Selection of tuning parameter in sparse linear regression

Alkomiet Belal   Supervisor: Doc. Ing. Václav Šmídl, Ph.D

[1]Department of Mathematical Engineering
Faculty of Nuclear Sciences and Physical Engineering

[2]Institute of Information Theory and Automation
Academy of Science of the Czech Republic

November, 2016

Introduction and prupose.

- Linear regression model: We are concerned with the problem of determination of a source term of atmospheric release of pollutant. Such problem can be viewed as a statistical solution of a linear regression equation : $y = Mx$

where: M sensitivity matrix $M \in R^{m \times n}$
y denotes observed data $(y_1, ..., y_m) \in R^m$
x is the unknown source term of the release $(x_1, ..., x_n) \in R^n$

- Variable selection: identifying the best subset among many variables to include in a model.
- Lasso : is a regression analysis method that performs both variable selection and regularization.

The objective of this work is to evaluate and compare three methods for selecting the optimal value of tuning parameterin terms of coefficient estimation accuracy and correct variable selection through simulation studies.

Outline

1. Validation.
   - ETEX experiment.
   - Cross Validation
   - L-curve
   - Variable selection stability
2. Our Results
   - L-curve
   - Cross Validation
   - Variable selection stability
   - Summary

Outline

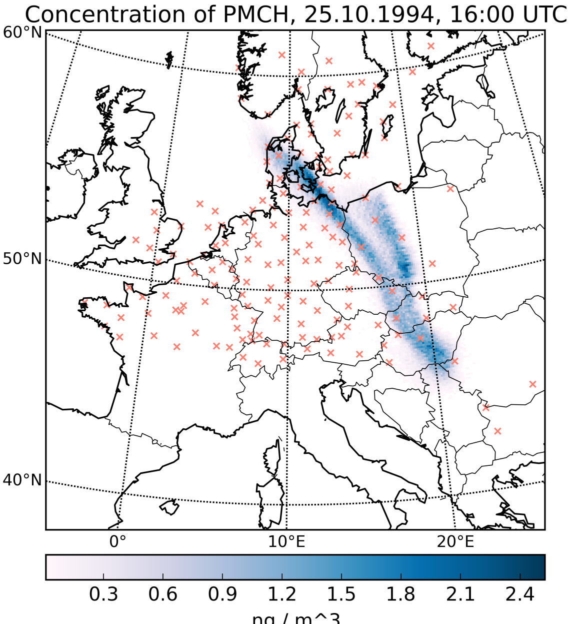
1. Validation.
   - ETEX experiment.
   - Cross Validation
   - L-curve
   - Variable selection stability

2. Our Results
   - L-curve
   - Cross Validation
   - Variable selection stability
   - Summary

The European tracer experiment (ETEX).

- Two releases of perfluorocarbon that took place in autumn of 1994 in north-western part of France.
- These releases were tracked across Europe using a network of 168 ground stations with limited airborne support .
- The aim of the experiment was to simulate an emergency response situation for meteorological modellers whose task was to create long-range dispersion prediction models in real time.

The European tracer experiment (ETEX).

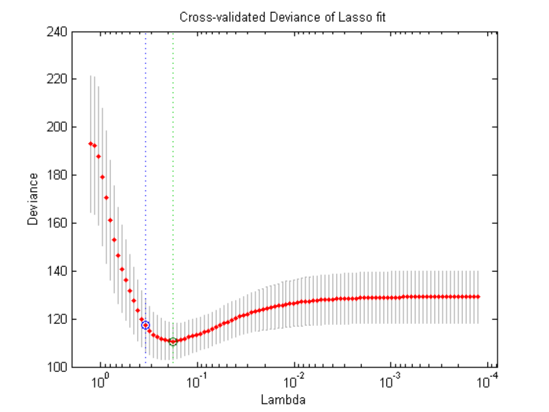


Outline

1. Validation.
   - ETEX experiment.
   - Cross Validation
   - L-curve
   - Variable selection stability
2. Our Results
   - L-curve
   - Cross Validation
   - Variable selection stability
   - Summary

Cross Validation.

- Primary method for estimating a tuning parameter λ.
- Divide the data into K roughly equal parts.



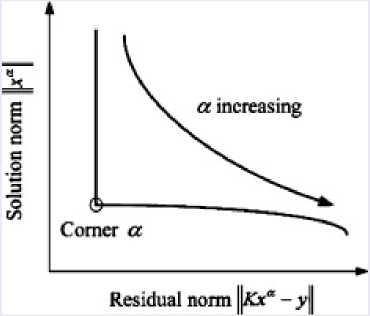
Outline

1. Validation.
    - ETEX experiment.
    - Cross Validation
    - L-curve
    - Variable selection stability

2. Our Results
    - L-curve
    - Cross Validation
    - Variable selection stability
    - Summary

L-curve

- The L-curve is a log-log plot of the norm of a regularized solution versus the norm of the corresponding residual norm.
- It is a convenient graphical tool for displaying the trade-off between the size of a regularized solution and its fit to the given data, as the regularization parameter varies.



α increasing
Corner α
Solution norm $\|x^\alpha\|$
Residual norm $\|Kx^\alpha - y\|$

Outline

1. Validation.
   - ETEX experiment.
   - Cross Validation
   - L-curve
   - Variable selection stability

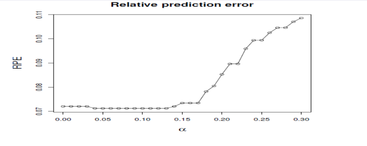
2. Our Results
   - L-curve
   - Cross Validation
   - Variable selection stability
   - Summary

Variable selection stability

- It is a tuning parameter selection criterion based on variable selection stability.
- The key idea is to select the tuning parameters so that the resultant penalized regression model is stable in variable selection.

$$\hat{\lambda} = min\left\{\lambda : \frac{\hat{S}(\Psi,\lambda,m)}{max_{\lambda\,'}\hat{S}(\Psi,\lambda\,',m)} \geq 1-\alpha n\right\} \qquad (1)$$

$\hat{S}(\Psi,\lambda,m) = \mathrm{B}^{-1}\sum_{\mathrm{b}=1}^{\mathrm{B}}\hat{S}^{*}(\Psi,\lambda,m)$: The average estimated variable selection stability .$\hat{S}^{*}(\Psi,\lambda,m) = k(A^{*b}_{1\lambda},A^{*b}_{2\lambda})$ the variable selection stability of $\Psi(:,\lambda)$

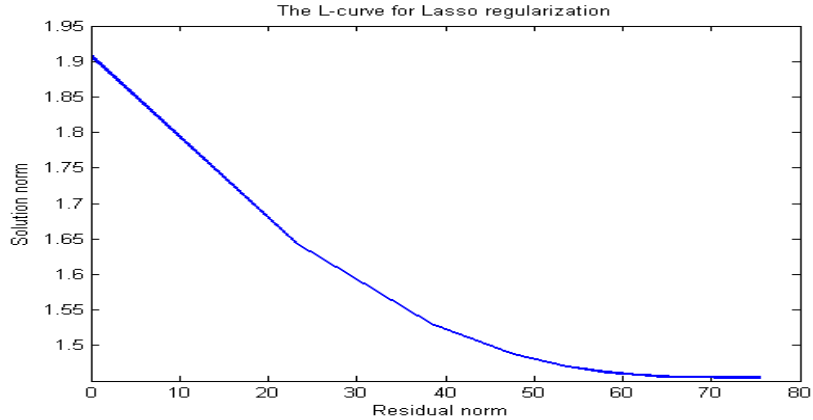


Outline

1. Validation.
    - ETEX experiment.
    - Cross Validation
    - L-curve
    - Variable selection stability
2. Our Results
    - L-curve
    - Cross Validation
    - Variable selection stability
    - Summary

L-curve

Figure:



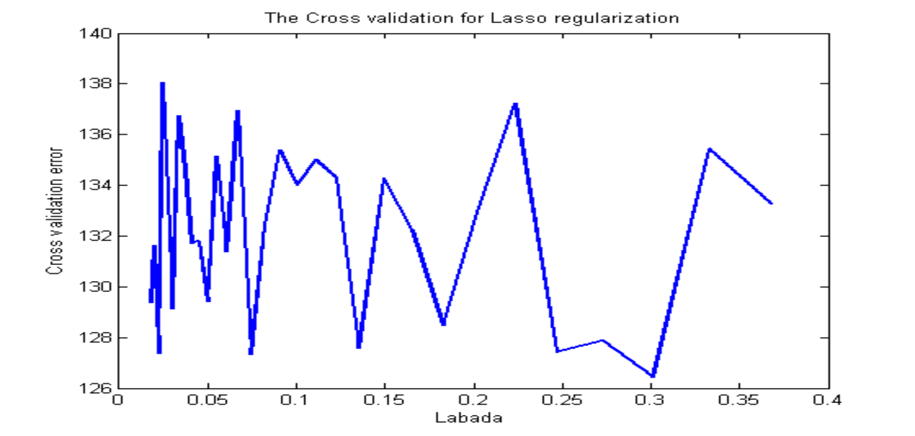
Outline

1. Validation.
   - ETEX experiment.
   - Cross Validation
   - L-curve
   - Variable selection stability

2. Our Results
   - L-curve
   - Cross Validation
   - Variable selection stability
   - Summary

Cross Validation

Figure:

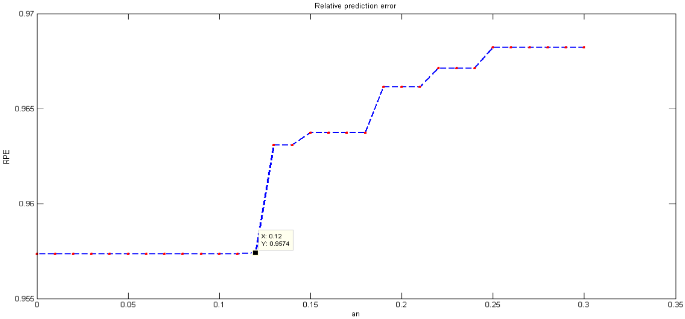


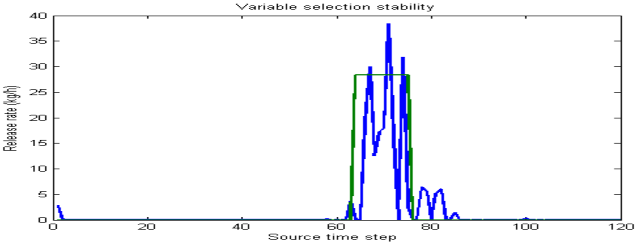
Outline

1. Validation.
   - ETEX experiment.
   - Cross Validation
   - L-curve
   - Variable selection stability

2. Our Results
   - L-curve
   - Cross Validation
   - Variable selection stability
   - Summary

Variable selection stability.





Outline

1. Validation.
   - ETEX experiment.
   - Cross Validation
   - L-curve
   - Variable selection stability

2. Our Results
   - L-curve
   - Cross Validation
   - Variable selection stability
   - Summary

Summary

- The L-curve criterion has its limitations, it does not work well when the solution is very smooth .
- While cross-validation can be computationally expensive in general, it is very easy and fast to compute, but It is also important to realise that it doesn't always work.if there are exact duplicate observations then leaving one observation out will not be effective. (still a topic of research).
- Variable selection stability is seen to give generally good results, but the question We have to investigate around. if there were a few variables of significantly large effects then any selection criterion selecting only these "big" variables would be stable !!

Thank you for your attention......