COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Factor Creation with Conditional Autoencoders

Research by: Andreas Tsampoulas

Supervised by: Dr. Ali Hirsa, Sikun Xu

ask².ai

Introduction

- Factor exposure analysis is a quantitative method that provides asset managers and investors the opportunity to better understand and evaluate the main components that drive their investment strategies.
- Fama-French model is an asset pricing model developed in 1992 that expands on the capital asset pricing model (CAPM) by adding size risk and value risk factors to the market risk factor in CAPM.
- In contrast to the Fama–French model, Gu, Kelly and Xiu [1], view risk variables as latent so they have developed a conditional autoencoder architecture, which is a deep neural network that incorporates a variety of time-varying asset attributes using a feedforward network to account for the nonlinear nature of return dynamics.
- The goal of this project is to try to create latent risk factors of mutual funds' asset returns based on the Conditional Autoencoder, by incorporating observable characteristics such as size characteristics, fundamentals, and alternative data, that can be used to identify regimes, predict the performance of the portfolio, and manage the portfolio's risk.

Architecture of the Conditional Autoencoder

- The conditional autoencoder proposed in [1] is an extension of the standard autoencoder architecture.
- The standard autoencoder is a neural network where the outputs try to approximate the input data, after passing through a small number of neurons in the hidden layer, creating a representation of the input (encoding), similar to what PCA does linearly.
- This representation is then mapped to the output layer (decoding), trying to accurately approximate the input data (Figure 1).

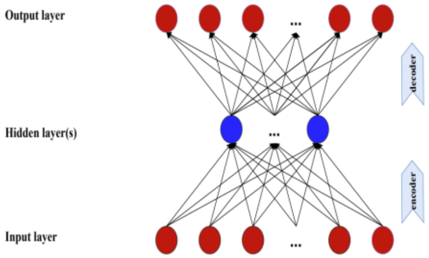

Figure 1: Standard Autoencoder [1]

- The conditional autoencoder allows for time-varying return distributions that incorporate asset characteristics.
- The architecture of the model is illustrated in Figure 2, where the output of the model, which in our case is the asset returns(top), is a function of both asset characteristics (left input) and, as in the case of the autoencoder, the individual asset returns (right input).
- The architecture permits for asset returns to be individual stock returns or portfolios that are formed from the stocks in the sample based on the asset characteristics. In this project mutual fund returns are used.
- The right side of this architecture is a traditional autoencoder, where the N individual asset returns are mapped to themselves. The first step is to translate the input of the N asset returns into K factor loadings for each period t. These K factor loadings are the latent factors we are trying to estimate.
- On the left side of the conditional autoencoder the K factor loadings (beta output) of N individual stocks is generated from the PxN characteristics (input), by passing through hidden layer(s) for each period t.
- The model output is the dot product of the NxK factor loadings on the left with the NxK factor premia on the right.

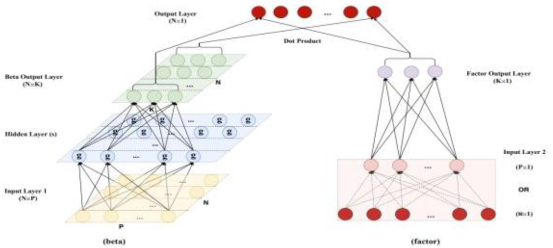

Figure 2: Conditional Autoencoder [1]

Data

- The focus of the project is to create factors that can explain the exposures of the daily returns of mutual funds.
- In Figure 3, we can see how the distribution of the exposures of each of the large, mid, and small cap funds to the Fama-French factors
- The Conditional Autoencoder was implemented on 92 large cap mutual funds between 2015-2022.

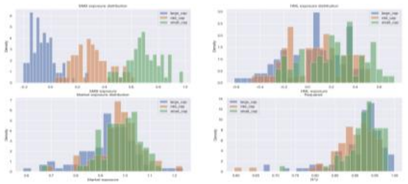

Figure 3: Distribution of exposures for large, mid, and small cap funds

- The observable characteristics of the mutual funds for the implementation were the revenue exposures of the large cap funds to 13 markets (Africa, Asia developed, Asia Emerging, Australia, Canada, Europe Emerging, Europe excluding Eurozone, Eurozone, Japan, Latin America, Middle East, UK, US).
- Following [1] we limit the influence of outliers, by rank-normalizing the characteristics to the [-1, 1] interval and set missing values to 0.

Results

- In Figure 4 we can see the graph of the conditional autoencoder architecture where the symbols of Figure 2 are replaced with the actual units used in the architecture.
- N is equal to 92 and P is equal to 13 and t is equal to 246468 (daily data from 2015 – 2022).
- On the right side of the architecture, the 92 funds are translated into K factor, which are the latent factors we are creating. In this project, we experimented with different numbers of factors ranging from 2 to 6 (Figure 4 shows 3 factors).
- The output of the model is the dot product of the factor loadings based on the asset characteristics on the asset characteristics on the left (92xK) and the factors of the returns of the mutual funds (Kx1).

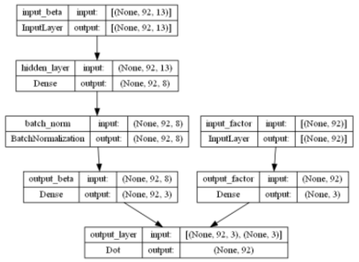

Figure 4: Graph of The Conditional Autoencoder Architecture

Table 1: Linear Regression results of the Conditional Autoencoder and the Fama-French factors against the each of the 92 mutual funds (2015-2022)

	Conditional Autoencoder					Fama-French model
Average of Metric	2 Factors	3 Factors	4 Factors	5 Factors	6 Factors	3 Factors
R squared (stdev)	0.926 (0.079)	0.933 (0.075)	0.941 (0.062)	0.947 (0.066)	0.943 (0.067)	0.916 (0.087)
Intercept Value	0.000	0.000	0.000	0.000	0.000	0.000
Intercept P-Value	0.506	0.471	0.463	0.474	0.275	0.508

- All scenarios in Table 1 perform very well, and in Figure 5 we can explore the stability of the factors by visualizing the exposures of one of the large cap funds ("B08833") throughout overlapping time frames (2015-2017, 2017-2019, 2019-2021, 2021-2022) for the case of the 3 and 5 Conditional Autoencoder factors and the 3 Fama-French factors.

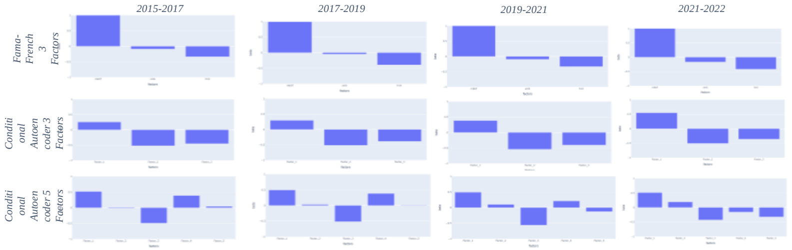

Figure 5: Factor Exposures throughout time of 3 and 5 Conditional Autoencoder and 3 Fama-French factors

Conclusion

- All in all, the flexibility of deep learning helped us compress complex data while losing little information and therefore counter the curse of dimensionality associated with rich datasets with many features and alternative data.
- Moreover, the conditional autoencoder architecture performed better than the Fama-French factors and successfully identified and created sensitive latent factors that can explain the exposures of the mutual funds' returns.

References

[1] Shihao Gu, B. T. (2019). *Autoencoder Asset Pricing Models.* Chicago Booth.