Linear regression, model averaging, and Bayesian techniques for predicting chemical activities from structure.

Jarad B. Niemi* and Gerald J. Niemi†

December 14, 2012

Abstract

A primary goal of quantitative structure-activity relationships (QSARs) and quantitative structure-property relationships (QSPRs) is to predict chemical activities from chemical structure. Chemical structure can be quantified in many ways resulting in hundreds, if not thousands, of measurements for every chemical. Chemical activities measures how the chemical interacts with other chemicals, e.g. toxicity, biodegradability, boiling point, and vapor pressure. Typically there are more chemical structure measurements than chemicals being measured, the so-called large-p, small-n problem. Here we review some of the statistical procedures that have been commonly used to explore these problems in the past and provide several examples of their use. Finally, we peek into the future to discuss two areas that we believe will see dramatically increased attention in the near future: model averaging and Bayesian techniques.

Keywords: regression, ridge regression, LASSO, elastic net, principal component analysis, Bayesian analysis, model selection, model averaging, k-means clustering, principal component regression, partial least squares, modeling, prediction, AIC, BIC, cross-validation

*Department of Statistics, Iowa State University, Ames, IA 50011 USA, niemi@iastate.edu
†Natural Resources Research Institute and Department of Biology, University of Minnesota, 5013 Miller Trunk Highway, Duluth, MN 55811 USA; gniemi@d.umn.edu

1 Introduction

The science of quantitative structure-activity relationships (QSAR) has a varied but relatively recent history [1]. The central tenet of structure-activity relationships (SAR) is that *form follows function* and this idea has probably been in existence for ages. Yet, the quantitative aspect of QSAR from a computer-age perspective is relatively young; perhaps only in existence for about 30 to 40 years. Hence, this is a very young area of science and ripe for opportunities and advancement. The age of computers and our ability to compile, quantify, and analyze information is unprecedented.

The basic theory of QSAR is that the structure of a chemical determines its activity [2, 3, 4]. The mystery of chemicals and of chemistry is how structure or substructures are related with activity. Any change in chemical structure (e.g., the addition of a methyl group or element) results in different chemical behavior. It is of great societal interest to predict how chemical activity changes with chemical structure. If we could do so, then more effective drugs can be developed as well as the development of more effective, but safer chemicals for societal use.

The emergence of computers has dramatically increased the use statistics to problems in chemistry. Before the modern computing age, the calculations for many of the statistical procedures were too time-consuming to perform by hand especially for large datasets. In addition, hand calculations are subject to considerable error.

Large datasets in QSAR and in the field of computational chemistry have emerged for both chemical structure and activity. For instance, many software programs are now available such as Molconn-Z [5], Polly [6], DRAGON[7], and CODESSA [8] that calculate measurements of chemical structure. A past limitation existed when chemical activity data were available, but there were few

structure measurements available to allow predictions of those activities. Similarly, in the past the exploration of activity-to-activity correlation approaches used in predictive pharmacology and toxicology failed because experimental activity data were unavailable. As the computer and information age has emerged, there are many additional databases available that are based on standardized endpoints such as toxicology [9, 10], mutagenicity[11], and chemical activities [12].

Today, statistical applications are common in chemistry and they have a variety of names such as chemometrics and pattern recognition. Here our primary goal is to 1) summarize several statistical techniques that have been used extensively in the past 30 years, 2) explore the recent use and potential for Bayesian statistical analysis in QSARs, and 3) provide examples of these statistical techniques in past QSAR studies. This chapter is not intended to be an exhaustive review of all statistical procedures used in QSARs, QSPRs, or other associated analyses on the relationships between chemical structure and their properties or activities.

2 A statistical goal

The databases that house chemical structure and property measurements are ever increasing; a fundamental issue is that there are typically more structure measurements than chemicals being measured. Even as more chemicals are added to the database, scientists will increase the number of ways we can measure them and the issue will remain. If we use p to refer to the number of different measurements taken and n to refer to the number of different chemicals in the database, then this situation is referred to as the large-p, small-n problem.

The QSAR goal discussed here is to use the structure and activities measure-

ments on a set of chemicals to predict the unknown activities of a different set of chemicals for which structure measurements are available. We concentrate on a single chemical activity at a time, although the methods below could be used individually for each activity that requires prediction. Further, we restrict our attention to predicting activities that are continuous, e.g. boiling point, as opposed to properties that are categorical, e.g. mutagenicity. The latter may be analyzed by methods such as logistic regression and discriminant analysis.

Throughout the following we will use the following notation:

- Y: a $n \times 1$ vector of chemical activity measurements and
- X: a $n \times p$ matrix of chemical structure measurements

where, typically, the first column of X is a vector of ones. The ith chemical has property measurement Y_i and structure measurements X_i, the ith row of X. We are then typically interested in predicting the chemical property measurements of a new chemical, Y^*, based on its chemical structure measurements, X^*.

3 Modeling

To predict continuous chemical activities from measurements of chemical structure, we focus on the linear regression model. Although other methods such as generalized additive models and recursive partitioning allow more flexibility, they require a larger sample size, n, which is often not available.

3.1 Multiple linear regression

Multiple linear regression defines a model that has the form $Y = X\beta + \epsilon$, where β is a set of unknown regression parameters and the random deviation has $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2 \mathrm{I}$ where I is an identity matrix of order n. The model says that the activity for chemical i is a linear combination of the structure measurements

and a random error, i.e. $Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i$ where, again, X_{i1} is often set equal to 1 to provide a model intercept. From this relationship, it should be clear that if $\beta_j = 0$, then X_{ij} does not affect the activity and therefore the jth structure measurement is not important for determining the activity in this model. The ordinary least squares (OLS) estimate for the vector parameter β is $\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$ which is found by minimizing the quantity $||Y - X\beta||_2 = (Y - X\beta)'(Y - X\beta)$. To then predict the property for a new chemical, we use $\hat{Y}^* = X^*\hat{\beta}_{OLS}$.

Niemi et al. [13] used multiple regression to develop a prediction model for octanol/water partition coefficient. The independent variables used were 70 variables algorithmically-derived from information content and from molecular connectivity indices [3]. The analysis used a best-subsets (see Section 4.2) regression model for a dataset of over 4,000 chemicals with measured values of octanol/water partition coefficients. Explained variation ranged from 63 to 90% among 14 different groups of chemicals; the groups were formed on the basis of the degree of hydrogen bonding. Both information content and molecular connectivity indices were equally as effective in the prediction equations.

This example of regression uses a combination of simple grouping of a large dataset of over 4,000 chemicals using a theoretical basis that degree of hydrogen bonding is related to octanol/water partitioning. There are a multitude of variations that have been applied to regression analysis in QSAR studies [14] and variations of multivariate techniques combined with regression for making predictions about chemical properties such as partial least squares regression [15, 16]. All of these approaches likely have merit because the chemical universe is diverse and simple changes in chemical structure can have profound changes in chemical properties. Most of these statistical approaches, however, should be viewed as exploratory techniques subject to extensive scrutiny, further exper-

imental testing, and ultimately the development of mechanistic understanding for the relationship between structure and activity.

One area of prediction using regression in QSAR that has received some scrutiny is the statistics of validation. Several authors [17, 18, 19] suggest that many publications present a naive q^2 and provide an improved means to present validation for a predictive relationship. Furthermore, in a comparative study of principal components regression, partial least squares, and ridge regression, ridge regression out-performed the other two [20].

Two important assumptions exist for multiple linear regression that typically make its direct use in QSAR studies dubious. The first is that the number of observations must be larger than the number of structure measurements (small-p, large-n). The second is that the structure measurements are uncorrelated which is questionable when many structure measurements are made. We now introduce two other statistical approaches: principal component regression and ridge regression that are useful for regression analysis when there exists a large-p, small-n problem and when the structure measurements are correlated.

3.2 Principal component regression

Principal component regression (PCR) is a two-step procedure that initially utilizes principal component analysis (PCA) to select principal components and then performs multiple regression using the selected principal components. Principal component analysis (PCA) is a multivariate statistical technique that uses correlations between and among variables to identify new components that are linear combinations of the original variables [21, 22]. PCA is part of a family of statistical procedures (e.g., factor analysis) that are used when there are a large number of variables, many of which are highly correlated. This is often the case with the algorithmically-derived variables used in QSAR such as regression

when collinearity among independent variables violates statistical assumptions. Furthermore, in datasets where the number of independent variables is large relative to the number of chemicals (n) available in the dataset, then spurious correlations can be a problem. A relevant solution is to use a dimension reduction procedure like PCA to reduce the number of independent variables by eliminating pairs of variables that are highly correlated or using the principal components as new uncorrelated, independent variables in the analysis. If the principal components are used, it is often difficult to interpret the results so calculations of the correlations between the original variables and the principal components are useful.

For instance, Basak et al. [23] used PCA for 151 topological indices for a training set of 220 compounds. About 60% of the variation in the 151 indices could be explained by the first principal component and more than 95% of the variation could be explained by the first 12 principal components. This indicated substantial redundancy among the topological indices. PCA allowed the number of independent variables to be reduced to 60 and subsequently used in further analysis of the dataset. In these cases where there are a large number of potential explanatory variables there is no option except to reduce the complexity of the problem by using a dimension reduction procedure like PCA or in combination with regression approaches [24].

Numerous additional examples of this type of procedure exist in the QSAR literature [25, 13, 1]. There are a wide variety of additional dimension reduction procedures available. As the name implies, their purpose is to reduce the dimensionality of the data to the essential and important dimensions. PCA is one of the most common forms and seeks to identify orthogonal factors that are very useful in analysis such as multiple regression that assumes orthogonality among the independent variables. More complex dimension reduction procedures use

various mathematical variations of factoring the independent variables or rotations of the factor axes to increase the interpretation of the resulting variables, e.g., varimax rotation.

This approach to principal components regression where PCA is run first on the independent variables alone followed by regression using top principal components is typically driven by a desire to eliminate multicolinearity in the independent variables. Since the PCA is run without regard to the dependent variable this leads to shortcomings of the PCR methodology. First, there is no reason to believe the top principal components are related to the dependent variable and thus elimination of lower components may eliminate the important relationships. Second, use of principal components as independent variables leads to an analysis that is hard to interpret. Third, PCA is useless in designed experiments since the principal components are determined entirely by the experimental design. To alleviate some of these shortcomings, [26] provides an approach to dimension reduction of the independent variables that generates a *sufficient reduction* of the these variables which depends on the observed dependent variable values.

3.3 Penalized Regression

3.3.1 Ridge regression

Ridge regression (RR) is an alternative option to PCR that does not require eliminating highly collinear structure measurements. The basic idea behind RR is to shrink the OLS regression coefficient estimates toward zero by adding a penalty for large coefficients. Rather than minimizing the quantity $||Y - X\beta||_2$ which results in the OLS estimates, ridge regression minimizes the quantity $||Y - X\beta||_2 + k||\beta||_2$ for a chosen $k \geq 0$. If $k = 0$, the OLS estimate is the RR estimate and no shrinkage is observed. In contrast, for $k > 0$ the RR estimate

is $\hat{\beta}_{RR} = (X'X + kI)^{-1}X'Y$ and as k increases the estimates for β get closer and closer to zero [24]. For predicting the activity of a new chemical, the OLS estimate is replaced with the RR estimate, i.e. $\hat{Y}^* = X^*\hat{\beta}_{RR}$. The choice of k is left to Section 4.

Many articles have utilized ridge regression for dealing with the large-p, small-n problem in the QSAR literature [24, 27, 17, 28, 29]. In particular, [30] used ridge regression to determine whether biodescriptors provide additional information over chemodescriptors in predicting eight toxicity measures in 14 halocarbons. The biodescriptors, which were obtained by exposing the halocarbons to hepatocytes and producing a two-dimensional electrophoresis gel, were found to provide additional information over the use of chemodescriptors alone.

Although ridge regression is gaining popularity much is still unknown about its theoretical properties in the $p >> n$ situation. For example, are the ridge regression estimators consistent, i.e. do they recover the truth as the number of observations increases? The difficulty here is that to ensure $p >> n$ when the number of observations increases, the number of independent variables must also increase.

3.3.2 LASSO

Ridge regression is one specific type of *regularized regression* which also includes LASSO (least absolute shrinkage and selection operator) [31, 32] and the elastic net [33]. LASSO minimizes the quantity $||Y - X\beta||_2 + k||\beta||_1$ where $||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$. So whereas ridge regression penalizes the square of the deviation of β from zero, LASSO penalizes the absolute value of the deviation from zero. The adaptive LASSO improves on the original by allowing adaptively determined weights for penalizing individual coefficients [34]. Group LASSO is an extension to LASSO for predefined groups of independent variables that are included or removed as a whole [35].

3.3.3 Elastic net

Unlike ridge regression, LASSO cannot select more structure measurements than observations and therefore [18] suggest it may not be appropriate for use in QSAR studies. The elastic net penalizes both the square and the absolute deviation from zero and therefore is somewhere between ridge regression and LASSO. The elastic net is a promising approach for selecting important structure measurements while still retaining predictive ability [18].

3.3.4 Additional penalized regression approaches

Other statistical techniques are available for dealing with the large-p, small-n problem, particularly PLS (partial least squares/projection to latent spaces) [36, 37]. The QSAR literature appears to be favoring the use of RR [24, 38], although at least one has suggested that using PCR and RR together is preferable [39]. More recent work has generalized LASSO for use in the $p >> n$ situation by combining a Bayesian regression approach with a loss function to set some coefficients to zero [40]. Another option that is closely related to LASSO and RR, is the horseshoe [41].

3.4 Clustering Techniques

In the modeling discussed above, we have implicitly assumed that all chemicals being analyzed are equally described by the one model that is chosen. Given the heterogeneity in chemical structures and activities, it is intuitive that certain chemical groups would follow one model while another would follow a quite different model. Therefore it seems reasonable to cluster chemicals into groups with similar structures. Statistical cluster analysis encompasses many different algorithms and methods for grouping objects, e.g., chemicals, of similar type into respective groups. In QSAR applications there are situations where the

chemical database may be relatively large and contain compounds of many different types, e.g., halogens, phenols, alkanes, etc. It may be difficult to find a statistical model that will produce satisfactory results when a database contains chemicals of many different types [25] or different modes of action [23]. Cluster analysis can be useful to *a priori* group chemicals into similar groups based on chemical structures or activities. Individual prediction models within a cluster can then subsequently be developed.

A common clustering technique is k-means clustering in which the user can determine *a priori* the number of clusters or one can iterate the analysis to determine an optimal number of clusters in an exploration of a dataset (see Section 4). Niemi et al. [42] used k-means clustering to explore the persistence or degradation of 287 chemicals tested with the standard biochemical oxygen demand (BOD) procedure. The 287 chemicals were derived from an extensive literature search of available BOD values, plus scrutiny of the quality of the BOD procedure used. The dataset was diverse and consisted of a wide variety of chemical groups, e.g., halogens, aldehydes, hydrocarbons, acids, and sulfonates. Fifty-four molecular connectivity indices were calculated and five chemical properties were either available or estimated. To reduce the dimensionality, PCA was used and resulted in eight principal components that explained more than 94% of the variation in the original data. The eight principal components were calculated in a k-means clustering algorithm that was iterated many times to identify an optimum number of clusters that provided the best discrimination of biodegradable and persistent chemicals. Once the analyses were completed, a series of structural features were identified that were associated with degradable and persistent chemicals. The overall model correctly classified 85% of the degradable chemicals and 94% of the persistent chemicals. In addition, several chemicals that were misclassified as degradable or persistent were retested. In

many cases, retesting of the chemicals indicated that the biodegradability model was correct and the original biodegradability test values were erroneous.

4 Model selection

It is often the goal of an analysis to choose one final model based on the data at hand. The choices to arrive at this final model are extensive including which structure measurements to include in the model, how many principal components to include, what the ridge regression penalty should be, and how to cluster chemicals. Here we discuss a number of statistical tools used to compare models for the ultimate goal of selecting one model for prediction purposes.

4.1 F-test

In some cases, the models under consideration are *nested*. Model A is nested in model B if the parameters in model B can be set to particular values to recover model A. Consider the two regression models:

$$\text{A: } Y_i = X_{i1}\beta_1 + \qquad \epsilon_i \qquad \text{and}$$
$$\text{B: } Y_i = X_{i1}\beta_1 + X_{i2}\beta_2 + \epsilon_i.$$

Model A is nested in model B since setting $\beta_2 = 0$ recovers model A. Often we are interested in determining whether model A or model B is preferable. Model B will always fit the data better than model A since it has an additional parameter. Unfortunately adding this additional parameter may simply fit noise and therefore harm our predictions. Therefore we need a way to distinguish when the model is fitting noise and when it is fitting signal.

A formal approach to compare two nested models is an F-test. This test determines whether the larger model is a statistically significant improvement

over the smaller model. If it is, this suggests the additional parameters involved in creating the larger model are likely modeling signal rather than noise.

This is the approach used in [24] where the structures are grouped into categories: topostructural (TSI) , topochemical (TCI), geometrical (3D), and semiempirical quantum chemical (QC) variables. Models were tested in hierarchical lists where TSI was added first followed by TCI, then 3D, and finally QC. The analysis showed that in most of the datasets incorporating all four categories provided the best model, each category provided a statistically significant improvement over the smaller model.

4.2 Akaike/Bayesian information criterion

Often we are not interested simply in nested models. Consider a simple example where there are two predictor variables and we consider the four models consisting of every combination of variable inclusions. Then the model that has only the first variable is not nested in the model that has only the second variable rendering the F-test ineffective. The most common approach to determining which variables to include is to use either Akaike Information Criterion (AIC) [43] or Bayesian Information Criterion (BIC) [44].

Both of these criteria put penalties on the number of parameters in a model and thereby encourage model parsimony. If the number of models is small enough, then the criterion can be computed for all models and the model with the best criterion, called the *best subsets* model, can be chosen [13]. Typically p is too large to enumerate all models in a reasonable amount of time and then the criterion is combined with a stepwise selection procedure to find a reasonable model [45, 46], but no guarantee is made that this procedure finds the best subset.

4.3 Cross-validation

The F-test, AIC, and BIC are useful tools when we are interested in which structure measurements to include, but these tools are not useful in determining the number of principal components to use, the ridge regression parameter, or how many clusters to use. A good approach for these choices is cross-validation [17].

Although many variants of cross-validation exists, we only describe leave-one-out cross-validation. This approach calculates the prediction sum of squares (PRESS) for each candidate model, e.g. each number of principal components. PRESS is calculated according to the following procedure:

1. For $i = 1, \ldots, n$
   (a) Fit the model while leaving out chemical i
   (b) Predict the property of chemical i, $\hat{Y}_i$
   (c) Calculate the squared prediction error for chemical i, $(Y_i - \hat{Y}_i)^2$
2. Sum all the squared prediction errors (PRESS)

The candidate model with lowest PRESS is chosen.

An alternative to this cross-validation approach is to separate the dataset into two groups: the training and hold-out testing data. All candidate models are fit using the training data and then a model is chosen based on performance among the testing data. Although computationally faster than cross-validation, this hold-out testing approach is only reliable when both the training and testing data are numerous [47]. Due to the small sample sizes typically available in QSAR studies, the cross-validation approach described here will be more reliable and less wasteful than a hold-out approach [17, 47].

5 Model averaging

The previous section outlined methodology for selecting **one** best model and assuming it is the true model to make predictions. But, as George Box once wrote [48]:

> *All models are wrong, but some are useful.*

It is often useful to interpret the best model to suggest a mechanism that describes the property being analyzed, but we should not pretend that this model is the truth. For the purposes of making a prediction, we should instead acknowledge our uncertainty about model truth and account for that uncertainty. This is exactly what model averaging does.

Suppose we consider a total of J models, e.g. if we have 10 structure measurements then we could consider the set of $J = 2^{10} = 1024$ models that includes all combinations of those measurements. Now suppose that our prediction for Y^* based on X^* from each model j is $\hat{Y}_j^*$, then the *model averaged prediction* is $\sum_{j=1}^{J} w_j \hat{Y}_j^*$ where w_j are model weights such that $\sum_{j=1}^{J} w_j = 1$.

One approach to determining these weights is to use the AIC values for each model [49]. Let AIC_j be the AIC value for model j, AIC_{min} be the minimum AIC among the J models, and $\Delta AIC_j = AIC_j - AIC_{min}$. Then the *Akaike weight* for each model is

$$w_j = \frac{e^{-\Delta AIC_j/2}}{\sum_{i=1}^{J} e^{-\Delta AIC_j/2}}.$$

A difficulty with the use of model averaging in practice is the number of possible models. If p is in the hundreds and we consider the models consisting of all combinations of predictors being in the model, then we have as many models as atoms in the universe, $10^{80} \approx 2^{266}$. It is infeasible to estimate the parameters, predict new values, and calculate the weight for all models. Fortunately, we can

approximate the model averaged prediction if we can find the models with large weight, w_j. Methods, such as shotgun stochastic search [50], are currently being developed to efficiently find these large weight models.

In [51], AIC model averaging was used in conjunction with PLS, PCR, and cross-validation to determine the key biological predictors responsible for generating a specific cytokine response. A specific difficulty for their study was the use of time-course measurements which provide a profile of ligand-induced changes in protein phosphorylation state and cytokine output response in macrophage-like RAW 264.7 cells. These time-course measurements are highly correlated and therefore when used as predictors can severely violate independence assumptions. Through the use of model averaging and variable selection techniques, the authors were able to relax this assumption and provide both a predictive and, possibly, mechanistic understanding of the cytokine reponse.

6 Bayesian statistics

The use of Bayesian statistics is increasing all fields of science including QSAR studies. An appealing aspect of Bayesian statistics is the coherence of all methodologies through the use of conditional probability and Bayes' rule [52] as in equation (1)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{1}$$

In all Bayesian analyses, A represents anything we don't know whereas B represents everything we know or assume. For example, B includes the data, e.g. measured chemical activities. In contrast, A represents model parameters, e.g. regression coefficients, or predictions, e.g. unmeasured chemical activities. The goal of a Bayesian analysis is to obtained the *posterior*, $P(A|B)$, based on the information provided in the *prior*, $P(A)$, the statistical model, $P(B|A)$, and

the *marginal likelihood*, $P(B)$. In this way, we can view the Bayesian approach as a formal mathematical tool to move from the information we have before an experiment is observed, i.e. the prior, to the information we have after an experiment is concluded, i.e. the posterior.

The interpretation of a Bayesian analysis is much different from the interpretation of a frequentist analysis. For example, in the model selection context a frequentist produces a p-value where a Bayesian produces a posterior model probability. The interpretation of the p-value is *the probability of observing a test statistic as or more extreme than that observed, if the null hypothesis is true* while a posterior model probability (for the null hypothesis) has the interpretation as *the probability that the null hypothesis is true given the data we observed*. Similarly for parameter uncertainty a frequentist produces a confidence interval where a Bayesian produces a credible interval. The interpretation of a $100(1-\alpha)\%$ confidence interval is *over repeated realizations of the data, the constructed confidence intervals will contain the true parameter* $100(1-\alpha)\%$ *of time time* while a $100(1-\alpha)\%$ credible interval has the interpretation *the probability the true parameter value is in the interval is* $100(1-\alpha)\%$. In both cases, the latter is a more natural interpretation (at least to us), but comes at the cost of requiring a prior distribution for parameters and, for model probabilities, a prior probability for models.

In the rest of this section, we will show the natural connection between previously mentioned techniques, e.g. regression, ridge regression, and model averaging, and Bayesian methods. For a more thorough review of Bayesian background and approaches please see [53, 54].

6.1 Bayesian regression

In the regression problem described in Section 3.1, we are typically interested in estimating the unknown parameters β and σ^2 based on the available data. Equation (2) provides a rewriting of Bayes' rule to utilize the notation previously introduced where lower case ps are now used since we are talking about continuous distributions.

$$p(\beta, \sigma^2|y) = \frac{p(y|\beta, \sigma^2)p(\beta, \sigma^2)}{p(y)} \tag{2}$$

In this statement, $p(y|\beta, \sigma^2)$ represents the regression model $y = X\beta + \epsilon$, $p(\beta, \sigma^2)$ represents prior information available concerning the model parameters, and $p(\beta, \sigma^2|y)$ represents the information available after analyzing the new data. It is common, albeit confusing, to eliminate X and the model itself from the conditional probability statements in equation (2) since neither of these are included in A and B of equation (1).

A convenient computational choice for the prior, $p(\beta, \sigma^2)$, is to choose a normal-inverse gamma prior for β and σ^2, specifically $p(\beta, \sigma^2) = p(\beta|\sigma^2)p(\sigma^2) = N(\beta; \beta_0, \sigma^2\Sigma_0)Ga(\sigma^{-1}; \alpha_0, \beta_0)$ where $N(a; b, c)$ represents a normal distribution for a with mean b and variance matrix c and $Ga(d; e, f)$ represents a gamma distribution with shape e and rate f. If we simultaneously let b, c, and Σ_0^{-1} approach zero, then we obtain the prior $p(\beta, \sigma^2) \propto \sigma^{-2}$ where the proportionality symbol is used to indicate that this is not a proper distribution since it does not integrate to one. Nonetheless, the posterior is a proper distribution and is

$$p(\beta, \sigma^2|y) = N(\beta; \hat{\beta}_{OLS}, \sigma^2(X'X)^{-1})Ga(\sigma^{-2}; n/2, b_n) \tag{3}$$

where $b_n = (y - X\hat{\beta}_{OLS})'y - X\hat{\beta}_{OLS})/2$. Therefore the posterior expectation of β, $E[\beta|y]$, is exactly the same as the ordinary least squares estimate.

6.1.1 Informative priors

For simplicity, assume now that σ^2 is known and we are interested in providing an informative prior for β. A computationally convenient choice will be a normal distribution with mean 0 and variance Σ_0. If we further assume that $\Sigma_0 = \tau^2 \mathrm{I}$, then the posterior expectation for β is $\hat{\beta}_{RR}$, the ridge regression estimate where $k = \sigma^2/\tau^2$ [55].

A computationally less convenient choice is the Laplace [56], also called the double exponential, prior distribution. If this prior, centered at zero, is used, then the posterior expectation for β is a LASSO estimate. If the prior is a mixture of a normal and Laplace prior both centered at zero, then the resulting posterior expectation for β is an elastic net estimate [57].

Rather than strictly providing better parameter estimates, informative priors can also be used to formally incorporate scientific knowledge. This was used in [58] to combine information across multiple experiments to build a predictive model of ligand-receptor binding affinities. It has been suggested that Bayesian regression be further explored for its benefit in decision making [59].

6.2 Bayesian prediction

To predict a new chemical activity from its structure, we use Y^* as unknown while Y is known. Utilizing the rules of probability, we arrive at the following prediction equation:

$$p(Y^*|Y) = \int p(Y^*|\beta, \sigma^2) p(\beta, \sigma^2|Y) d\beta d\sigma^2.$$

This equation describes the entire distribution for our prediction for Y^* which can be helpful in understanding how much uncertainty we have in the predicted point estimate. The point estimate is found by taking the expectation and using

the law of iterated expectations:

$$E[Y^*|Y] = E[E[Y^*|\beta, \sigma^2, Y]] = E[X^*\beta|Y] = X^*\hat{\beta}$$

where $\hat{\beta}$ will be the point estimate for β for the model under consideration, e.g. for ridge regression, it is $\hat{\beta}_{RR}$. Therefore, to obtain a point estimate under the Bayesian approach we have exactly the same two-step process: 1) estimate the parameters in the model and 2) predict the new data point based on those estimates.

6.3 Bayesian model averaging

As discussed in Section 5, there is no reason to presume that the one model we have selected is actually the true model and predictions can be improved if, rather than selecting a single model, all models are entertained as possibilities and our prediction is based on a weighted average over all these models. The Bayesian approach provides a formal derivation of this approach called *Bayesian model averaging* [60, 61] which we outline here.

Using the laws of probability, we have

$$p(Y^*|Y) = \sum_{j=1}^{J} p(Y^*|M_j)P(M_j|Y)$$

where the upper case P is used since this is an actual probability. To find a point estimate for Y^*, we calculate its expectation

$$E[Y^*|Y] = \sum_{j=1}^{J} E[Y^*|M_j]P(M_j|Y).$$

The expectation for each model is calculated according to the previous section, i.e. estimate the parameters and then predict Y^* based on those estimates.

Therefore this approach is exactly consistent with the model averaging approach in Section 5 if we set $w_j = P(M_j|Y)$.

To find the posterior model probability $P(M_j|Y)$, we use Bayes' rule

$$P(M_j|Y) = \frac{p(Y|M_j)P(M_j)}{p(Y)} \tag{4}$$

where $P(M_j)$ is our prior probability for model j, $p(Y|M_j)$ indicates how well our data is described by that model, and $p(Y) = \sum_{i=1}^{p} p(Y|M_j)P(M_j)$ assures that the posterior probability over all models sums to unity. Bayesian model averaging in regression models can be accomplished using the `BMA` package [62] in the statistical software `R` [63].

7 Summary

In this article, we covered the use of linear regression techniques for continuous-valued activites in QSAR and suggested model averaging and Bayesian approaches as possible future directions to extend the use of these techniques. We would be remiss not to mention that there are several other approaches to dealing with the large-p, small-n problem including PLS [36, 37] and Bayesian neural networks [64, 65, 66]. Bayesian neural networks can provide extremely good predictive power under cross-validation scrutiny, but we prefer the interpretability afforded regression models which can lead to mechanistic understanding of how structure affects activity. A number of authors have tried to compare these different methods [67, 68, 69, 70]. We also did not cover the vast field of binary- or categorical-valued activities [71, 72, 73], but even there the idea of Bayesian model averaging has improved predictive power [72].

Statistical analysis and particularly multivariate statistics provide the mathematical chemist with a powerful arsenal of tools to improve our understanding

of SARs. Here we have provided several examples of their applications to problems in QSAR. With the recent emergence of the Internet and outstanding search engines, there is an extensive amount of information describing and illustrating the use of these statistical techniques. However, it is important to recognize that much of the information has not been peer-reviewed and we urge the reader to seek standard textbooks and the vast peer-reviewed literature that has developed and been accepted by the scientific community. In addition to the Internet, there are many excellent statistical packages that are now available for most of the standard, classical statistical tests such as regression, PCA, and clustering techniques as well as their many variations. Many of the manuals that come with these statistical packages are also well-documented. Exceptions for available software still apply to many of the Bayesian approaches, but this is likely to improve in the future.

In this brief review of some older statistical techniques and some new approaches, we have tried to provide a flavor for how many of the complex problems in SAR can be simplified with the use of multivariate statistics. However, statistics is in itself a vast field of science and we certainly cannot do it justice in a brief review. It is incumbent upon the scientist to clearly articulate the question(s) he/she seeks to address and fully understand the potential statistical techniques that could address the question(s). We strongly encourage the scientist to also seek professional advice from a statistician and include a statistician in team approaches to solving these complex problems in QSAR. Moreover, it is wise to include or consult a statistician in the start of a project rather than expecting one to fix a problem or analyze data after it has been gathered.

References

[1] Basak, S.C. et al. Chemo-bioinformatics based mathematical descriptors and their applications in computational drug design. *Current Computer-Aided Drug Design*, **2010**, *6*(4), 223–224.

[2] Hansch, C. Quantitative structure-activity relationships in drug design. *Drug Design*, **1971**, *1*, 271–342.

[3] Basak, S.C.; Grunwald, G.D.; Niemi, G.J., Use of graph-theoretic and geometrical molecular descriptors in structure-activity relationships, In *From Chemical Topology to Three-dimensional Geometry.* Balabah, Plenum Press, New York **1997**.

[4] Cronin, M.T.D., Quantitative structure-activity relationships (QSARs) applications and methodology, In Puzyn, Tomasz; Leszczynski, Jerzy; Cronin, Mark T., editors, *Recent Advances in QSAR Studies* volume 8 of *Challenges and Advances in Computational Chemistry and Physics* , pp. 3–11. Springer Netherlands **2010**.

[5] eduSoft LC, VAAshland, Molconn-Z 4.10 Manual, http://www.edusoft-lc.com/molconn/manuals/400/ **2011**.

[6] Basak, S.C.; Gieschen, D.P.; Harriss, D.K.; Magnuson, V.R. Physicochemical and topological correlates of the enzymatic acetyltransfer reaction. *Journal of Pharmaceutical Sciences*, **1983**, *72*(8), 934–937.

[7] Todeschini, R.; Consonni, V. DRAGON-software for the calculation of molecular descriptors, version 1.0 for Windows, milano chemometrics and QSAR research group. *Freely available at http://www.vcclab.org/lab/edragon/*, **2000**.

[8] Katritzky, A.R.; Lobanov, V.S.; Karelson, M. CODESSA software. *University of Florida, SemiChem, Shawnee, KS*, **1994**, , p. 211.

[9] Veith, G.D. On the nature, evolution and future of quantitative structure-activity relationships (QSAR) in toxicology. *SAR and QSAR in Environmental Research, 15*, **2004**, *5*(6), 323–330.

[10] Schultz, T.W.; Carlson, R.E.; Cronin, M.T.; Hermens, J.L.; Johnson, R.; O'Brien, P.J.; Roberts, D.W.; Siraki, A.; Wallace, K.D.; Veith, G.D. A conceptual framework for predicting the toxicity of reactive chemicals: modeling soft electrophilicity. *SAR and QSAR in Environmental Research*, **2006**, *17*(4), 413–428.

[11] Basak, S.C.; Mills, D. Prediction of mutagenicity utilizing a hierarchical QSAR approach. *SAR and QSAR in Environmental Research*, **2001**, *12*(6), 481.

[12] Mackay, D., *Handbook of physical-chemical properties and environmental fate for organic chemicals: Introduction and hydrocarbons* volume 1, CRC Press/Taylor & Francis **2006**.

[13] Niemi, G.J.; Basak, S.C.; Grunwald, G.; Veith, G.D. Prediction of octanol/water partition coefficient (KOW) with algorithmically derived variables. *Environmental Toxicology and Chemistry*, **1992**, *11*(7), 893–900.

[14] Yap, C.W.; Li, H.; Ji, Z.L.; Chen, Y.Z. Regression methods for developing QSAR and QSPR models to predict compounds of specific pharmacodynamic, pharmacokinetic and toxicological properties. *Mini reviews in medicinal chemistry*, **2007**, *7*(11), 1097–1107.

[15] Wold, S.; Sjöström, M.; Eriksson, L., *Partial least squares projections to latent structures (PLS) in chemistry*, Wiley Online Library **2006**.

[16] Wold, S.; Eriksson, L.; Kettaneh, N., *PLS in Data Mining and Data Integration*, Springer **2010**.

[17] Hawkins, D.M.; Basak, S.C.; Mills, D. Assessing model fit by cross-validation. *Journal of Chemical Information and Computer Sciences*, **2003**, *43*(2), 579–586.

[18] Kraker, J.J.; Hawkins, D.M.; Basak, S.C.; Natarajan, R.; Mills, D. Quantitative structure-activity relationship (qsar) modeling of juvenile hormone activity: Comparison of validation procedures. *Chemometrics and Iintelligent Laboratory Systems*, **2007**, *87*(1), 33–42.

[19] Basak, S.C.; Mills, D.; Hawkins, D.M.; Kraker, J.J., Proper statistical modeling and validation in QSAR: A case study in the prediction of rat fat-air partitioning, In *AIP Conference Proceedings* volume 963 , p. 548 **2007**.

[20] Basak, S.C.; Mills, D.; Hawkins, D.M. Characterization of dihydrofolate reductases from multiple strains of plasmodium falciparum using mathematical descriptors of their inhibitors. *Chemistry & Biodiversity*, **2011**, *8*(3), 440–453.

[21] Pearson, K. Liii. on lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, **1901**, *2*(11), 559–572.

[22] Jolliffe, I. Principal component analysis. *Encyclopedia of Statistics in Behavioral Science*, **2002**.

[23] Basak, S.C.; Grunwald, G.D.; Host, G.E.; Niemi, G.J.; Bradbury, S.P. A comparative study of molecular similarity, statistical, and neural methods for predicting toxic modes of action. *Environmental Toxicology and Chemistry*, **1998**, *17*(6), 1056–1064.

[24] Hawkins, D.M.; Basak, S.C.; Shi, X. QSAR with few compounds and many features. *Journal of Chemical Information and Computer Sciences*, **2001**, *41*(3), 663–670.

[25] Basak, S.C.; Magnuson, V.R.; Niemi, G.J.; Regal, R.R. Determining structural similarity of chemicals using graph-theoretic indices. *Discrete Applied Mathematics*, **1988**, *19*(1-3), 17 – 44.

[26] Cook, R.D. Fisher lecture: Dimension reduction in regression. *Statistical Science*, **2007**, *22*(1), 1–26.

[27] Basak, SC; Mills, D.; Hawkins, DM; El-Masri, HA Prediction of tissue-air partition coefficients: A comparison of structure-based and property-based methods. *SAR and QSAR in Environmental Research, 13*, **2002**, *7*(8), 649–665.

[28] Hawkins, D.M.; Basak, S.C.; Mills, D. QSARs for chemical mutagens from structure: ridge regression fitting and diagnostics. *Environmental Toxicology and Pharmacology*, **2004**, *16*(1-2), 37–44.

[29] Basak, S.C.; Mills, D.; Hawkins, D.M.; Kraker, J.J. Quantitative structure–activity relationship (QSAR) modeling of human blood: Air partitioning with proper statistical methods and validation. *Chemistry & Biodiversity*, **2009**, *6*(4), 487–502.

[30] Hawkins, D.M.; Basak, S.C.; Kraker, J.; Geiss, K.T.; Witzmann, F.A. Combining chemodescriptors and biodescriptors in quantitative structure-activity relationship modeling. *Journal of chemical information and modeling*, **2006**, *46*(1), 9–16.

[31] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **1996**, *58*, 267–288.

[32] Hastie, T.; Tibshirani, R.; Friedman, J., *The elements of statistical learning: data mining, inference and prediction*, Springer 2 edition **2009**.

[33] Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **2005**, *67*(2), 301–320.

[34] Zou, H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **2006**, *101*(476), 1418–1429.

[35] Meier, L.; Van De Geer, S.; Bühlmann, P. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **2008**, *70*(1), 53–71.

[36] Dunn, W. J.; Wold, S.; Edlund, U.; Hellberg, S.; Gasteiger, J. Multivariate structure-activity relationships between data from a battery of biological tests and an ensemble of structure descriptors: The PLS method. *Quantitative Structure-Activity Relationships*, **1984**, *3*(4), 131–137.

[37] Roy, P.P.; Roy, K. On some aspects of variable selection for partial least squares regression models. *QSAR & Combinatorial Science*, **2008**, *27*(3), 302–313.

[38] Al-Hassan, Y.M.; Al-Kassab, M.M. A Monte Carlo comparison between ridge and principal components regression methods. *Applied Mathematical Sciences*, **2009**, *3*(42), 2085 – 2098.

[39] Vigneau, E.; Devaux, MF; Qannari, EM; Robert, P. Principal component regression, ridge regression and ridge principal component regression in spectroscopy calibration. *Journal of chemometrics*, **1997**, *11*(3), 239–249.

[40] Bondell, Howard; Reich, Brian, Consistent high-dimensional bayesian vari-

able selection via penalized credible regions, Accepted for publication in the Journal of the American Statistical Association.

[41] Carvalho, C.M.; Polson, N.G.; Scott, J.G. The horseshoe estimator for sparse signals. *Biometrika*, **2010**, *97*(2), 465–480.

[42] Niemi, G.J.; Veith, G.D.; Regal, R.R.; Vaishnav, D.D. Structural features associated with degradable and persistent chemicals. *Environmental Toxicology and Chemistry*, **1987**, *6*(7), 515–527.

[43] Akaike, H. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, **1974**, *19*(6), 716–723.

[44] Schwarz, G. Estimating the dimension of a model. *The annals of statistics*, **1978**, *6*(2), 461–464.

[45] Rose, R.M.; St. J. Warne, M.; Lim, R.P. Quantitative structure–activity relationships and volume fraction analysis for nonpolar narcotic chemicals to the australian cladoceran *Ceriodaphnia cf. dubia*. *Archives of Environmental Contamination and Toxicology*, **1998**, *34*(3), 248–252.

[46] Vighi, M.; Migliorati, S.; Monti, G.S. Toxicity on the luminescent bacterium *Vibrio fischeri* (beijerinck). i: QSAR equation for narcotics and polar narcotics. *Ecotoxicology and Environmental Safety*, **2009**, *72*(1), 154–161.

[47] Hawkins, D.M. The problem of overfitting. *Journal of chemical information and computer sciences*, **2004**, *44*(1), 1–12.

[48] Box, G.E.P.; Draper, N.R., *Empirical model-building and response surfaces*, John Wiley & Sons **1987**.

[49] Burnham, K.P.; Anderson, D.R., *Model selection and multimodel inference: a practical information-theoretic approach*, Springer Verlag **2002**.

[50] Hans, C.; Dobra, A.; West, M. Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association*, **2007**, *102*(478), 507–516.

[51] Wu, Y.; Johnson, G.; Gomez, S. Data-driven modeling of cellular stimulation, signaling, and output response in RAW 264.7 cells. *Journal of Molecular Signaling*, **2008**, *3*, 1–14, 10.1186/1750-2187-3-11.

[52] Bayes, M.; Price, M. An Essay towards solving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFRS. *Philosophical Transactions*, **1763**, *53*, 370.

[53] Armstrong, N.; Hibbert, D.B. An introduction to bayesian methods for analyzing chemistry data:: Part 1: An introduction to bayesian theory and methods. *Chemometrics and Intelligent Laboratory Systems*, **2009**, *97*(2), 194–210.

[54] Hibbert, D.B.; Armstrong, N. An introduction to Bayesian methods for analyzing chemistry data: Part II: A review of applications of Bayesian methods in chemistry. *Chemometrics and Intelligent Laboratory Systems*, **2009**, *97*(2), 211–220.

[55] Lindley, D.V.; Smith, A.F.M. Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B (Methodological)*, **1972**, *34*(1), 1–41.

[56] Norton, R.M. The double exponential distribution: Using calculus to find a maximum likelihood estimator. *The American Statistician*, **1984**, *38*(2), 135–136.

[57] Li, Q.; Lin, N. The Bayesian elastic net. *Bayesian Analysis*, **2010**, *5*(1), 151–170.

[58] Murray, C.W.; Auton, T.R.; Eldridge, M.D. Empirical scoring functions. II. the testing of an empirical scoring function for the prediction of ligand-receptor binding affinities and the use of bayesian regression to improve the quality of the model. *Journal of Computer-Aided Molecular Design*, **1998**, *12*, 503–519, 10.1023/A:1008040323669.

[59] Sahlin, U.; Filipsson, M.; Öberg, T. A risk assessment perspective of current practice in characterizing uncertainties in qsar regression predictions. *Molecular Informatics*, **2011**, *30*(6-7), 551–564.

[60] Raftery, A.E.; Madigan, D.; Hoeting, J.A. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, **1997**, *92*(437), 179–191.

[61] Hoeting, J.A.; Madigan, D.; Raftery, A.E.; Volinsky, C.T. Bayesian model averaging: A tutorial. *Statistical Science*, **1999**, *14*(4), 382–401.

[62] Raftery, A.; J.Hoeting; Volinsky, C.; I.Painter; Yeung, K.Y., *BMA: Bayesian Model Averaging* **2011**, R package version 3.14.1.

[63] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing Vienna, Austria **2011**, ISBN 3-900051-07-0.

[64] Frank, R.; Winkler, D.A. A quantitative structure-activity relationships model for the acute toxicity of substituted benzenes to tetrahymena pyriformis using Bayesian-regularized neural networks. *Chemical research in toxicology*, **2000**, *13*(6), 436–440.

[65] Qin, Y.; Deng, H.; Yan, H.; Zhong, R. An accurate nonlinear QSAR model for the antitumor activities of chloroethylnitrosoureas using neural networks. *Journal of Molecular Graphics and Modelling*, **2011**, *29*(6), 826–833.

[66] Jalali-Heravi, M.; Mani-Varnosfaderani, A. QSAR modelling of integrin antagonists using enhanced bayesian regularised genetic neural networks. *SAR and QSAR in Environmental Research*, **2011**, *22*(3-4), 293–314.

[67] Eriksson, L.; Jaworska, J.; Worth, A.P.; Cronin, M.T.D.; McDowell, R.M.; Gramatica, P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs. *Environmental Health Perspectives*, **2003**, *111*(10), 1361.

[68] Yao, X.J.; Panaye, A.; Doucet, J.P.; Zhang, R.S.; Chen, H.F.; Liu, M.C.; Hu, Z.D.; Fan, B.T. Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression. *Journal of chemical information and computer sciences*, **2004**, *44*(4), 1257–1266.

[69] Nandi, Sisir; Vracko, Marjan; Bagchi, Manish C. Anticancer activity of selected phenolic compounds: QSAR studies using ridge regression and neural networks. *Chemical Biology & Drug Design*, **2007**, *70*(5), 424–436.

[70] Basak, S.C.; Mills, D. Quantitative structure-activity relationships for cycloguanil analogs as PfDHFR inhibitors using mathematical molecular descriptors. *SAR and QSAR in Environmental Research, 21*, **2010**, *3*(4), 215–229.

[71] McDowell, R.M.; Jaworska, J.S. Bayesian analysis and inference from QSAR predictive model results. *SAR and QSAR in environmental research*, **2002**, *13*(1), 111–125.

[72] Angelopoulos, N.; Hadjiprocopis, A.; Walkinshaw, M.D. Bayesian model averaging for ligand discovery. *Journal of chemical information and modeling*, **2009**, *49*(6), 1547–1557.

[73] Bender, A. Bayesian methods in virtual screening and chemical biology. *Methods in Molecular Biology*, **2011**, *672*, 175.