

SIMULTANEOUS ROBUST ESTIMATION AND VARIABLE SELECTION

Lauren McCann
Roy Welsch

Massachusetts Institute of Technology

email: rwelsch@mit.edu

30 March 2006

1

Background – Model

$$y = X\beta + \varepsilon$$

n observations (rows)

p variables (columns)

- p could be greater than n
- What are the important variables (features)?
- Explanation versus prediction
- What about “bad” data?
- Robustly center and scale (median and MAD)?

2

Background – Outliers

- Vertical (or additive) – (x_i, y_i) does not follow linear pattern of the majority, but x_i not outlying
- Leverage point – x_i is outlying
 - good if (x_i, y_i) follows majority
 - bad if (x_i, y_i) does not
- Regular – none of the above
- Should we always address all outlier problems at once? Local models?
- How bad is real life?

3

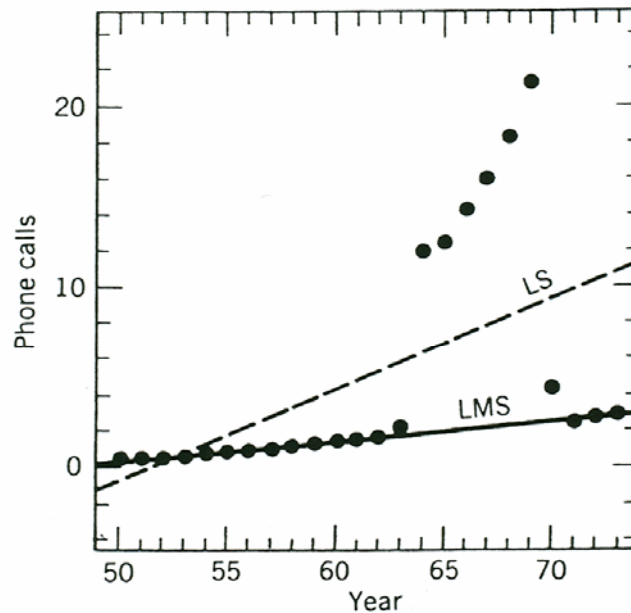
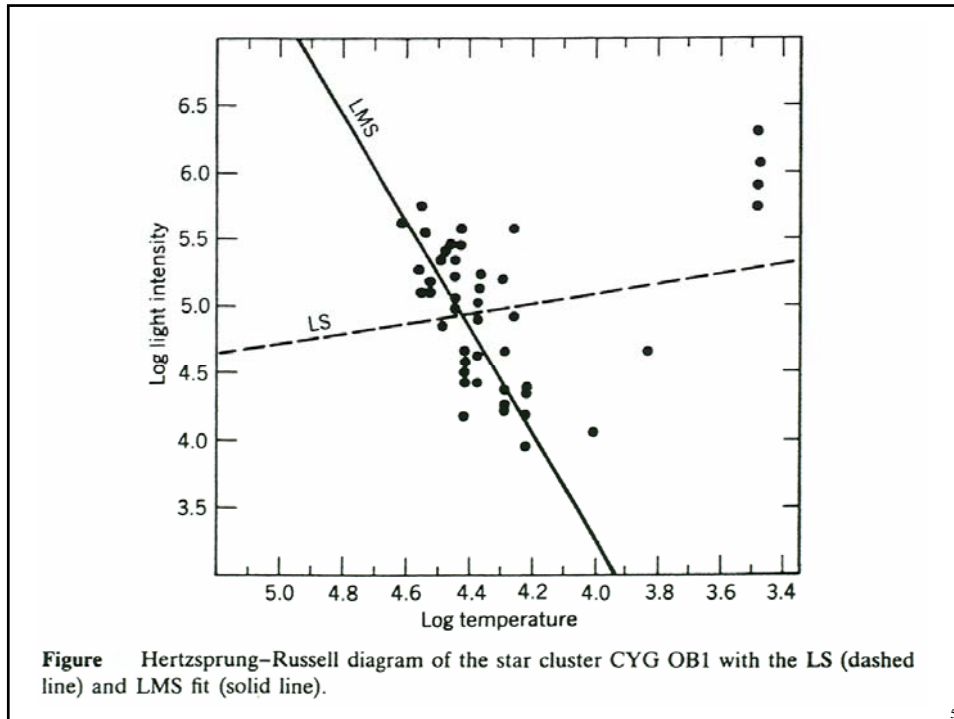


Figure Number of international phone calls from Belgium in the years 1950–1973 with the LS (dashed line) and LMS fit (solid line).

4



Background – Robust

- M-estimation for additive outliers
- Bounded-influence – addresses leverage, but breakdown like $1/p$.
- MM-estimation – efficient and high breakdown, but needs a high breakdown start like LTS.
- Least trimmed squares (LTS)

Let $r^2_{(i)}$ be ordered squared residuals. Find

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^h r^2_{(i)}(\boldsymbol{\beta})$$

with breakdown about $(n-h)/n$.

Background – LTS

- LTS hard to compute
- Randomly find many subsets of size p (with centered data). Get coefficients from LS fit.
- Choose coefficients that give lowest LTS.
- Not a global optimum, but works quite well in practice.
- Fast version available (Rousseuw and Van Driessen, 2000), uses concentration steps.
- Not perfect, see Hawkins and Olive (2002).

7

Background – Selection

- All possible subsets – each one robustly fit but very expensive computationally.
- “Robustify” C_p , etc. (Within sample prediction error.)
- Forward or backward robust search (S+).
- Bounded-influence with selection by cross-validation (Ronchetti, Field, and Blanchard, 1997).
- Can we get close to all possible subsets and still have high breakdown?
 - row sampling required?
 - column sampling?

8

Background – Machine Learning

- Random Forests (RF) – build optimal models (trees) but choose random subsets of (splitting) variables
- Bagging (bootstrap aggregation) – average over models fit on subsamples
- Boosting – iterative improvement
- Stability – continuous dependence on the data
- Training, validation, and test data sets (samples)

9

Background – Regularization (Penalties)

- LS

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2$$

- LASSO (shrink to δ) – hard zeros, if $\delta_j = 0$.

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^p |\beta_j - \delta_j|$$

- RIDGE ($p > n$ not a problem)

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^p (\beta_j - \delta_j)^2$$

10

Background – Regularization (Penalties)

- Elastic net – both L1 and L2 penalties to get hard zeros and $p > n$
- Other loss functions, etc.

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^n \rho(y_i - \mathbf{x}_i^T \boldsymbol{\beta})$$

11

Background – LARS

- “LAR(S) builds a regression model in piecewise linear forward steps, assessing explanatory variables one at a time; each step is taken along the equiangular direction between the set of explanators. The step size is less greedy than classical forward stepwise regression, smoothly blending in new variables rather than adding them discontinuously.” (Efron, et. al. 2004)
- LARS is a fast (same order as LS) forward selection algorithm that can be used, for example, for the L1 penalty (LASSO).

12

Plan – Rows

- Draw random row samples of size $p = \max$ number of variables.
- Draw random row samples of size $p + ?$
- Use fast LTS with limited concentration steps on full model, retain $k (\geq p)$ rows with smallest residuals. Compare to Atkinson and Riani (2000) forward (row) search.
- Other?

13

Plan – Columns

- Random sample of columns (“random forests”). Fit by LS. Judge quality of model by robust prediction scores (MAD) on validation sets. Retain M “best” models for further analysis. We use the bagging idea of selecting a variable if it appears 50% (or, perhaps, more) of the time in the M models.
- Fast forward search via LARS. Generates p models (1, 2, . . . , p variables). Choose and retain models as above. Modification uses the best result for each model from LARS fit or LS fit. If happen to have good subset of the data, why not use LS?

14

Plan – Variables to Keep

- Given M best models need to decide which variables to retain. Count number of times a variable appears in M models. If no variables needed, this count is binomial with parameters M and $1/2$. Use the Bonferroni inequality to account for multiple comparisons and develop a critical value to decide on retention.
- For M best models, compute for each variable

$$\frac{|\text{median}(b_j)|}{\text{mad}(b_j)}$$

and retain the j th variable, if this value is large.

- Model with lowest average prediction.

15

Plan – “Boosting”

- Penalty methods shrink coefficient estimates toward the “prior,” δ . Given the many subsamples for a high breakdown (or even a randomly cross-validated non high breakdown method), we can use an initial set of samples to determine if the prior might not be zero (or δ) and set it to a revised value, say median (b_j) , where the median is taken over the subsamples in the initial set and

$$\frac{|\text{median}(b_j)|}{\text{mad}(b_j)}$$

exceeds some threshold. Otherwise, the prior remains at zero.

- This could be iterated more than once until average prediction score for the best model fails to improve.

16

Benchmarks

We used examples contained in the Ronchetti, et. al. (1997) paper for simulation comparisons.

Error distributions:

- e1 standard normal
- e2 93% standard normal, 7% $N(0, 5)$
- e3 slash
- e4 90% standard normal, 10% $N(30, 1)$

Explanatory variables:

- without leverage all columns generated from uniform [0,1]
- leverage like uniform, but two leverage points – 3s in all columns but the first and another with 5s in all columns but the first

17

Structural models:

All models had six variables (median centered and MAD scaled) plus the intercept with five non-zero coefficients (t value about 6) and $n = 60$.

LS-CV refers to results using least squares with all possible subsets and cross-validation (Shao, 1993).

BIF-CV refers to results obtain by Ronchetti, et. al. (1997) using bounded influence estimation and all possible subsets with cross-validation.

For each case, we used 200 simulation runs.

18

Other Ideas

- To address additive (vertical outliers) adjoin $n \times n$ identity matrix to $n \times p$ \mathbf{X} matrix (DV). Selection method can then choose real explanatory variables as well as rows to delete. Without external row sampling, this is not high breakdown. Since $n + p > n$, penalty methods essential.
 - Can be combined with “boosting” priors (DVB).
 - Row sampling for high breakdown? Remove dummy columns with all zeros.

19

LARS Simulations

LARSD-T	Dummy variables added ($n/4$), t -stat (.05/6) to select real variables.
LARSD-T2	Like LARSD-T but t -stat also used to select dummies to keep.
LARS-S6-CV	Lets LARS select variables on (2000) samples of size 6. Cross-validated with MAD prediction on remaining samples. Total models now 12,000.
LARS-S8-CV	Like LARS-S6-CV but samples of size 8.
LARSD-S6-CV	Like LARS-S6-CV but dummy variables added.
LARSD-S8-CV	Like LARSD-S6-CV but samples of size 8.

20

Other Simulations

- RF6 Columns chosen randomly (random forests) with row samples of size p , the maximum number of variables (= 6).
- RF15 Like RF6 but 15 row samples used.
- DVNS Dummy variables with no sampling. t -statistics used with ridge regression.
- DVBNS Like DVNS, but first pass used to get some prior information for parameter shrinkage.

21

LARS Results

Method	Without Leverage				Leverage				
	e1	e2	e3	e4	e1	e2	e3	e4	
LS-CV	188	47	0	3	186	44	0	1	
BIF-CV	157	160	10	167	169	166	7	172	
LARSD-T	178	167	28	152	178	186	41	154	*
LARSD-T2	185	181	12	150	187	182	9	154	
LARS-S6-CV	132	128	8	124	167	143	12	157	
LARS-S8-CV	169	156	10	156	182	174	24	175	*
LARSD-S6-CV	130	122	3	109	167	152	14	155	
LARSD-S8-CV	161	149	17	140	172	168	36	168	

22

Other Results

	Without Leverage				With Leverage			
	e1	e2	e3	e4	e1	e2	e3	e4
LS-CV	188	47	0	3	186	44	0	1
BIF-CV	157	160	10	167	169	166	7	172
DVNS	164	166	10	141	80	86	45	66
DVBNS	178	180	18	168	114	129	62	107
RF6	76	54	0	62	92	74	1	97
RF15	148	138	3	120	176	173	13	159

23

The Khan-VanAelst-Zamar Method

- LARS uses the correlation matrix and is therefore not robust. Khan et.al. (2005) propose replacing the correlation matrix with a (fast) robust version. They considered several Winsorized approaches for correlation estimation – plug-in (P) and cleaning (C).
- Doing this first requires treating outliers before selection, something we are trying to avoid, if possible.

24

KVZ Simulation

- KVZ use the Ronchetti et.al. simulation design with the same error distributions and non-leverage design matrix. The leverage case has just one bad row – (5,5,3,3,3,3). There are only three non-zero coefficients with values 7, 5, 3.
- The performance measures are:
 Exact (E) – percentage of times non-zero variables are chosen first and in their true order (7,5,3).
 Global (G) – percentage of times non-zero variables chosen first, but any order is allowed.

25

KVZ Comparisons

Method		Without Leverage				Leverage				
		e1	e2	e3	e4	e1	e2	e3	e4	
LARS	E	97	86	11	8	0	1	1	2	
LARS	G	100	89	26	24	0	2	5	7	
WP	E	96	97	58	78	92	85	46	59	
WP	G	99	99	77	89	94	86	61	68	
WC	E	96	98	54	82	96	94	52	83	
WC	G	99	99	76	92	98	96	71	92	
LARSD-T	E	95	95	61	83	96	95	65	85	*
LARSD-T	G	100	100	80	88	100	99	85	91	*

26

KVZ Conclusions

- LARSD-T (dummy variables and t -statistics for selection) works well and makes some gains compared to KVZ for slash (e3) and additive outlier (e4) cases.
- LARSD-T is a simple way to achieve robust selection where leverage outliers are not a major concern.
- Sparse nature of the dummy variable matrix adds little computational complexity to LARS itself.

27

Conclusions

- No universal solution, yet.
- How to add back data from the start?
- Good prediction vs. selection.
- Fast selection methods with feature weights built in – TREES, SVM.
- Statistical efficiency issues.

28

References

- Atkinson, A. C. and Riani, M., (2000) *Robust Diagnostic Regression Analysis*, Springer Verlag, New York.
- Breiman, L., (1996) Bagging predictors, *Machine Learning*, **24**(2), 123-140.
- Breiman, L., (2001) Random forests, *Machine Learning*, **45**(1), 5-32.
- Duan, K-B., Rajapakse, J., Azuaje, F., Multiple SVM-RFE for Gene Selection in Cancer Classification with Expression Data, *IEEE Transactions on Nanobioscience*, **4** (3), 228-234.
- Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., (2003) Least Angle Regression, *Annals of Statistics*, **32**, 407-499.
- Hawkins, D.M. and Olive, D.J., (2002) Inconsistency of Resampling Algorithms for High-Breakdown Regression Estimators and a New Algorithm, *Journal of the American Statistical Association*, **97**, 136-159.

29

- Morgenthaler, S., Welsch, R.E. and Zenide, A., (2004) Algorithms for Robust Model Selection in Linear Regression, *Theory and Applications of Recent Robust Methods*, edited by M. Hubert, G. Pison, A. Struyf and S. Van Aelst, Series: Statistics for Industry and Technology, Birkhauser, Basel, 195–206.
- Ronchetti, E., Field, C. and Blanchard, W., (1997) Robust Linear Model Selection by Cross-Validation, *Journal of the American Statistical Association* **92**, 1017–1032.
- Rousseuw, P. J. and Van Driessen, K., (2000) An Algorithm for Positive-Breakdown Regression Based on Concentration Steps, in *Data Analysis: Scientific Modeling and Practical Application*, edited by W. Gaul, O. Opitz, and M. Schader, Springer Verlag, New York, 335-346.
- Shao, J., (1993) Linear Model Selection by Cross-Validation, *Journal of the American Statistical Association*, **88**, 486–494.
- Torkkola, K. and Tuv, E., (2005) Ensembles of Regularized Least Squares Classifiers for High-Dimensional Problems in *Feature Extraction, Foundations and Applications*, edited by I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, Springer, (in press).
- Zou, H. and Hastie, T., (2005) Regularization and variable selection via the elastic net, *J.R. Statist. Soc. B*, **67**, Part 2, 301-320.

30