Random Sample Reliability

Gabriel Okasa and Kenneth A. Younge
gabriel.okasa@epfl.ch kenneth.younge@epfl.ch

Introduction

Researchers frequently test and improve *model fit* by holding the sample constant and varying the model. We propose *Random Sample Reliability* (RSR) as a computational method to assess and improve *sample fit* by holding the model constant and varying the sample. RSR is a new method for re-sampling data to estimate reliability of the fit of observations within a sample for a given model. We propose an RSR-based method of data *annealing* to check the sensitivity of regression results across the least reliable observations in a sample, and we define an RSR-based method of weighted regression to combine *model fit* and *sample fit* into a more robust analysis.

Motivation

- Complement the model fit paradigm by acknowledging sample fit
- Combine robust approaches from statistics and machine learning
- Develop computational method for estimation of data reliability

RSR Method

- Employ re-sampling approach for reliability scoring of data [1]
- Apply weighting approach for robust regression estimation [2]
- Introduce annealing procedure for sensitivity analysis

Reliability Scoring

Algorithm 1: Random Sample Reliability (RSR)

Input: Data $Z{=}(X, Y)$, Samples S, Model $M(x)$, Loss $\mathcal{L}\big(y, M(x)\big)$
Output: Reliability Score $\hat{\psi}(x)$

begin
  for $s = 1$ **to** S **do**
    Sample Z_s^* of size $\dim(X) + 1$ w/o replacement;
    Estimate the model $M(x; Z_s^*)$ via OLS;
    Evaluate loss $\hat{\gamma}_s(x) = \mathcal{L}\big(y, \hat{M}(x; Z_s^*)\big)$ out-of-bag;
  end
  Average over the losses $\hat{\Gamma}^S(x) = \frac{1}{S}\sum_{s=1}^{S} \hat{\gamma}_s(x)$;
  Create reliability score $\hat{\psi}(x) = \frac{\hat{\Gamma}^S(x) - \max_i\{\hat{\Gamma}^S(x)\}}{\min_i\{\hat{\Gamma}^S(x)\} - \max_i\{\hat{\Gamma}^S(x)\}}$;
end

Robust Fitting

- Re-estimate the regression model weighted by reliability scores
- Apply bootstrapping for inference to reflect estimation uncertainty

Sensitivity Annealing

- Sort reliability scores in increasing order $\hat{\psi}_{(1)}(x) \leq \cdots \leq \hat{\psi}_{(N)}(x)$
- Re-estimate model by sequentially dropping the least reliable data
- Plot the coefficient path to assess the sensitivity to unreliable data

Simulation

- Stylized setting with synthetic data contaminated by outliers
- Check the properties of the RSR algorithm for reliability scoring
- Test the performance against OLS, RANSAC and Huber estimator

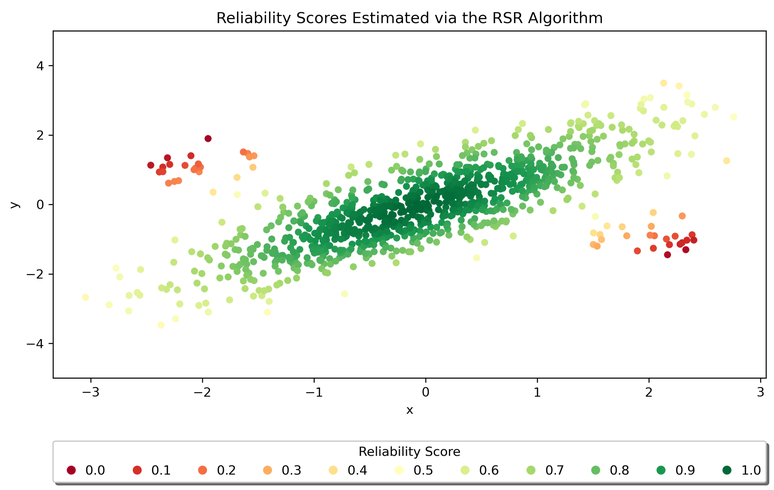

Figure 1: Reliability Scoring of Data Points.

Setting

- Linear DGP with symmetric distribution of outliers with leverage
- Estimation of the linear model of the form: $Y = \alpha + X\beta + \epsilon$

Results

	MSE	MAE	SD
OLS	0.0649	0.2538	0.0218
Huber	0.0137	0.1147	0.0229
RANSAC	0.0185	0.1068	0.1361
RSR	0.0042	0.0602	0.0243

Table 1: Simulation Results for Effect Estimation.

- Weighting by reliability scores effectively reduces estimation bias
- RSR weighted fit outperforms competing robust regression methods

Application

- Replication of field experiment on economics of charitable giving [3]
- Estimation of the effect of mail solicitations on donation amount
- Perform sensitivity analysis via the RSR annealing procedure

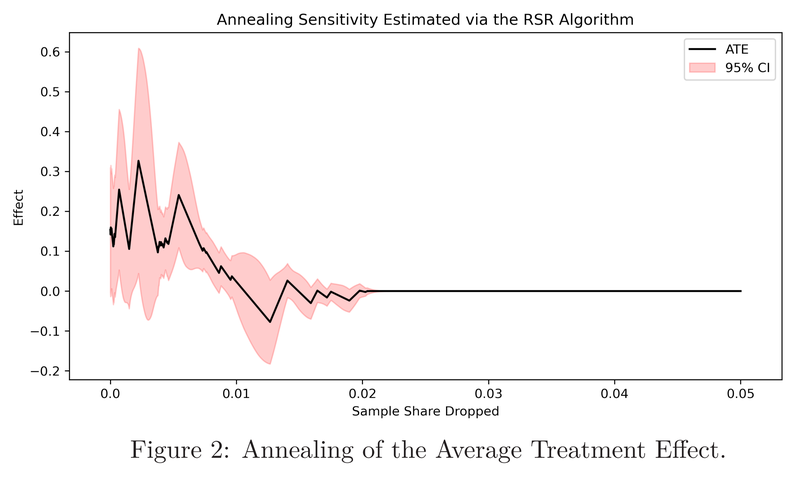

Figure 2: Annealing of the Average Treatment Effect.

Insights

- The average treatment effect is highly sensitive to unreliable data
- No evidence of an effect for the 98% of sample with high reliability
- Effect heterogeneity as potential explanation for the 2% least reliable

Code

- Python implementation of RSR provided in the samplefit library
- Available on GitHub: `https://github.com/okasag/samplefit`

References

[1] Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 24(6), 381–395.

[2] Huber, P. J. (1964). Robust Estimation of a Location Parameter. The Annals of Mathematical Statistics, 35(1), 73–101.

[3] Karlan, D., & List, J. A. (2007). Does price matter in charitable giving? Evidence from a large-scale natural field experiment. American Economic Review, 97(5), 1774–1793.

Conclusion

We combine ideas from statistics and machine learning to develop *Random Sample Reliability* (RSR) – a new re-sampling approach to estimate reliability of data based on the specified model. RSR entails three aspects: *Scoring* estimates reliability scores for every data point in a sample; *Annealing* tests the sensitivity of the estimation results to a sequential removal of the most unreliable observations; and *Fitting* estimates a weighted regression that down-weights the unreliable data. Simulation results reveal favourable performance of RSR for robust estimation and an application of RSR in an empirical study provides new insights about the studied phenomenon.