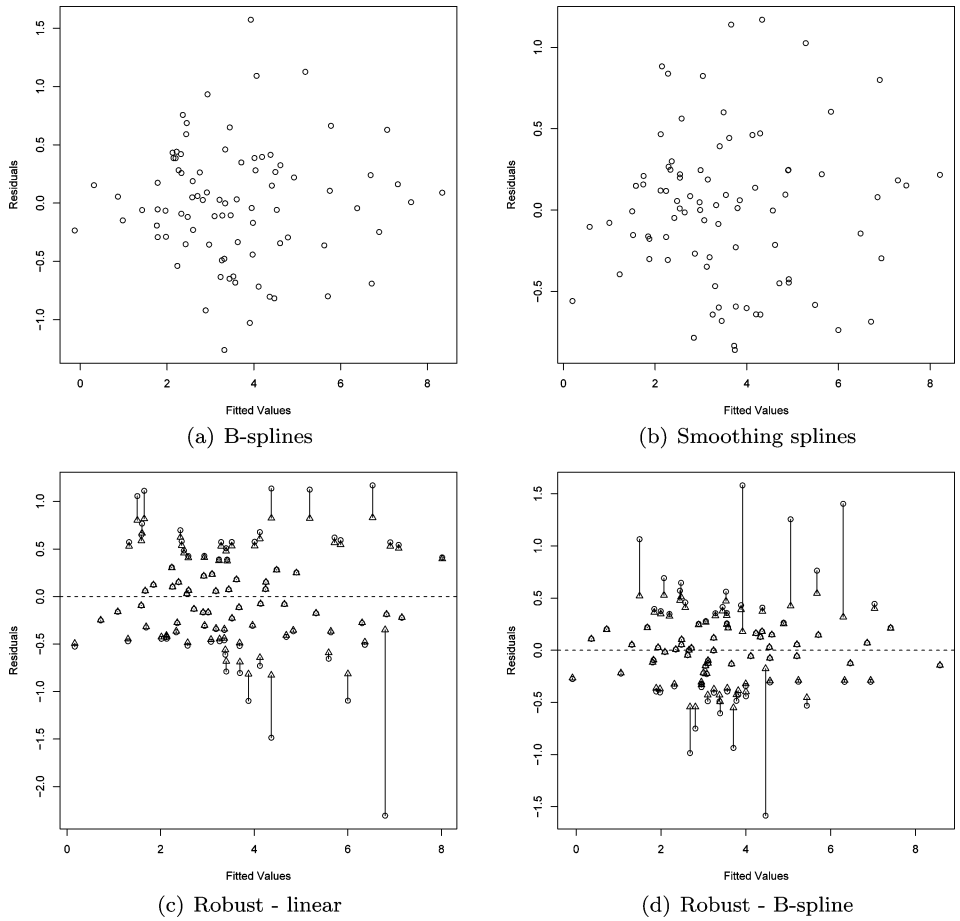
Thus, our recommended model is the robustly fitted model which includes ESXL, SASA, FOSA, DIPL, and HBDN. This model has an estimated conditional expected loss of prediction of 0.125 which represents a 48% improvement over the Duffy and Jorgensen model. In terms of the chemical relevance of our chosen variables, electrostatic interactions are modeled using DIPL, as was the case for the Duffy and Jorgensen model. Our additional terms incorporate explicitly the Lennard-Jones interactions of the solutes (ESXL), together with additional terms reflecting this and cavity formation in the solvent. The presence of the HBDN term is harder to rationalize, although from the reported correlation structure of the variables,[1] HBDN is correlated with both the Coulomb and Lennard-Jones intermolecular interactions.

3.2. Octanol/Gas. We fitted a linear regression model to identify the contributing explanatory variables and provide



(a) Residual plot (b) Normal probability plot

Figure 4. Regression diagnostics for the multiple linear regression model for log *P*(*octanol*/*gas*) reported by Duffy and Jorgensen.[1]



(a) B-splines (b) Smoothing splines

(c) Robust - linear (d) Robust - B-spline

Figure 5. Residual plots for the models fitted to the *logP*(*octanol*/*gas*) data. In the case of robust regression the effect of the weighting on the residuals is shown.

a good description of the data for the log *P*(*octanol*/*gas*) response. Backward elimination starting with all variables produced a four variable model including SASA, ESXC, FOSA, and HBDN with an estimated conditional expected loss of prediction of 0.491, the same model reported by Duffy and Jorgensen.[1] The diagnostics (Figure 4(a),(b)) suggest that this model does not fit well as the residual plot is fan shaped rather than an even horizontal band of points, and there is a discrepant point (off the line) in the lower tail of the normal probability plot. This point corresponds to aniline, and if this molecule is excluded and the model fitting process repeated, the same final model is selected at the end of the search with no marked improvement in model diagnostics. This suggests that a different approach is needed.

The models were extended to include B-splines to investigate whether they capture curvature in the data. GAMs with smoothing splines were also considered, and the diagnostic plots for two models identified by this approach are given in Figure 5(a),(b), which suggest improvements in the fit of the model. The contributions of the smoothing terms in a GAM to its four variables are shown in Figure 6. These plots show the contribution of each descriptor to the fitted values as the descriptor varies over its range. The ESXC relationship does not appear to be dramatically nonlinear, but the tests of this variable recommend using a smoother rather than linear term to represent this explanatory variable. Plots for the other three variables show clear nonlinearity and more clearly support the use of smoothing splines.

Regression models were also fitted to the log *P*(*octanol*/*gas*) response using robust methods, and these suggested a small group of molecules that have a reduced impact on the

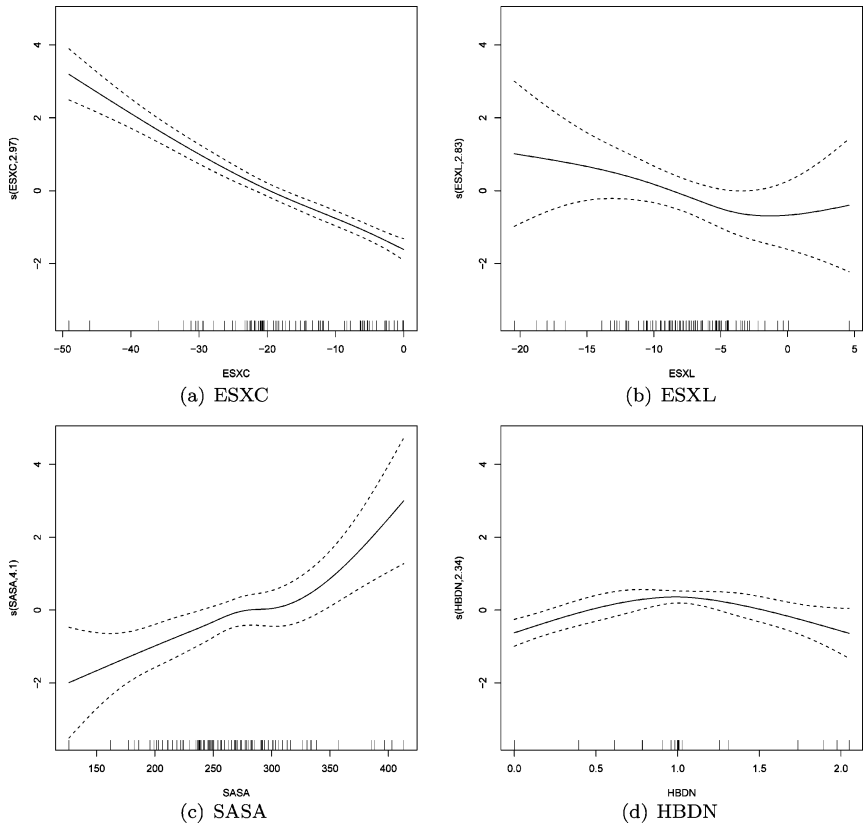


Figure 6. GAM plots for the model using smoothing splines for four variables.

model. Diagnostics for robustly fitted models using linear and B-spline representations of the explanatory variables are shown in parts (c) and (d) of Figure, respectively. Simplification of the full models in these two cases suggest that there are four contributing variables—ESXC, ESXL, FOSA, and HBDN with estimated expected conditional losses of prediction of 0.397 for the linear model and 0.246 for the B-spline model.

An overall summary of models that we consider for this response are given in Table 2, and as in section 3.1 cross-validation chooses the GAMs as the best models and the linear models with B-splines perform poorly. When we computed bootstrap estimates of expected conditional loss of prediction, the GAMs again were shown to be the best of the nonrobust models, but, using B-splines and robust parameter estimation, we also achieved a good fit to the data and the most accurate predictions. The contributing variables chosen by the different classes of models are very similar, with ESXC and HBDN being common to all models and a surface area term (either SASA or FOSA) contributing to a good description of the data. A reduction in estimated conditional loss of prediction of about 50% over the Duffy and Jorgensen (DJ) model is achieved when we consider different models which deal with the curvature and discrepant molecules for the log *P*(*octanol*/*gas*) response. In terms of the variables selected, these are very similar and, in some cases, identical to those reported by Duffy and Jorgensen.[1] Interestingly, of the four variables in our preferred model, three are also present in the preferred model for solvation in hexadecane, suggesting that we are capturing in a consistent way part of the physics of nonaqueous solvation.

3.3. Water/Gas. The same group of molecules that have experimental data for log *P*(*octanol*/*gas*) also have experimental log *P*(*water*/*gas*) values. Starting with a full linear model and removing explanatory variables until the model can no longer be simplified left four variables, ESXC, FOSA, the number of amines, and the number of nitro and acid groups with an estimated conditional expected loss of prediction of 0.564. This is a similar model to that reported by Duffy and Jorgensen.[1] Diagnostic plots for this model indicated that the fit to the data is reasonable, with only a couple of points with large fitted values that are separate from the majority of the molecules.

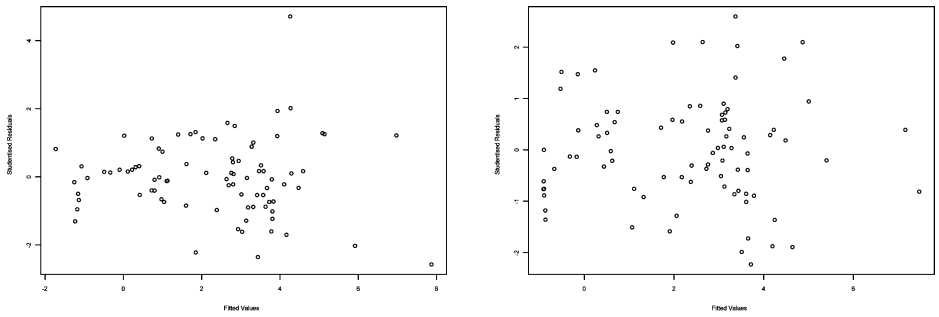
B-splines and cubic polynomials were also used to investigate whether the model for this response could be improved. The diagnostic plots were poor when a full model was used, but these problems disappeared when the number of explanatory variables was reduced (this is shown by two residual plots in parts (a) and (b) in Figure 7(,respectively). FOSA is an important contributing variable in the model when linear terms are used to describe the contributing variables but is replaced by SASA when cubic or B-spline terms were used to increase the flexibility of the model. Different combinations of cubics, B-splines, and linear terms were considered for the contributing variables and reported in Table 3.

Regression models were also fitted robustly to the data to investigate whether they provided an improved description of the log *P*(*water*/*gas*) data. In the case of linear terms for the explanatory variables, the same four variables were identified as contributing to the model. Differences between model parameters for the robust and nonrobust fits occur for the number of amines and the number of nitro and acid groups variables. Inclusion of B-splines, or cubic polynomials in the case of ARSA and HBDN, further improved the fit to the data.

Table 2: Models for the Response log *P*(*Octanol*/*Gas*) Based on a Variety of Methods, Identified by the Code in the First Column (lm, gam, rlm)

		cross-validation		bootstrap		
call	model formula	l-o-o	10-fold	av	10% tr	df
lm	ESXC+ESXL+SASA+FOSA+ARSA+DIPL+HBDN+HBAC+INME+ INST+amine+nitro acid	0.411	0.472	0.330	0.331	13
lm	ESXC+SASA+FOSA+HBDN+amine	0.409	0.435	0.406	0.406	6
lm	ESXC+SASA+FOSA+HBDN	0.454	0.464	0.491	0.491	5
lm	bs(ESXC,4)+bs(SASA,4)+bs(FOSA,4)+bs(HBDN,3)	0.381	0.594	0.302	0.303	16
lm	bs(ESXC,4)+bs(SASA,4)+bs(HBDN,3)+bs(FOSA,3)	0.374	0.578	0.297	0.298	15
lm	bs(ESXC,4)+bs(SASA,4)+bs(FOSA,3)	0.560	0.609	0.592	0.593	12
gam[a]	s(ESXC,6)+s(ESXL,6)+s(SASA,6)+s(FOSA,6)+s(ARSA,6)+ s(HBDN,6)+amine+nitro acid	0.333	0.552	0.207	0.207	15.36
gam[a]	s(ESXC,6)+s(ESXL,6)+s(SASA,6)+s(HBDN,6)+amine+nitro acid+FOSA	0.353	0.377	0.197	0.197	14.60
gam	s(ESXC,6)+s(SASA,6)+s(HBDN,6)+amine+nitro acid+FOSA	0.318	0.343	0.213	0.213	12.22
rlm	ESXC+ESXL+SASA+FOSA+ARSA+HBDN+INME+INST+amine+nitro acid			0.286	0.286	11
rlm	ESXC+ESXL+SASA+FOSA+ARSA+HBDN+INST+amine+nitro acid			0.289	0.289	10
rlm	ESXC+ESXL+SASA+FOSA+HBDN+INST+amine+nitro acid			0.295	0.295	9
rlm	ESXC+ESXL+SASA+FOSA+HBDN+amine+nitro acid			0.298	0.298	8
rlm	ESXC+ESXL+SASA+FOSA+HBDN+amine			0.321	0.322	7
rlm	ESXC+ESXL+FOSA+HBDN+amine			0.359	0.360	6
rlm	ESXC+ESXL+FOSA+HBDN			0.397	0.397	5
rlm	ESXC+ESXL+HBDN			0.498	0.498	4
rlm	bs(ESXC,4)+bs(ESXL,4)+bs(SASA,4)+bs(FOSA,4)+bs(ARSA,3)+ bs(HBDN,3)+bs(INME,4)+amine+nitro acid			0.194	0.195	29
rlm	bs(ESXC,4)+bs(ESXL,4)+bs(SASA,4)+bs(FOSA,4)+bs(ARSA,3)+ bs(HBDN,3)+bs(INME,4)+amine			0.253	0.255	28
rlm[a]	bs(ESXC,4)+bs(ESXL,4)+bs(SASA,4)+bs(FOSA,4)+bs(HBDN,3)+ bs(INME,4)+amine			0.185	0.187	25
rlm[a]	bs(ESXC,4)+bs(ESXL,4)+bs(SASA,4)+bs(FOSA,4)+bs(HBDN,3)+amine			0.191	0.194	21
rlm[a]	bs(ESXC,4)+bs(ESXL,4)+bs(FOSA,4)+bs(HBDN,3)+amine			0.224	0.226	17
rlm	bs(ESXC,4)+bs(ESXL,4)+bs(FOSA,4)+bs(HBDN,3)			0.243	0.246	16
rlm	bs(ESXC,4)+bs(ESXL,4)+bs(HBDN,3)			0.275	0.279	12

[a] Denotes good models.



(a) Full 12 variable model (b) 5 variable model

Figure 7. Regression diagnostics for the linear regression model using B-splines for *logP*(*water*/*gas*) based on all variables and a good subset.

Competing models were compared using bootstrap estimates of conditional expected loss of prediction. Results for a range of possible models are shown in Table 3. The model reported by Duffy and Jorgensen[1] with ESXC, FOSA, amine, and nitro acid has an estimated conditional expected loss of prediction of 0.564. The fit can be improved by using B-splines or cubic polynomials which reduces the estimate to 0.417 when ESXC, SASA, HBAC, amine, and nitro acid are included. The robustly fitted regression models offer further reductions in the conditional expected loss of prediction, and they are made more flexible by handling curvature via the B-splines. For the B-spline model with ESXC, SASA, FOSA, HBDN, HBAC, and amine we get an estimate of 0.296, which is 48% better than the DJ model. Note that the number of nitro and acid groups does not appear to be important when B-splines are used in the robustly fitted regression models, but it is included in all the other models. This suggests that the variables capture the behavior of the discrepant observations in this data set.

As noted by Duffy and Jorgensen,[1] we find ESXC to be a very important predictor for log *P*(*water*/*gas*), reflecting the dominance of electrostatics in aqueous solvation. Cavity formation and the hydrophobic effect are modeled through the surface area terms, and the presence of both hydrogen bonding variables reflects the very important role of hydrogen bonds in hydration.

3.4. Octanol/Water. The fourth response has the largest number of observations in the Duffy and Jorgensen[1] data—196 molecules that are supplemented by an extra 48 molecules. The data can be divided into organic molecules, drugs, and aromatic heterocycles, and we consider fitting models to the organic molecules, drugs, and then all the molecules together. There are 20 aromatic heterocycles so we do not try to fit a model to this group of molecules.

Table 3: Models for the Response log *P*(*Water*/*Gas*) Based on a Variety of Methods, Identified by the Code in the First Column (lm, gam, rlm)

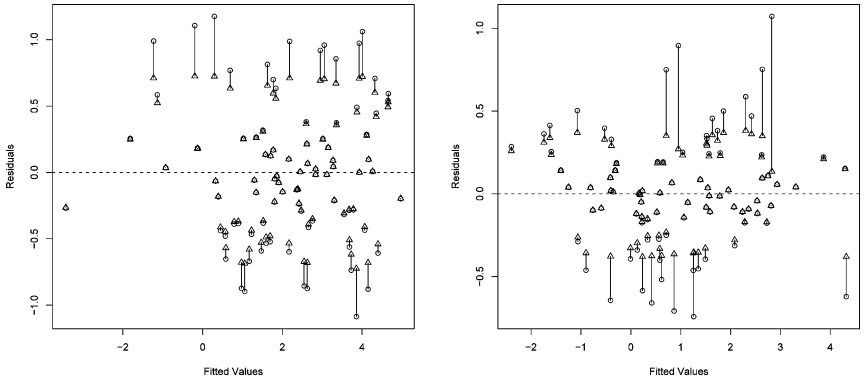
		cross-validation		bootstrap		
call	model formula	l−o-o	10-fold	av	10% tr	df
lm	bs(ESXC,4)+amine+nitro acid	0.561	0.568	0.535	0.535	7
lm	bs(ESXC,3)+amine+nitro acid	0.630	0.644	0.619	0.619	6
lm	bs(ESXC,4)+bs(SASA,4)+amine+nitro acid	0.557	0.548	0.495	0.497	11
lm	bs(ESXC,4)+bs(SASA,4)+bs(HBAC,4)+amine+nitro acid	0.812	0.738	0.417	0.417	15
lm	bs(ESXC,4)+bs(SASA,3)+amine+nitro acid	0.546	0.552	0.490	0.492	10
lm	bs(ESXC,3)+bs(SASA,3)+amine+nitro acid	0.695	0.709	0.628	0.629	9
lm	bs(ESXC,5)+bs(SASA,3)+amine+nitro acid	0.559	0.566	0.484	0.486	11
lm	bs(ESXC,4)+SASA+amine+nitro acid	0.575	0.595	0.547	0.548	8
lm	ESXC+FOSA+amine+nitro acid	0.591	0.602	0.564	0.565	5
lm	ESXC+ESXL+SASA	1.129	1.160	1.640	1.641	4
rlm	ESXC+SASA+FOSA+ARSA+DIPL+HBDN+HBAC+INST+amine+nitro acid			0.446	0.446	11
rlm	ESXC+SASA+FOSA+ARSA+HBDN+HBAC+INST+amine+nitro acid			0.430	0.431	10
rlm	ESXC+SASA+FOSA+HBDN+HBAC+INST+amine+nitro acid			0.432	0.432	9
rlm	ESXC+SASA+FOSA+HBDN+HBAC+amine+nitro acid			0.457	0.457	8
rlm	ESXC+SASA+FOSA+HBAC+amine+nitro acid			0.520	0.521	7
rlm	ESXC+FOSA+HBAC+amine+nitro acid			0.538	0.539	6
rlm	ESXC+FOSA+amine+nitro acid			0.574	0.575	5
rlm	bs(ESXC,4)+bs(SASA,4)+bs(FOSA,4)+bs(DIPL,4)+bs(HBDN,3)+ bs(HBAC,4)+bs(INME,4)+amine+nitro acid			0.238	0.238	30
rlm[a]	bs(ESXC,4)+bs(SASA,4)+bs(FOSA,4)+bs(HBDN,3)+bs(HBAC,4)+ bs(INME,4)+amine+nitro acid			0.241	0.242	26
rlm[a]	bs(ESXC,4)+bs(SASA,4)+bs(FOSA,4)+bs(HBDN,3)+bs(HBAC,4)+ amine+nitro acid			0.268	0.270	22
rlm[a]	bs(ESXC,4)+bs(SASA,4)+bs(FOSA,4)+bs(HBDN,3)+bs(HBAC,4)+amine			0.296	0.296	21
rlm	bs(ESXC,4)+bs(SASA,4)+bs(FOSA,4)+bs(HBDN,3)+bs(HBAC,4)			0.495	0.495	20
rlm	bs(ESXC,4)+bs(SASA,4)+bs(HBDN,3)+bs(HBAC,4)			0.514	0.515	16
rlm	bs(ESXC,4)+bs(SASA,4)+bs(HBDN,3)			0.692	0.692	12
rlm	bs(ESXC,4)+bs(HBDN,3)			0.759	0.759	8

[a] Denotes good models.

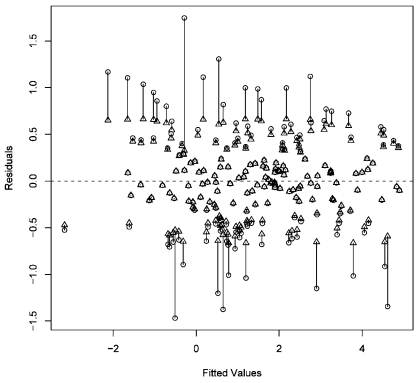
First we consider the 87 organic molecules that have experimental values for log *P*(*octanol*/*water*). Variable selection leads to a model with four explanatory variables—ESXL, HBAC, the number of amines, and the number of nitro and acid groups which has an estimated conditional expected loss of prediction of 0.149. Diagnostic plots for this model suggested that there may be curvature present, so we considered using B-splines for the explanatory variables. This analysis suggested that ESXL should be replaced by ESXC and SASA included and represented by B-splines for an estimated conditional expected loss of prediction of 0.113. We also fitted GAMs to the organic molecule data and found that smoothing splines are required to describe the contribution of ESXC, but linear terms can be used for three other variables as well as the discrete variables counting the number of amines, nitro, and acid groups. Here the estimated conditional expected loss of prediction is 0.069. Robust regression fits suggested using five variables in the model which has the same descriptors as the other models (residual plot in Figure 8), but these fits were not as good as the best GAMs. A summary of models we fitted to the data is given in Table 4. Estimated conditional expected loss of prediction suggested that robustly fitted models perform slightly better than the models with linear terms and that the GAMs are an obvious improvement in terms of prediction on all of the other types of model, with a bootstrap estimate of prediction error of 0.069, a 53% improvement over the DJ model.

We then considered the group of drugs (89 molecules) and fitted linear models to these data. Diagnostics for a five variable model with SASA, DIPL, HBAC, amine, and nitro acid with an estimated conditional expected prediction loss of 0.370 are good with only sucrose standing out from the rest of the residuals. Using B-splines to represent the explanatory variables did not substantially improve the fit of the model, but GAMs with more flexibility for the smoothing splines did improve the fit. The impact of using smoothing splines is shown in Figure 9 for the model where ESXL, FOSA, DIPL, and INST are described using smoothers, and HBAC, SASA, amine, and nitro acid are linear. Three of the variables have obvious nonlinearity, INST is less clear-cut as only 1.5 degrees of freedom are used by its smoothing spline so perhaps a linear term could be used. Robustly fitted regression models were also considered, and although these improve on the nonrobustly fitted models with linear or B-spline terms, the GAMs appear to be the best choice for this response and the drugs. Table 5 lists models that we considered for the drugs, and the estimated conditional expected loss of prediction support the use of GAMs. The best GAM has estimated conditional expected loss of prediction of 0.190 which is a 48% improvement on the best linear model.

Finally we considered the combined group of 196 molecules (drugs, organics, and aromatic heterocycles) and fitted models to this larger set of data. We obtained good fit with a model using linear terms in SASA, DIPL, HBAC, amine, and nitro acid (estimated conditional expected loss of prediction of 0.660) and mild further improvements from extending the model to use B-splines for describing the effect of the contributing variables on log *P*(*octanol*/*water*). GAMs also fitted the data well but at the cost of greater complexity of the model. Robust regression fits were also used; the robust regression fit to the model with ESXL, SASA, DIPL, HBDN, HBAC, INST, amine, and nitro acid gave a good



(a) Residual plot - drugs (b) Residual plot - organic



(c) Residual plot - rest

Figure 8. Regression diagnostics for models based on different subsets of the octanol/gas data.

Table 4: Models for the Organic Molecules for Response log *P*(*Octanol*/*Water*) Based on a Variety of Methods, Identified by the Code in the First Column (lm, gam, rlm)

		cross-validation		bootstrap		
call	model formula	l-o-o	10-fold	av	10% tr	df
lm	ESXC+ESXL+SASA+FOSA+ARSA+DIPL+HBDN+HBAC+INME+INST+amine+nitro acid	0.156	0.144	0.114	0.114	13
lm	ESXL+HBAC+amine+nitro acid	0.144	0.136	0.149	0.149	5
lm	SASA+HBAC+amine+nitro acid	0.193	0.174	0.207	0.207	5
lm	bs(ESXC,4)+bs(SASA,4)+bs(DIPL,4)+bs(HBAC,4)+bs(INST,4)+amine+nitro acid	2.508	1.528	0.081	0.081	23
lm	bs(ESXC,4)+bs(SASA,4)+bs(HBAC,4)+amine+nitro acid	0.667	0.796	0.113	0.113	15
gam[a]	s(ESXC,5)+s(SASA,5)+s(ARSA,5)+s(DIPL,5)+s(HBAC,5)+s(INST,5)+amine+nitro acid	0.172	0.146	0.056	0.056	19.53
gam[a]	s(ESXC,5)+s(SASA,5)+s(ARSA,5)+s(HBAC,5)+s(INST,5)+amine+nitro acid	0.148	0.130	0.063	0.063	15.70
gam[a]	s(ESXC,5)+s(HBAC,5)+s(INST,5)+amine+nitro acid+SASA+ARSA	0.125	0.136	0.055	0.055	15.70
gam	s(ESXC,5)+amine+nitro acid+SASA+ARSA+HBAC	0.134	0.117	0.069	0.069	9.91
rlm	ESXC+ESXL+SASA+ARSA+HBDN+HBAC+INME+INST+amine+nitro acid			0.101	0.101	11
rlm	ESXC+ESXL+SASA+ARSA+HBDN+HBAC+INME+amine+nitro acid			0.101	0.101	10
rlm	ESXL+SASA+ARSA+HBDN+HBAC+INME+amine+nitro acid			0.100	0.101	9
rlm	ESXL+SASA+HBDN+HBAC+INME+amine+nitro acid			0.103	0.103	8
rlm	ESXL+SASA+HBAC+INME+amine+nitro acid			0.108	0.109	7
rlm	ESXL+SASA+HBAC+amine+nitro acid			0.114	0.115	6
rlm	ESXL+HBAC+amine+nitro acid			0.140	0.140	5
rlm	bs(ESXC,4)+bs(SASA,4)+bs(FOSA,4)+bs(ARSA,3)+bs(DIPL,4)+bs(HBDN,3)+bs(HBAC,4)+bs(INME,4)+nitro acid			0.116	0.116	32
rlm	bs(ESXC,4)+bs(SASA,4)+bs(FOSA,4)+bs(ARSA,3)+bs(DIPL,4)+bs(HBAC,4)+bs(INME,4)+nitro acid			0.090	0.090	29
rlm	bs(ESXC,4)+bs(SASA,4)+bs(ARSA,3)+bs(DIPL,4)+bs(HBAC,4)+bs(INME,4)+nitro acid			0.091	0.091	25
rlm	bs(ESXC,4)+bs(SASA,4)+bs(ARSA,3)+bs(HBAC,4)+bs(INME,4)+nitro acid			0.110	0.111	21
rlm	bs(ESXC,4)+bs(SASA,4)+bs(ARSA,3)+bs(HBAC,4)+nitro acid			0.114	0.114	17
rlm	bs(ESXC,4)+bs(SASA,4)+bs(HBAC,4)+nitro acid			0.157	0.157	14

[a] Denotes good models.

description of the data with an estimated conditional loss of prediction of 0.226. SASA and HBAC appear in all the different types of models, and the number of amines and the number of nitro and acid groups also contribute to the fit. Table 6 provides a summary of models considered, and the estimated conditional expected loss of prediction support using robust regression to describe the data. The importance of SASA and HBAC is totally consistent with the models of Duffy and Jorgensen. The former reflects the importance of van der Waals interactions in the octanol solvent, while

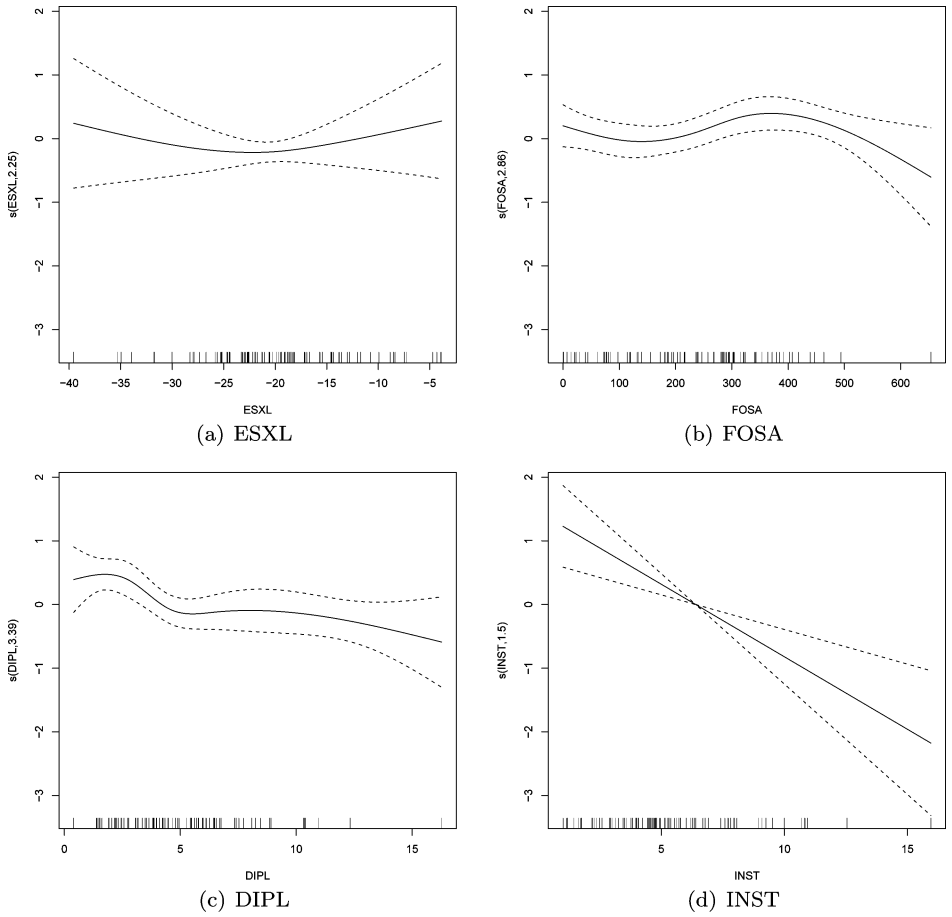


Figure 9. GAM plots for the model related to the drugs.

Table 5: Models for the Drugs for Response log *P*(*Octanol*/*Water*) Based on a Variety of Methods, Identified by the Code in the First Column (lm, gam, rlm)

		cross-validation		bootstrap		
call	model formula	l-o-o	10-fold	av	10% tr	df
lm	ESXC+ESXL+SASA+FOSA+ARSA+DIPL+HBDN+HBAC+INME+INST+amine+nitro acid	0.324	0.319	0.268	0.268	13
lm	SASA+HBAC+INST+amine+nitro acid	0.321	0.306	0.327	0.327	6
lm	SASA+DIPL+HBAC+amine+nitro acid	0.348	0.326	0.370	0.370	6
lm	ESXL+HBAC+amine+nitro acid	0.458	0.44	0.580	0.580	5
lm	bs(SASA,4)+bs(DIPL,4)+bs(HBAC,4)+bs(INST,4)+amine+nitro acid	1.842	1.179	0.235	0.235	19
lm	bs(SASA,4)+bs(HBAC,4)+bs(INST,4)+amine+nitro acid	1.472	0.871	0.306	0.307	15
lm	bs(SASA,4)+bs(INST,4)+amine+nitro acid+bs(HBAC)	0.567	0.485	0.305	0.306	14
lm	bs(SASA,4)+amine+nitro acid+bs(HBAC)+bs(INST)	0.872	0.695	0.324	0.325	13
lm	bs(INST,4)+amine+nitro acid+bs(HBAC)+bs(SASA)	0.462	0.518	0.310	0.310	13
gam[a]	s(ESXL,5)+s(SASA,5)+s(FOSA,5)+s(DIPL,5)+s(HBAC,5)+s(INST,5)+amine+nitro acid	0.365	0.378	0.139	0.138	19.66
gam[a]	s(ESXL,5)+s(SASA,5)+s(FOSA,5)+s(DIPL,5)+s(INST,5)+amine+nitro acid+HBAC	0.262	0.302	0.130	0.130	19.66
gam	s(ESXL,5)+s(FOSA,5)+s(DIPL,5)+s(INST,5)+amine+nitro acid+HBAC+SASA	0.253	0.309	0.190	0.190	14.99
rlm	ESXC+ESXL+SASA+FOSA+ARSA+DIPL+HBAC+INST+amine+nitro acid			0.280	0.280	11
rlm	ESXC+SASA+FOSA+ARSA+DIPL+HBAC+INST+amine+nitro acid			0.276	0.276	10
rlm	ESXC+SASA+ARSA+DIPL+HBAC+INST+amine+nitro acid			0.279	0.279	9
rlm	SASA+ARSA+DIPL+HBAC+INST+amine+nitro acid			0.287	0.287	8
rlm	SASA+DIPL+HBAC+INST+amine+nitro acid			0.294	0.294	7
rlm	SASA+HBAC+INST+amine+nitro acid			0.345	0.345	6
rlm[a]	bs(ESXL,4)+bs(SASA,4)+bs(FOSA,4)+bs(DIPL,4)+bs(HBDN,3)+bs(HBAC,4)+bs(INME,4)+amine+nitro acid			0.147	0.148	30
rlm	bs(ESXL,4)+bs(SASA,4)+bs(FOSA,4)+bs(DIPL,4)+bs(HBAC,4)+bs(INME,4)+amine+nitro acid			0.181	0.182	27
rlm	bs(ESXL,4)+bs(SASA,4)+bs(DIPL,4)+bs(HBAC,4)+bs(INME,4)+amine+nitro acid			0.199	0.200	23
rlm	bs(ESXL,4)+bs(SASA,4)+bs(HBAC,4)+bs(INME,4)+amine+nitro acid			0.261	0.262	19
rlm	bs(SASA,4)+bs(HBAC,4)+bs(INME,4)+amine+nitro acid			0.331	0.333	15
rlm	bs(SASA,4)+bs(INME,4)+amine+nitro acid			0.477	0.478	11

[a] Denotes good models.

Table 6: Models for the Combined Group for Response log *P*(*Octanol*/*Water*) Based on a Variety of Methods, Identified by the Code in the First Column (lm, gam, rlm)

		cross-validation		bootstrap		
call	model formula	l-o-o	10-fold	av	10% tr	df
lm	ESXC+ESXL+SASA+FOSA+ARSA+DIPL+HBDN+HBAC+INME+INST+amine+nitro acid	0.309	0.290	0.626	0.626	13
lm	SASA+DIPL+HBAC+amine+nitro acid	0.295	0.286	0.660	0.660	6
lm	bs(ESXC,4)+bs(SASA,4)+bs(FOSA,4)+bs(HBAC,4)+bs(INME,4)+bs(INST,4)+amine+nitro acid	0.399	0.402	0.431	0.432	27
lm	bs(SASA,4)+bs(HBAC,4)+bs(INME,4)+bs(INST,4)+amine+nitro acid	0.307	0.308	0.534	0.534	19
lm	bs(SASA,4)+bs(HBAC,4)+bs(INME,4)+amine+nitro acid+bs(INST)	0.380	0.378	0.591	0.591	18
lm	bs(SASA,4)+bs(HBAC,4)+bs(INST,4)+amine+nitro acid+bs(INME)	0.357	0.351	0.578	0.578	18
gam	s(ESXC)+s(ESXL)+s(SASA)+s(HBDN)+s(HBAC)+s(INST)+amine+nitro acid	0.963	1.052	0.330	0.330	38.50
gam	s(ESXC)+s(ESXL)+s(HBDN)+s(HBAC)+s(INST)+amine+nitro acid+SASA	1.212	0.410	0.490	0.491	21.68
gam	s(HBDN)+s(HBAC)+s(INST)+amine+nitro acid+SASA+s(ESXL,20)	0.402	0.321	0.581	0.581	12.55
rlm[a]	ESXL+SASA+FOSA+ARSA+DIPL+HBDN+HBAC+INST+amine+nitro acid			0.225	0.225	11
rlm[a]	ESXL+SASA+ARSA+DIPL+HBDN+HBAC+INST+amine+nitro acid			0.224	0.225	10
rlm[a]	ESXL+SASA+DIPL+HBDN+HBAC+INST+amine+nitro acid			0.226	0.226	9
rlm	ESXL+SASA+DIPL+HBAC+INST+amine+nitro acid			0.237	0.237	8
rlm	ESXL+SASA+DIPL+HBAC+amine+nitro acid			0.239	0.239	7
rlm	SASA+DIPL+HBAC+amine+nitro acid			0.260	0.261	6
rlm[a]	bs(ESXC,4)+bs(ESXL,4)+bs(SASA,4)+bs(FOSA,4)+bs(HBAC,4)+bs(INME,4)+bs(INST,4)+amine+nitro acid			0.199	0.200	31
rlm[a]	bs(ESXL,4)+bs(SASA,4)+bs(FOSA,4)+bs(HBAC,4)+bs(INME,4)+bs(INST,4)+amine+nitro acid			0.213	0.214	27
rlm	bs(SASA,4)+bs(FOSA,4)+bs(HBAC,4)+bs(INME,4)+bs(INST,4)+amine+nitro acid			0.262	0.263	23
rlm	bs(SASA,4)+bs(HBAC,4)+bs(INME,4)+bs(INST,4)+amine+nitro acid			0.292	0.293	19
rlm	bs(SASA,4)+bs(HBAC,4)+bs(INME,4)+amine+nitro acid			0.320	0.321	15

[a] Denotes good models.

the latter discriminates between the hydrogen bonding abilities of the octanol and water systems. The additional variables we find to be significant reflect further the subtleties of the solvation of molecules by these solvents. That additional terms related to the numbers of certain chemical groups are required and may reflect either an inadequacy of the molecular mechanics derived descriptors for the chemical functionalities, perhaps due to the charge model, issues with the experimental data for these compounds, perhaps due to multiple protonation states, or a failure to capture some specific aspects of the solvation physics in the molecular mechanics descriptors.

4. MODEL BUILDING

Our approach to model building is to begin with graphical exploration of the data, including exploring the relationship between descriptors before fitting models. We then consider a simple class of models (such as multiple regression models) with all or a large set of possible descriptors (called the full model). We use diagnostic plots to assess the fit of the model and, when necessary, to suggest modifications to the model. Generally, we try to modify the model (by applying transformations to some variables or by including further descriptors) but note the valuable lesson learned in subsection 3.3 about the effect of overfitting on diagnostics. We then try to simplify the model (i.e. obtaining a small set of interesting models with fewer terms) by applying an eclectic mix of information theoretic criterion and formal testing.

We found evidence of curvature and discrepant observations in our analyses so we proceeded to explore these by widening the class of models and using robust fitting procedures. In each case, we used diagnostics and model selection methods to obtain a small set of interesting models with as few descriptors in them as possible, consistent with maintaining good fit.

We compared the different subsets of models by using cross-validation and bootstrap methods to estimate the conditional expected loss of prediction of all the models in each subset. We found the cross-validation estimates to be highly variable and preferred the bootstrap estimates. We selected as our final model the smallest model (fewest descriptors in the simplest form) with an estimated conditional expected loss of prediction close to the minimum achieved over the models under consideration.

5. CONCLUSION

Analysis of the data from Duffy and Jorgensen[1] indicates that improved models can be identified by dealing with curvature and discrepant observations for all four responses. We initially consider linear terms for the explanatory variables, followed by B-splines, and then smoothing splines via the GAM framework to identify the contributing variables. Model diagnostics are used to find problems with fitted models and to suggest ways to handle these problems. We also use robust regression fitting with Tukey M-estimators to reduct the impact of influential observations on the models in an attempt to describe the majority of the data with good accuracy while accepting that some molecules are not well described by our parsimonious regression models. Model selection uses a mixture of formal testing for significance of variables and an information criterion to reduce the number of explanatory variables as far as possible. We also consider resampling methods such as bootstrapping and cross-validation to calculate the conditional expected loss of prediction to compare different models and different fitting methods.

In general, models were identified that yielded an improvement of approximately 50% in the conditional expected loss of prediction over those originally reported by Duffy and Jorgensen. In most cases, robust regression yielded the

best-performing models, by down-weighting discrepant observations. The identification of these molecules is useful since it may reflect a deficiency either in the experimental observation that we are attempting to reproduce or in the descriptors being used in the fit. Alternatively, these molecules may interact with solvent in a fashion that is not captured by the model. It was also found that cross-validation approaches to estimating the reliability of the statistical models were often unreliable and that bootstrapping methods are to be preferred.

ACKNOWLEDGMENT

We are grateful to the EPSRC (GR/R67729) for supporting this work. We would like to thank Professor Bill Jorgensen and Dr. Erin Duffy for their generous provision of the original data from their publication.

REFERENCES AND NOTES

(1) Duffy, E. M.; Jorgensen, W. L. Prediction of Properties from Simulations: Free Energies of Solvation in Hexadecane, Octanol, and Water. *J. Am. Chem. Soc.* **2000**, *122*, 2878−2888.

(2) Rousseeuw, P. J.; Leroy, A. M. *Robust Regression & Outlier Detection;* Wiley: 1987.

(3) Hampel, F. R.; Ronchetti, E. M.; Rousseeuw, P. J.; Stahel, W. A. *Robust Statistics: The Approach Based on Influence Functions;* John Wiley: New York, 1986.

(4) Staudte, R. G.; Sheather, S. J. *Robust Estimation & Testing;* John Wiley: New York, 1990.

(5) Hastie, T.; Tibshirani, R. Generalized Additive Models. *Statistical Science* **1986**, 297−318.

(6) Hastie, T. J.; Tibshirani, R. J. *Generalized Additive Models;* Chapman & Hall: 1990.

(7) R Development Core Team. *R: A language and environment for statistical computing*; R Foundation for Statistical Computing: Vienna, Austria, 2004; ISBN 3-900051-00-3.

(8) Wood, S. N. Modelling and smoothing parameter estimation with multiple quadratic penalties. *J. R. Stat. Soc. B* **2000**, *62*, 413−428.

(9) Breiman, L.; Spector, P. Submodel selection and evaluation in regression. The x-random case. *Int. Stat. Rev.* **1992**, *60*, 291−319.

(10) Shao, J. Bootstrap Model Selection. *J. Am. Stat. Assoc.* **1996**, *91*, 655−665.

(11) Wisnowski, J. W.; Simpson, J. R.; Montgomery, D. C.; Runger, G. C. Resampling methods for variable selection in robust regression. *Comput. Stat. Data Anal.* **2003**, *43*, 341−355.

(12) Davison, A. C., Hinkley, D. V., Eds.; *Bootstrap Methods and their Application;* Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press: 1997.

CI050056I