BRJ - RE02

Autoencoder Asset Pricing Models

In this paper, the authors develop a novel conditional asset pricing model (conditional as it can be tuned according to the desired risk profile). The underlying model is a neural autoencoder that allows to reduce dimensionality as PCA does but with the advantage of being able to also capture arbitrary complex non-linear relationships thanks to the neural network structure (composition of non-linear activation functions). In addition, following the example of Kelly, Pruitt, and Su (KPS, 2019) with IPCA, this auto-encoding scheme uses information in data covariates to guide dimension reduction.

More in details, the paper start with an extended methodology section that includes a literature review of autoencoders theory and standard machine learning techniques (e.g. regularization, training/validation/test split, optimization through stochastic gradient descent, batch normalisation). While a quick introduction of these ML techniques is more than justified, I consider the level of detail to which autoencoders are explained excessive, especially given the economic nature of the paper and the already abundant computer science literature on the topic. The section is concluded by a clear explanation on how the covariates were included in the autoencoder model and how the can be used for the particular case of asset pricing.

The rest of the paper is dedicated to the description of the datasets used and the experimental results. While the dataset is very lare, spanning more than 60 of stocks markets of 3 major exchanges (NYSE, AMEX, and NASDAQ), the study would have benefited from using more diverse data by including information form European and asian stock exchanges (e.g. SHG, Euronext or JPX).

The extensive list of experiments is thoroughly reported in tables and graphs, clearly showing the superiority of the author's model compared to previously existing ones. The set of chosen baselines is appropriate as it consist in PCA (standard techniques that autoencoders generalise) and IPCA which is a very strong baseline and allows to showcase the good performance achieved by autoencoders. It would have been interesting to see more statistics about the variation of prediction performance over time to see how quickly the model can adapt to unexpected events like the 2007-8 financial crisis.

Finally the authors strengthen their results by running a series of Monte Carlo simulations and reconfirmed their previous finding. This is a smart additional experiment that allows to strengthen the study without the need of additional data.

Overall the paper is well structured and clearly explained. It builds upon existing solid mathematical and machine learning foundations and applies this external literature to the particular case of asset pricing. As previously mentioned there is nevertheless space for improvement by cutting down on some technicalities that can already be found in the existing literature and by enlarging the geographic scope of the experiments.