How shrinkage can be used for robust methods

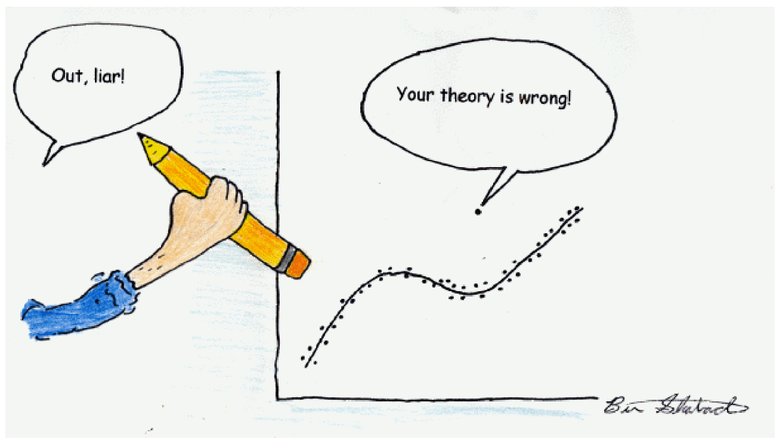
Elisa Cabana[1]

[1] IMDEA Networks Institute, Madrid, Spain: elisa.cabana@imdea.org

3rd BYMAT Conference, December 1-3, 2020

Why robust?

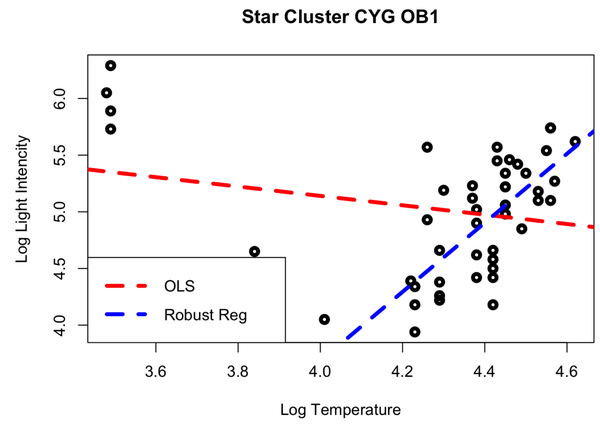
"Outliers are observations resulting from a secondary process, which differs from the background distribution."



Examples of applications:

- Medicine & Health: Neuroscience, epidemic outbreaks.
- Finance: Portfolio optimization, credit card fraud detection.
- Telco & Networks: Social Nets, fake news, mobile fraud detection.

Several data analysis techniques are influenced by the presence of outliers, due to masking or swamping effects. For example, OLS in linear regression:



Classical methods to detect outliers can also be inaccurate if they depend on non-robust estimators. For example, the classical Mahalanobis distance for a multivariate sample $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}, \mathbf{x}_i \in \mathbb{R}^p$:

$$MD(\mathbf{x}_i) = ((\mathbf{x}_i - \hat{\boldsymbol{\mu}})\hat{\Sigma}^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^t)^{1/2}\,,$$

where $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ are the classical sample mean vector and sample covariance matrix estimators, respectively.

Our proposal: shrinkage for robust estimation

A shrinkage estimator is a trade-off:

$$\hat{E}_{Sh} = (1-\eta)\hat{E} + \eta\hat{T}$$

- Frequently used in Finance and Portfolio optimization.
- Reduces the estimation error, while maintaining robustness.
- Results in a positive definite and well-conditioned scatter matrix.

Proposed robust estimators for centrality and scatter in Cabana, E., Lillo, R. E. and Laniado, H., 2019:

$$\hat{\boldsymbol{\mu}}_{Sh} = (1-\eta_\mu)\hat{\boldsymbol{\mu}}_{MM} + \eta_\mu\nu_\mu\boldsymbol{e}\,,$$

where $\hat{\boldsymbol{\mu}}_{MM}$ is the robust multivariate L-1 median.

$$\hat{\Sigma}_{Sh} = (1-\eta_\Sigma)\hat{S}_{Sh(MM)} + \eta_\Sigma\nu_\Sigma I\,,$$

where $\hat{S}_{Sh(MM)}$ is the robust comedian estimator, defined as $\hat{S}_{Sh(MM)} = 2.198 \cdot (med((\mathbf{x}_{\cdot j} - (\hat{\boldsymbol{\mu}}_{Sh})_j)(\mathbf{x}_{\cdot k} - (\hat{\boldsymbol{\mu}}_{Sh})_k)),\ \ j,k = 1, ..., p.$

The unknown parameters $\eta_\mu, \nu_\mu, \eta_\Sigma, \nu_\Sigma$ are optimally and robustly estimated, by reducing the quadratic error.

Proposed Robust Mahalanobis distance based on shrinkage (Cabana, E., Lillo, R. E. and Laniado, H., 2019):

$$RMD_{Sh}(\mathbf{x}_i) = ((\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{Sh})\hat{\Sigma}_{Sh}^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{Sh})^t)^{1/2}\,.$$

Proposed robust regression based on shrinkage (Cabana, E., Lillo, R. E. and Laniado, H., 2020):

For $\mathbf{z} = (\mathbf{x}, y)$, define a weight function:

$$w_i = I(RMD^2_{Sh}(\mathbf{z}_i) \leq \chi^2_{p+1,0.975})\,.$$

Obtain the Shrinkage reweighted estimators for the mean and covariance matrix:

$$\hat{\mathbf{t}}_n = \frac{\sum_{i=1}^n w_i\mathbf{z}_i}{\sum_{i=1}^n w_i}, \qquad \hat{\mathbf{C}}_n = \frac{\sum_{i=1}^n w_i(\mathbf{z}_i - \hat{\mathbf{t}}_n)(\mathbf{z}_i - \hat{\mathbf{t}}_n)^t}{\sum_{i=1}^n w_i}\,.$$

Shrinkage Reweighted Regression (SRR) estimators:

$$\hat{\boldsymbol{\beta}}^{SRR} = (\hat{\mathbf{C}}_{n_{xx}})^{-1}(\hat{\mathbf{C}}_{n_{xy}}), \qquad \hat{\alpha}^{SRR} = (\hat{\mathbf{t}}_n)_y - (\hat{\boldsymbol{\beta}}^{SRR})^t(\hat{\mathbf{t}}_n)_x$$

Results and Real data examples

Simulations:

- Multivariate Normal data.
- Heavy tailed and skewed multivariate distributions.
- Contaminated and correlated data.

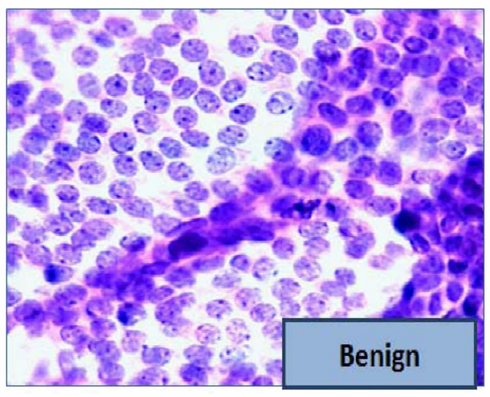
Hiper-parameters in play:

- High dimension and large sample size.
- Types of outliers.
- Level of contamination.

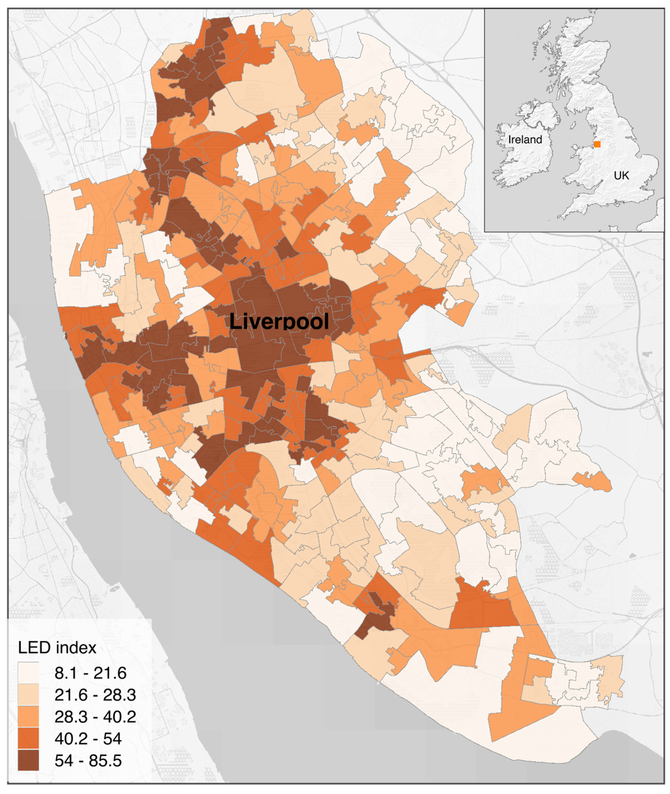
Results:

- Improved performance compared with other robust alternatives, even in high dimension or high contamination.
- Manages well the presence of correlation and deviation from normality.
- Approximately affine equivariant and high empirical breakdown value.
- Competitive computational time.

Outlier detection in Cancer data



Poverty areas detection based on satellite technologies



References

[1] Cabana, E., Lillo, R. E., Laniado, H. *Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators.* Statistical Papers. 2019.
[2] Cabana, E., Lillo, R. E., Laniado, H. *Robust regression based on shrinkage with application to Living Environment Deprivation.* Stochastic Environmental Research and Risk Assessment. 2020.
[3] Cabana, E., Lillo, R. E. *Robust Multivariate Control Chart based on Shrinkage for Individual Observations.* Submitted in Journal of Quality Technology. 2020.