# 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**

# Table of contents

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

# Outline

1 The data and the variables

2 The models

3 Results

4 Conclusion

# The data

- 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.

| 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 [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

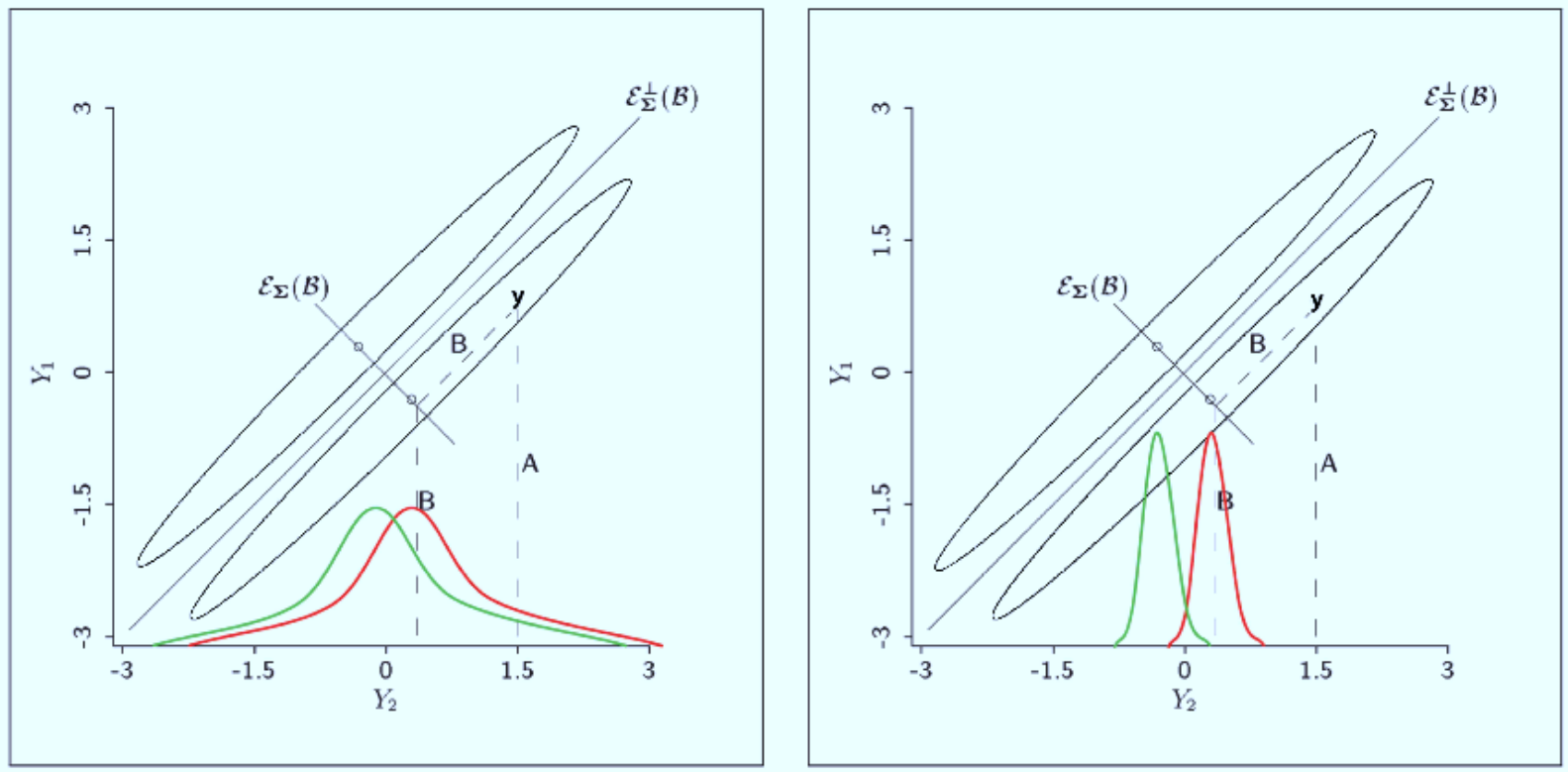
4 Conclusion

# Envelope regression model

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

- Due to **Cook, Li and Chiaromonte, 2010** [1].
- $\mathbf{Y} \in \mathbb{R}^{r \times n}$ 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 $\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0$.

# Graphical illustration of envelope model




**(Source: Stat 8932 class notes, R. Dennis Cook)**

# Envelope regression model for our data: basics

- 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)

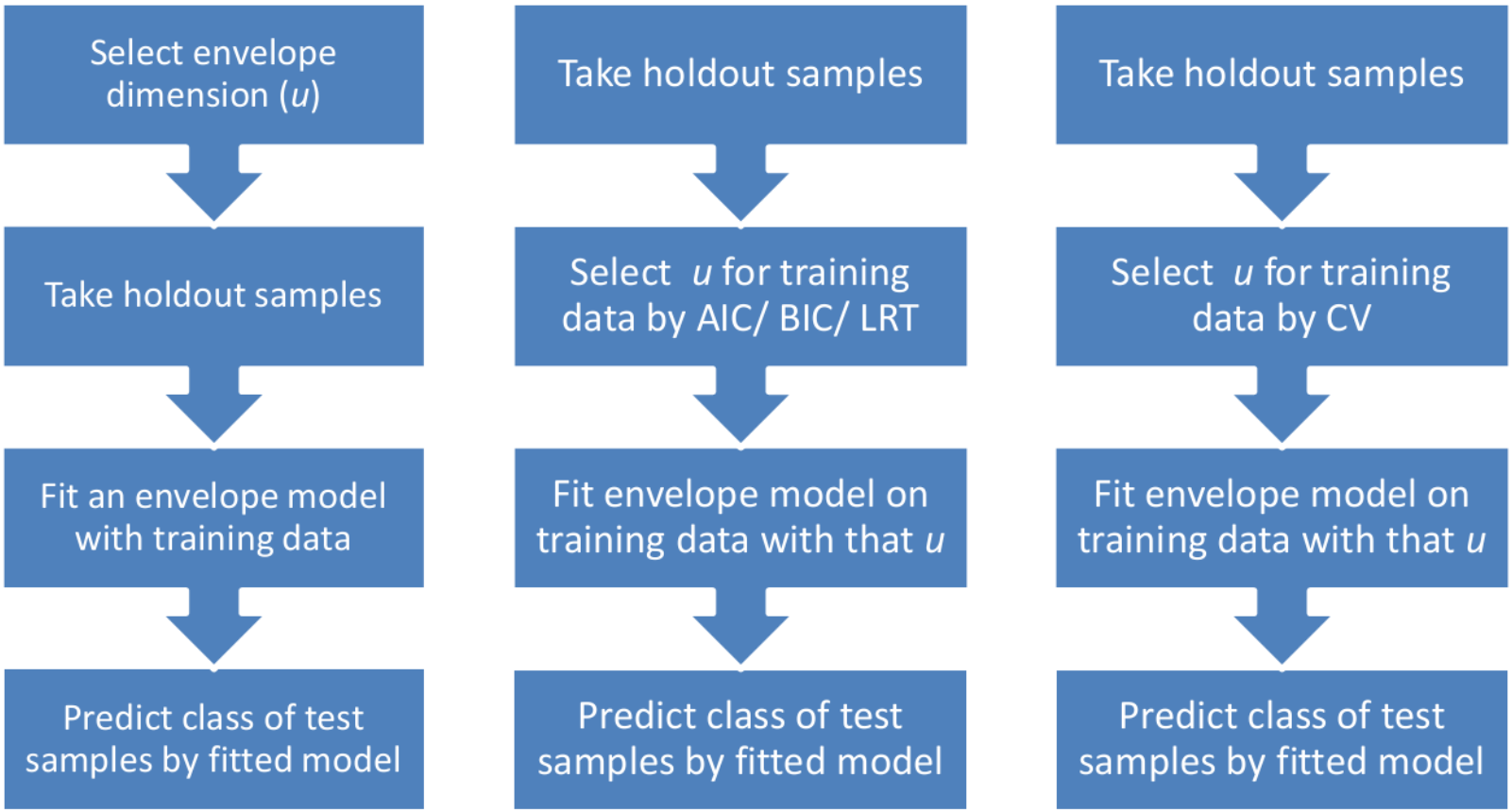
$$\mathbf{X} = \mathbf{YBV}^T + \mathbf{FV}^T + \mathbf{E}$$

- 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 $\mathbf{F} \sim \mathcal{N}_q(\mathbf{0}, \mathbf{\Sigma_f}), \mathbf{E} \sim \mathcal{N}_\mathbf{p}(\mathbf{0}, \sigma_\mathbf{e}^\mathbf{2}\mathbf{I_p})$ are random error matrices s.t. $\mathbf{\Sigma} = \mathbf{V\Sigma}_f\mathbf{V}^T + \sigma_e^2\mathbf{I}_p$.
- A modified EM algorithm is used to obtain the unknown parameters $\theta = (\mathbf{B}, \mathbf{V}, \mathbf{\Sigma}_f, \sigma_e^2)$.
- The vector of mutagenicity status is now used as the supervision data matrix $\mathbf{Y}$, while the data on 308 predictors is the matrix $\mathbf{X}$.

# Prediction through Linear Discriminant Analysis

- **Envelope model**- Estimate the envelope basis, say $\hat{\mathbf{\Gamma}}$, reduce the matrix of predictors by multiplying it with the basis and then apply Fisher's Linear Discriminant Analysis on $\hat{\mathbf{\Gamma}}^T\mathbf{Y}$.
- **supSVD**- Here the notations are reversed and $\mathbf{X}$ is our $508 \times 307$ data matrix. After obtaining the loading matrix $\mathbf{V}$, we transform the data matrix as: $\mathbf{U} = \mathbf{XV}$, and apply LDA on $\mathbf{U}$, taking $\mathbf{Y}$ as the 0/1 class variable.
- Correct classification percentages are obtained through cross-validation on the full sample.

# Naïve CV vs. Two-fold CV



# Outline


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

# 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.

# Performance of envelopes and supSVD in prediction

| Model description | Type of predictors in model | No. of predictors | Correct classification % | | |
|---|---|---|---|---|---|
| | | | Total | Mutagens | Non-mutagens |
| Ridge regression[2] | TS+TC | 298 | **76.97** | **83.98** | **69.84** |
| Ridge regression[2] | TS+TC+3D+QC | 307 | **77.17** | **84.38** | **69.84** |
| Ridge regression after variable selection[4] | 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** |
| SupSVD LDA 90% cutoff | TS | 103 (5) | **59.45** | **70.31** | **48.41** |
| | TC | 195 (37) | **70.47** | **76.56** | **64.29** |
| | TS+TC | 298 (32) | **68.90** | **75.39** | **62.30** |
| | TS+TC+3D+QC | 307 (34) | **70.47** | **77.73** | **63.09** |
| SupSVD LDA 95% cutoff | TS | 103 (8) | **60.04** | **67.58** | **52.38** |
| | TC | 195 (51) | **72.44** | **78.13** | **66.67** |
| | TS+TC | 298 (48) | **70.47** | **78.91** | **61.90** |
| | TS+TC+3D+QC | 307 (51) | **71.06** | **78.91** | **63.09** |

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

# Outline


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

# 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
- Logistic Envelope Regression.
- supSVD a potential plausible approach because of its general framework and computational stability and applicability in $n << p$ scenario.

# 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., Li, B., and Chiaromonte, F.

Envelope models for parsimonious and efficient multivariate linear regression.

*Stat. Sinica 20* (2010), 927–1010.

Hawkins, D., Basak, S., and Mills, D.

QSARs for chemical mutagens from structure: ridge regression fitting and diagnostics.

*Environ. Toxicol. Pharmacol. 16* (2004), 37–44.

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.

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. 9* (2013), 463–471.

Soderman, J.

*CRC Handbook of Identified Carcinogens and Noncarcinogens: Carcinogenicity-Mutagenicity Database*.

CRC Press, Boca Raton, FL, 1982.

# THANK YOU!