ML for Macroeconomic

Refresher on Variable Selection

- Variable selection — known as “feature selection” in the ML literature
- The Lasso and applications of random forests are then two examples of the variety of variable selection techniques that have been developed over the years.
- *Different procedures have different operating characteristics, i.e., give rise to different bias-variance tradeoffs as we vary their respective tuning parameters. In fact, depending on the problem setting, the bias-variance tradeoff provided by best-subset selection may be more or less useful than the tradeoff provided by the lasso*
- *Now, what Hastie et al. are referring to is that “best-subset selection” naturally has a very low bias, but, in noisy data, methods such as the Lasso can often deliver better results*

because they trade-in some bias for a substantial reduction in variance. Indeed, this type of tradeoff is the central motivation of this article, as we compare the Lasso to another technique entailing efficient variance reduction: RFs. Before proceeding to the simulation study, let me also offer a very brief description of Lasso and RF.

- ***The Lasso*** *belongs to the class of Shrinkage Methods with the special property that, for an appropriate choice of the penalty term, coefficients are shrunken exactly to zero. Since the penalty term shrinks the size of all coefficients, a* ***relaxed Lasso*** *has been developed, where control of non-zero coefficients is separated from overall shrinkage.*
- *The defining feature of* ***RFs*** *is the aggregation of decorrelated decision trees, which is particularly effective for the reduction in variance. While RF, unlike the Lasso, does not naturally perform variable selection, Genuer et al. (2010) developed a stepwise ascending variable introduction strategy for this*

*purpose and implemented it in the very good **VSURF** package, available on the R-CRAN library.*

Simulation Study

- *Given the dependence of the bias-variance trade-off on the problem structure, it is key for our simulation study to emulate our real-world application, i.e. a cross-country study of economic growth*
- *We make data quality the focal point of our comparison and instrumentalise it through the Signal-to-Noise Ratio (**SNR**). We follow Hasties et al. (2017) and consider ten values of SNR ranging from 0.05 to 6 on a log scale — for comparison the corresponding values of the "Proportion of Variance Explained" are reported, as well. Importantly, as the authors emphasise, studies are often too optimistic about signal clarity and argue that, in the real world, observational data tends to have **log(SNR) < 1**!*

- *As a result of this SNR scale, a simulation study calculates 10 x #sim synthetic data sets and applies the Lasso, relaxed Lasso and RF to each of them.*
- *The data sets are sparse so that only 5 out of 50 covariates influence the outcome. A Toeplitz matrix is used for the variance-covariance Matrix, so that variable's relative position determines the correlation structure. Note that we choose a linear DGP, which needs to be taken into account when comparing the results of parametric to non-parametric methods. Finally, the number of simulations is set to 100.*