Laboratory Sessions

Building Models. Multivariate Regression

Exercise 6 Develop correlation models using environmental data

Develop correlation models between the concentration of persistent organic pollutants measured in fish samples and morphological and geographical parameters defining the lakes and fish samples were obtained (as a test of the global distillation theory for the distribution of contaminants in the earth)

The problem

The presence of organochlorine compounds (OCs)—namely hexachlorobenzene (HCB), hexachlorocyclohexanes (HCH), polychlorobiphenyls (PCBs #28, 52, 101, 118, 138, 153 and 180), dichlorodiphenyltrichloroethane (DDT) and dichlorodiphenyldichloroethane (DDE)—was examined in various fish tissues (muscle and liver) sampled in 23 mountain lakes in Europe. The dependence of the concentration of these organochlorine compounds on geographical parameters (altitude, longitude, latitude and temperature) and physiological parameters (lipid content, age, weight and size) is assessed using principal components analysis (PCA) and partial least squares (PLS).

In both cases, accumulation of organochlorine compounds in fish tissues in mountain lakes were shown to depend significantly on altitude and latitude. Therefore, the results support the global distillation theory and can be applied to a description of the altitudinal profiles of these compounds in mountain lakes.


The data set

The data set is stored in the file lakes.mat which contains the matrix $X$ with the values of 8 variables describing fish samples obtained in 22 mountain lakes in different places of Europe (see publications). These 8 variables are: 'lipid' (lipid content of fish specimens), 'Weight' (weight of fish specimens), 'Length' (length of fish specimens), 'Age' (age of fish specimens), 'Height' (height of the lake), 'Temp' (annual average temperature of the water lake), 'Lati' (latitude geo coordinates of the lake), 'Long' (longitude geo coordinates of the lake). Apart from these 8 variables, the following 12 organic chemical compounds were measured (average values of several fish samples of every lake): DDE, DDT, HCB, a-HCH, g-HCH, PCB28, PCB52, PCB101, PCB101, PCB138, PCB153, PCB180

Name of variables: meanHCB meanaHCH meanddt meanpcb101 meanpcb138 meanpcb180 meanppcb52 meanX meandde meangHCH meanpcb118 meanpcb153 meanpcb28 (namevarall gives the name of all the variables, namevarorg,
the name of the measured organic compounds, namevarpar the name of the physiological and geographical parameters.

Name of lakes are given in lakenames character variables (see map in the paper; lake Milchsee in the map number 14 was eliminated, therefore numbers 14-22 are 15-23 in the map)

The questions

1) Built a data matrix with all variables (22,20), 22 samples in which a total number of 20 variables are measured (geographic, morphological and chemical)

2) Preprocess the data (autoscaling)

3) Create a PCA model. How many components are necessary to explain the experimental data variance? Do they identify significant trends?

4) Interpret loadings plot. Are there correlation between variables? How can be interpreted these correlations?

5) Interpret scores plots. Can you deduce any type of distribution between the samples from different lakes (according to their geography, mean temperature, …)?

6) What type of correlations can be investigated?

7) Develop correlation models between the concentration of each of the contaminants accumulated in fish and the physiological and geographical parameters.

8) Cross validation and number of components

9) Interpret the models looking at the following PLS outputs: pls loadings, pls weights and VIPs for the more significant components

10) Compare prediction versus actual values

11) Can the theory of global distillation be assessed with the proposed data set?
Exercice 7 Develop correlation models using toxicological data

Develop coorelation models between the concentration of persistent organic pollutants measured in two different rivers and different biochemical responses

The problem

Environmental factors affecting aquatic invertebrate communities were assessed using Daphnia magna in situ bioassays and biological indices based on community assemblages of benthic macroinvertebrates. Investigations were carried out in two heavily industrialized and urbanized river basins from the NE of Spain (Llobregat and Besos). Measures of energy consumption (i.e. algal grazing), and of specific biochemical responses (biomarkers) were conducted on individuals transplanted upstream and downstream from effluent discharges of sewage treatment plants.

Principal Component and Partial Least Square Projections to Latent Structures regression analyses were used to investigate the possible correlations among the measured responses in D. magna, in in situ bioassays and in macroinvertebrate assemblages on one side and different (20) environmental variables: seven of them including habitat degradation, suspended solids, nitrogenous, d conductivity related parameters levels of organophosphorus compounds and of polycyclic aromatic hydrocarbons.

See publication: Combined use of Daphnia magna in situ bioassays, biomarkers and biological indices to diagnose and identify environmental pressures on invertebrate communities in two Mediterranean urbanized and industrialized rivers (NE Spain) Joana Dam´ asioa,e, Roma` Tauler a, Elisabeth Teixi do´ b, Maria Rieradevall c, Narcis Prat c, Maria Carmen Rivad, Amadeu M.V.M. Soarese, Carlos Barataa, Aquatic Toxicology 87 (2008) 310–320

The data set

The data set is stored in the file llob_besos.mat which contains the matrix data has the values of 27 variables describing 28 river samples from Besos and Loobregat sites. The first 6 variables, feeding (FED), catalasa activity (CAT), glutathione S-transferase activity (GST”), acetylcholinesterase activity (ACHE), Iberian Monitoring Working Party Biological Index (IBMWP) and the Iberian Average Score per taxon (IASPT) are used as possible toxicological parameters indicating the quality of effluent river waters. The rest of variables (see below) are used as environmental parameters: a) water physicochemical parameters including flow rate (l/s), temperature (T; °C), pH, conductivity (S/cm), sulphates (SO4, mg/l), chlorides (Cl, mg/l), dissolved oxygen (O2, mg/l), suspended solids (SS, mg/l), N-ammonium (NH4, mg/l), N-nitrites (NO2,mg/l), N-nitrates (NO3,mg/l), P-phosphates (PO4,mg/l); and b) concentration of pollutants: Cu and Zn, triazines (TRZ) polycyclic aromatic hydrocarbons (PAH) organochlorate pesticides (OCL) organophosphorous compounds (OPs) alklyphenol ethoxilates (APEs); quality indexes: Ecological Quality Index (QBR) and Fluvial Habitat Index (IHF, HI).
The Llobregat and Besos river basins (NE Spain, see map in paper below) include an area of 6400km² and main channel flows of 0.1–12m³/s. Like other Mediterranean systems, the natural resources of Llobregat and Besos’ river basins have been greatly affected by human activities such as agriculture, urbanization. Salinity increased as a result of mining activities and an intensive water use for human consumption (supplying water urban areas including Barcelona city), which together have severely deteriorated the ecological status of the main rivers and tributaries since 1970s. During 1990s the construction of sewage treatment plants (STPs) and of salt collectors substantially improved the chemical and biological quality of water, thereby allowing the survival of fish and invertebrate species in middle and lowland reaches. Deployment sites comprised eleven and eight points along the Llobregat and Besos’ river systems, respectively (map in Fig. 1 paper). Stations were selected to include clean upstream(sites L1, L2, L3, L4, L5, B1, B3, B3), polluted middle (site L6, L8, B5, B6) and downstream reaches (sites L9, L10, L11, B4, B7, B8).

Sample names are given in sampnames variable, which gives a code that can be located in the map of Figure 1 paper (below) clean upstream(sites L1, L2, L3, L4, L5, B1, B3, B3), polluted middle (site L6, L8, B5, B6) and downstream reaches (sites L9, L10, L11, B4, B7, B8).

The questions

1) Missing values in matrix data. How to deal with missing values? How to deal with below the detection limit values?
2) Data pretreatment. Explore the use of the log data pretreatment. Advantages and disadvantages
3) Data pretreatment: autoscaling
4) Investigate correlations among all the variables and samples: Built PCA models for all the variables and samples. How many components are sufficient?
5) Plot loadings. Check the correlations among variables, specially between the biochemical parameters and the rest.
6) Plot scores. Do you observe clusters? Can Besos and Llobregat river samples be distinguished?. Perform cluster dendograma analysis (program cluster)
7) Built PLS models among the more relevant biochemical parameters and the rest of variables.
8) Cross validation of the model. How many components are needed?
9) Explore loadings and weights vectors. Interpret
10) Explore regression vectors. Interpret
11) Calculate VIPs and decide about the more influential variables
12) Compare predicted vs actual vales and validate the models
13) What information can be obtained of all these analysis
Exercise 8 Develop regression models for analytical calibration and prediction

Develop a calibration/prediction model for mixtures of gasolines using inverse linear prediction methods (MLR, PCR and PLSR)

Compare inverse multivariate linear regression methods (like MLR, PCR or PLS) with classical least squares (CLS)

The problem

Test different multivariate calibration methods using near infrared (NIR) spectroscopic data on 30 pseudo gasoline samples which are a mixture of 5 components. Inverse calibration models can be used for calibration in the presence of interferences. Compare inverse least squares procedures with classical least squares procedures. Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR) are specially suitable multivariate calibration methods for this type of data. Classical Least Squares (CLS) calibration needs input information about the five analytes and allows the recovery of the pure response of them. CLS cannot be used with incomplete calibration information (in the presence of unknown interferences).

The data set

nir_data set contains the measured near infrared absorbance spectra on the 30 pseudo gasoline samples. The data set consists of two 30 by 401 matrices of spectra, spec1 and spec2, of the absorbances of 30 samples measured on two different spectrometers, a 30 by 5 conc matrix containing the concentration of the 5 components in each sample mixture, and a 1 by 401 lamda vector containing the wavelengths corresponding to the spectra.

The questions

Develop classical least squares, CLS, models for the prediction of all analytes on the first 30 samples (spec1) from nir_data and use to predict the other 30 samples (spec2). Plot your estimates of the pure component spectra. Develop the models using all analytes and only one analyte and the other as interferents.

Develop inverse prediction least squares, MLR-ILS models for the prediction of one analyte in the presence of interferences. Develop the calibration models for the first 30 samples (spec1) from nir_data and use them to predict the other 30 samples (spec2).

Do the same for PCR and PLSR methods for the 5 analytes and compare results. What data pretreatment is used and why? How many factors are used? Calculate the values of RMSEP (root-mean square error of prediction) for the remaining samples for each of the analytes. Report for RMSEC, RMSECV and RMSEP

Predict the concentration of the five analytes in the final 10 samples. Calculate and compare RMSEP values obtained by the different methods. Do you observe differences? Explain.
Chemometrics Course

Software

Multivariate Regression

`pcr.m`

Principal Components Regression for multivariate y-block

```
[b,ssq,t,p,eigs]=pcr(x,y,pc,\text{\textit{out}})
```

Input:

- `x` is the matrix of predictor variables (x-block)
- `y` is the vector or matrix of predicted variables (y-block)
- `pc` maximum number of principal components
- `\text{\textit{out}}` optional input, when set to 0 suppresses echoing of results

Output:

- `b` matrix of regression coefficients for each number of principal components, where each block of `ny` rows (\`ny` is the number of y-block variables) corresponds to the PC model for that number of principal components
- `ssq` captured variance information for each number of principal components
- `t` x-block of scores
- `p` x-block of loadings
- `eigs` eigenvalues of the x-block covariance matrix

`pcr1.m`

Principal Components Regression for univariate y

```
[t,p,b]=pcr1(x,y,pc)
```

PCR1 Principal components regression for univariate y. The inputs are the matrix of predictor variables (x), vector of predicted variable (y), and maximum number of principal components to consider (pc). The outputs are the x-block scores (t), the x-block loadings (p) and the matrix of regression coefficients (b) for each number of principal components, where each row corresponds to the % PCR model for that number of principal components.

`pls.m`
Partial Least Squares Regression for univariate or multivariate y-block

\[ [b,ssq,p,q,w,t,u,bin] = \text{pls}(x,y,lv,\text{out}) \]

**Inputs**
- \( x \) matrix of predictor variables (x-block)
- \( y \) is the vector or matrix of predicted variables (y-block)
- \( lv \) maximum number of latent variables
- \( out \) optional input, when set to 0 suppresses echoing of results

**Outputs**
- \( b \) matrix of regression coefficients for each number of latent variables, where each block of \( ny \) rows (\( ny \) is the number of y-block variables) corresponds to the PLS model for that number of latent variables
- \( ssq \) captured variance information for each number of principal components
- \( p \) x-block of loadings
- \( q \) y-block of loadings
- \( w \) x-block latent variables weights
- \( t \) x-block of scores
- \( u \) y-block of scores
- \( bin \) internal regression vectors (between \( t \) and \( u \))

**plspred.m**

Predictions based on existing PLS model

\[ \text{ypred} = \text{plspred}(x,bin,p,q,w,lv) \]

**Inputs** are the matrix of predictor variables (\( x \)), the PLS model inner-relation coefficients (\( bin \)), the x-block loadings (\( p \)), the y-block loadings (\( q \)), the x-block weights (\( w \)), and the number of latent variables to use in prediction (\( lv \)). The output is the vector or matrix of the predicted values (ypred).

**Crossval.m**

Crossvalidation for PCA, MLR, PCR and PLS

\[ [\text{press},\text{compress},\text{rmsecv},\text{rmsec}] = \text{crossval}(x,y,\text{rm},\text{cvm},lv,\text{split},\text{iter},\text{mc},\text{out},\text{osc}) \]

**Input:**
- \( x \) predictor variables (x-block)
- \( y \) predicted variables (y-block), \( y=[] \), empty for \( rm='PCA' \)
rm regression method ('PCA' for PCA, 'mlr' for MLR, 'pcr' for PCR, 'nip' for NIPALS PLS, 'sim' for SIMPLS)
cvm cross-validation method ('loo' for leave-one-out, 'vet' for venetian blinds, 'con' for contiguous data sets, 'rnd' for random subsets)
Lv, maximum number of latent variables

split, number of times to split the data
iter, number of times to perform cross-validation (required for cvm='rnd')
mc suppresses mean centering when set to zero; mean-centering when set to one
out suppresses output plots when set to zero; give plots when set to one
osc number of orthogonal signal correction components; osc=[nocomp,iter,tol], nocomp is the number of OSC components, iter number of iterations and tol is the tolerance

Outputs

press predicted error sum of squares for each subset
compress cumulative press
rmsecv, root mean square error of cross validation
rmsec, root mean square error of calibration
when out is different from zero, plots of rmsecv and rmsec are given

Examples

```matlab
[press, compress] = crossval(x,y,'nip','loo',10);
[press, compress] = crossval(x,y,'pcr','vet',10,3);
[press, compress] = crossval(x,y,'nip','con',10,5);
[press, compress] = crossval(x,y,'sim','rnd',10,3,20);
[press, compress] = crossval(x,y,[],'pca','loo',10);
[press, compress] = crossval(x,y,[],'pca','vet',10,3);
[press, compress] = crossval(x,y,[],'pca','con',10,5);
```
**lsreg2.m**

Comparison of two data sets (vectors) displaying their correlation

**lsreg2 (x,y)**

x and y are the two vectors to compare

Useful function for model validation to compare the quality of the predictions versus the actual values. It can be used in internal validation (calibration step) and in external validation (with data not used to built the model). It gives statistics of the best line which ideally should have a slope of one, an offset of zero and a correlation coefficient close to 1

Calculates the slope, offset, correlation coefficient and standard deviation of residuals of the best line fitted by least squares regression

\[ y = \text{slope} \times x + \text{offset} \]

It also calculates other statistics (using internal function pe2.m)

\[ \text{dev}_i = y_i - \hat{y}_i \quad i=1,...,N \]

\[ \text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} \text{dev}_i^2}{N}} \]

\[ \text{bias} = \frac{\sum_{i=1}^{N} \text{dev}_i}{N} \]

\[ \text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (\text{dev}_i - \text{bias})^2}{N-1}} \]

\[ \text{erel} = 100 \times \sqrt{\frac{\sum_{i=1}^{N} \text{dev}_i^2}{\sum_{i=1}^{N} y_i^2}} \]
**vip.m**

Calculate Variable Importance in Projection from regression model.

**vip_scores = vip(model)**

Variable Importance in Projection (VIP) scores estimate the importance of each variable in the projection used in a PLS model and is often used for variable selection or to study the importance of a variable in a multivariate correlation model. A variable with a VIP Score close to or greater than 1 (one) can be considered important in given model. Variables with VIP scores significantly less than 1 (one) are less important and might be good candidates for exclusion from the model.

The input is a PLS model structure (model).

The output (vip_scores) is a set of column vectors equal in length to the number of variables included in the model. It contains one column of VIP scores for each column of the original calibration y-block.

Another possibility is to use the **vipr.m** function

**vipscores = vipr(T,P,w,b,ny,nx,labels);**

- T scores (nsamples,1)
- P loads (nvar,1)
- w weights (nvar,1)
- b final regression vector (nvar,1)
- ny number of ys
- nx number of xs or nr of samples
- labels of variables