A Metalearning Study for Robust Regression

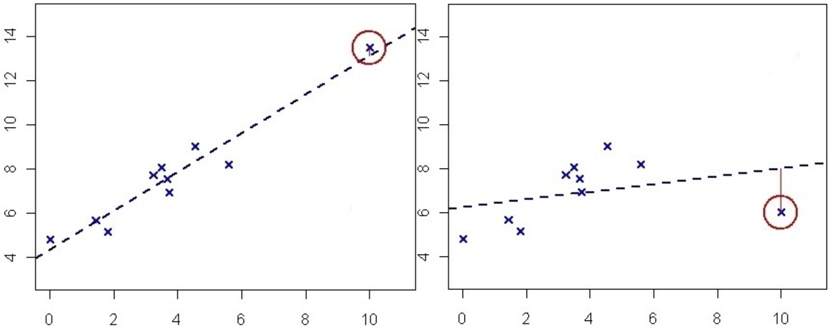
Jan Kalina and Barbora Peštová

Institute of Computer Science of the Czech Academy of Sciences

A metalearning study for robust regression

- **Robustness**
- Principles of metalearning
- A standard study
- An advanced study

Outliers in linear regression



- Outliers vs. leverage points
- Outlier detection: masking and swamping effects

Regression methods

- Parametric regression models
  - Linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + e_i, \quad i = 1, \ldots, n$$

  - Nonlinear regression model
  - Generalized linear models

- Nonparametric regression (regression curve estimation, function approximation)
  - Regression trees
  - Multilayer perceptrons
  - Support vector regression
  - Kernel-based methods (kernel estimation of regression curve)
  - Regularization networks

Advantages of parametric regression

- No overfitting,
- Comprehensibility,
- Diagnostic tools and remedies,
- Efficient computation,
- Modifications for a high dimension (LASSO),
- Modifications robust to outliers,
- Available hypothesis tests,
- Confidence interval for parameter estimates,
- Confidence band (region) for the whole regression curve (or line).

The concept of robustness

Robust statistics

- Sensitivity of standard methods
- Contaminated normal distribution
- Breakdown point
- Not robustness with respect to the model (data distribution)
- Robustification of standard methods
- Asymptotic theory
- Confusion with other robustness concepts (robust algorithm, robust against overfitting)
- Which robust method should be used?

1. Huber P.J. *Robust statistics.* Wiley, New York, 1981.
2. Hampel F.R., Rousseeuw P.J., Ronchetti E.M., Strahel W.A. *Robust Statistics: The approach based on influence functions*. Wiley, New York, 1986.
3. Rousseeuw P.J., Leroy A.M. *Robust regression and outlier detection.* Wiley, New York, 1987.
4. Jurečková J., Sen P.K., Picek J. *Methodology in robust and nonparametric statistics.* CRC Press, Boca Raton, 2013.

Regression M-estimators

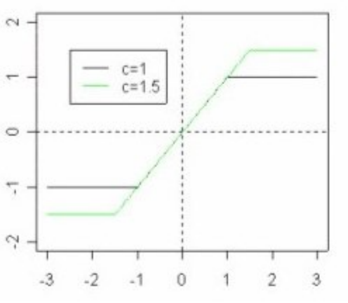
Recall least squares for the model $Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + e_i$:

$$\min \sum_{i=1}^{n} u_i^2 \Longleftrightarrow \sum_{i=1}^{n} \mathbf{X}_i u_i = \mathbf{0}, \quad \text{where } u_i = Y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}$$

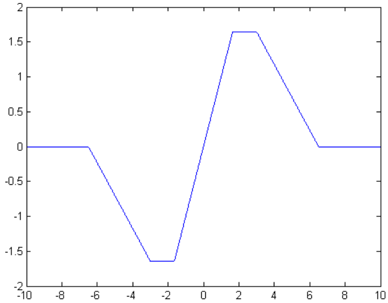
M-estimator:

$$\sum_{i=1}^{n} \mathbf{X}_i \psi(u_i) = \mathbf{0}$$

Huber's ψ:



Hampel's ψ:



Least trimmed squares (LTS)

- $$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + e_i, \quad i = 1, \ldots, n$$
- Residuals for a fixed $\mathbf{b} = (b_0, b_1, \ldots, b_p)^T \in \mathbb{R}^{p+1}$:

  $$u_i(b) = Y_i - b_0 - b_1 X_{i1} - \cdots - b_p X_{ip}, \quad i = 1, \ldots, n$$
- Squared residuals arranged in ascending order:

  $$u_{(1)}^2(\mathbf{b}) \leq u_{(2)}^2(\mathbf{b}) \leq \cdots \leq u_{(n)}^2(\mathbf{b}).$$
- $h =$ trimming constant
- LTS estimator

  $$\mathbf{b}_{LTS} = \arg\min \sum_{i=1}^{h} u_{(i)}^2(b) \quad \text{over} \quad \mathbf{b} = (b_0, b_1, \ldots, b_p)^T \in \mathbb{R}^{p+1}$$
- Properties

Least weighted squares regression (LWS)

Residuals for a fixed value of $\mathbf{b} = (b_1, \ldots, b_p)^T \in \mathbb{R}^p$:

$$u_i(\mathbf{b}) = y_i - b_1 X_{i1} - \cdots - b_p X_{ip}, \quad i = 1, \ldots, n.$$

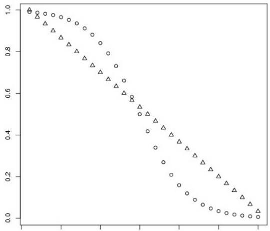
We arrange squared residuals in ascending order:

$$u^2_{(1)}(\mathbf{b}) \leq u^2_{(2)}(\mathbf{b}) \leq \cdots \leq u^2_{(n)}(\mathbf{b}).$$

The **least weighted squares** (LWS) estimator:

$$\mathbf{b}^{LWS} = \arg\min \sum_{i=1}^{n} w_i u^2_{(i)}(\mathbf{b}) \quad \text{over} \quad \mathbf{b} = (b_1, \ldots, b_p)^T \in \mathbb{R}^p,$$

where $w_1, \ldots, w_n$ are data-dependent (adaptive) weights, or fixed weights:



Least weighted squares regression (LWS)

Definition, basic properties, algorithm:

- Víšek J.Á. (2002): The least weighted squares I,II. *Bulletin of the Czech Econometric Society* 15/2002, 31 – 58; 16/2002, 1 – 28.
- Čížek P. (2011): Semiparametrically weighted robust estimation of regression models. *Computational Statistics and Data Analysis* **55**, 774 – 788.

- High efficiency for normal distribution.
- High breakdown point for contaminated normal distribution (high robustness against noise or influential outliers in the data).
- Local robustness (to small changes in the center of the data).

Original idea: Metalearning for linear regression estimators

- Which regression method is the most suitable for a particular data set?
- 24 publicly available data sets suitable for linear regression
- Our expectations

A metalearning study for robust regression

- Robustness
- **Principles of metalearning**
- A standard study
- An advanced study

1. Brazdil P., Giraud-Carrier C., Soares C., Vilalta E. (2009): *Metalearning: Applications to data mining*. Springer, Berlin.
2. Rice, J.R. (1976): The algorithm selection problem. Advances in Computers **15**, 65 – 118.
3. Smith-Miles K., Baatar D., Wreford B., Lewis R. (2014): Towards objective measures of algorithm performance across instance space. *Computers and Operations Research* **45**, 12 – 24.
4. Smith-Miles K.A. (2009): Cross-disciplinary perspectives on meta-learning for algorithm selection. *ACM Computing Surveys* **41**, Article 6.

Motivation and description

- Empirical approach for (black-box) comparing methods (classification, optimization)
- Lack of guidelines for **method selection**
- Which algorithm is likely to perform best for my problem?
- On which types of data sets does a method (algorithm) outperform its competitors?
- Why a method works on a particular data set? Which features are the most relevant?
- Attempt to generalize information across data sets
- A data set (instance) viewed as a point in a high-dimensional space
- Method selection is a learning (classification) task, learning to learn, metaknowledge
- Attempt to generalize information from other data sets
- Primary learning = base learning
- Learn prior knowledge from previously analyzed data sets and exploit it for a given data set

Framework P-A-F-Y-S (Smith-Miles, 2009)

- P: data sets
  - A small set
  - A too large number leads to overfitting (Brazdil et al., 2009)
  - Real data sets (simulated data sets are biased)
  - Some metadata publicly available
- A: algorithms
  - Fully automatic, including finding suitable parameters
- F: features of the data sets
  - How many
  - Relevant for the model selection
  - Their choice requires to understand the primary task
  - Examples of typical features
- Y: prediction measure
  - Should be computed using cross validation
- S: metalearning method over **metadata**
  - Methods: classification (k-NN with Euclidean distance, naïve Bayes), clustering, self-organizing maps, PCA, tree-like rules ...
  - Sometimes: ordering of methods, regression, prediction of performance

Advantages of metalearning

- Additional knowledge from previously analyzed data sets
- No theoretical analysis needed
- Clear, simple
- Comprehensible
- Feasible
- Popular in computer science

Limitations

- Crucial to find suitable features
- No discussion of robustness
- Not much can be said in general
- Particular tasks, tailor-made approaches
- Comparing particular versions of algorithms
- There should be many data sets
- Each method is a set of methods with various parameters, the approach requires many decisions
- Metametalearning

A metalearning study for robust regression

- Robustness
- Principles of metalearning
- **A standard study**
- An advanced study

Description of the standard study: P-A-F-Y-S

- P: data sets
  - 24 publicly available data sets (not a too small number)
  - Continuous response, continuous regressors
  - Clean & pre-processed data, missing values
- A: algorithms
  - Least squares, Huber's M-estimator, Hampels's M-estimator, LTS $(h = \lfloor 0.5n \rfloor,\ h = \lfloor 0.75n \rfloor)$
- F: features of the data sets
- Y: prediction measure
  - Mean square prediction error
    $$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,$$
    where
    $$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}, \quad i = 1, \ldots, n$$
  - Autovalidation
  - Leave-one-out cross validation
- S: metalearning method (classification to 5 groups)

Data sets and selected 9 features

Data set	Feature								
	1	2	3	4	5	6	7	8	9
	n	*p*						Outliers	
Aircraft	23	4	0.17	0.69	0.21	3.02	0.88	0.04	0.07
Ammonia	21	3	0.14	0.82	−0.19	3.11	0.91	0	0.18
Auto MPG	392	4	0.01	0	0.71	4.05	0.71	0.03	0
Cirrhosis	46	4	0.09	0.11	−0.21	2.07	0.81	0	0.61
Coleman	20	5	0.25	0.15	0.51	5.09	0.91	0.05	0.33
Delivery	25	2	0.08	0.27	0.03	3.07	0.96	0.04	0.00
Education	50	3	0.06	0.93	0.26	2.71	0.59	0.02	0.00
Electricity	16	3	0.19	0.22	0.78	3.84	0.92	0.06	0.13
Employment	16	6	0.38	0.48	0.42	2.44	1.00	0	0.87
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Aim: exploit the knowledge for new data sets

Selected 9 features of the data sets

1. The number of observations n
2. The number of variables p
3. The ratio n/p
4. Normality of residuals (p-value of Shapiro-Wilk test)
5. Skewness of residuals
6. Kurtosis of residuals
7. Coefficient of determination R^2,
8. Percentage of outliers (estimated by the LTS) – important!
9. Heteroscedasticity (p-value of Breusch-Pagan test)

Results of primary learning

Data	Autovalidation					Leave-one-out				
set	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
Aircraft	1	3	2	5	4	5	3	4	1	2
Ammonia	1	3	2	5	4	5	3	4	1	2
Auto MPG	1	3	2	5	4	5	3	4	2	1
Cirrhosis	2	3	1	4	5	2.5	1	2.5	5	4
Coleman	1	2	4	5	3	1	2	4	5	3
Delivery	1	2	5	4	3	5	4	2	3	1
Education	1	3	2	5	4	5	1	3	4	2
Electricity	1	3	2	5	4	2	3	1	5	4
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

- Ranks according to the mean prediction error (possible ties)
- (1) Least squares
- (2) Huber's M-estimator
- (3) Hampels's M-estimator
- (4) LTS with $h = \lfloor 0.5n \rfloor$
- (5) LTS with $h = \lfloor 0.75n \rfloor$
- Leave-5-out: slightly different

Results of metalearning

Method	Autovalidation	Leave-one-out
LDA	0.67	0.29
SVM (linear)	0.71	0.38
SVM (polynomial)	0.58	**0.42**
SVM (radial)	0.58	**0.42**
SVM (sigmoid)	0.50	0.38
k-NN (k=1)	1.00	0.29
k-NN (k=3)	0.58	0.29
k-NN (k=5)	0.54	0.33

- Methods (and their principles):
    - LDA: linear discriminant analysis
    - SVM: support vector machine
    - *k*-NN: *k*-nearest neighbor
- Which variables are the most relevant?

A metalearning study for robust regression

- Robustness
- Principles of metalearning
- A standard study
- **An advanced study**

Description of the advanced study: P-A-F-Y-S

- P: data sets
  - Omit data sets with a too large n or p
  - 21 data sets
- A: algorithms
  - The same
  - Possibly reduce their number
- F: features of the data sets
  - Include outlyingness of X
- Y: prediction measure
  - Trimmed mean square prediction error (TMSPE) for a given h

$$\frac{1}{n}\sum_{i=1}^{h}(Y_i - \hat{Y}_i)^2$$

- S: metalearning method
  - Unprecedented interpretation
  - Only leave-one-out cross validation
  - Robust classification: MWCD-LDA requires more observations and assigns weights to individual observations

Primary learning: Results

	The best method		
Data set	MSE	TMSPE ($h = 0.9n$)	TMSPE ($h = 0.5n$)
Ammonia	4	4	5
Auto MPG	1	2	5
Cirrhosis	1	1	2
Delivery	4	3	2
Education	2	4	4
Electricity	4	2	4
⋮	⋮	⋮	⋮
Overall	LS	M-estimators	LTS

1	Least squares
2	Huber
3	Hampel
4	LTS ($h = \lfloor 0.75n \rfloor$)
5	LTS ($h = \lfloor 0.5n \rfloor$)

Selected 10 features of the data sets

1. The number of observations n,
2. The number of variables p,
3. The ratio n/p,
4. Normality of residuals (p-value of Shapiro-Wilk test),
5. Skewness,
6. Kurtosis,
7. Coefficient of determination R^2,
8. Percentage of outliers (estimated by the LTS)
9. Heteroscedasticity (p-value of Breusch-Pagan test)
10. Donoho-Stahel outlyingness measure of X

Selected 10 features of the data sets

Data set	Feature									
	1	2	3	4	5	6	7	8	9	10
	n	*p*								
1	23	4	0.17	0.69	0.21	3.02	0.88	0.04	0.07	0
2	46	4	0.09	0.11	−0.21	2.07	0.81	0	0.61	0
3	20	5	0.25	0.15	0.51	5.09	0.91	0.05	0.33	0
4	25	2	0.08	0.27	0.03	3.07	0.96	0.04	0.00	0.08
5	50	3	0.06	0.93	0.26	2.71	0.59	0.02	0.00	0
6	16	3	0.19	0.22	0.78	3.84	0.92	0.06	0.13	0
7	16	6	0.38	0.48	0.42	2.44	1.00	0	0.87	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Results of metalearning

- Different data from before
- Classification correctness in a leave-one-out cross validation study
- 5 groups (correctly classified data sets):
- LDA, SVM, k-NN (various methods are the best)
- 10 variables: noise prefered to signal
- No effect of standardization

Number of variables	MSE	TMSPE ($h = 0.9$)	TMSPE ($h = 0.5$)
10	0.38	0.43	0.33
9	0.38	0.52	0.33
8	0.43	0.48	0.33
7	0.48	**0.52**	0.29
6	0.48	0.48	0.33
5	0.48	0.43	0.29
4	0.48	0.33	0.33
3	0.48	0.43	0.38
2	0.48	0.43	0.38
1	**0.48**	0.33	**0.38**

Results of metalearning

Classification to 5 groups vs. 3 groups vs. 2 groups

Number of	MSE			TMSPE (0.9)			TMSPE (0.5)		
variables	5	3	2	5	3	2	5	3	2
10	0.38	0.57	0.43	0.43	0.57	0.67	0.33	0.62	0.62
9	0.38	0.57	0.57	0.52	0.71	0.71	0.33	0.62	0.76
8	0.43	0.57	0.67	0.48	0.76	0.76	0.33	0.71	0.86
7	0.48	0.62	0.67	**0.52**	**0.67**	0.81	0.29	0.67	0.86
6	0.48	0.67	0.67	0.48	0.76	0.86	0.33	0.76	0.76
5	0.48	0.71	0.67	0.43	0.67	0.81	0.29	**0.76**	0.81
4	0.48	0.71	0.67	0.33	0.67	0.81	0.33	0.67	0.81
3	0.48	0.71	**0.76**	0.43	0.67	0.86	0.38	0.71	0.86
2	0.48	**0.71**	0.71	0.43	**0.76**	**0.86**	0.38	0.71	**0.86**
1	**0.48**	0.57	0.71	0.33	0.67	0.81	**0.38**	0.71	0.71

- The best variables are considered. Which are these?
- Effect of robust prediction error
- Effect of reducing the number of groups
- PCA is suboptimal

A closer look at interpretation

- TMSPE, $h = 0.5n$
- 2 groups: LTS vs. rest (LS & M-estimators)
- The best single variable: Heteroscedasticity (p-value of Breusch-Pagan test)
- Classification performance with LDA $15/21 = 0.71$
- How LDA is performed?
- $p < 0.4 \quad \Longrightarrow \quad$ classify to LTS
- $p > 0.4 \quad \Longrightarrow \quad$ classify to rest
- Breusch-Pagan test sensitive to violations of normality, its p-value arbitrary due to data contamination

	Truly best method		
	LTS	Rest	$\sum$
$p < 0.4$	11	4	15
$p > 0.4$	2	4	6
$\sum$	13	8	21

A closer look at interpretation

- TMSPE, $h = 0.9n$
- 2 groups: LTS vs. rest (LS & M-estimators)
- The best single variable: Normality of residuals (p-value of Shapiro-Wilk test)
- Classification performance with LDA $17/21 = 0.81$
- How LDA is performed?
- $p > 0.695 \quad \Longrightarrow \quad$ classify to LTS
- $p < 0.695 \quad \Longrightarrow \quad$ classify to rest
- Unequal groups

	Truly best method		
	Rest	LTS	$\sum$
$p < 0.695$	12	3	15
$p > 0.695$	1	5	6
$\sum$	13	8	21

Appendix: Sensitivity of Metalearning

Description of the sensitivity study: P-A-F-Y-S

- P: data sets
  - 24 publicly available data sets
- A: algorithms
  - Least squares, Hampels's M-estimator, LTS ($h = \lfloor 0.75n \rfloor$), LWS with linear weights
- F: 9 features of the data sets
- Y: prediction measure
  - Mean square prediction error

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

  - Leave-one-out cross validation
- S: metalearning method (classification to 4 groups)

Data contamination

- Each measured value will be denoted as X_{ijk}
  - i corresponds to a particular data set
  - j to an observation within this data set
  - k to a particular variable
- Replace X_{ijk} by $X_{ijk} + \varepsilon_{ijk}$, where ε's are (mutually) independent random variables independent on the given data
  - ε_{ijk} is generated from normal distribution $\mathrm{N}(0, s\hat{\sigma}^2_{ijk})$
  - $\hat{\sigma}^2_{ijk}$ is an estimated variance of the j-th variable within the i-th data set
  - s is a chosen constant

1. Local contamination. Each observation in each data set is contaminated by a slight noise, i.e. with a small s.
2. Global contamination. A small percentage of observations is contaminated by severe noise, while the remaining ones are retained. Particularly, $c \cdot 100$ % of the values are randomly chosen for each data set across all relevant features for a given (and rather large) s.

Results of primary learning

			Best method						
				Local contam.			Global contam.		
			Raw	with $s =$			with $s = 9$ and $c =$		
Data set		$\hat{\sigma}^2$	data	0.1	0.2	0.3	0.06	0.12	0.18
1	Aircraft	57.8	3	3	3	3	4	3	3
2	Ammonia	8.9	4	4	3	3	4	4	4
3	Auto MPG	17.9	3	3	3	3	3	4	3
4	Cirrhosis	103	1	2	3	3	1	3	3
5	Coleman	3.2	1	1	1	1	1	1	1
6	Delivery	9.7	2	3	3	2	3	3	3
7	Education	1537	2	2	2	3	2	4	3
8	Electricity	0.85	2	2	2	2	2	2	4
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

- (1) least squares,
- (2) Hampel's M-estimator,
- (3) LTS with $h = \lfloor 0.75n \rfloor$,
- (4) LWS with linearly decreasing weights

Results of metalearning

	Best method						
		Local contam.			Global contam.		
Classification	Raw		with $s =$		with $s = 9$ and $c =$		
Method	data	$s = 0.1$	$s = 0.2$	$s = 0.3$	0.06	0.12	0.18
SVM (linear)	**0.38**	**0.38**	**0.38**	**0.38**	**0.38**	**0.38**	**0.38**
LDA	0.29	0.29	0.29	0.25	0.17	0.29	**0.38**
MWCD-LDA	0.33	0.33	0.33	0.33	0.29	0.33	0.33
k-NN (k=1)	0.29	0.25	0.21	0.25	0.29	0.33	0.29
k-NN (k=3)	0.29	0.29	0.25	0.25	0.33	0.29	0.25
k-NN (k=5)	0.33	0.33	0.33	0.29	**0.38**	0.33	**0.38**

Results of leave-one-out cross validation:

- The LWS estimator (with only simple weights) turns out to be the best method for some data sets, which is a novel argument in favor of the method.
- MWCD-LDA together with SVM classifier are the only methods not mislead by the contamination.

What contributes to the sensitivity of metalearning

- The problem itself is unstable and the whole process should be robustified
- The choice of (very different) data sets.
- Difficult (and unreliable) extrapolation for a very different (outlying) data set.
- The prediction measure. In our case, PMSE is very vulnerable to outliers.
- The number of algorithms/methods. If their number is larger than very small, we have the experience that learning the classification rule becomes much more complicated and less reliable.
- The classification methods for the metalearning task depend on their own parameters or selected approach, which is another source of uncertainty and thus instability.
- Solving the metalearning method (S) by classification tools increases the vulnerability as well, because only the best regression estimator is chosen ignoring information about the performance of other estimators.
- Model selection is unstable.
- The process of metalearning itself is too automatic so the influence of outliers is propagated throughout the process and the user cannot manually perform an outlier detection or deletion.

Limitations of metalearning

- Fully automatic approach would not find the reason for the over-optimistic results (black-box)
- Choice of data sets
- Various dimensionality
- Features
  - How many (e.g. p-value depends on n)
  - Relevant ones are typically ignored
- Number of algorithms (methods)
- Association vs. causality
- Some classifiers for the metalearning depend on their own parameters or selected approach (e.g. Naïve Bayes)
- Bad extrapolation for a very different data set
- Perhaps for only a specific task

Future work

Robustification of metalearning:

- Particular task: Extraction of rules (which itself is very unstable)
- Regression (perhaps ordinal regression)
- Some classifiers give also ranking of methods
- Use the whole vector of ranks
- Estimate the prediction performance
- Ensemble classification improves stability and robustness to noise
- Ensembles can be viewed actually as metalearning
- Robustness was introduced to learning by Breiman

$\Longrightarrow$ THANK YOU FOR YOUR ATTENTION $\Longleftarrow$