Modelling Free Energies of Solvation Data

Ralph Mansson[a], Alan Welsh[b] and Jonathan Essex[c]
[a]School of Mathematics, University of Southampton
[b]Centre for Mathematics and its Applications, ANU
[c]School of Chemistry, University of Southampton

June 2, 2005

S3RI
Southampton Statistical
Sciences Research Institute

Summary of Presentation

- Introduction to free energy of solvation data.
- Classes of models considered for the data.
- Techniques for selecting subsets of important variables.
- An example from data of Duffy and Jorgensen (2000).
- Conclusions and future work.

Free Energy of Solvation Data I

Calculation of free energies of binding can be beneficial in rational drug design. There are theoretical approaches, e.g. free energy perturbation and thermodynamic integration, that can provide accurate estimates for free energies of binding. These techniques require computationally intensive calculations which make these methods impractical for work involving large numbers of molecules.

Other faster methods have been proposed to improve the efficiency of these simulations, e.g. linear interaction energy which uses a single variable equation to describe the change in energy. Subsequent work has increased the number of variables considered from modelling this relationship.

Free Energy of Solvation Data II

Duffy and Jorgensen (2000) ran Monte Carlo simulations on approximately 200 organic solutes with the overall aim of predicting partition coefficients, $logP$. The value of these coefficients measures the degree to which a compound is partitioned between two immiscible solvents. In this study the partition coefficient was modelled based on twelve descriptors calculated by simulations.

Response: $logP$ for octanol and water determined experimentally.

Variables: Descriptors based on surface are of the compound (four variables), changes in energy between bound and unbound states, hydrogen bonding and other features of the compounds.

Statistical Models

No prior model was assumed for the data, and we aim to derive empirically a useful approximation to the unknown relationship between partition coefficient and a set of simulated descriptors. The range of models we consider is not exhaustive and is based on previous positive experience in deriving models.

The data set has twelve descriptors and the analysis aims to select a subset of these descriptors that provides a good representation of variability in $logP$. It is likely that there are multiple models with similar predictive ability so we will not report only a single model or be too concerned with interpret the variables selected. We will discuss issues related to variable selection later.

In general the model is written in the form

$$Y_i = \beta_0 + \sum_{j=1}^{p} f_j(X_{ij}) + \epsilon_i \qquad \epsilon_i \sim \text{ independent } N(0, \sigma^2)$$

Southampton Statistical Sciences Research Institute

Statistical Models - Linear Regression

A simple model includes linear terms to describe the relationship between the response Y_i and the descriptors X_{ij}. We will use the Wilkinson-Rogers notation to describe our models, and a p variable linear model is given by

$$Y \sim X_1 + X_2 + \cdots + X_p.$$

This model is linear in the parameters and not the variables. We can choose Y to be any specified transformation of the response, and the Xs can be transformations of one or more descriptors. This flexibility can be used to represent some of the relationships by cubic polynomials or B-splines. In particular we consider single knot B-splines for the Duffy and Jorgensen data.

Statistical Models - GAMs

Generalized Additive Models have been given prominence by Hastie & Tibshirani (1990) and other researchers. Smooth functions are used to represent each variable and for a p variable model can be specified as

$$Y \sim s(X_1) + s(X_2) + \cdots + s(X_p)$$

and the degrees of freedom for each smooth are chosen simultaneously as part of the model fitting when using the implementation of Wood (2000). We can check whether a linear term can be used in the place of a smoother or if a term can be excluded from the model completely.

Effect plots are useful for describing the contribution of each variable to the fitted values and are produced for each smoother.

Statistical Models - Robust Regression I

Curvature in a regression model may be due to the linear approximation being poor or it might be due to discrepant observations. When a data set has discrepant observations we can consider robust parameter estimation for linear models. We can also include B-splines within the robust regression framework to make the model more flexible. For example

$$Y \sim X_1 + bs(X_2) + \cdots + X_p$$

A robust regression model aims to provide a good fit to the majority of the data while downweighting the influence of a small part of the data set. When considering diagnostics to assess model fit we use weighted residuals to take this into account.

Statistical Models - Robust Regression II

We use robust regression models with M-estimators which minimise a criterion function, $\rho(\epsilon_i)$, of the residual sum of squares

$$\sum_{i=1}^{N} \rho(Y_i - \mathbf{X}_i^T \boldsymbol{\beta})$$

and we choose the Tukey bisquare function for ρ, but there are other choices for the criterion function. To minimise this criterion funciton we solve the equations

$$\sum_{i=1}^{N} \psi(Y_i - \mathbf{X}_i^T \boldsymbol{\beta}) \mathbf{X}_i^T = 0$$

numerically using iteratively reweighted least squares where ψ, the influence function, is the derivative of the criterion function.

Statistical Models - Diagnostics I

It is important to use diagnostics to assess the fit of a model. When we have selected a model we need to evaluate the appropriateness of the

- form of the expectation function
- additivity of the error
- constant variance
- normality
- independence.

These assumptions are usually interrelated and investigated simultaneously.

Statistical Models - Diagnostics II

We investigate these assumptions informally using graphical displays based on the fitted values for the model, $\hat{Y}_i$, and the residuals

$$\epsilon_i = Y_i - \hat{Y}_i.$$

A plot of residuals against fitted values $(\hat{Y}_i, \epsilon_i)$ allows a check for the expectation function, $E(Y|X)$, additivity of the errors, and constant variance. We hope for a horizontal, even band of points where there are no vertical patterns centred on the x-axis. Different patterns indicate departures from the model assumptions.

We check for normality by plotting ordered residuals against the expected normal quantiles. The ideal plot is a straight line. To investigate dependence we need to know the order in which the data was collected.

Model Selection

Traditional approaches to model selection are based on a combination of stepwise selection and formal tests to identify the best subset of the descriptors. A criterion is used, e.g. AIC or Mallows' Cp, to add or remove descriptors.

Stepwise Selection: Options include forward, backwards, or best subsets. These procedures often retain a large number of variables in the model and can be combined with information criterion to select a parsimonious model.

Formal Tests: We can test the contribution of a descriptor or subset of the descriptors using t-tests and F-tests respectively.

We also consider resampling methods to calculate estimates of average prediction error for different classes of model. The number of parameters in the model needs to be incorporated in the selection procedure.

Model Selection - Cross-validation

This is a method of evaluating a model by dividing the data set into parts, and in turn using each part to test the model fitted to the remaining data. Cross-validation chooses the model with the best average predictive ability.

Leave-one-out cross-validation is frequently used which fits N models with one observation excluded and uses these models to predict the single observation. Other popular approaches are to divide the data into 5 or 10 (approximately) equal-sized parts, which is also computationally less expensive. 5-fold and 10-fold cross-validation have been shown to perform better for model selection in simulation studies.

There have been concerns about the ability of cross-validation to discriminate between models and its tendency to overfitting. A penalty term to deal with the number of model parameters has been proposed recently.

Model Selection - Bootstrapping I

An idea for statistical inference where training sets are created by resampling with replacement from the original data, which implies that observations can occur multiple times in the bootstrap sample. There are two choices - bootstrapping residuals or pairs, for the X fixed and random cases respectively.

In the case of bootstrapping residuals it is recommended to adjust the residuals, especially in the case of robust regression where it is possible to repeatedly sample discrepant observations. Davison and Hinkley (1997) proposed a leverage correction for the robust case which reduces some of these problems.

Model Selection - Bootstrapping II

Boostrapping an estimate of average prediction error for a given model follows these steps.

- Generate the bootstrap data either with residuals or pairs.
- Estimate the model parameters based on this new data.
- Use the model to compute the prediction error for the current sample.

These calculations are performed for B bootstrap samples.

Octanol/Water Example I

There are 196 compounds with experimental data for the $logP$ coefficient of octanol/water. The data are a combination of drugs, organic molecules, and aromatic heterocycles. The data set has 10 descriptors that are continuous and 2 descriptors that take on discrete values.

We use R for exploratory data analysis, to fit different models, and to produce estimates of prediction error via resampling methods. The **mgcv** implementation of GAMs of Wood (2000) is used with cross-validation selecting the smoothness for each descriptor and effect plots casn be used to investigate curvature. For robust regression we use the **rlm** function in the MASS library of Venables and Ripley (2002) with our implementation of a function to exclude model terms using a robust test. Bootstrapping is run using functions from the **boot** library in R.

Octanol/Water Example II

We considered three groups of molecules and selected a good subset of descriptors for each one.

Linear: Decent fit to the data and we can improve on the fit by dealing with curvature using single knot B-splines for the descriptors.

GAM: These models using smoothing splines also fit the data well but at the cost of added complexity (more degrees of freedom).

Robust: Reduce the influence of some of the compounds gives us good models using either linear or B-spline terms, but without the same complexity of the GAMs.

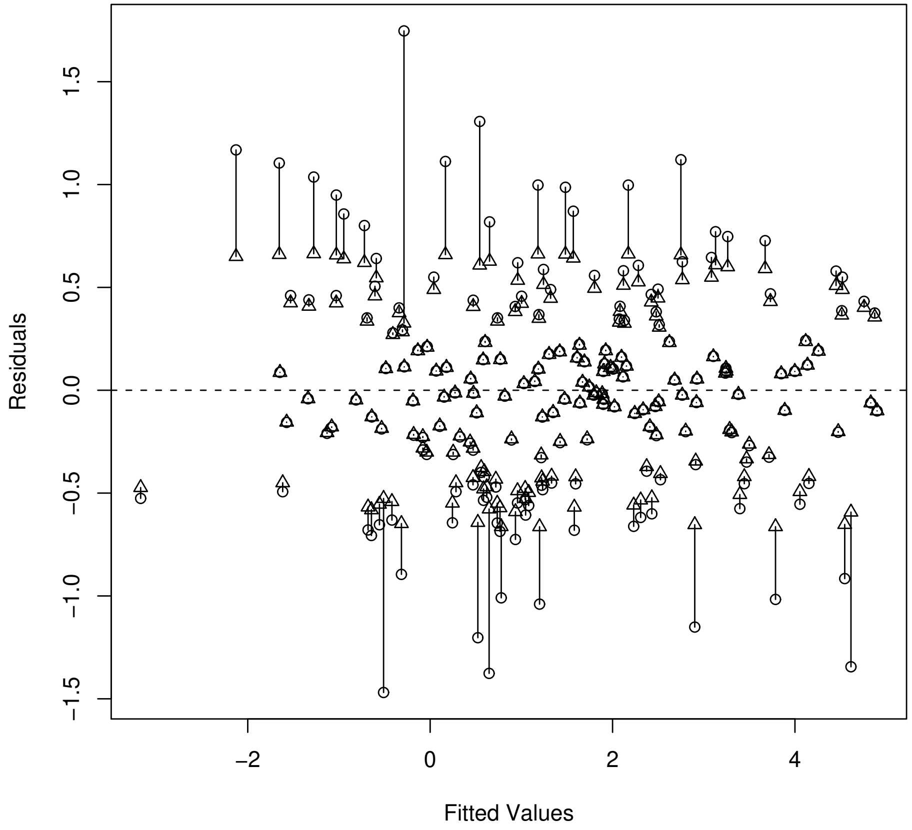
Octanol/Water Example III

Some good models for the data:

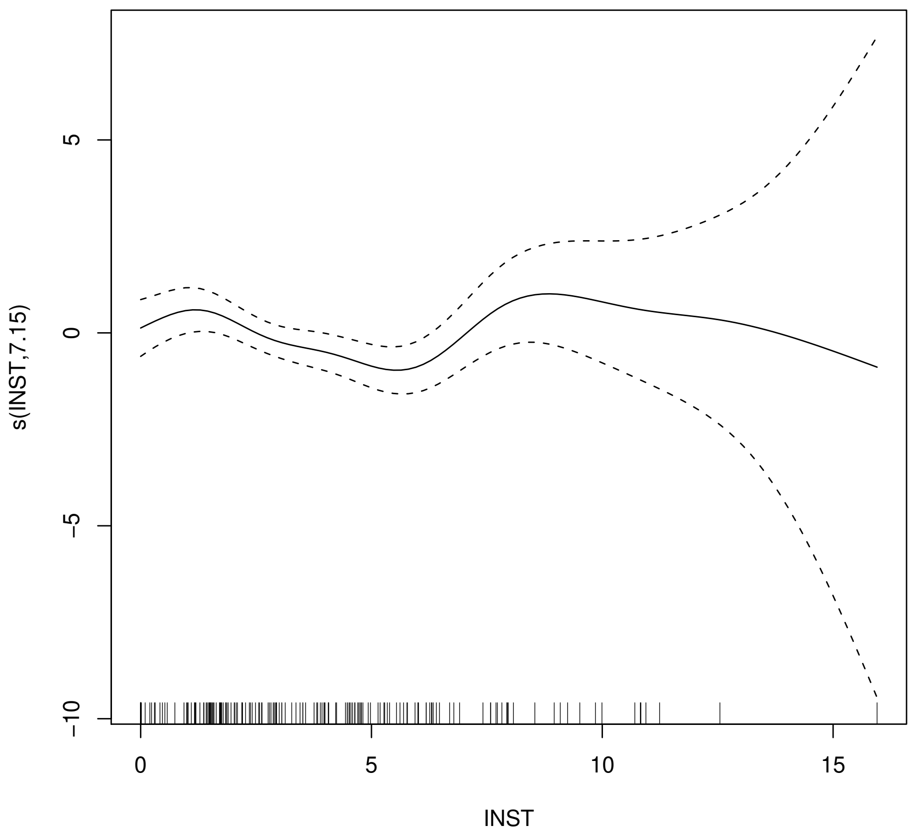
Model	P.E.	df
lm $\sim$ SASA+DIPL+HBAC+amine+nitro.acid	0.660	6
gam $\sim$ s(ESXC)+s(ESXL)+s(SASA)+s(HBDN)		
+s(HBAC)+s(INST)+amine+nitro.acid	0.330	38.5
rlm $\sim$ ESXL+SASA+DIPL+HBAC+amine+nitro.acid	0.228	7

Prediction error (P.E.) is estimated using 200 bootstrap samples.

Example - Robust Regression Residual Plot



Example - GAM Effect Plot



Southampton Statistical Sciences Research Institute

Conclusions

There is likely to be more than one model that is appropriate for a set of data. It is prudent to report a range of models with similar predictive power rather than concentrate on a single $best$ model.

It is sensible to use linear regression models where possible due to their simplicity. For the type of data we have encountered there are complications that can be handled well using smoothing splines and/or robust regression to reduce the influence of discrepant observations.

Variable selection is a difficult problem that has been approached in different ways. We recommend a combination of automatic selection and information based criterion to reduce the dimensionality of a model. Cross-validation and bootstrapping have been proposed to compute prediction error for a given model.

Future Work

Further sets of data will be generated using simulation by colleagues in the School of Chemistry at Southampton. The aim is to use the experience of analysing the Duffy and Jorgensen data with the new data.

The issue of variable selection needs to be investigated in more detail to compare the current techniques with a view towards providing practical advice for modellers.

Another area of interest is GAMs and the robust parameter estimation. A combination of the two methods would be useful tool for model building.