Statistical Inference when Fitting Simple Models to High Dimensional Data

Lukas Steinberger[1] **and Hannes Leeb**[1]

[1] *University of Vienna, Department of Statistics and OR*

Abstract

We study linear subset regression in the context of the high-dimensional overall model $y = \theta' Z + u$ with univariate response y and a $d-$vector of random regressors Z, independent of u. Here, 'high-dimensional' means that the number n of available observations may be much less than d. We consider simple linear submodels where y is regressed on a set of p regressors given by $X = B'Z$, for some $d \times p$ matrix B with $p \leq n$. The corresponding simple model, i.e., $y = \gamma' X + v$, can be justified by imposing appropriate restrictions on the unknown parameter θ in the overall model; otherwise, this simple model can be grossly mis-specified. We show that the least-squares predictor obtained by fitting the simple linear model is typically close to the Bayes predictor $E[y|X]$ in a certain sense, uniformly in $\theta \in \mathbb{R}^d$, provided only that d is large. Moreover, we establish the asymptotic validity of the standard $F-$test on the surrogate parameter which realizes the best linear population level fit of X on y, in an appropriate sense. On a technical level, we extend recent results from [4] on conditional moments of projections from high-dimensional random vectors; see also [1, 2, 3].

Keywords: High-dimensional models, mis-specified model, regression analysis, prediction, F-test.
AMS subject classifications: 62F05, 62J05

Bibliography

[1] Diaconis, P. and Freedman, D. (1984). Asymptotics of Graphical Projection Pursuit. *The Annals of Statistics* **12**, 3, 793–815.

[2] Dümbgen, L. and Conte-Zerial, P. (2012). On Low-Dimensional Projections of High-Dimensional Distributions. In *From Probability to Statistics and Back: High-Dimensional Models and Processes. A Festschrift in Honor of Jon Wellner*, M. Banerjee; F. Bunea; J. Huang; V. Koltchinskii; and M.H. Maathius, Eds. IMS Collections, Vol. **9**, 91–104.

[3] Hall, P. and Li, K.-C. (1993). On almost Linearity of Low Dimensional Projections from High Dimensional Data. *The Annals of Statistics* **21**, 2, 867–889.

[4] Leeb, H. (2013). On the Conditional Distributions of Low-Dimensional Projections from High-Dimensional Data. *The Annals of Statistics,* forthcoming.