Modeling Mutagenicity Status of a Diverse Set of Chemical Compounds by Envelope Methods

Subho Majumdar

Stat 8932 project summary
School of Statistics,University of Minnesota- Twin Cities
e-mail: majum010@umn.edu

1 Introduction

Envelope models [1] provide a flexible setup for dimension reduction in multivariate data analysis. They can work in conjunction with multivariate linear regression to produce estimates with reduced variance, and envelope-reduced data can also be used for class prediction by linear discriminant analysis. In this project we apply envelope methods on a Chemometrics dataset to assess their performance in estimation and prediction.

The data for this project were taken from the CRC Handbook of Identified Carcinogens and Noncarcinogens [2]. The response variable is 0/1 mutagen status obtained from *Ames test of mutagenicity*. The 508 compounds- 256 mutagens and 252 non-mutagens were classified as mutagen (scored 1) if its Ames score exceeded a certain cutoff, non-mutagen (scored 0) otherwise. The chemical compounds in this dataset come from diverse chemical classes, for example, aliphatic alkanes, Monocyclic and polycyclic compounds and Amines.

We have 4 types of descriptors for each chemical compound, namely (1) Topostructural (TS)- define the molecular topology, i.e. connectedness of atoms within a molecule (103 descriptors), (2) Topochemical (TC)- have information on atom and bond types (195 descriptors),(3) 3-dimensional (3D)- define 3-dimensional aspects of the overall molecular structure (3 descriptors) and (4) Quantum-Chemical (QC)- electronic aspects of molecular structure (6 descriptors).

Previous works on this dataset have focused on prediction through ridge regression models [3] as well as variable selection through an iterative algorithm [4]. Here we shall use envelope regression models for estimating the effects of the predictors on mutagenicity, as well as employ discriminant analysis for the purpose of mutagenicity prediction.

2 Methods

2.1 Envelope Regression model

The basic structure of the model [1], for $i = 1, 2, ..., n$, is:

$$\mathbf{Y}_i = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$$

$$\text{with } \boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}^T$$

where $\mathbf{Y} \in \mathbb{R}^{r\times n}$ multivariate response vector, $\mathbf{X} \in \mathbb{R}^{p\times n}$ non-stochastic predictors $\boldsymbol{\alpha} \in \mathbb{R}^r$ is the intercept and $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$ the matrix of regression coefficients, both being unknown. Finally $\boldsymbol{\Gamma} \in \mathbb{R}^{r\times u}, \boldsymbol{\Gamma}_0 \in \mathbb{R}^{r\times(r-u)}$ are the semi-orthogonal basis matrices of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ and its orthogonal complement, respectively, with $\mathcal{B} = \text{span}(\boldsymbol{\beta})$ and $0 \le u \le r$ being the dimension of the envelope. Also $\boldsymbol{\Omega} = \boldsymbol{\Gamma}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^T, \boldsymbol{\Omega}_0 = \boldsymbol{\Gamma}_0\boldsymbol{\Sigma}\boldsymbol{\Gamma}_0^T$ are the coordinate matrices corresponding to $\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0$.

For estimation purposes, we use the 0/1 mutagenicity status as univariate predictor and the data on 308 descriptors as multivariate responses.

2.2 Envelope Linear Discriminant Analysis

To assess the predictive performance of envelope models, we first estimate the envelope basis as per 3.1, say $\hat{\boldsymbol{\Gamma}}$, reduce the matrix of predictors by multiplying it with the basis and then applying Fisher's Linear Discriminant Analysis [5] on $\hat{\boldsymbol{\Gamma}}^T\mathbf{Y}$.

2.3 Tackling rank-deficiency

Because of the rank-deficient nature of the original data envelope methods cannot be applied to the actual variables. Instead we first do Principal Component Analysis on the matrix of predictors $\mathbf{Y}$, take the minimum number of PCs (k) that explain $\ge 90\%$ of the total variance in a loading matrix $\mathbf{L} \in \mathbb{R}^{r\times k}$ and apply the methods in 3.1 and 3.2 on the transformed predictors $\mathbf{L}^T\mathbf{Y}$ instead of $\mathbf{Y}$.

Set of descriptors	No. of PCs	Envelope dim (u)	% var explained by			Envelope gain ratios for		
			PC1	PC2	PC3	PC1	PC2	PC3
TS	7	3	70.43	10.35	2.60	25.91	36.17	2.10
TC	8	4	75.89	6.52	2.42	15.40	35.26	1.00
TS + TC	13	6	70.27	7.94	2.21	10.40	37.99	1.22
Full	15	11	58.19	7.60	5.98	1.00	1.00	1.00

Table 1. Summary of envelope models for different sets of descriptors

For the envelope regression model, we get back the original coefficient estimates and their standard errors by back-transformation on their principal components counterparts. If $\mathbf{b} \in \mathbb{R}^k$ is the envelope estimate of coefficients of k principal components and $\hat{\nu}_1, ..., \hat{\nu}_k$ their standard errors, then $\mathbf{Lb}$ gives coefficient estimates in the original scale, and $\sum_{i=1}^{k} l_{ji}^2 \hat{\nu}_i^2$ the variance of the j^{th} coefficient, with $j = 1, 2, ..., r$.

3 Summary of results

All analyses were done on MATLAB version R2010a [6]. A hierarchical approach is taken for building the model, first using only TS and TC descriptors to build envelope models, then using TS+TC and finally the full set of descriptors.

The first 3 models in Table 1 show significant gains due to envelopes, especially for the first 2 principal components. For the full set of predictors, the iterative algorithm did not converge for the default tolerance values of the objective and gradient functions, and the results here are from a model obtained with higher tolerance values.

Significance of individual predictors were obtained for each model by t-ratios. For the 1-variable models, linearly correlated predictors tend to be simultaneously significant, but this behavior is is less observed in the combined model (TS+TC).

Prediction using envelope LDA was done by leave-one-out cross-validation. While doing cross-validation here it is imperative in each step to first separate the holdout sample, then use other samples to predict the envelope dimension (u) each time and use that u to build predictive model for that holdout sample. Compared to the previous two studies [3][4], the envelope LDA performed worse for mutagen class prediction of the compounds.

4 Conclusion

Although the efficacy of envelope methods has been demonstrated for estimation purposes, its performance in prediction is not good. Possible reasons for this are high ratio of material to immaterial variation, heteroskedasticity caused by diverse chemical classes among compounds and variation of scales between different types of variables. A more detailed formulation of the envelope model keeping these issues in mind should improve the predictive performance. Also, a logistic envelope regression model could be a new approach towards estimation and prediction.

Acknowledgment

I thank Prof. Dennis Cook for his guidance and valuable inputs throughout the project and Xin Zhang for sharing his MATLAB codes for logistic envelope analysis.

References

1. Cook, R.D.; Li B.; Chiaromonte F. Envelope models for parsimonious and efficient Multivariate Linear Regression. *Stat. Sinica*, **2010**, 20, 927-1010.
2. Soderman, J.V. *CRC Handbook of Identified Carcinogens and Noncarcinogens: Carcinogenicity-Mutagenicity Database*, CRC Press: Boca Raton, FL, **1982**.
3. Hawkins, D.M.; Basak, S.C.; Mills, D. QSARs for chemical mutagens from structure: ridge regression fitting and diagnostics. *Environ. Toxicol. Pharmacol.*, **2004**, 16, 37-44.
4. Majumdar S.; Basak S.C.; Grunwald G.D. Adapting Interrelated Two-Way Clustering Method for Quantitative Structure-Activity Relationship (QSAR) Modeling of Mutagenicity/ Non-Mutagenicity of a Diverse Set of Chemicals. *Curr. Comput. Aided Drug Des.*, **2013**, 9, 000-000.
5. Fisher, R.A. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eug.*, **1936**, 7, 179-188.
6. Mathworks Inc. MATLAB Version 7.10 (R2010a), **2010**.