EVaRegression: A novel approach to Stable Regression

Hamza Tazi Bouardi[1] and Pierre-Henri Ramirez[1]

[1]Master of Business Analytics, Operations Research Center, Massachusetts Institute of Technology

I. Motivation

The goal of the present project is to extend the Stable Regression (or CVaR Regression) [1] to a new coherent risk measure. First, the CVaR **is too conservative** (1), as it gives a lot of importance to the worst errors. This can be very dangerous when outliers – which we believe always exist in practice – are present in the data. Indeed, through this approach we would be amplifying their impact on our regression. Also, the CVaR regression gives non-zero weights only to the worst errors, thus neglecting the others, that can be arbitrarily low. Therefore this objective function is **unable to distinguish two regressions producing the same "worst errors" but different "small errors"** (2). Indeed, we only know that these errors are lesser or equal than the best individual error among those selected for the regression. Moreover, its formulation in [1] undoubtedly makes the problem **intractable at a very large scale** (3) for convex functions other than the absolute value. Even using the dualized version, in the general case where we won't be doing a linear regression, the constraints won't be tractable ones. We claim that (1), (2) and (3) could be overcome by replacing the CVaR by another measure of its family, *i.e.* a coherent risk measure [2] - the Entropic Value at Risk (EVaR).

II. Model

The Entropic Value at Risk [5] is a coherent risk measure and the tightest possible upper bound obtained from Chernoff's inequality for the CVaR [5]. The EVaR is defined as follows:

$$\boxed{\mathrm{EVaR}_{1-\alpha}(Z) = \inf_{s\geq 0} \frac{1}{s}\log\left(\frac{M_Z(s)}{\alpha}\right)}$$

where M_Z is the moment-generating function defined for Z a real valued random variable as $t \in \mathbb{R} \mapsto M_Z(t) = \mathbb{E}(e^{tZ})$. In our case, if we call ℓ the loss function, f the prediction function, Z corresponds to the random variable defined by $Z_i = \ell(f(X_i), Y_i)$. Therefore, our optimization problem, with a regularization term $\Gamma(f)$ that depends on the estimator, can be written as follows:

$$\mathcal{L}_{\mathrm{EVaR}}(\alpha) = \min_{f\in\mathcal{F}, s>0} s\log\left(\frac{1}{n\alpha}\sum_{i=1}^{n}\exp\left(\frac{\ell(f(X_i,Y_i)) + \lambda\Gamma(f)}{s}\right)\right)$$

$$\iff \boxed{\mathcal{L}_{\mathrm{EVaR}}(\alpha) = \min_{f\in\mathcal{F}, s>0} s\log\left(\frac{1}{n\alpha}\sum_{i=1}^{n}\exp\left(\frac{\ell(f(X_i,Y_i))}{s}\right)\right) + \lambda\Gamma(f)}$$

III. Solving Methods

We focused on the particular cases where $\ell_1(f_w(X_i), Y_i) = |w^\top X_i - Y_i|$, $\ell_2(f_w(X_i), Y_i) = \left(w^\top X_i - Y_i\right)^2$, and $\Gamma(f) = ||w||_2^2$ where $w \in \mathbb{R}^p$ are the regression parameters. We tried several methods to solve this convex minimization problem:

- First, we attempted to use a cutting planes algorithm unsuccessfully. We also unsuccesfully tried using commercial solvers (Mosek) and express this objective as an exponential cone constraint.
- Then, we applied a constant stepsize gradient descent (GD) algorithm which ended up being effective but fairly slow.
- Finally, we implemented Nesterov's accelerated gradient descent (NGD) which proved to be very efficient and faster than classic gradient descent: we chose this method for experimentations.

IV. Preprocessing and Model Selection

The experiments were conducted on three real world datasets [6] of variable sizes. The data was **randomly splitted** into 70% train and 30% test. We **normalized** every feature based on the training data's mean and standard deviation to avoid divergence of gradient descent due to exponential overflow. We **tuned the hyperparameters** (regularization constant, step sizes) by splitting the train data into **train/validation with ratio** α, with α in $\{0.7, 0.6, 0.5\}$. That split was either random (Classic Ridge Regression), or given by the worst errors for train (Ridge Stable and EVaR Regressions) after the fitting.

V. Empirical Results Comparison

We compare the test MSE of Ridge Regularized Stable Regression and Classic Ridge Regression with the Ridge Regularized ℓ_1 and ℓ_2-EVaR regressions. The EVaR Regressions were fitted either with a random start (RS) or with a warm start (WS) given by the Classic Ridge Regression.

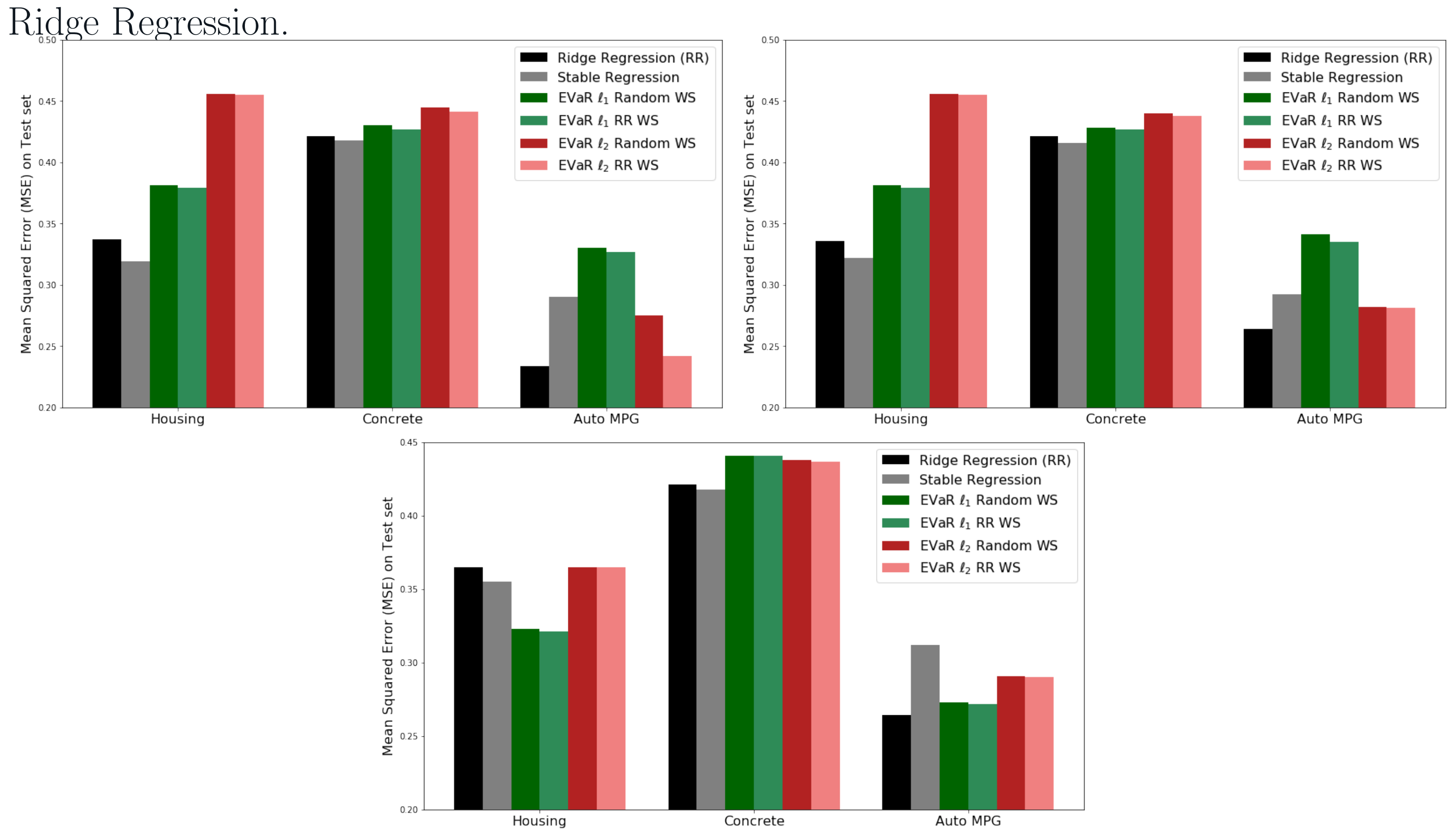

Fig. 1: Test MSE of all methods for $\alpha_1 = k_1/n = 0.5$ (**Left**), $\alpha_2 = k_2/n = 0.6$ (**Right**) and $\alpha_3 = k_3/n = 0.7$ (**Bottom**)

- The ℓ_1 EVaR seems to perform better than the ℓ_2 EVaR on these datasets
- There is no significant score difference between Stable and EVaR regressions, except for the Concrete dataset
- ℓ_1 and ℓ_2 EVaR Regressions perform uniformly better than Stable Regression on the Auto MPG dataset

For the last observation, our hypothesis was that EVaR regressions perform better on datasets containing a fair amount of noise and/or outliers, which is the case for the Auto MPG dataset, which motivated part **VII**.

VI. Other property of EVaR

The EVaR also has one additional desirable property, as it is what we call a **strongly monotone** risk measure. A risk measure ρ has such property if for (X, Y), a pair of real-valued random variables which verify the following conditions:

- (i) $X \geq Y$ (ii) $\mathbb{P}(X > Y) > 0$
- (iii) $\operatorname{ess\,sup}(X) > \operatorname{ess\,sup}(Y)$ or $\operatorname{ess\,sup}(X) = \operatorname{ess\,sup}(Y) = +\infty$

We have that $\rho(X) < \rho(Y)$. This property enables EVaR to distinguish two regressions producing the same "worst errors" but different "small errors", while the CVaR cannot as it is not a strongly monotone risk measure.

VII. Sensitivity Analysis

We introduced some noise $N \sim \mathcal{N}(0, 1.5)$ on 5% of randomly selected rows for each of the three datasets and re-fitted all the models. The EVaR regressions were fitted using the Classic Ridge Regression Warm Start, as we have seen in part **V** that it tends to perform better than random warm start.

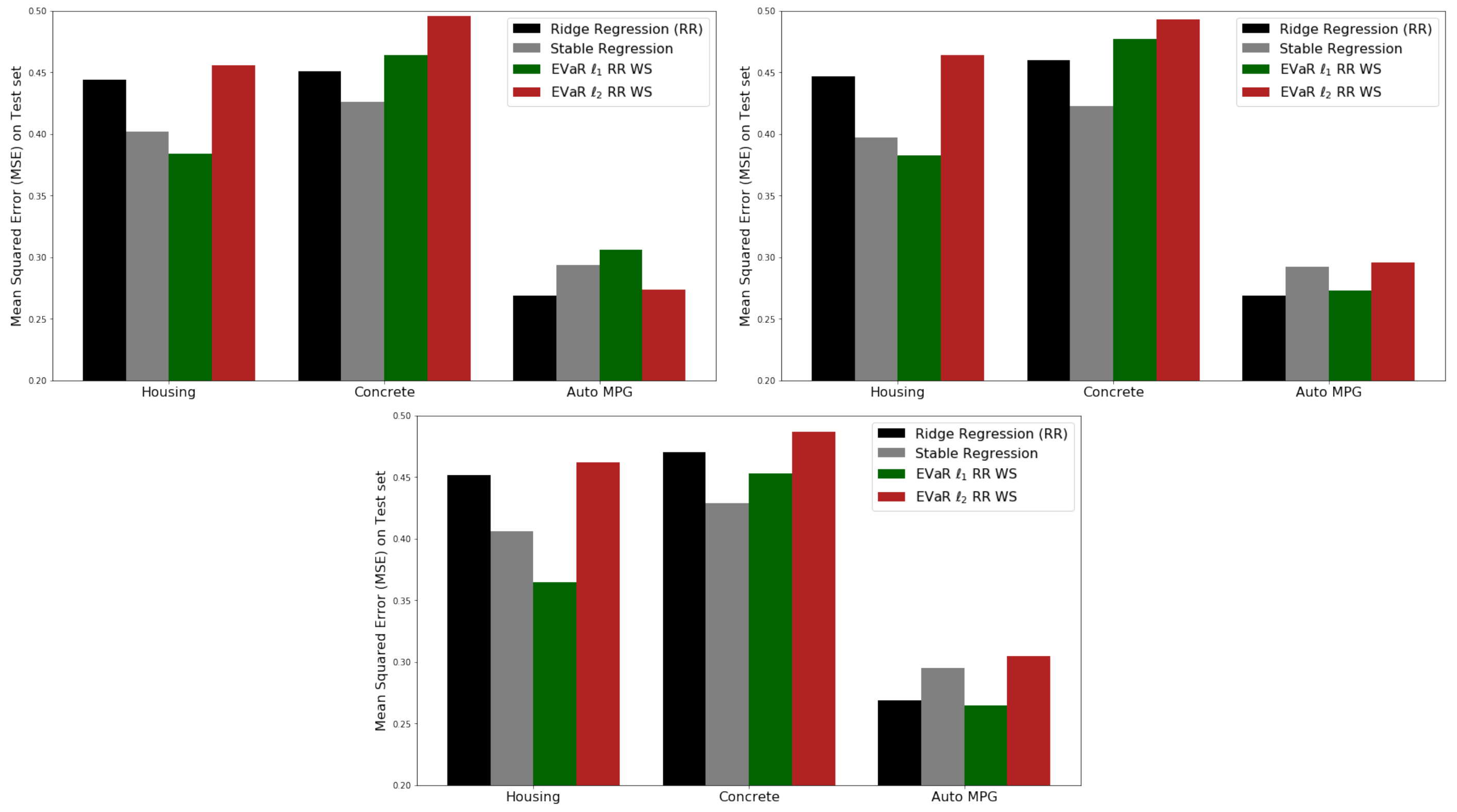

Fig. 2: Test MSE of all methods with added noise $N \sim \mathcal{N}(0, 1.5)$ on 5% of the data for $\alpha_1 = k_1/n = 0.5$ (**Left**), $\alpha_2 = k_2/n = 0.6$ (**Right**) and $\alpha_3 = k_3/n = 0.7$ (**Bottom**)

- The ℓ_1-EVaR outperforms the Stable regression in two data sets out of three, with an average MSE improvement of 6.1% on Housing and 4.2% on Auto MPG, proving to some extent that we solve point (1) using EVaR instead of CVaR.
- Again, the ℓ_1 EVaR consistently performs better that the ℓ_2 EVaR regression: an interpretation could be that the exponential in the EVaR makes the squared errors on noisy data much more predominant than the absolute value of the errors, therefore more difficult to minimize.

VIII. Acknowledgements

We would like to thank Ryan Cory-Wright and Michael Lingzhi Li for their support and their help in both defining the scope of this project and in its development.

IX. References

References

[1] D. Bertsimas and J, Dunn. *Machine Learning Under a Modern Optimization Lens*, Dynamic Ideas LLC, 2019, 333-366.

[2] P. Artzner, F. Delbaen, J-M. Eber and D. Heath. *Coherent Measures of Risk*. Mathematical Finance, Vol 9, No 3 (July 1999), 203-228.

[3] R. Kohavi, *A study of cross-validation and bootstrap for accuracy estimation and model selection*. International Joint Conference on Artificial Intelligence (IJCAI), 1995.

[4] C. Acerbi, D. Tasche. *Expected Shortfall: a natural coherent alternative to Value at Risk*. Economic Notes. 31 (2): 379–388, 2002.

[5] A. Ahmadi-Javid. *Entropic Value-at-Risk: A New Coherent Risk Measure*. J Optim Theory Appl, 155: 1105, 2012.

[6] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.