

多溴联苯醚气相色谱保留时间的模型预测与选择

报告人: 于淼

环境化学与生态毒理学国家重点实验室
中国科学院生态环境研究中心

2014 年 04 月 20 日

1 研究背景

- 环境问题
- 色谱定性原理

2 统计学习理论

- 统计建模理论
- 统计建模思想

3 PBDEs 气相色谱保留时间预测

- 问题描述
- 模型选择
- 网络应用



1 研究背景

- 环境问题
- 色谱定性原理

2 统计学习理论

- 统计建模理论
- 统计建模思想

3 PBDEs 气相色谱保留时间预测

- 问题描述
- 模型选择
- 网络应用



环境污染物的定性问题

污染物名称	种类
多氯联苯 (PCBs)	209
多溴联苯醚 (PBDEs)	209
单取代多氯联苯	837
单取代多溴联苯醚	837
短链氯化石蜡	...

- 环境污染物溯源
- 环境污染物迁移转化
- 污染物定性是环境过程研究的基础



1 研究背景

- 环境问题
- 色谱定性原理

2 统计学习理论

- 统计建模理论
- 统计建模思想

3 PBDEs 气相色谱保留时间预测

- 问题描述
- 模型选择
- 网络应用



从保留指数到模型

- 色谱定性依赖分离过程的分子间作用力
- 色谱柱与参考物描述定性范围
- 根据参考物的性质与其保留时间关系构建模型 $Y = f(x)$
- 根据模型与待测物的性质预测保留时间定性



线性溶剂能关系

线性溶剂能关系 (LSER) 模型就是将溶质分配状况与溶质极性、偶极矩、氢键及分子量进行多元线性回归的模型

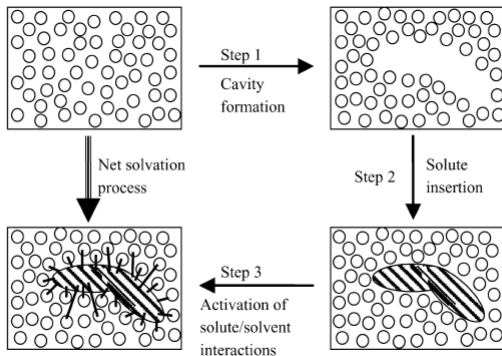
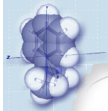


Fig. 5. Model of the solvation process: Step 1, solvent cavity formation. Step 2, solute insertion. Step 3, activating solute/solvent interactions.

(Mark Vitea et al. 2006)



分子结构的计算描述



Home
About
Partners
Software
Articles
Servers
Download
Web Services
How to cite?
Contact

Virtual Computational Chemistry Laboratory

[Home](#) [About](#) [Partners](#) [Software](#) [Articles](#) [Servers](#) [Download](#) [Web Services](#) [How to cite?](#) [Contact](#)

Welcome to the E-Dragon home page!

start the program

[mirror connection](#)

E-DRAGON can analyse max 149 molecules and max 150 atoms per molecule. Current version is Dragon 5.4 from 28 March 2006.

E-DRAGON is the electronic remote version of the well known software **DRAGON**, which is an application for the calculation of molecular descriptors developed by the **Milano Chemometrics and QSAR Research Group** of Prof. R. Todeschini. These descriptors can be used to evaluate molecular structure-activity or structure-property relationships, as well as for similarity analysis and highthroughput screening of molecule databases.

DRAGON provides more than 1,600 molecular descriptors that are divided into 20 logical blocks. The user can calculate not only the simplest atom type, functional group and fragment counts, but also several topological and geometrical descriptors. The first release of DRAGON dates back to 1997. Updates and inclusions of new molecular descriptors are regularly made in order to advance research in QSAR.

To run DRAGON the user needs molecular structure files previously obtained by other specific molecular modelling software. The most common molecular file formats are accepted. In E-DRAGON the accepted molecular structure files SMILES, SDF (MDL) or MOL2 (Sybyl) files. DRAGON requires 3D optimised structures with **HYDROGENS**. In E-DRAGON, if the 3D atom coordinates are not available for molecules, the user can calculate them on-line using **CORINA**, provided by **Molecular Networks GMBH**.

E-Dragon was developed as a result of collaboration between Dr. Tetko, Prof. Todeschini's and Prof. Gasteiger's teams.

<http://www.vcclab.org>

ON-LINE SOFTWARE

ALOGPS 2.1

ASNN

E-BABEL

PNN

PCLIENT

E-DRAGON 1.0

PLS

UFS

SPC

- 分子描述符的计算获取已不再困难
- 实验数据的获取仍主要依赖文献



多元线性回归的局限性

- 高维诅咒
 - 样本数少于变量数
 - 方程不可解
 - 数据空间稀疏
- 变量选择
 - 全局最优解计算上不合算
 - 局部最优解
- 模型选择
 - 理论模型存在共线性
 - 黑箱模型无法对理论探索提供线索



1 研究背景

- 环境问题
- 色谱定性原理

2 统计学习理论

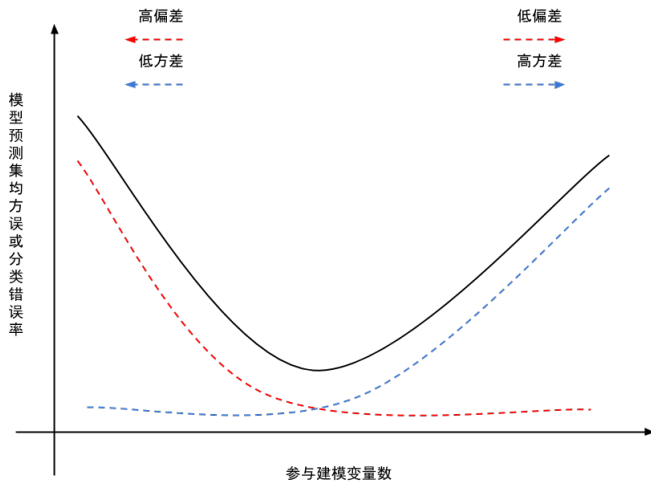
- 统计建模理论
- 统计建模思想

3 PBDEs 气相色谱保留时间预测

- 问题描述
- 模型选择
- 网络应用



Bias-Variance Tradeoff



1 研究背景

- 环境问题
- 色谱定性原理

2 统计学习理论

- 统计建模理论
- 统计建模思想

3 PBDEs 气相色谱保留时间预测

- 问题描述
- 模型选择
- 网络应用



统计建模思想

- ① 数据集分为训练集与检测集
- ② 模型参数要通过在训练集上进行交叉检验确定
- ③ 过拟合可通过模型惩罚项来降低
- ④ 数据规律可通过模拟来探索或验证
- ⑤ 模型可根据研究目的分层或嵌套



1 研究背景

- 环境问题
- 色谱定性原理

2 统计学习理论

- 统计建模理论
- 统计建模思想

3 PBDEs 气相色谱保留时间预测

- 问题描述
- 模型选择
- 网络应用



问题描述

- 目的：构建 PBDEs 气相色谱保留时间预测模型并探索变量间关系
- 数据获取
 - 209 种 PBDEs 分子结构经构型优化后通过 E-dragon 计算分子描述符
 - 删除常量分子描述符
 - 实验数据来自文献 (Wei hua *et al.* 2010) 中 180 种 PBDEs 在 DB-5ms 柱的色谱保留时间
 - 数据标准化处理
- 数据分割
 - 随机选取 150 个分子作为训练集参与建模
 - 模型效果通过剩余 30 个数据验证



1 研究背景

- 环境问题
- 色谱定性原理

2 统计学习理论

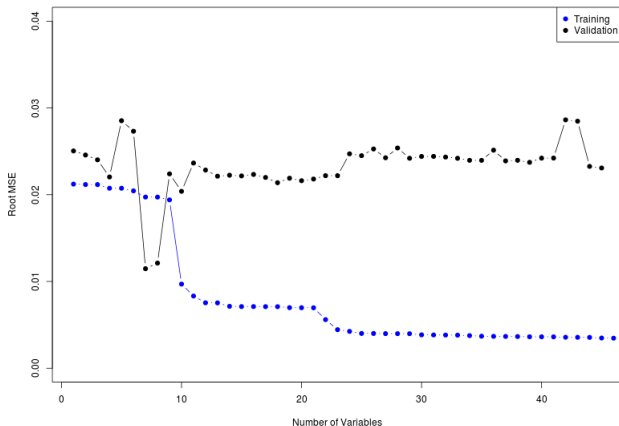
- 统计建模理论
- 统计建模思想

3 PBDEs 气相色谱保留时间预测

- 问题描述
- **模型选择**
- 网络应用



多元线性回归



● 过拟合严重且交叉检验失效



- 岭回归

- 不同于最小二乘估计对 RSS 的最小化, 岭回归最小化 $RSS + \lambda \sum_{j=1}^p \beta_j^2$, 其中 λ 为调谐参数, 后面一项为收缩惩罚, 是个 L_2 范数, 使参数估计逼近 0, 选择合适 λ 很重要, 可用交叉检验来实现
- 岭回归的调谐参数与范数收缩状况可看作最小均方误的函数来表现 bias-variance tradeoff
- 岭回归适用于全变量参与建模的情况



lasso 与岭回归

- 岭回归

- 不同于最小二乘估计对 RSS 的最小化, 岭回归最小化 $RSS + \lambda \sum_{j=1}^p \beta_j^2$, 其中 λ 为调谐参数, 后面一项为收缩惩罚, 是个 L_2 范数, 使参数估计逼近 0, 选择合适 λ 很重要, 可用交叉检验来实现
- 岭回归的调谐参数与范数收缩状况可看作最小均方误的函数来表现 bias-variance tradeoff
- 岭回归适用于全变量参与建模的情况

- Lasso

- Lasso 最小化 $RSS + \lambda \sum_{j=1}^p |\beta_j|$, 收缩惩罚为 L_1 范数, λ 亦可用交叉检验来实现
- Lasso 可以通过软边界直接收缩到 0 实现变量选择



lasso 与岭回归

● 岭回归

- 不同于最小二乘估计对 RSS 的最小化, 岭回归最小化 $RSS + \lambda \sum_{j=1}^p \beta_j^2$, 其中 λ 为调谐参数, 后面一项为收缩惩罚, 是个 L_2 范数, 使参数估计逼近 0, 选择合适 λ 很重要, 可用交叉检验来实现
- 岭回归的调谐参数与范数收缩状况可看作最小均方误的函数来表现 bias-variance tradeoff
- 岭回归适用于全变量参与建模的情况

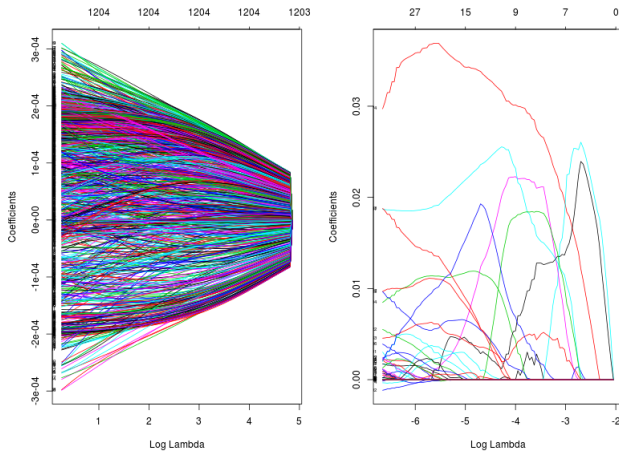
● Lasso

- Lasso 最小化 $RSS + \lambda \sum_{j=1}^p |\beta_j|$, 收缩惩罚为 L_1 范数, λ 亦可用交叉检验来实现
- Lasso 可以通过软边界直接收缩到 0 实现变量选择

● 贝叶斯视角下, 岭回归与 lasso 关于线性模型系数的先验分布是不同的

- 岭回归为高斯分布, 接近 0 时平坦, 后验概率等同最优解
- Lasso 为拉普拉斯分布, 接近 0 时尖锐, 先验概率系数接近 0, 后验概率不一定为稀疏向量

lasso 与岭回归



主成分回归与偏最小二乘回归

- 主成分回归

- 前提是自变量间不独立，将 p 个自变量向量投影到 M 维空间 ($M < p$)
- 各自变量在主成分方向上方差最大
- 实际为无监督算法，得到主成分后作为新变量进行最小二乘回归，主成分个数的选择影响模型效果
- 岭回归疑似为主成分回归的连续版，两者都需要标准化，效果也相近



主成分回归与偏最小二乘回归

- 主成分回归

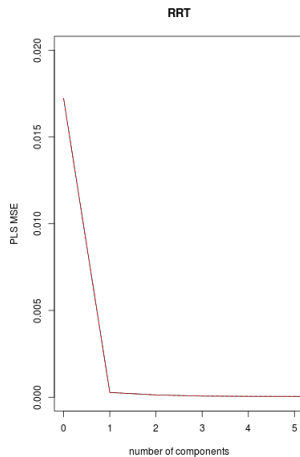
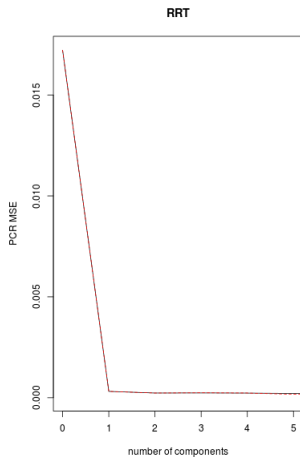
- 前提是自变量间不独立，将 p 个自变量向量投影到 M 维空间 ($M < p$)
- 各自变量在主成分方向上方差最大
- 实际为无监督算法，得到主成分后作为新变量进行最小二乘回归，主成分个数的选择影响模型效果
- 岭回归疑似为主成分回归的连续版，两者都需要标准化，效果也相近

- 偏最小二乘回归

- 第一个投影方向为因变量与自变量回归方向，后续投影是对残差投影方向的回归，重复得到监督学习的效果
- 化学学科常用



主成分回归与偏最小二乘回归

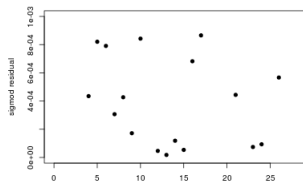
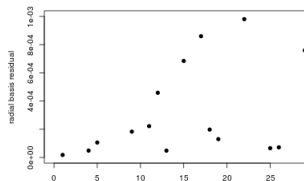
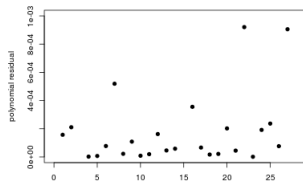
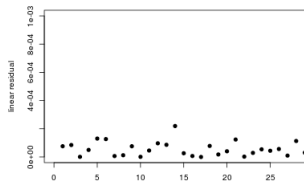


支持向量机回归

- 超平面: p 维空间里 $(p-1)$ 维子空间
- 支持向量: 形成最大边界分类器所需要的边界点, 用以支持最大边界超平面
- 支持向量分类器: 有些情况不存在超平面, 需要求一个软边界来适配最多的分类
- 回归是分类的连续版
- 使用内积的核函数计算上简单且等价与高维空间超平面分类
- 不同核函数暗示不同的变量响应关系



支持向量机回归



1 研究背景

- 环境问题
- 色谱定性原理

2 统计学习理论

- 统计建模理论
- 统计建模思想

3 PBDEs 气相色谱保留时间预测

- 问题描述
- 模型选择
- 网络应用



```
# install.packages('shiny')  
library(shiny)  
# you need connect to the internet  
runGitHub("shinyBDE", "yufree")
```

Database of PBDEs' RRTs on DB-5ms


Select the index of PBDEs congener and the prediction model


Index of PBDEs:

1 47 209

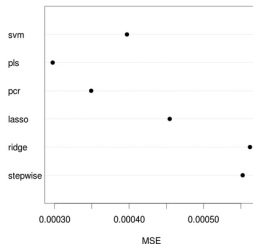
Choose a model to predict RRT

Partial least squares regression

 shiny is a product of RStudio

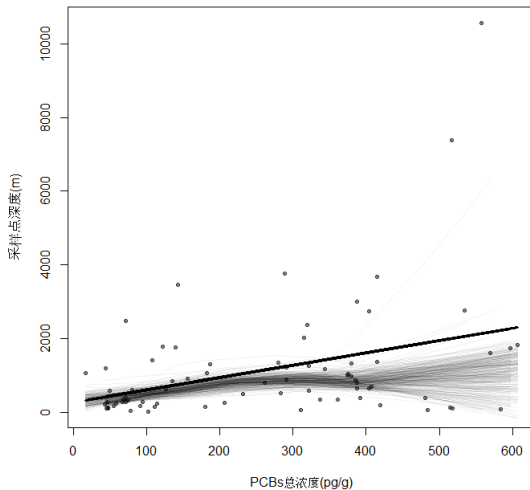
 shinyBDE is a product of Miao YU

The RRT of BDE-47 is 0.3958 and in Wei's paper is 0.398



普吉特海湾 PCBs 数据

双变量关系探索



总结与展望

- 统计学习模型对于色谱理论探索性研究有较好的指导性
- 以预测为目的的模型开发对于实际问题的解决很有必要
- 环境科学研究应该充分利用网络资源与开放数据



总结与展望

- 统计学习模型对于色谱理论探索性研究有较好的指导性
- 以预测为目的的模型开发对于实际问题的解决很有必要
- 环境科学研究应该充分利用网络资源与开放数据



Thank You

