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

Subho Majumdar

Stat 8054 project summary
School of Statistics,University of Minnesota- Twin Cities
e-mail: majum010@umn.edu

Abstract. We apply envelope methods on a dataset consisting of 508 diverse chemicals for two different purposes: estimating effects of the 307 predictors on mutagenicity of compounds, and employing discriminant analysis for the purpose of mutagenicity prediction. 'Two-deep' leave-one-out cross-validation is used for the purpose of prediction, meaning that envelope dimension of the predictive model is selected separately for each holdout compound. Rank deficiency in the data is tackled in two ways: by applying envelope methods as well as Supervised SVD on the first few principal components of the predictor matrix. Finally predictive performance of envelope models are compared with those of two earlier papers on the same dataset.

Keywords : Envelope models; Supervised SVD; Dimension reduction; Two-deep cross-validation; Chemometrics; QSAR

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.

Supervised Singular-Value Decomposition (SupSVD) is a recently proposed method [4] that provides a general framework, from which principal component analysis/ factor analysis and reduced-rank regression arise as special cases. In this a latent structure of a fixed dimension lower than the actual number of parameters is assumed and it is estimated using both the actual data and supervision data using a modified EM algorithm.

Previous works on this dataset have focused on prediction through ridge regression models [3] as well as variable selection through an iterative algorithm [5]. Here we shall use hierarchical envelope regression and SupSVD models for estimating the effects of the predictors on mutagenicity, as well as employ discriminant analysis on transformed data for the purpose of mutagenicity prediction.

2 Data

The data for this project were taken from the CRC Handbook of Identified Carcinogens and Non-carcinogens [7]. 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:

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*)

3 Methods

3.1 Envelope regression model

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

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

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

where $\mathbf{Y} \in \mathbb{R}^{n\times r}$ multivariate response vector, $\mathbf{X} \in \mathbb{R}^{n\times p}$ 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 $0 \leq u \leq r$ 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 307 descriptors as multivariate responses.

3.2 Supervised Singular-Value Decomposition (SupSVD)

This recently proposed model [4] assumes a latent lower-dimensional structure for the matrix of predictors $\mathbf{X} \in \mathbb{R}^{n\times p}$ and estimates it using a supervision data matrix $\mathbf{Y} \in \mathbb{R}^{n\times r}$. The assumed model is:

$$\mathbf{X} = \mathbf{Y}\mathbf{B}\mathbf{V}^T + \mathbf{F}\mathbf{V}^T + \mathbf{E}$$

where $\mathbf{B} \in \mathbb{R}^{r\times q}$ is the multivariate matrix of coefficients, $\mathbf{V} \in \mathbb{R}^{p\times q}$ full-rank loading matrix. Here $0 \leq q \leq r$ is the dimension of the underlying space of latent parameters, and $\mathbf{F} \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{f}), \mathbf{E} \sim \mathcal{N}_\mathbf{p}(\mathbf{0}, \sigma_\mathbf{e}^\mathbf{2}\mathbf{I}_\mathbf{p})$ are random error matrices ($\boldsymbol{\Sigma}_f \in \mathbb{R}^{q\times q}$ diagonal). Hence the overall covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ can be decomposed as $\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Sigma}_f\mathbf{V}^T + \sigma_e^2\mathbf{I}_p$. The modified EM algorithm to obtain the unknown parameters $\boldsymbol{\theta} = (\mathbf{B}, \mathbf{V}, \boldsymbol{\Sigma}_f, \sigma_e^2)$ is given in the appendix.

The vector of mutagenicity status is now used as the supervision data matrix $\mathbf{Y}$, while the data on 308 predictors is the matrix $\mathbf{X}$ (notice that $\mathbf{X}$ and $\mathbf{Y}$ are reversed from 3.1). The structure of the envelope model might seem similar to the SupSVD model considering the facts that both of them consider decompositions of the sample covariance matrix $\boldsymbol{\Sigma}$, but they arise from different underlying assumptions. While envelope models try to obtain the smallest reducing subspace of $\boldsymbol{\Sigma}$ that contains the matrix of coefficients $\mathbf{B}$, SupSVD can be seen as a generalization of the factor analysis/ PCA model, in which the data matrix itself is a result of a transformation on the underlying space of unknown independent predictors.

3.3 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 $\geq 90\%$ of the total variance in a loading matrix $\mathbf{L} \in \mathbb{R}^{r \times k}$ and apply the methods in 3.1 on the transformed predictors $\mathbf{L}^T\mathbf{Y}$ instead of $\mathbf{Y}$. Due to the substantially less computational burden of the SupSVD algorithm, both the 90% and 95% cutoffs from PCA were considered as the dimension of the space of latent predictors i.e. q.

3.4 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 apply Fisher's Linear Discriminant Analysis [2] on $\hat{\boldsymbol{\Gamma}}^T\mathbf{Y}$.

For SupSVD the notations are reversed and $\mathbf{X}$ is our 508×307 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.

4 Results

4.1 Envelope estimation

All envelope 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. To take care of the difference in magnitude across predictors, each entry x in the data matrix $\mathbf{X}$ is transformed as: $x \mapsto \log(x + C)$, where $C = -\lfloor x \rfloor$ if $x < 0$. and $C = 1$ otherwise. After this we standardize $\mathbf{X}$ and $\mathbf{Y}$.

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 $j = 1, 2, ..., r$.

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 single variable-type models, linearly correlated predictors tend to be simultaneously significant, but this behavior is is less observed in the combined model (TS+TC) (**Table 2**).

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

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

Table 2. Significance of predictors in envelope models

Case	Mean error(SE)	
	SupSVD	PCA
I. SupSVD	0.0768 (0.034)	0.0794 (0.031)
II. PCA	0.0803 (0.032)	0.0768 (0.034)

Table 3. $MSE_{\mathbf{V}}$ means and standard errors for the two simulation scenarios

4.2 SupSVD analysis

The efficacy of this method in estimation of the loading matrix $\mathbf{V}$ in the supervised setup is demonstrated by the following simulation. We set $n = 1000, p = 10, r = 2, q = 3$. We generate the entries of the 1000×2 matrix $\mathbf{Y}$ as iid $\text{Ber}(1, 0.5)$. The loading matrix $\mathbf{V}$ is obtained by taking p iid samples from $\mathcal{N}_r(\mathbf{0}, \mathbf{I_r})$ and orthogonalizing the columns. Finally we take $\boldsymbol{\Sigma}_f$ =diag$(10, 7, 4), \sigma_e^2 = 2$, and generate $\mathbf{F}, \mathbf{E}$. We choose $\mathbf{B}$ to for two scenarios:

1. r rows of $\mathbf{B}$ as iid samples from $\mathcal{N}_q(\mathbf{0}, \mathbf{I_q})$. Thus $\mathbf{X} = \mathbf{YBV}^T + \mathbf{FV}^T + \mathbf{E}$ and this is the typical SupSVD scenario.
2. Take $\mathbf{B} = \mathbf{0}$. Then we have $\mathbf{X} = \mathbf{FV}^T + \mathbf{E}$ and the model becomes an unsupervised factor analysis/PCA model.

For each of the two cases, we generate 1000 samples and apply SupSVD and PCA on each of them. The error in estimating $\mathbf{V}$ can be calculated for both the methods:

$$MSE_{\mathbf{V}} = \frac{1}{pq}\|\mathbf{V} - \hat{\mathbf{V}}\|_{\mathbb{F}}^2$$

For PCA, $\hat{\mathbf{V}}$ is taken as the first q columns of the original PCA loading matrix obtained. The results are summarized in **Table 3**. In presence of supervision data, SupSVD turns out to be more accurate than unsupervised PCA, but this is not the case when $\mathbf{B} = \mathbf{0}$.

For each of the 4 subsets of the data: TS, TC, TS+TC and full, we did PCA on the data and take the first few components that explain 90% (or 95%) of the total variance, take the number of components as the underlying dimension and perform the SupSVD analysis. The

#Component	1	2	3	4	5	6	7	8
Before	75.425	8.247	5.01	3.135	2.642	1.366	1.159	0.908
After	75.186	8.187	4.922	2.992	2.476	1.303	1.108	0.856

Table 4. First 8 components of TS data, before and after SupSVD

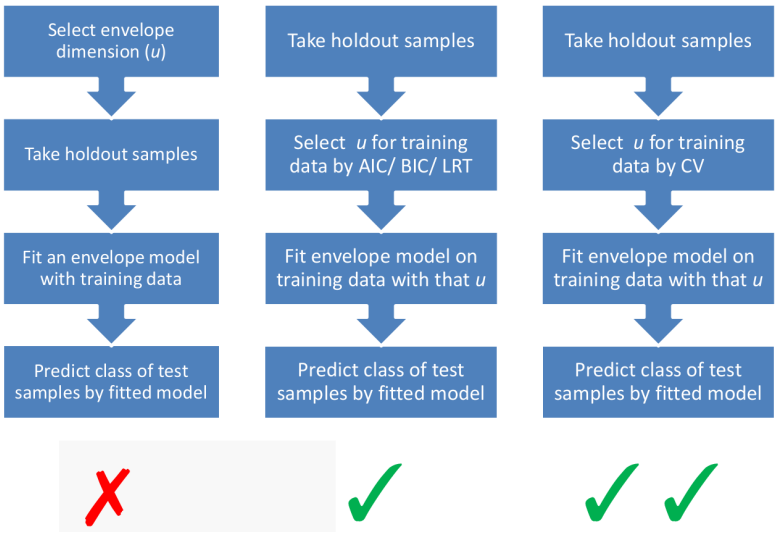

Fig. 1. Naïve CV vs. Two-fold CV

diagonal entries of the matrix $\boldsymbol{\Sigma}_f$ lists the variance of the latent components. These can be compared to variances of the same principal components of the data to study the effect due to predictors. For example, **Table 4** lists the top 8 PC's that explain 95% of the total variances in the TS data, and compares it to the diagonal entries of the $\boldsymbol{\Sigma}_f$ obtained.

4.3 Predictive performance of models

Prediction using envelope LDA was done by leave-one-out cross-validation. While doing cross-validation here it is imperative not to fix the dimension of the envelope u beforehand and then take holdout samples to build envelope models. Since information from the holdout compound has already been used to obtain u, this overestimates the predictive power of the method (hence called Naïve CV, see **Figure 1**). Instead in each step one should 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. Since we are aiming for prediction, one should select the envelope dimension each time by CV as well (Two-fold CV), but due to larger computational burden we select it by AIC/BIC each time.

Table 5 summarizes the predictive performance of all the methods (for SupSVD, numbers in brackets indicate dimension of the latent variable space). Envelope reduction after transformation by first few principal components seems to perform the worst in terms of prediction. SupSVD performs better, and for a larger number of latent components prediction performance is slightly better, but it still can't beat previous analyses.

Model description	Type of predictors in model	No. of predictors	Correct classification %		
			Total	Mutagens	Non-mutagens
Ridge regression[3]	TS+TC	298	76.97	83.98	69.84
Ridge regression[3]	TS+TC+3D+QC	307	77.17	84.38	69.84
Ridge regression after variable selection[5]	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 5. Prediction performance of various models

5 Conclusion

Looking at the scatterplot of points for the first 5 latent components in the data on TS variables only (see Appendix 2), we can see that a linear approach to discriminant analysis for this data might not be a good idea itself. Here 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. Finally, the supervised SVD has the potential to be a plausible approach towards classification and supervised data analysis because of its general framework and more computational stability compared to envelopes.

Acknowledgment

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

References

1. Cook, R., Li, B., and Chiaromonte, F. Envelope models for parsimonious and efficient multivariate linear regression. *Stat. Sinica 20* (2010), 927–1010.
2. Fisher, R. The use of multiple measurements in taxonomic problems. *Annals of Eug. 7* (1936), 179–188.
3. 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.
4. Li, G., Yang, D., Shen, H., and Nobel, A. Supervised Singular Value Decomposition and its asymptotic properties. *Technometrics Submitted*.

5. 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.
6. MATLAB. *version 7.10.0 (R2010a).* The MathWorks Inc., Natick, Massachusetts, 2010.
7. Soderman, J. *CRC Handbook of Identified Carcinogens and Noncarcinogens: Carcinogenicity-Mutagenicity Database.* CRC Press, Boca Raton, FL, 1982.

Appendix 1: Algorithm for SupSVD parameter estimation

We start with the model as defined in 3.2, and our goal is to estimate the parameter $\boldsymbol{\theta} = (\mathbf{B}, \mathbf{V}, \boldsymbol{\Sigma}_f, \sigma_e^2)$. Define $\mathbf{U} = \mathbf{YB} + \mathbf{F}$. Then using the following Expectation-Maximization-Standardization (EMS) algorithm one can obtain the parameter estimates iteratively (as described in [4]). It is proved that the likelihood function $\mathcal{L}(\mathbf{X}|\boldsymbol{\theta})$ increases in each iteration, hence in the stopping criterion we subtract the likelihood function of the current iterate from the one of the next iterate.

Algorithm 1 The EMS Algorithm for Parameter Estimation under the SupSVD Model

procedure SupSVD(data $\mathbf{X} \in \mathbb{R}^{n\times p}$, supervision data $\mathbf{Y} \in \mathbb{R}^{n\times r}$)
  Set initial values for the parameters $\boldsymbol{\theta}^{(0)} = (\mathbf{B}^{(0)}, \mathbf{V}^{(0)}, \boldsymbol{\Sigma}_f^{(0)}, \sigma_e^{2(0)}); i = 0$
  while $\mathcal{L}(\mathbf{X}|\boldsymbol{\theta}^{(i+1)}) - \mathcal{L}(\mathbf{X}|\boldsymbol{\theta}^{(i)}) >$threshold **do**
    E step: Calculate
      $\mathrm{E}_{\mathbf{U}}(\mathbf{U}|\mathbf{X}, \boldsymbol{\theta}^{(i)}) = \boldsymbol{\Theta}_{\mathbf{U}|\mathbf{X}}^{(i)} = \left[\mathbf{YB}^{(i)}\left(\sigma_e^{2(i)}\boldsymbol{\Sigma}_f^{(i)^{-1}}\right) + \mathbf{XV}^{(i)}\right]\left[\mathbf{I}_q + \sigma_e^{2(i)}\boldsymbol{\Sigma}_f^{(i)^{-1}}\right]^{-1}$
      $\mathrm{Var}_{\mathbf{U}}(\mathbf{U}|\mathbf{X}, \boldsymbol{\theta}^{(i)}) = \boldsymbol{\Omega}_{\mathbf{U}|\mathbf{X}}^{(i)} = \left[\boldsymbol{\Sigma}_f^{(i)^{-1}} + \sigma_e^{-2(i)}\mathbf{I}_q\right]^{-1}$
    M step: Calculate unconstrained maximizers
      $\hat{\mathbf{B}} = (\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\boldsymbol{\Theta}_{\mathbf{U}|\mathbf{X}}^{(i)}$
      $\hat{\mathbf{V}} = \mathbf{X}^T\boldsymbol{\Theta}_{\mathbf{U}|\mathbf{X}}^{(i)}\left[n\boldsymbol{\Omega}_{\mathbf{U}|\mathbf{X}}^{(i)} + \boldsymbol{\Theta}_{\mathbf{U}|\mathbf{X}}^{(i)T}\boldsymbol{\Theta}_{\mathbf{U}|\mathbf{X}}^{(i)}\right]^{-1}$
      $\widehat{\boldsymbol{\Sigma}_f} = \frac{1}{n}\left[n\boldsymbol{\Omega}_{\mathbf{U}|\mathbf{X}}^{(i)} + \boldsymbol{\Theta}_{\mathbf{U}|\mathbf{X}}^{(i)T}\boldsymbol{\Theta}_{\mathbf{U}|\mathbf{X}}^{(i)} + \hat{\mathbf{B}}^T\mathbf{Y}^T\mathbf{Y}\hat{\mathbf{B}} - \hat{\mathbf{B}}^T\hat{\mathbf{Y}}^T\boldsymbol{\Theta}_{\mathbf{U}|\mathbf{X}}^{(i)} - \boldsymbol{\Theta}_{\mathbf{U}|\mathbf{X}}^{(i)T}\mathbf{Y}\hat{\mathbf{B}}\right]$
      $\widehat{\sigma_e^2} = \frac{1}{np}\left[\mathrm{tr}(\mathbf{X}^T\mathbf{X}) + 2\,\mathrm{tr}\left(\boldsymbol{\Theta}_{\mathbf{U}|\mathbf{X}}^{(i)}\hat{\mathbf{V}}^T\mathbf{X}^T\right) + n\,\mathrm{tr}\left(\hat{\mathbf{V}}^T\hat{\mathbf{V}}\boldsymbol{\Omega}_{\mathbf{U}|\mathbf{X}}^{(i)}\right) + \mathrm{tr}\left(\boldsymbol{\Theta}_{\mathbf{U}|\mathbf{X}}^{(i)}\hat{\mathbf{V}}^T\mathbf{V}\boldsymbol{\Theta}_{\mathbf{U}|\mathbf{X}}^{(i)T}\right)\right]$
    S step: Calculate next iterates
      Get $\mathbf{V}^{(i+1)}, \boldsymbol{\Sigma}_f^{(i+1)}$ from eigen-decomposition of $\hat{\mathbf{V}}\widehat{\boldsymbol{\Sigma}_f}\hat{\mathbf{V}}^T$: $\mathbf{V}^{(i+1)}\boldsymbol{\Sigma}_f^{(i+1)}\mathbf{V}^{(i+1)T} = \hat{\mathbf{V}}\widehat{\boldsymbol{\Sigma}_f}\hat{\mathbf{V}}^T$
      Reorder columns of $\mathbf{V}^{(i+1)}$, rows/columns of $\boldsymbol{\Sigma}_f^{(i+1)}$ as per decreasing column norm of $\mathbf{XV}^{(i+1)}$
      $\mathbf{B}^{(i+1)} = \hat{\mathbf{B}}\hat{\mathbf{V}}^T\mathbf{V}^{(i+1)}$
      $\sigma_e^{2(i+1)} = \widehat{\sigma_e^2}$
    Set $i \leftarrow i + 1$
  end while

Appendix 2: Scatterplot of latent components

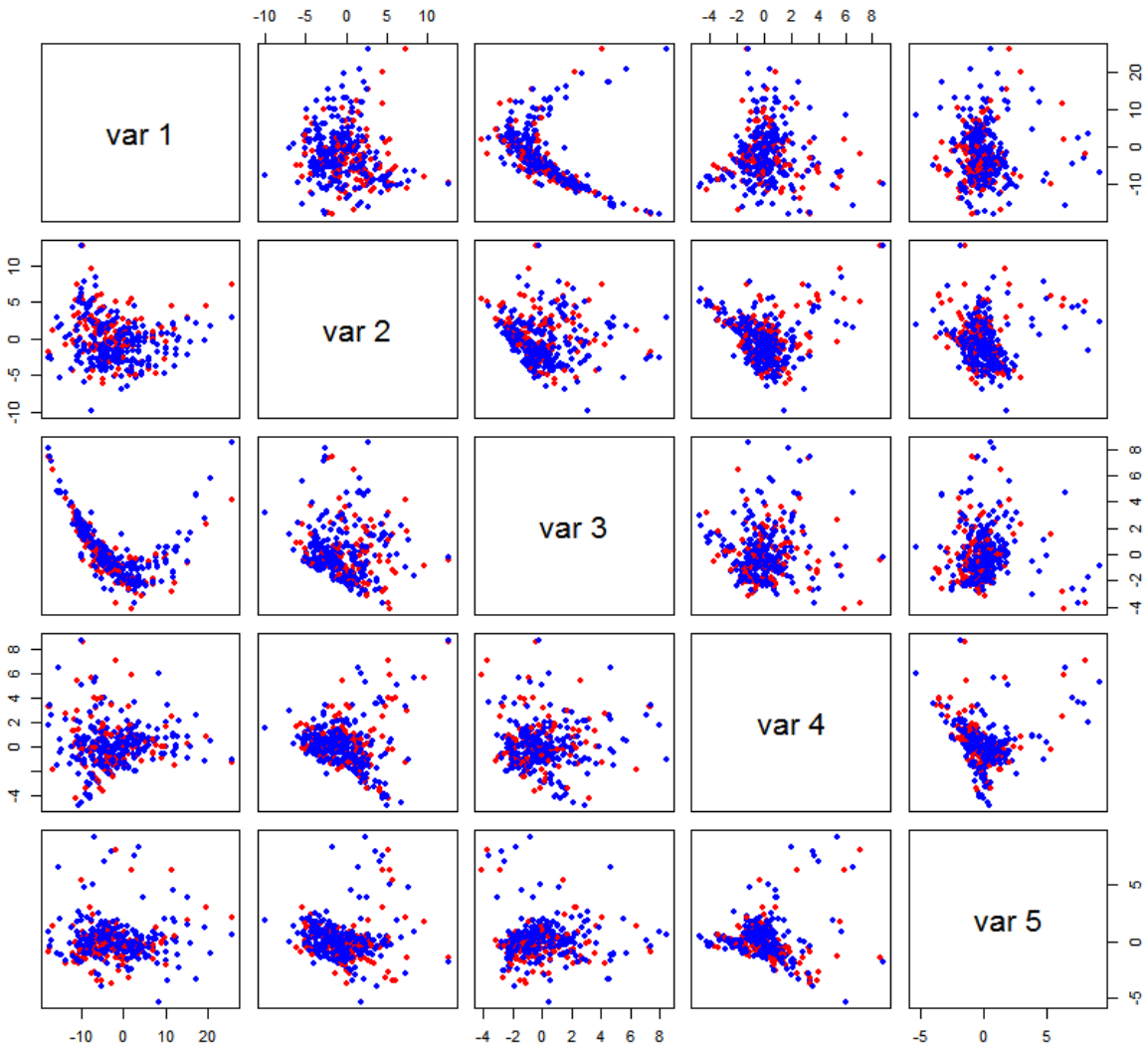

Fig. 2. Plot of first 5 latent components for TS data, 90% cutoff (red = mutagen, blue = non-mutagen)