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

Subho Majumdar

School of Statistics, University of Minnesota

Table of contents

Outline

The data

- The data were taken from the CRC Handbook of Identified Carcinogens and Non-carcinogens.
- The response variable is 0/1 mutagen status obtained from *Ames test of mutagenicity*. A chemical compound was classified as mutagen (scored 1) if its Ames score exceeded a certain cutoff, non-mutagen (scored 0) otherwise.
- Total 508 compounds- 256 mutagens and 252 non-mutagens.
- The dataset is diverse, meaning that chemical compounds belong to different chemical classes, some fairly different from each other, like Alkanes and Amines.

Chemical Class	Number of Compounds
Aliphatic alkanes, alkenes, alkynes	124
Monocyclic compounds	260
Monocyclic carbocycles	186
Monocyclic heterocycles	74
Polycyclic compounds	192
Polycyclic carbocycles	119
Polycyclic heterocycles	73
Nitro compounds	47
Nitroso compounds	30
Alkyl halides	55
Alcohols, thiols	93
Ethers, sulfides	38
Ketones, ketenes, imines, quinones	39
Carboxylic acids, peroxy acids	34
Esters, lactones	34
Amides, imides, lactams	36
Carbamates, ureas, thioureas, guanidines	41
Amines, hydroxylamines	143
Hydrazines, hydrazides, hydrazones, traizines	55
Oxygenated sulfur and phosphorus	53
Epoxides, peroxides, aziridines	25

Description of variables

Four types of variables:

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).
4. **Quantum-Chemical (QC)**- Electronic aspects of molecular structure (6 descriptors).

Previous work

- Use of **Ridge Regression** to build a predictive model of mutagenicity (*Hawkins et al, 2004*). The 0/1 mutagenicity score was used as response variable since 1 corresponds to a higher mutagenicity score and 0 corresponds to a lower one.

- **Variable selection** on a larger set of predictors by adapting a supervised clustering algorithm previously used on high-dimensional genetic data (*Majumdar et al, 2013*).

Outline

The Envelope regression model (*Cook, Li and Chiaromonte, 2010*)

$$\mathbf{Y}_i = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \text{ with } \boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}^T$$
$$i = 1, 2, ..., n$$

- $\mathbf{Y} \in \mathbb{R}^{r \times n}$ multivariate response vector, $\mathbf{X} \in \mathbb{R}^{p \times n}$ *non-stochastic predictors*.
- $\boldsymbol{\alpha} \in \mathbb{R}^r$ intercept, $\boldsymbol{\beta} \in \mathbb{R}^{r \times p}$ matrix of regression coefficients: both unknown.
- $\boldsymbol{\Gamma} \in \mathbb{R}^{r \times u}, \boldsymbol{\Gamma}_0 \in \mathbb{R}^{r \times (r-u)}$ semi-orthogonal basis matrices of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ and its orthogonal complement, respectively, with $\mathcal{B} = \text{span}(\boldsymbol{\beta})$ and $0 \leq u \leq r$ being the dimension of the envelope.
- $\boldsymbol{\Omega} = \boldsymbol{\Gamma}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^T, \Omega_0 = \boldsymbol{\Gamma}_0\boldsymbol{\Sigma}\boldsymbol{\Gamma}_0^T$ coordinate matrices corresponding to $\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0$.

Enveloping the predictors

- *Stochastic predictors*, distributed as i.i.d. $N(\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Sigma}_{\mathbf{X}})$.
- $\boldsymbol{\Sigma}_{\mathbf{X}} = \boldsymbol{\Phi}\boldsymbol{\Omega}\boldsymbol{\Phi}^T + \boldsymbol{\Phi}_0\boldsymbol{\Omega}_0\boldsymbol{\Phi}_0^T$, with $\boldsymbol{\Phi} \in \mathbb{R}^{p\times q}, \boldsymbol{\Phi}_0 \in \mathbb{R}^{p\times(p-q)}, 0 \leq q \leq p$ being semi-orthogonal basis matrices of the envelope of span($\boldsymbol{\beta}^T$) and its orthogonal complement, respectively.
- $\boldsymbol{\Omega} = \boldsymbol{\Phi}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Phi}^T, \Omega_0 = \boldsymbol{\Phi}_0\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Phi}_0^T$ coordinate matrices corresponding to $\boldsymbol{\Phi}, \boldsymbol{\Phi}_0$.

Envelope regression model for our data: basics

- log-transformed data.
- the predictors taken as multivariate response, and the 0/1 mutagenicity status taken as the single predictor, and then envelope regression models are obtained.
- Hierarchical approach to observe the effect of adding different classes of predictors: separate envelope models fit on data with only TS, only TC, TC + TS and full set of predictors.

Envelope regression model for our data: rank-deficiency

- Even after log-transformation the data remains rank-deficient in nature, so envelope models cannot be fit on the actual data.
- For each set of variables, at first Principal Component Analysis is done on the data-matrix, and then envelopes are fit on the minimum number of principal components that explain $\geq 90\%$ of the total variation.
- The coefficient estimates for original variables and their asymptotic standard errors are then obtained from their envelope counterparts by back-transformation.

Getting back coefficient estimates and standard errors

Suppose $\mathbf{Y} \in \mathbb{R}^{r \times n}$ is the original matrix of multivariate responses, and $\mathbf{L} \in \mathbb{R}^{r \times k}$ is the PC loading matrix. Thus the transformed predictors are $\mathbf{L}^T\mathbf{Y}$. $\mathbf{b} \in \mathbb{R}^{k \times 1}$ is the envelope estimator of coefficients of PCs in our case, so that for $i = 1, ..., k$

$$b_i \sim N(\mathbf{l}_i^T \beta, \nu_i^2)$$

with ν_i^2 being the i-th diagonal element of $\mathbf{L\Sigma L}^T$, and $\mathbf{l}_i$ the i-th column of $\mathbf{L}$.

Thus the vector $\mathbf{Lb}$ gives the estimate of coefficients of original variables, and $\sum_{i=1}^{k} l_{ji}^2 \hat{\nu}_i^2$ estimates the variance of its j-th coordinate, $j = 1, ..., r$.

Results: Variance reduction by envelopes

In all the envelope models, there were massive gains in terms of variation. The gains were especially large for the first 2 principal components.

Set of descriptors	No. of PCs	Envelope dimension (u)	% variance 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

Note:

- With default tolerances of objective and gradient function in `env` the algorithm did not converge in 1000 iterations. For this reason they were set to `1e-7` and `1e-4`.
- As far as other PCs of full model were concerned, PCs 9, 11, 13 and 15 gave 1.26, 1.96, 1.88 and 1.5-fold gains, respectively.

Results: significant predictors

Set of descriptors	No. of descriptors		Significant PCs
	Total	Significant	
TS	103	51	4, 5, 7
TC	195	98	1, 3, 8
TS + TC	298	89	4, 7, 11
Full	307	56	2, 3, 8, 10, 14

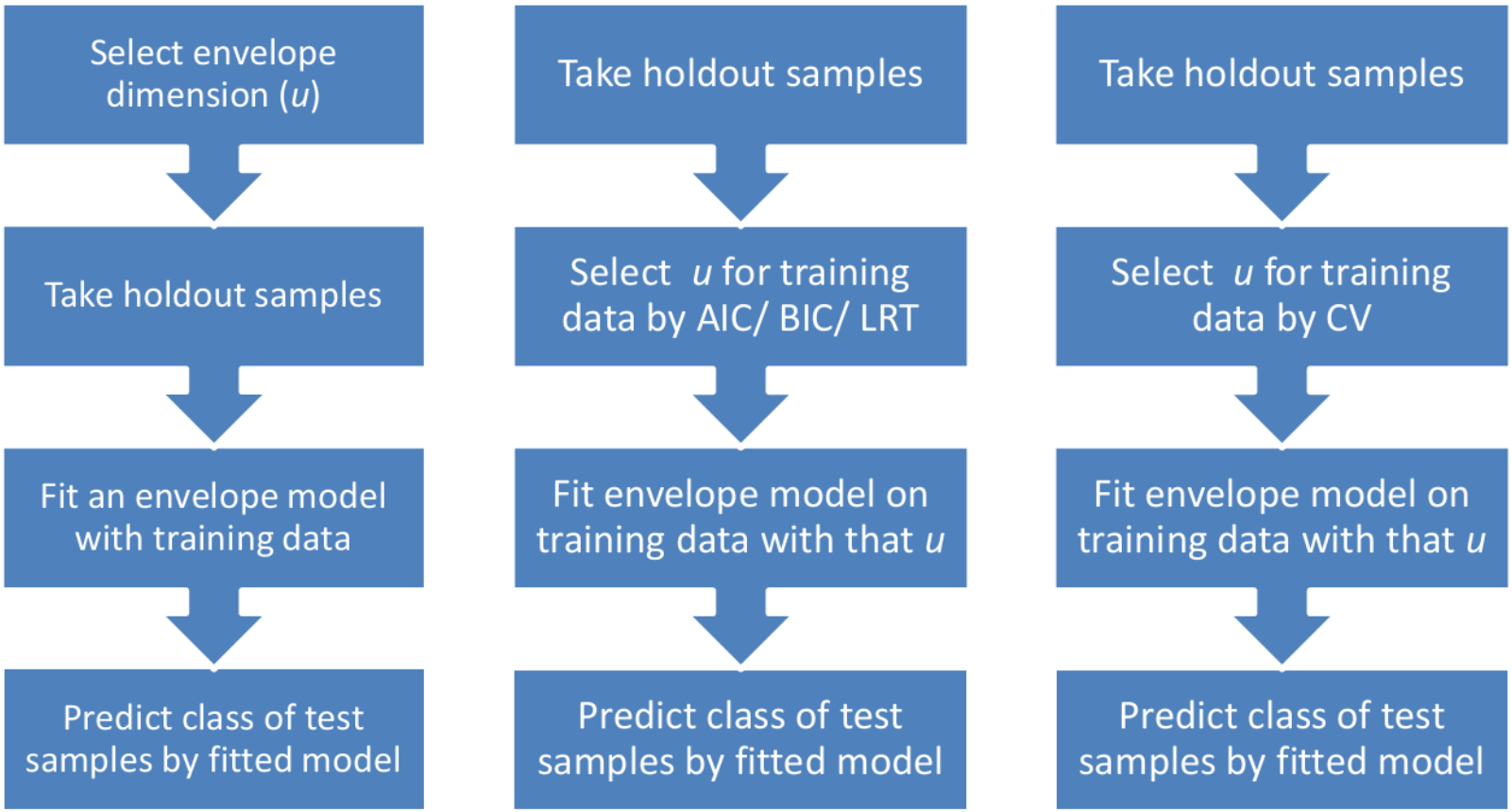
- For envelope models on only TS or only TC data, correlated predictors have a tendency to all have significant t-ratios.
- When the two types of variables are combined, there is less linear dependency among significant predictors.

Outline

Procedure: Envelope Linear Discriminant Analysis

1. Use the envelope basis matrix $\hat{\mathbf{\Gamma}}$ to reduce the predictors.
2. Then use the transformed predictors $\hat{\mathbf{\Gamma}}^T \mathbf{L}^T \mathbf{Y}$ to do Linear Discriminant Analysis, and use that rule to predict class of new samples.
3. Correct classification percentages are obtained through cross-validation on the full sample.
4. Leave-one-out CV in place of k-fold as predictions can vary across different samples of folds because of diverse nature of the dataset.

Naïve CV vs. Two-fold CV



Performance of envelopes in prediction

Model description	Type of predictors in model	No. of predictors	Correct classification %		
			Total	Mutagens	Non-mutagens
Ridge regression	TS+TC	298	76.97	83.98	69.84
Ridge regression	TS+TC+3D+QC	307	77.17	84.38	69.84
Ridge regression after variable selection	TS+TC+AP	203	78.35	84.38	72.22
Envelope LDA	TS	103	57.09	65.63	48.41
	TC	195	58.27	69.92	46.43
	TS+TC	298	60.24	69.14	51.19

Outline

Discussion

- For estimation, envelope models performed really well in conjunction with PCA for rank-deficient data, offering heavy gains for the major principal components over OLS.
- Possible reason for the poor performance in prediction:
  - High material to immaterial variation ratio
  - Heteroskedasticity caused by diverse chemical classes among compounds
  - Variation of scales between different types of variables
- **3D plot of PCs in MATLAB**
- A more detailed formulation should improve the predictive performance of envelope models.
- Logistic Envelope Regression.

Acknowledgements

- Prof. Dennis Cook, for his guidance and valuable inputs.
- Henry Zhang, for providing his codes for logistic envelope regression.
- Greg Grunwald, UofM-Duluth for providing the dataset.

References

Cook, R.D.; Li B.; Chiaromonte F. Envelope models for parsimonious and efficient Multivariate Linear Regression. *Stat. Sinica*, **2010**, 20, 927-1010.

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.

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.

THANK YOU!