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

Subho Majumdar

School of Statistics, University of Minnesota
Motivation

- Predictive analysis of data in Chemistry
- Generation of *in silico* models to predict activities of chemical compounds
- Application in drug development to reduce cost of manufacturing derivatives of chemicals
- **Specific problem** Binary class prediction in heterogeneous multivariate data (e.g. mutagen/ non-mutagen, curative effect of drug): *dimension reduction*
1 The data and the variables
2 The models
3 Results
4 Conclusion
1. The data and the variables

2. The models

3. Results

4. Conclusion
The data were taken from the CRC Handbook of Identified Carcinogens and Non-carcinogens [5].

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 fairly different from each other, like Alkanes and Amines.
<table>
<thead>
<tr>
<th>Chemical Class</th>
<th>Number of Compounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aliphatic alkanes, alkenes, alkynes</td>
<td>124</td>
</tr>
<tr>
<td>Monocyclic compounds</td>
<td>260</td>
</tr>
<tr>
<td>Monocyclic carbocycles</td>
<td>186</td>
</tr>
<tr>
<td>Monocyclic heterocycles</td>
<td>74</td>
</tr>
<tr>
<td>Polycyclic compounds</td>
<td>192</td>
</tr>
<tr>
<td>Polycyclic carbocycles</td>
<td>119</td>
</tr>
<tr>
<td>Polycyclic heterocycles</td>
<td>73</td>
</tr>
<tr>
<td>Nitro compounds</td>
<td>47</td>
</tr>
<tr>
<td>Nitroso compounds</td>
<td>30</td>
</tr>
<tr>
<td>Alkyl halides</td>
<td>55</td>
</tr>
<tr>
<td>Alcohols, thiols</td>
<td>93</td>
</tr>
<tr>
<td>Ethers, sulfides</td>
<td>38</td>
</tr>
<tr>
<td>Ketones, ketenes, imines, quinones</td>
<td>39</td>
</tr>
<tr>
<td>Carboxylic acids, peroxy acids</td>
<td>34</td>
</tr>
<tr>
<td>Esters, lactones</td>
<td>34</td>
</tr>
<tr>
<td>Amides, imides, lactams</td>
<td>36</td>
</tr>
<tr>
<td>Carbamates, ureas, thioureas, guanidines</td>
<td>41</td>
</tr>
<tr>
<td>Amines, hydroxylamines</td>
<td>143</td>
</tr>
<tr>
<td>Hydrazines, hydrazides, hydrazones, triazines</td>
<td>55</td>
</tr>
<tr>
<td>Oxygenated sulfur and phosphorus</td>
<td>53</td>
</tr>
<tr>
<td>Epoxides, peroxides, aziridines</td>
<td>25</td>
</tr>
</tbody>
</table>
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).

Previous work

- Use of **Ridge Regression** to build a predictive model of mutagenicity [2]. 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 [4].
Outline

1. The data and the variables
2. The models
3. Results
4. Conclusion
Envelope regression model

\[ Y_i = \alpha + \beta X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \Sigma) \text{ with } \Sigma = \Gamma \Omega \Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T \]

\[ i = 1, 2, \ldots, n \]

- Due to Cook, Li and Chiaromonte, 2010 [1].

- \( Y \in \mathbb{R}^{r \times n} \) multivariate response vector, \( X \in \mathbb{R}^{p \times n} \) non-stochastic predictors.

- \( \alpha \in \mathbb{R}^r \) intercept, \( \beta \in \mathbb{R}^{r \times p} \) matrix of regression coefficients: both unknown.

- \( \Gamma \in \mathbb{R}^{r \times u}, \Gamma_0 \in \mathbb{R}^{r \times (r-u)} \) semi-orthogonal basis matrices of \( E_\Sigma(B) \) and its orthogonal complement, respectively, with \( B = \text{span}(\beta) \) and \( 0 \leq u \leq r \) being the dimension of the envelope.

- \( \Omega = \Gamma \Sigma \Gamma^T, \Omega_0 = \Gamma_0 \Sigma \Gamma_0^T \) coordinate matrices corresponding to \( \Gamma, \Gamma_0 \).
Graphical illustration of envelope model

(Source: Stat 8932 class notes, R. Dennis Cook)
log-transformed data.

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.

Data rank deficient, so PCA was performed on data and envelope model was built on first few PCs that explained 90% (or 95%) of total variation.
Supervised Singular-Value Decomposition (SupSVD)

\[ \mathbf{X} = \mathbf{YBV}^T + \mathbf{FV}^T + \mathbf{E} \]

- Due to Li et al, 2014 [3].

- Matrix of predictors \( \mathbf{X} \in \mathbb{R}^{n \times p} \), supervision data matrix \( \mathbf{Y} \in \mathbb{R}^{n \times r} \).

- \( \mathbf{B} \in \mathbb{R}^{r \times q} \) is the multivariate matrix of coefficients, \( \mathbf{V} \in \mathbb{R}^{p \times q} \) full-rank loading matrix.

- \( 0 \leq q \leq r \) the dimension of the underlying space of latent parameters, and \( \mathbf{F} \sim \mathcal{N}_q(\mathbf{0}, \Sigma_f) \), \( \mathbf{E} \sim \mathcal{N}_p(\mathbf{0}, \sigma_e^2 \mathbf{I}_p) \) are random error matrices s.t. \( \Sigma = \mathbf{V} \Sigma_f \mathbf{V}^T + \sigma_e^2 \mathbf{I}_p \).

- A modified EM algorithm is used to obtain the unknown parameters \( \theta = (\mathbf{B}, \mathbf{V}, \Sigma_f, \sigma_e^2) \).

- The vector of mutagenicity status is now used as the supervision data matrix \( \mathbf{Y} \), while the data on 308 predictors is the matrix \( \mathbf{X} \).
- **Envelope model** - Estimate the envelope basis, say $\hat{\Gamma}$, reduce the matrix of predictors by multiplying it with the basis and then apply Fisher’s Linear Discriminant Analysis on $\hat{\Gamma}^T Y$.

- **supSVD** - Here the notations are reversed and $X$ is our $508 \times 307$ data matrix. After obtaining the loading matrix $V$, we transform the data matrix as: $U = XV$, and apply LDA on $U$, taking $Y$ as the 0/1 class variable.

- Correct classification percentages are obtained through cross-validation on the full sample.
Method of Cross-validation

**Naïve CV vs. Two-fold CV**

- **Naïve CV**
  - Select envelope dimension ($u$)
  - Take holdout samples
  - Fit an envelope model with training data
  - Predict class of test samples by fitted model

- **Two-fold CV**
  - Take holdout samples
  - Select $u$ for training data by AIC/ BIC/ LRT
  - Fit envelope model on training data with that $u$
  - Predict class of test samples by fitted model

- Two-fold CV is better than naïve CV.
In all the envelope models, there were massive gains in terms of variation. The gains were especially large for the first 2 principal components.

<table>
<thead>
<tr>
<th>Set of descriptors</th>
<th>No. of PCs</th>
<th>Envelope dimension (u)</th>
<th>% variance explained by</th>
<th>Envelope gain ratios for</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>PC1</td>
<td>PC2</td>
</tr>
<tr>
<td>TS</td>
<td>7</td>
<td>3</td>
<td>70.43</td>
<td>10.35</td>
</tr>
<tr>
<td>TC</td>
<td>8</td>
<td>4</td>
<td>75.89</td>
<td>6.52</td>
</tr>
<tr>
<td>TS + TC</td>
<td>13</td>
<td>6</td>
<td>70.27</td>
<td>7.94</td>
</tr>
<tr>
<td>Full</td>
<td>15</td>
<td>11</td>
<td>58.19</td>
<td>7.60</td>
</tr>
</tbody>
</table>

**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.
### Table: Comparison of predictive performance of various models

<table>
<thead>
<tr>
<th>Model description</th>
<th>Type of predictors in model</th>
<th>No. of predictors</th>
<th>Correct classification %</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total</td>
<td>Mutagens</td>
<td>Non-mutagens</td>
</tr>
<tr>
<td>Ridge regression[2]</td>
<td>TS+TC</td>
<td>298</td>
<td>76.97</td>
<td>83.98</td>
<td>69.84</td>
</tr>
<tr>
<td>Ridge regression[2]</td>
<td>TS+TC+3D+QC</td>
<td>307</td>
<td>77.17</td>
<td>84.38</td>
<td>69.84</td>
</tr>
<tr>
<td>Ridge regression after variable selection[4]</td>
<td>TS+TC+AP</td>
<td>203</td>
<td>78.35</td>
<td>84.38</td>
<td>72.22</td>
</tr>
<tr>
<td>Envelope LDA</td>
<td>TS</td>
<td>103 (5)</td>
<td>59.45</td>
<td>70.31</td>
<td>48.41</td>
</tr>
<tr>
<td></td>
<td>TC</td>
<td>195 (37)</td>
<td>70.47</td>
<td>76.56</td>
<td>64.29</td>
</tr>
<tr>
<td></td>
<td>TS+TC</td>
<td>298 (32)</td>
<td>68.90</td>
<td>75.39</td>
<td>62.30</td>
</tr>
<tr>
<td></td>
<td>TS+TC+3D+QC</td>
<td>307 (34)</td>
<td>70.47</td>
<td>77.73</td>
<td>63.09</td>
</tr>
<tr>
<td>SupSVD LDA 90% cutoff</td>
<td>TS</td>
<td>103 (8)</td>
<td>60.04</td>
<td>67.58</td>
<td>52.38</td>
</tr>
<tr>
<td></td>
<td>TC</td>
<td>195 (51)</td>
<td>72.44</td>
<td>78.13</td>
<td>66.67</td>
</tr>
<tr>
<td></td>
<td>TS+TC</td>
<td>298 (48)</td>
<td>70.47</td>
<td>78.91</td>
<td>61.90</td>
</tr>
<tr>
<td></td>
<td>TS+TC+3D+QC</td>
<td>307 (51)</td>
<td>71.06</td>
<td>78.91</td>
<td>63.09</td>
</tr>
</tbody>
</table>

**Table:**

- **Model description**: Description of the model used for prediction.
- **Type of predictors in model**: Specific types of predictors used in the model.
- **No. of predictors**: Total number of predictors used in the model.
- **Correct classification %**: Accuracy of classification into mutagens and non-mutagens.
- **Total**: Overall accuracy.
- **Mutagens**: Accuracy for mutagens.
- **Non-mutagens**: Accuracy for non-mutagens.
Outline

1. The data and the variables
2. The models
3. Results
4. Conclusion
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

Logistic Envelope Regression.

supSVD a potential plausible approach because of its general framework and computational stability and applicability in $n << p$ scenario.
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.
Cook, R., Li, B., and Chiaromonte, F.
Envelope models for parsimonious and efficient multivariate linear regression.  

Hawkins, D., Basak, S., and Mills, D.
QSARs for chemical mutagens from structure: ridge regression fitting and diagnostics.  

Li, G., Yang, D., Shen, H., and Nobel, A.
Supervised Singular Value Decomposition and its asymptotic properties.  
Technometrics Submitted.

Majumdar, S., Basak, S., and Grunwald, G.

Soderman, J.
CRC Handbook of Identified Carcinogens and Noncarcinogens: Carcinogenicity-Mutagenicity Database.  
THANK YOU!