Data-Driven Construction of Financial Factor Models*

Hassan T. Anis and Roy H. Kwon

Factor models are used to model asset (expected) returns as functions of various factors. Instead of simply using the observable returns in computations, factor models attempt to decompose risk and return into latent factors representing the systematic and idiosyncratic components driving price movements. To construct a factor model, two tasks have to be performed: *feature selection*, selecting a small subset given a large number of factors to overcome the curse of dimensionality and overfitting in regression, and *feature engineering*, determining the interactions between the factors.

Since the publication of the seminal work of Fama and French (1993), hundreds of papers have been published containing many factors that attempt to explain the cross-section of expected returns. Green et al. (2013) list 330 stock-level predictive signals that have been published in the academic literature. Financial firms may use an even larger set. This area of research has been very active as of late, trying to find new factors that add explanatory power to asset pricing, relative to the existing 'factor zoo'. Yet, recently, concern has been expressed about whether all these factors have explanatory power. Some works, like Harvey et al. (2016) and Harvey and Liu (2018), propose sound evaluation frameworks for assessing the efficacy of new (and existing) factors. Additionally, given the large number of existing factors, fears of data-mining have grown. Specifically, the concern is that if a (linear) regression model is provided with a set of factors, including some irrelevant or redundant subsets, that the model will try to fit all factors indiscriminately or generate spuriously high significance levels (Bryzgalova, 2015). The result would be a small in-sample error and terrible out-of-sample performance.

This work proposes a unified, data-driven framework to construct factor models that tackle the aforementioned shortcomings of traditional factor models. It is concerned with the process of constructing the factor model, rather than how the factors themselves are constructed. The process presented here produces sparse factor models, that do not require the modeler to a-priori explicitly specify which subset of factors to use nor what their interactions should be; it is completely data-driven. The framework is a systematic, two-step process of dimensionality re-

*Submitted & under review by INFORMS journal.

duction and nonlinear transformation that produces parsimonious, general factor models. Thus, it balances the bias–variance tradeoff in a data-driven fashion to achieve an overall framework that first reduces the variance of the problem by limiting the input space and then decreases the bias error by generalizing the problem to included nonlinear interactions.

The first stage reduces the dimensionality of the overall factor model by, given p initial factors, selecting $k << p$ factors in the first stage. A variety of linear, LASSO- and MIP-based dimensionality reduction methods, as well as autoencoding (Hinton and Salakhutdinov, 2006), an unsupervised, neural network-based framework that performs nonlinear dimensionality reduction, are compared. By explicitly performing dimensionality reduction, the first stage of this framework forces the overall model to be sparse. This is a highly desirable characteristic as it 'helps reduce [a model's] intrinsic complexity at little cost of statistical efficiency' (Fan et al., 2011), thus, reducing the overall model variance and limiting overfitting.

The second stage takes in the factors from the first stage and constructs nonlinear factor models. While sparse *linear* models in economics are 'generally biased' (Fan et al., 2011), the proposed framework reduces said bias by generalizing the factor model to include nonlinear terms using neural-based architectures. By moving beyond linear factor models, the restrictive assumption of factor independence is removed. The result is a comparison of three-factor models with increasing nonlinearity: a model with linear dimensionality reduction and affine factor combination, a model with linear dimensionality reduction and nonlinear factor combination, and a model with nonlinear dimensionality reduction and nonlinear factor combination. All three models are data-driven as the set of included factors and their interactions with one another are not decided by the modeler a-priori.

Experiments using daily asset return data and factors validate the use of MIP-based best subset selection (BSS) as proposed by Bertsimas et al. (2016) for linear feature selection as it results in parsimonious sets of factors that behave in accordance with economic expectations. It also avoids redundant or irrelevant factors with insignificant factor loadings and has the best out-of-sample performance in terms of R^2_{adj} compared to other linear methods. For the BSS problem with a Least Absolute Deviation objective, a heuristic is introduced that leverages simple norm properties, leading to significant computational speedups at training time. Finally, computational results show that the second stage nonlinearity introduced by the deep FNNs yields statistically significant improvements in accuracy, while the first stage nonlinear dimensionality reduction leads to minor, statistically insignificant out-of-sample accuracy improvements compared to ones with linear factor selection.

References

D. Bertsimas, A. King, R. Mazumder, et al. Best subset selection via a modern optimization lens. *The annals of statistics*, 44(2):813–852, 2016.

S. Bryzgalova. Spurious factors in linear asset pricing models. *LSE manuscript*, 1, 2015.

E. F. Fama and K. R. French. Common risk factors in the returns on stocks and bonds. *J. of Financial Economics*, 33(1):3–56, 1993.

J. Fan, J. Lv, and L. Qi. Sparse high-dimensional models in economics. *Annu. Rev. Econ.*, 3 (1):291–317, 2011.

J. Green, J. R. Hand, and X. F. Zhang. The supraview of return predictive signals. *Review of Accounting Studies*, 18(3):692–730, 2013.

C. R. Harvey and Y. Liu. Lucky factors. *Available at SSRN 2528780*, 2018.

C. R. Harvey, Y. Liu, and H. Zhu. ... and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–68, 2016.

G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.