Market Efficiency in the Age of Big Data

Ian Martin†
London School of Economics and CEPR

Stefan Nagel‡
University of Chicago, NBER, CEPR, and CESifo

October 2019

PRELIMINARY AND INCOMPLETE
NOT FOR PUBLIC DISTRIBUTION

Modern investors face a high-dimensional prediction problem: thousands of observable variables are potentially relevant for forecasting. We reassess the conventional wisdom on market efficiency in light of this fact. In our model economy, which resembles a typical machine learning setting, \( N \) assets have cash flows that are a linear function of \( J \) firm characteristics, but with uncertain coefficients. Risk-neutral Bayesian investors impose shrinkage (ridge regression) or sparsity (Lasso) when they estimate the \( J \) coefficients of the model and use them to price assets. When \( J \) is comparable in size to \( N \), returns appear cross-sectionally predictable using firm characteristics to an econometrician who analyzes data from the economy ex post. A factor zoo emerges even without \( p \)-hacking and data-mining. Standard in-sample tests of market efficiency reject the no-predictability null with high probability, despite the fact that investors optimally use the information available to them in real time. In contrast, out-of-sample tests retain their economic meaning.

---

†London School of Economics; i.w.martin@lse.ac.uk
‡University of Chicago, Booth School of Business; stefan.nagel@chicagobooth.edu