Fundamental Dimension

Adam D. Clark-Joseph and Daniel Graves

Research Objective

We propose to implement a novel estimation technique to empirically investigate financial-market complexity in different universes of equities. (We introduce the estimator from a theory perspective in a recent companion paper, Clark-Joseph, Graves, and Mendel 2018.) The technique is a method of determining the appropriate number of factors for nonlinear factor-structures in high-dimensional time-series data. This number of factors, which we term "fundamental dimension," is essentially the limit for dimensionality reduction, and as such characterizes "complexity" in some sense.

At long horizons, the cross-section of asset returns exhibits a strong, linear, factor structure. As Fama and French observed, a significant fraction of the long-horizon variation in cross-sectional returns can be explained by a small number of factors. Collectively, the individual returns on thousands of different assets constitute a very high-dimensional object, but much of this high-dimensional object turns out to be merely a transformation of the handful of factors—a very low-dimensional object. As documented in Chinco, Clark-Joseph, and Ye (2017), the cross-section of returns also exhibits a "sparse" structure of sorts at *short* time horizons.

Classic asset-pricing models, such as Ross' Arbitrage Pricing Theory, offer a rationale for the observed linearity of the factor structure at long time horizons, but this rationale for linearity becomes increasingly tenuous as we consider progressively shorter horizons, where other economic mechanisms progressively dominate. Because nonlinearity can occlude underlying sparsity, rigorous analysis for short horizons requires disentangling sparsity from linearity. To date, this disentanglement task appears to be impossible in the fully general case, but our new technique enables us to perform the task in the special case of non-stationary data. By shifting our attention from returns to price levels, we can exploit this special case.

Analyzing the fundamental dimension of different universes of equities will offer insight into the underlying structural complexity of various stock markets. In other words, it will provide an empirical measure of the number of different processes driving price movement within various sets of equities, and, because we focus on short horizons, these processes admit an informational interpretation. We will investigate how the fundamental dimension relates to other measures of "broadness" of an index, as well as diversification of a portfolio of equities. Exploration of fundamental dimension's variation over time across a fixed universe of equities may also yield quantitative insights into changes in the complexity of information diffusion processes over time, and with respect to significant market events. It is our hope that a taxonomic classification of different indices based on the fundamental dimension of their constituents will elucidate some of the causes for differing volatility and price dynamics across indices.

Empirical/Theoretical Strategy

Our estimation method is a substantial generalization of the techniques for determining the appropriate number of factors in linear factor-models, in that it places only minimal restrictions on the functional form of the mapping from underlying factors to observed prices. Consequently,

we can assume—innocuously—that the values of the underlying factors follow a random walk on a lattice of integers of some fixed dimension (which will be the "fundamental dimension").

The key observation is that the joint behavior of the individual constituent prices will inherit certain features from random-walk properties of the underlying factors. In particular, the high-dimensional joint-price process will inherit the recurrence behavior of the low-dimensional factor process. For a simple random walk on a lattice of dimension D, the probability distribution of the number of recurrences (returns to the starting-point) observed during a sample-interval of length T depends on D. Computing the empirical relative-frequency distribution therefore enables us to make an estimate of D.

We will examine intraday prices for US, German, and Japanese equities. Treating the time-*t* prices of the N different equities as a single point in N-dimensional space, we will compute the approximate distribution of recurrences in sample-intervals of length T, for a range of T's. From these recurrence distributions, we will estimate fundamental dimension. We will investigate the fundamental dimension of a variety of broad stock index universes (S&P 500, Russell 2000, DJIA) and narrow-based or approximately narrow equity universes (DJTA, XLE). In addition, we will explore the distributional properties of fundamental dimension over a variety of randomly constructed equity subsets.

As intraday Japanese equity data has no time overlap with US or German intraday equity data, and German equity data overlaps with the main US trading session for only two hours daily, the two international datasets offer the opportunity of independent testing of fundamental dimension results from US data, and may elucidate whether fundamental dimension is idiosyncratic to regional markets, time-variant based on an unobserved process influencing all three regions, or constant across all regions and various periods.

Data Description

The research dataset includes US, German, and Japanese equity data from 2016 and 2017. US equity data consists of quote and NBBO data (bid/offer/trades) for equities in the NYSE TAQ database. All data is stored in compressed form, with US data having a size of 9.9TB. Penny stocks and equities that experience fewer than twenty executed trades per hour on average are eliminated. We only examine equities that have data for all of 2016-2017 so as not to confound the dimensionality with other events like IPOs. Timestamps are by the microsecond for US data. German and Japanese equity data will be studied with the same methodology and comparable datasets, using the significant non-overlap between European and US data, and complete lack of overlap between Japanese and German data, to provide confirmation of the robustness of results.

Need for XSEDE Resources

We estimate that computations for each round-the-clock day of trading data will require roughly 30 core-hours of processing, so we expect computations for our two-year sample to use about 15,000 total core-hours. Fortunately, computations for individual days in our sample can be run in parallel, so parallel-computing resources from XSEDE would enable us to complete our analysis in a matter of weeks. We use Matlab for code development, but can convert to C/C++ for deployment if appropriate. One of the authors on this proposal, Clark-Joseph, has used similar parallelization with XSEDE resources extensively in prior research.