ESSENTIAL REGRESSION

FLORENTINA BUNEA
DEPARTMENT OF STATISTICS AND DATA SCIENCE
CORNELL UNIVERSITY

We introduce the Essential Regression model, which provides an alternative to the ubiquitous K-sparse high dimensional linear regression on p variables. While K-sparse regression assumes that only K components of the observable X directly influence Y, Essential Regression allows for all components of X to influence Y, but mediated through a K-dimensional random vector Z. The mediator Z is unobservable, and made *interpretable* via a modeling assumption through which each component of Z is given the physical meaning of a small group of the X-variables.

Formally, E-Regression is a new type of latent factor regression model, in which the unobservable factors Z influence linearly both the response Y and the covariates X. Its novelty consists in the conditions that give Z interpretable meaning as well as render the regression coefficients $\beta \in R^K$ relating Y to Z – along with other important parameters of the model – identifiable.

The interpretability of the latent factors Z in Essential Regression allows us to provide conceptually new inferential tools for regression in high dimensions. In particular, K-dimensional inference at the Z level is a viable alternative to existing approaches that benefits from the greater simplicity and lower dimensionality of the "essence" Z when compared to X. It is furthermore well known that inference performed in K-sparse regression – after consistent support recovery and estimation of K – is valid only for the large regression coefficients of Y on X, which makes this approach problematic in practice. In contrast, inference performed in regression at the lower resolution level given by Z is uniform over the space of the regression coefficients $\beta \in R^K$, although it is performed after estimating consistently K and the subset of the X-variables that explain Z. For this we construct computationally efficient estimators of β, derive their minimax rate, and show that they are minimax-rate optimal in Euclidean norm for every sample size n. We show that the component-wise estimates of β are asymptotically normal, with small asymptotic variance. This is a new addition to the literature in factor models, in which the inference problem is under-explored.

Prediction of Y from X under E-Regression complements, in the low signal to noise ratio regime, the battery of methods developed for prediction under other factor model specifications. Similarly to other methods, it is particularly powerful when p is large, with further refinements made possible by the Essential Regression model specifications.

E-Regression also provides a statistical framework for analysis in regression with clustered predictors, with or without overlap. This allows us to address possible inferential questions in post clustering-inference, and subsequently provide guidelines regarding the use and misuse of cluster averages as very popular dimension reduction devices in high dimensional regression.