Deep Portfolio Optimization via Distributional Prediction of Residual Factors

Kentaro Imajo, [1] **Kentaro Minami,** [1] **Katsuya Ito,** [1] **Kei Nakagawa,** [2]
[1] Preferred Networks, Inc.
[2] Nomura Asset Management Co., Ltd.
{imos, minami, katsuya1ito}@preferred.jp, k-nakagawa@nomura-am.co.jp

Abstract

Recent developments in deep learning techniques have motivated intensive research in machine learning-aided stock trading strategies. However, since the financial market has a highly non-stationary nature hindering the application of typical data-hungry machine learning methods, leveraging financial inductive biases is important to ensure better sample efficiency and robustness. In this study, we propose a novel method of constructing a portfolio based on predicting the distribution of a financial quantity called residual factors, which is known to be generally useful for hedging the risk exposure to common market factors. The key technical ingredients are twofold. First, we introduce a computationally efficient extraction method for the residual information, which can be easily combined with various prediction algorithms. Second, we propose a novel neural network architecture that allows us to incorporate widely acknowledged financial inductive biases such as amplitude invariance and time-scale invariance. We demonstrate the efficacy of our method on U.S. and Japanese stock market data. Through ablation experiments, we also verify that each individual technique contributes to improving the performance of trading strategies. We anticipate our techniques may have wide applications in various financial problems.

1 Introduction

Developing a profitable trading strategy is a central problem in the financial industry. Over the past decade, machine learning and deep learning techniques have driven significant advances across many application areas (Devlin et al. 2019; Graves, Mohamed, and Hinton 2013), which inspired investors and financial institutions to develop machine learning-aided trading strategies (Wang et al. 2019; Choudhry and Garg 2008; Shah 2007). However, it is believed that forecasting the nature of financial time series is essentially a difficult task (Krauss, Do, and Huck 2017). In particular, the well-known *efficient market hypothesis* (Malkiel and Fama 1970) claims that no single trading strategy could be permanently profitable because the dynamics of the market changes quickly. At this point, the financial market is significantly different from stationary environments typically assumed by most machine learning/deep learning methods.

Generally speaking, a good way for deep learning methods to adapt quickly to the given environment is to introduce a network architecture that reflects a good inductive bias for the environment. The most prominent examples for such architectures include convolutional neural networks (CNNs) (Krizhevsky, Sutskever, and Hinton 2012) for image data and long short-term memories (LSTMs) (Hochreiter and Schmidhuber 1997) for general time series data. Therefore, a natural question to ask is what architecture is effective for processing financial time series.

In finance, researchers have proposed various trading strategies and empirically studied their effectiveness. Hence, it is reasonable to seek architectures inspired by empirical findings in financial studies. In particular, we consider the following three features in the stock market.

1.1 Hedging Exposures to Common Market Factors

Many empirical studies on stock returns are described in terms of *factor models* (e.g., (Fama and French 1992, 2015)). These factor models express the return of a certain stock i at time t as a linear combination of K factors plus a residual term:

$$r_{i,t} = \sum_{k=1}^{K} \beta_i^{(k)} f_t^{(k)} + \epsilon_{i,t}. \tag{1}$$

Here, $f_t^{(1)}, \ldots, f_t^{(K)}$ are the common factors shared by multiple stocks $i \in \{1, \ldots, S\}$, and the residual factor $\epsilon_{i,t}$ is specific to each individual stock i. Therefore, the common factors correspond to the dynamics of the entire stock market or industries, whereas the residual factors convey some firm-specific information.

In general, if the return of an asset has a strong correlation to the market factors, the asset exhibits a large exposure to the risk of the market. For example, it is known that a classical strategy based on the *momentum* phenomenon (Jegadeesh and Titman 1993) is correlated to the Fama–French factors (Fama and French 1992, 2015), which exhibited negative returns around the credit crisis of 2008 (Calomiris, Love, and Peria 2010; Szado 2009). On the other hand, researchers found that trading strategies based only on the residual factors can be robustly profitable because such strategies can hedge out the time-varying risk exposure to

the market factor (Blitz, Huij, and Martens 2011; Blitz et al. 2013). Therefore, to develop trading strategies that are robust to structural changes in the market, it is reasonable to consider strategies based on residual factors.

A natural way to remove the market effect and extract the residual factors is to leverage linear decomposition methods such as principal component analysis (PCA) and factor analysis (FA). In the context of training deep learning-based strategies, it is expected to be difficult to learn such decomposition structures only from the observed data. One possible reason is that the learned strategies can be biased toward using information about market factors because the effect of market factors are dominant in many stocks (Pasini 2017). Hence, in order to utilize the residual factors, it is reasonable to implement a decomposition-like structure explicitly within the architecture.

1.2 Designing Architectures for Scale Invariant Time Series

When we address a certain prediction task using a neural network-based approach, an effective choice of neural network architecture typically hinges on *patterns* or *invariances* in the data. For example, CNNs (LeCun et al. 1999) take into account the shift-invariant structure that commonly appears in image-like data. From this perspective, it is important to find invariant structures that are useful for processing financial time series.

As candidates of such structures, there are two types of invariances known in financial literature. First, it is known that a phenomenon called volatility clustering is commonly observed in financial time series (Lux and Marchesi 2000), which suggests an invariance structure of a sequence with respect to its volatility (i.e., amplitude). Second, there is a hypothesis that sequences of stock prices have a certain time-scale invariance property known as the *fractal structure* (Peters 1994). We hypothesize that incorporating such invariances into the network architecture is effective at accelerating learning from financial time series data.

1.3 Constructing Portfolios via Distributional Prediction

Another important problem is how to convert a given prediction of the returns into an actual trading strategy. In finance, there are several well-known trading strategies. To name a few, the momentum phenomenon (Jegadeesh and Titman 1993) suggests a strategy that bets the current market trend, while the mean reversion (Poterba and Summers 1988) suggests another strategy that assumes that the stock returns moves toward the opposite side of the current direction. However, as suggested by the construction, the momentum and the reversal strategies are negatively correlated to each other, and it is generally unclear which strategy is effective for a particular market. On the other hand, modern portfolio theory (Markowitz 1952) provides a framework to determine a portfolio from distributional properties of asset prices (typically means and variances of returns). The resulting portfolio is unique in the sense that it has an optimal trade-off of returns and risks under some predefined conditions. From this perspective, distributional prediction of returns can be useful to construct trading strategies that can automatically adapt to the market.

1.4 Summary of Contributions

- We propose a novel method to extract residual information, which we call the *spectral residuals*. The spectral residuals can be calculated much faster than the classical factor analysis-based method without losing the ability to hedge out exposure to the market factors. Moreover, the spectral residuals can easily be combined with any prediction algorithms.
- We propose a new system for distributional prediction of stock prices based on deep neural networks. Our system involves two novel neural network architectures inspired by well-known invariance hypotheses on financial time series. Predicting the distributional information of returns allows us to utilize the optimal portfolio criteria offered by modern portfolio theory.
- We demonstrate the effectiveness of our proposed methods on real market data.

The supplementary material can be found in the extended version of our paper available on arXiv (Imajo et al. 2020). In the supplementary material, we also include appendices which contain detailed mathematical formulations and experimental settings, theoretical analysis, and additional experiments.

2 Preliminaries

2.1 Problem Setting

Our problem is to construct a time-dependent portfolio based on sequential observations of stock prices. Suppose that there are S stocks indexed by symbol i. The observations are given as a discrete time series of stock prices $\boldsymbol{p}^{(i)} = (p_1^{(i)}, p_2^{(i)}, \ldots, p_t^{(i)}, \ldots)$. Here, $p_t^{(i)}$ is the price of stock i at time t. We mainly consider the *return* of stocks instead of their raw prices. The return of stock i at time t is defined as $r_t^{(i)} = p_{t+1}^{(i)}/p_t^{(i)} - 1$.

A *portfolio* is a (time-dependent) vector of weights over the stocks $\boldsymbol{b}_t = (b_t^{(1)}, \ldots, b_t^{(i)}, \ldots, b_t^{(S)})$, where $b_t^{(i)}$ is the volume of the investment on stock i at time t satisfying $\sum_{i=1}^{S} |b_t^{(i)}| = 1$. A portfolio $\boldsymbol{b}_t$ is understood as a particular trading strategy, that is, $b_t^{(i)} > 0$ implies that the investor takes a long position on stock i with amount $|b_t^{(i)}|$ at time t, and $b_t^{(i)} < 0$ means a short position on the stock. Given a portfolio $\boldsymbol{b}_t$, its overall return R_t at time t is given as $R_t := \sum_{i=1}^{S} b_t^{(i)} r_t^{(i)}$. Then, given the past observations of individual stock returns, our task is to determine the value of $\boldsymbol{b}_t$ that optimizes the future returns.

An important class of portfolios is the *zero-investment portfolio* defined as follows.

Definition 1 (Zero-Investment Portfolio)**.** A zero-investment portfolio is a portfolio whose buying position and selling position are evenly balanced, i.e., $\sum_{i=1}^{S} b_t^{(i)} = 0$.

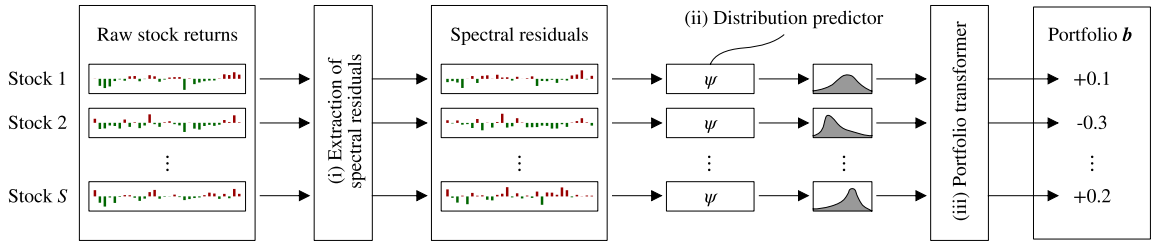

Figure 1: Overview of the proposed system. Our system consists of three parts: (i) the extraction layer of residual factors (ii) a neural network-based distribution predictor and (iii) transformation to the optimal portfolio.

In this paper, we restrict our attention to trading strategies that output zero-investment portfolio. This assumption is sensible because a zero-investment portfolio requires no equity and thus encourages a fair comparison between different strategies.

In practice, there can be delays between the observations of the returns and the actual execution of the trading. To account for this delay, we also adopt the delay parameter d in our experiments. When we trade with a d-day delay, the overall return should be $R_t := R_t^d = \sum_{i=1}^{S} b_t^{(i)} r_{t+d}^{(i)}$.

2.2 Concepts of Portfolio Optimization

According to modern portfolio theory (Markowitz 1952), investors construct portfolios to maximize expected return under a specified level of acceptable risk. The standard deviation is commonly used to quantify the risk or variability of investment outcomes, which measures the degree to which a stock's annual return deviates from its long-term historical average (Kintzel 2007).

The Sharpe ratio (Sharpe 1994) is one of the most referenced risk/return measures in finance. It is the average return earned in excess of the risk-free rate per unit of volatility. The Sharpe ratio is calculated as $(R_\mathrm{p} - R_\mathrm{f})/\sigma_\mathrm{p}$, where R_p is the return of portfolio, σ_p is the standard deviation of the portfolio's excess return, and R_f is the return of a risk-free asset (e.g., a government bond). For a zero-investment portfolio, we can always omit R_f since it requires no equity (Mitra 2009).

In this paper, we adopt the Sharpe ratio as the objective for our portfolio construction problem. Since we cannot always obtain an estimate of the total risk beforehand, we often consider sequential maximization of the Sharpe ratio of the next period. Once we predict the mean vector and the covariance matrix of the population of future returns, the optimal portfolio $\boldsymbol{b}^*$ can be solved as $\boldsymbol{b}^* = \lambda^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, where λ is a predefined parameter representing the relative risk aversion, $\boldsymbol{\Sigma}$ is the estimated covariance matrix, and $\boldsymbol{\mu}$ is the estimated mean vector (Kan and Zhou 2007). [1] Therefore, predicting the mean and the covariance is essential to construct risk averse portfolios.

[1]Note that $\boldsymbol{b}^*$ is derived as the maximizer of $\boldsymbol{b}^\top\boldsymbol{\mu} - \frac{\lambda}{2}\boldsymbol{b}^\top\boldsymbol{\Sigma}\boldsymbol{b}$, where $\boldsymbol{b}^\top\boldsymbol{\mu}$ and $\boldsymbol{b}^\top\boldsymbol{\Sigma}\boldsymbol{b}$ are the return and the risk of the portfolio $\boldsymbol{b}$, respectively.

3 Proposed System

In this section, we present the details of our proposed system, which is outlined in Figure 1. Our system consists of three parts. In the first part (i), the system extracts residual information to hedge out the effects of common market factors. To this end, in Section 3.1, we introduce the *spectral residual*, a novel method based on spectral decomposition. In the second part (ii), the system predicts future distributions of the spectral residuals using a neural network-based predictor. In the third part (iii), the predicted distributional information is leveraged for constructing optimal portfolios. We will outline these procedures in Section 3.2. Additionally, we introduce a novel network architecture that incorporates well-known financial inductive biases, which we explain in Section 3.3.

3.1 Extracting Residual Factors

As mentioned in the introduction, we focus on developing trading strategy based on the residual factors, i.e., the information remaining after hedging out the common market factors. Here, we introduce a novel method to extract the residual information, which we call the *spectral residual*.

Definition of the spectral residuals First, we introduce some notions from portfolio theory. Let $\boldsymbol{r}$ be a random vector with zero mean and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{S\times S}$, which represents the returns of S stocks over the given investment horizon. Since $\boldsymbol{\Sigma}$ is symmetric, we have a decomposition $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top$, where $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_S]$ is an orthogonal matrix and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_S)$ is a diagonal matrix of the eigenvalues. Then, we can create a new random vector as $\hat{\boldsymbol{r}} = \boldsymbol{V}^\top\boldsymbol{r}$ such that the coordinate variables $\hat{r}_i = \boldsymbol{v}_i^\top\boldsymbol{r}$ are mutually uncorrelated. In portfolio theory, $\hat{r}_i$s are called *principal portfolios* (Partovi and Caputo 2004). Principal portfolios have been utilized in "risk parity" approaches to diversify the exposures to the intrinsic source of risk in the market (Meucci 2009).

Since the volatility (i.e., the standard deviation) of the i-th principal portfolio is given as $\sqrt{\lambda_i}$, the raw return sequence has large exposure to the principal portfolios with large eigenvalues. For example, the first principal portfolio can be seen as the factor that corresponds to the overall market (Meucci 2009). Therefore, to hedge out common market factors, a natural idea is to discard several principal portfolios with largest eigenvalues. Formally, we define the spectral residuals as follows.

Definition 2. Let C ($< S$) be a given positive integer. We define the spectral residual $\tilde{\boldsymbol{\epsilon}}$ as a vector obtained by projecting the raw return vector $\boldsymbol{r}$ onto the space spanned by the principal portfolios with the smallest $S - C$ eigenvalues.

In practice, we calculate the empirical version of the spectral residuals as follows. Given a time window $H > 1$, we define a windowed signal $\boldsymbol{X}_t$ as $\boldsymbol{X}_t := [\boldsymbol{r}_{t-H}, \ldots, \boldsymbol{r}_{t-1}]$. We also denote by $\tilde{\boldsymbol{X}}_t$ a matrix obtained by subtracting empirical means of row vectors from $\boldsymbol{X}_t$. By the singular value decomposition (SVD), $\tilde{\boldsymbol{X}}_t$ can be decomposed as

$$\tilde{\boldsymbol{X}}_t = \boldsymbol{V}_t \underbrace{\mathrm{diag}(\sigma_1, \ldots, \sigma_S)\boldsymbol{U}_t}_{\text{principal portfolios}},$$

where $\boldsymbol{V}_t$ is an $S \times S$ orthogonal matrix, $\boldsymbol{U}_t$ is an $S \times H$ matrix whose rows are mutually orthogonal unit vectors, and $\sigma_1 \geq \cdots \geq \sigma_S$ are singular values. Note that $\boldsymbol{V}_t^\top \tilde{\boldsymbol{X}}_t$ can be seen as the realized returns of the principal portfolios. Then, the (empirical) spectral residual at time s ($t-H \leq s \leq t-1$) is computed as

$$\tilde{\boldsymbol{\epsilon}}_s := \boldsymbol{A}_t \boldsymbol{r}_s, \tag{2}$$

where $\boldsymbol{A}_t$ is the projection matrix defined as

$$\boldsymbol{A}_t := \boldsymbol{V}_t \mathrm{diag}(\underbrace{0, \ldots, 0}_{C}, \underbrace{1, \ldots, 1}_{S-C}) \boldsymbol{V}_t^\top.$$

Relationship to factor models Although we defined the spectral residuals through PCA, they are also related to the generative model (1), and thus convey information about "residual factors" in the original sense.

In the finance literature, it has been pointed out that trading strategies depending only on the residual factor $\boldsymbol{\epsilon}_t$ in (1) can be robust to structural changes in the overall market (Blitz, Huij, and Martens 2011; Blitz et al. 2013). While estimating parameters in the linear factor model (1) is typically done by factor analysis (FA) methods (Bartholomew, Knott, and Moustaki 2011), the spectral residual $\tilde{\boldsymbol{\epsilon}}_t$ is not exactly the same as the residual factors obtained from the FA. Despite this, we can show the following result that the spectral residuals can hedge out the common market factors under a suitable condition.

Proposition 1. Let $\boldsymbol{r}$ be a random vector in $\mathbb{R}^S$ generated according to a linear model $\boldsymbol{r} = \boldsymbol{B}\boldsymbol{f} + \boldsymbol{\epsilon}$, where $\boldsymbol{B}$ is an $S \times C$ matrix, and $\boldsymbol{f} \in \mathbb{R}^C$ and $\boldsymbol{\epsilon} \in \mathbb{R}^S$ are zero-mean random vectors. Assume that the following conditions hold:

- $\mathrm{Var}(f_i) = 1$ and $\mathrm{Var}(\epsilon_k) = \sigma > 0$.
- The coordinate variables in $\boldsymbol{f}$ and $\boldsymbol{\epsilon}$ are uncorrelated, that is, $\mathbb{E}[f_i f_j] = 0$, $\mathbb{E}[\epsilon_k \epsilon_\ell] = 0$, and $\mathbb{E}[f_i \epsilon_k] = 0$ hold for any $i \neq j$ and $k \neq \ell$.

Then, we have the followings.

(i) The spectral residual $\tilde{\boldsymbol{\epsilon}}$ defined in (2) is uncorrelated from the common factor $\boldsymbol{f}$.

(ii) The covariance matrix of $\tilde{\boldsymbol{\epsilon}}$ is given as $\sigma^2 \boldsymbol{A}_{\mathrm{res}}$. Under a suitable assumption[2], this can be approximated as a diagonal matrix, which means the coordinates variables of the spectral residual $\boldsymbol{\epsilon}_i$ ($i \in \{1, \ldots, S\}$) are almost uncorrelated.

[2]See Appendix B for a precise statement.

In the above proposition, the first statement (i) claims that the spectral residual can eliminate the common factors without knowing the exact residual factors. The latter statement (ii) justifies the diagonal approximation of the predicted covariance, which will be utilized in the next subsection. For completeness, we provide the proof in Appendix B. Moreover, the assumption that $\boldsymbol{\epsilon}$ is isotropic can be relaxed in the following sense. If we assume that the residual factor $\boldsymbol{\epsilon}$ is "almost isotropic" and the common factors $\boldsymbol{B}\boldsymbol{f}$ have larger volatility contributions than $\boldsymbol{\epsilon}$, we can show that the linear transformation used in the spectral residual is close to the projection matrix eliminating the market factors. Since the formal statement is somewhat involved, we give the details in Appendix B.

Besides, the spectral residual can be computed significantly faster than the FA-based methods. This is because the FA typically requires iterative executions of the SVD to solve a non-convex optimization problem, while the spectral residual requires it only once. Section 4.2 gives an experimental comparison of running times.

3.2 Distributional Prediction and Portfolio Construction

Our next goal is to construct a portfolio based on the extracted information. To this end, we here introduce a method to forecast future distributions of the spectral residuals, and explain how we can convert the distributional features into executable portfolios.

Distributional prediction Given a sequence of past realizations of residual factors, $\tilde{\epsilon}_{i,t-H}, \ldots, \tilde{\epsilon}_{i,t-1}$, consider the problem of predicting the distribution of a future observation $\tilde{\epsilon}_{i,t}$. Our approach is to learn a functional predictor for the conditional distribution $p(\tilde{\epsilon}_{i,t} \mid \tilde{\epsilon}_{i,t-H}, \ldots, \tilde{\epsilon}_{i,t-1})$. Since our final goal is to construct the portfolio, we only use predicted means and covariances, and we do not need the full information about the conditional distribution. Despite this, fitting symmetric models such as Gaussian distributions can be problematic because it is known that the distributions of returns are often skewed (Cont 2000; Lin and Liu 2018). To circumvent this, we utilize quantile regression (Koenker 2005), an off-the-shelf nonparametric method to estimate conditional quantiles. Intuitively, if we obtain a sufficiently large number of quantiles of the target variable, we can reconstruct any distributional properties of that variable. We train a function ψ that predicts several conditional quantile values, and convert its output into estimators of conditional means and variances. The overall procedure can be made to be differentiable, so we can incorporate it into modern deep learning frameworks.

Here, we provide the details of the aforementioned procedure. First, we give an overview for the quantile regression objective. Let Y be a scalar-valued random variable, and $\boldsymbol{X}$ be another random variable. For $\alpha \in (0, 1)$, an α-th conditional quantile of Y given $\boldsymbol{X} = \boldsymbol{x}$ is defined as

$$y(\boldsymbol{x}; \alpha) := \inf\{y' : P(Y \leq y' \mid \boldsymbol{X} = \boldsymbol{x}) \geq \alpha\}.$$

It is known that $y(\boldsymbol{x}; \alpha)$ can be found by solving the follow-

ing minimization problem

$$y(\boldsymbol{x};\alpha) = \operatorname*{argmin}_{y' \in \mathbb{R}} \mathbb{E}[\ell_\alpha(Y, y') \mid \boldsymbol{X} = \boldsymbol{x}],$$

where $\ell_\alpha(y, y')$ is the pinball loss defined as

$$\ell_\alpha(y, y') := \max\{(\alpha - 1)(y - y'), \alpha(y - y')\}.$$

For our situation, the target variable is $y_t = \tilde{\epsilon}_{i,t}$ and the explanatory variable is $\boldsymbol{x}_t = (\tilde{\epsilon}_{i,t-H}, \ldots, \tilde{\epsilon}_{i,t-1})^\top$. We want to construct a function $\psi : \mathbb{R}^H \to \mathbb{R}$ that estimate the conditional α-quantile of y_t. To this end, the quantile regression tries to solve the following minimization problem

$$\min_{\psi} \widehat{\mathbb{E}}_{y_t, \boldsymbol{x}_t}[\ell_\alpha(y_t, \psi(\boldsymbol{x}_t))].$$

Here, $\widehat{\mathbb{E}}_{y_t, \boldsymbol{x}_t}$ is understood as taking the empirical expectation with respect to y_t and $\boldsymbol{x}_t$ across t. We should note that a similar application of the quantile regression to forecasting conditional quantiles of time series has been considered in (Biau and Patra 2011).

Next, let $Q > 0$ be a given integer, and let $\alpha_j = j/Q$ $(j = 1, \ldots, Q-1)$ be an equispaced grid of quantiles. We consider the problem of simultaneously estimating α_j-quantiles by a function $\psi : \mathbb{R}^H \to \mathbb{R}^{Q-1}$. To do this, we define a loss function as

$$\mathcal{L}_Q(y_t, \psi(\boldsymbol{x}_t)) := \sum_{j=1}^{Q-1} \ell_{\alpha_j}(y_t, \psi_j(\boldsymbol{x}_t)),$$

where $\psi_j(\boldsymbol{x}_t)$ is the j-th coordinate of $\psi(\boldsymbol{x}_t)$.

Once we obtain the estimated $Q-1$ quantiles $\tilde{y}_t^{(j)} = \psi_j(\boldsymbol{x}_t)$ $(j = 1, \ldots, Q-1)$, we can estimate the future mean of the target variable y_t as

$$\hat{\mu}_t := \hat{\mu}(\tilde{\boldsymbol{y}}_t) = \frac{1}{Q-1} \sum_{j=1}^{Q-1} \tilde{y}_t^{(j)}. \tag{3}$$

Similarly, we can estimate the future variance by the sample variance of $\tilde{y}_t^{(j)}$

$$\hat{\sigma}_t := \hat{\sigma}(\tilde{\boldsymbol{y}}_t) = \frac{1}{Q-1} \sum (\tilde{y}_t^{(j)} - \hat{\mu}_t)^2 \tag{4}$$

or its robust counterpart such as the median absolute deviation (MAD).

Portfolio construction Given the estimated means and variances of future spectral residuals, we finally construct a portfolio based on optimality criteria offered by modern portfolio theory (Markowitz 1952). As mentioned in Section 2.2, the formula for the optimal portfolio requires the means and the covariances of the returns. Thanks to Proposition 1-(ii), we can approximate the covariance matrix of the spectral residual by a diagonal matrix. Precisely, once we calculate the predicted mean $\hat{\mu}_{t,j}$ and the variance $\hat{\sigma}_{t,j}^2$ of the spectral residual at time t, the weight for j-th asset is given as $\hat{b}_j := \lambda^{-1} \hat{\mu}_{t,j} / \hat{\sigma}_{t,j}^2$.

In the experiments in Section 4, we compare the performances of zero-investment portfolios. For trading strategies that do not output zero-investment portfolios, we apply a common transformation to portfolios to be centered and normalized. As a result, the eventual portfolio does not depend on the risk aversion parameter λ. See Appendix A.1 for details.

3.3 Network Architectures

For the model of the quantile predictor ψ, we introduce two architectures for neural network models that take into account scale invariances studied in finance.

Volatility invariance First, we consider an invariance property on amplitudes of financial time series. It is known that financial time series data exhibit a property called volatility clustering (Mandelbrot 1997). Roughly speaking, volatility clustering describes a phenomenon that large changes in financial time series tend to be followed by large changes, while small changes tend to be followed by small changes. As a result, if we could observe a certain signal as a financial time series, a signal obtained by positive scalar multiplication can be regarded as another plausible realization of a financial time series.

To incorporate such an amplitude invariance property into the model architectures, we leverage the class of positive homogeneous functions. Here, a function $f : \mathbb{R}^n \to \mathbb{R}^m$ is said to be positive homogeneous if $f(a\boldsymbol{x}) = af(\boldsymbol{x})$ holds for any $\boldsymbol{x} \in \mathbb{R}^n$ and $a > 0$. For example, we can see that any linear functions and any ReLU neural networks with no bias terms are positive homogeneous. More generally, we can model the class of positive homogeneous functions as follows. Let $\tilde{\psi} : S^{H-1} \to \mathbb{R}^{Q-1}$ be any function defined on the $H-1$ dimensional sphere $S^{H-1} = \{\boldsymbol{x} \in \mathbb{R}^H : \|\boldsymbol{x}\| = 1\}$. Then, we obtain a positive homogeneous function as

$$\psi(\boldsymbol{x}) = \|\boldsymbol{x}\| \tilde{\psi}\left(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}\right). \tag{5}$$

Thus, we can convert any function class on the sphere into the model of amplitude invariant predictors.

Time-scale invariance Second, we consider an invariance property with respect to time-scale. There is a well-known hypothesis that time series of stock prices have *fractal structures* (Peters 1994). The fractal structure refers to a self-similarity property of a sequence. That is, if we observe a single sequence in several different sampling rates, we cannot infer the underlying sampling rates from the shape of downsampled sequences. The fractal structure has been observed in several real markets (Cao, Cao, and Xu 2013; Mensi et al. 2018; Lee et al. 2018). See Remark 1 in Appendix A.2 for further discussion on this property.

To take advantage of the fractal structure, we propose a novel network architecture that we call *fractal networks*. The key idea is that we can effectively exploit the self-similarity by applying a single common operation to multiple subsequences with different resolutions. By doing so, we expect that we can increase sample efficiency and reduce the number of parameters to train.

Here, we give a brief overview of the proposed architecture, while a more detailed explanation will be given in Ap-

pendix A.2. Our model consists of (a) the resampling mechanism and (b) two neural networks ψ_1 and ψ_2. The input-output relation of our model is described as the following procedure. First, given a single sequence $\boldsymbol{x}$ of stock returns, the resampling mechanism $\mathrm{Resample}(\boldsymbol{x}, \tau)$ generates a sequence that corresponds to sampling rates specified by a scale parameter $0 < \tau \leq 1$. We apply $\mathrm{Resample}$ procedure for L different parameters $\tau_1 < \ldots < \tau_L = 1$ and generate L sequences. Next, we apply a common non-linear transformation ψ_1 modeled by a neural network. Finally, by taking the empirical mean of these sequences, we aggregate the information on different sampling rates and apply another network ψ_2. To sum up, the overall procedure can be written in the following single equation

$$\psi(\boldsymbol{x}) = \psi_2 \left(\frac{1}{L} \sum_{i=1}^{L} \psi_1(\mathrm{Resample}(\boldsymbol{x}, \tau_i)) \right). \qquad (6)$$

4 Experiments

We conducted a series of experiments to demonstrate the effectiveness of our methods on real market data. In Section 4.1, we describe the details of the dataset and some common experimental settings used throughout this section. In Section 4.2, we test the validity of the spectral residual by a preliminary experiment. In Section 4.3, we evaluate the performance of our proposed system by experiments on U.S. market data. We also conducted similar experiments on Japanese market data and obtained consistent results. Due to the space limitation, we provide the entire results for Japanese market data in Appendix E.

4.1 Dataset Description and Common Settings

U.S. market data For U.S. market data, we used the daily prices of stocks listed in S&P 500 from January 2000 to April 2020. We used data before January 2008 for training and validation and the remainder for testing. We obtained the data from Alpha Vantage [3].

We used opening prices because of the following reasons. First, the trading volume at the opening session is larger than that at the closing session (Amihud and Mendelson 1987), which means that trading with opening prices is practically easier than trading with the closing prices. Moreover, a financial institution cannot trade a large amount of stocks during the closing period because it can be considered as an illegal action known as "banging the close".

Common experimental settings We adopted the delay parameter $d = 1$ (i.e., one-day delay) for updating portfolios. We set the look-back window size as $H = 256$, i.e., all prediction models can access the historical stock prices upto preceding 256 business days. For other parameters used in the experiments, see Appendix C.3 as well.

Evaluation metrics We list the evaluation metrics used throughout our experiments.

- *Cumulative Wealth (CW)* is the total return yielded from the trading strategy: $\mathrm{CW}_T := \prod_{i=1}^{T}(1 + R_t)$.

[3] https://www.alphavantage.co/

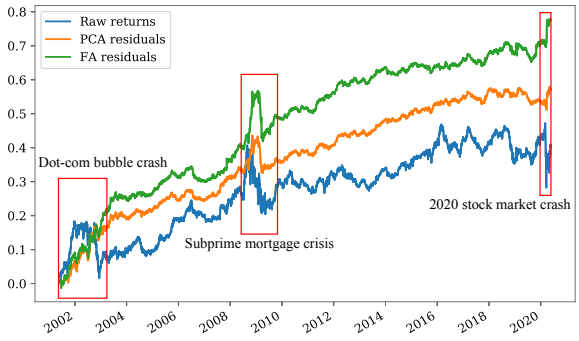

Figure 2: Cumulative returns of reversal strategies over raw returns, the FA residuals, and the spectral residuals. The reversal-based strategies are more robust against financial crises.

- *Annualized Return (AR)* is an annualized return rate defined as $\mathrm{AR}_t := (T_\mathrm{Y}/T) \sum_{i=1}^{T} R_t$, where T_Y is the average number of holding periods in a year.
- *Annualized Volatility (AVOL)* is annualized risk defined as $\mathrm{AVOL}_T := ((T_\mathrm{Y}/T) \sum_{i=1}^{T} R_t^2)^{1/2}$.
- *Annualized Sharpe ratio (ASR)* is an annualized risk-adjusted return (Sharpe 1994). It is defined as $\mathrm{ASR}_T := \mathrm{AR}_T / \mathrm{AVOL}_T$.

As mentioned in Section 2.2, we are mainly interested in ASR as the primary evaluation metric. AR and AVOL are auxiliary metrics for calculating ASR. While CW represents the actual profits, it often ignores the existence of large risk values. In addition to these, we also calculated some evaluation criteria commonly used in finance: Maximum Draw-Down (MDD), Calmar Ratio (CR), and Downside Deviation Ratio (DDR). For completeness, we provide precise definitions in Appendix C.1.

4.2 Validity of Spectral Residuals

As suggested in 3.1, the spectral residuals can be useful to hedge out the undesirable exposure to the market factors. To verify this, we compared the performances of trading strategies over (i) the raw returns, (ii) the residual factors extracted by the factor analysis (FA), and (iii) the spectral residuals.

For the FA, we fit the factor model (1) with $K = 30$ by the maximum likelihood method (Bartholomew, Knott, and Moustaki 2011) and extracted residual factors as the remaining part. For the spectral residual, we obtain the residual sequence by subtracting $C = 30$ principal components from the raw returns. We applied both methods to windowed data with length $H = 256$.

In order to be agnostic to the choice of training algorithms, we used a simple reversal strategy. Precisely, for the raw return sequence $\boldsymbol{r}_t$, we used a deterministic strategy obtained simply by normalizing the negation of the previous observation $-\boldsymbol{r}_{t-1}$ to be a zero-investment portfolio (see Appendix C.2 for the precise formula). We defined reversal strategies over residual sequences in similar ways.

Figure 2 shows the cumulative returns of reversal strategies performed on the Japanese market data. We see that the

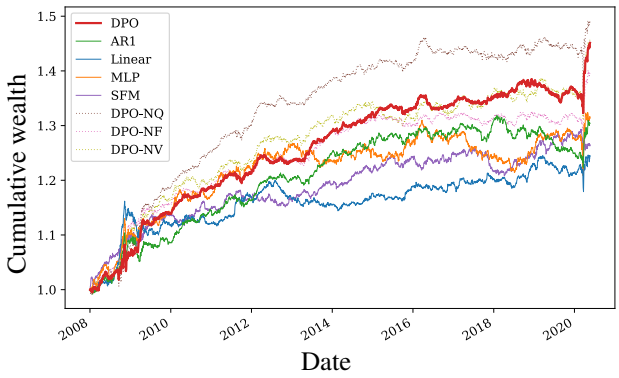

Figure 3: The Cumulative Wealth in U.S. market.

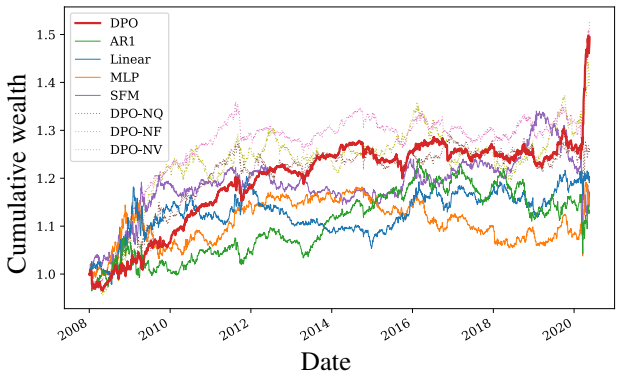

Figure 4: The Cumulative Wealth in U.S. market data without spectral residual extraction.

reversal strategy based on the raw returns is significantly affected by several well-known financial crises, including the dot-com bubble crush in the early 2000s, the 2008 subprime mortgage crisis, and the 2020 stock market crush. On the other hand, two residual-based strategies seem to be more robust against these financial crises. The spectral residual performed similarly to the FA residual in cumulative returns. Moreover, in terms of the Sharpe ratio, the spectral residuals ($= 2.86$) performed better than the FA residuals ($= 2.64$).

Remarkably, the spectral residuals were calculated much faster than the FA residuals. In particular, we calculated both residuals using the entire dataset which contains records of all the stock prices for $5,000$ days. For the PCA and the FA, we used implementations in the scikit-learn package (Pedregosa et al. 2011), and all the computation were run on 18 CPU cores of Intel Xeon Gold 6254 Processor (3.1 GHz). Then, extracting the spectral residuals took approximately 10 minutes, while the FA took approximately 13 hours.

4.3 Performance Evaluation of the Proposed System

We evaluated the performance of our proposed system described in Section 3 on U.S. market data. Corresponding results for Japanese market data are provided in Appendix E.

Baseline Methods We compare our system with the following baselines: (i) `Market` is the uniform buy-and-hold strategy. (ii) `AR(1)` is the $\mathrm{AR}(1)$ model with all coefficients being -1. This can be seen as the simple reversal strategy. (iii) `Linear` predicts returns by ordinary linear regression based on previous H raw returns. (iv) `MLP` predicts returns by a multi-layer perceptron with batch normalization and dropout (Pal and Mitra 1992; Ioffe and Szegedy 2015; Srivastava et al. 2014). (v) `SFM` is one of state-of-the-art stock price prediction algorithms based on the State Frequency Memory RNNs (Zhang, Aggarwal, and Qi 2017).

Additionally, we compare our proposed system (`DPO`) with some ablation models, which are similar to `DPO` except for the following points.

- DPO with No Quantile Prediction (`DPO-NQ`) does not use the information of the full distributional prediction, but instead it outputs conditional means trained by the L_2 loss.
- DPO with No Fractal Network (`DPO-NF`) uses a simple multi-layer perceptron instead of the fractal network.
- DPO with No Volatility Normalization (`DPO-NV`) does not use the normalization (5) in the fractal network.

	ASR↑	AR↑	AVOL↓	DDR↑	CR↑	MDD↓
Market	+0.607	**+0.130**	0.215	+0.939	+0.263	0.496
AR(1) on SRes	+0.858	+0.021	0.025	+1.470	+0.295	0.072
Linear on SRes	+0.724	+0.017	0.024	+1.262	+0.298	0.059
MLP on SRes	+0.728	+0.022	0.030	+1.280	+0.283	0.077
SFM on SRes	+0.709	+0.019	0.026	+1.211	+0.323	0.058
DPO-NQ	+1.237	+0.032	0.026	+2.169	+0.499	0.063
DPO-NF	+1.284	+0.027	**0.021**	+2.347	+0.627	**0.042**
DPO-NV	+1.154	+0.030	0.026	+2.105	+0.562	0.053
DPO (Proposed)	**+1.393**	+0.030	**0.021**	**+2.561**	**+0.656**	0.045

Table 1: Performance comparison on U.S. market. All the methods except for `Market` are applied to the spectral residuals (SRes).

Performance on real market data Figure 3 shows the cumulative wealth (CW) achieved in U.S. market. Table 1 shows the results for the other evaluation metrics presented in Section 4.1. For parameter C of the spectral residual, we used $C = 10$, which we determined sorely from the training data (see Appendix D.2 for details). Overall, our proposed method `DPO` outperformed the baseline methods in multiple evaluation metrics. Regarding the comparison against three ablation models, we make the following observations.

1. **Effect of the distributional prediction**. We found that introducing the distributional prediction significantly improved the ASR. While `DPO-NQ` achieved the best CW, `DPO` performed better in the ASR. It suggests that, without the variance prediction, `DPO-NQ` tends to pursue the returns without regard to taking the risks. Generally, we observed that `DPO` reduced the AVOL while not losing the AR.

	ASR↑	AR↑	AVOL↓	DDR↑	CR↑	MDD↓
AR(1)	+0.212	+0.011	0.051	+0.355	+0.067	0.160
Linear	+0.304	+0.016	0.052	+0.485	+0.127	0.125
MLP	+0.261	+0.013	0.048	+0.424	+0.103	0.122
SFM	+0.264	+0.014	0.051	+0.428	+0.079	0.171
DPO-NQ	+0.405	+0.020	0.048	+0.655	+0.172	0.114
DPO-NF	+0.854	**+0.034**	0.040	+1.472	+0.436	0.078
DPO-NV	+0.542	+0.029	0.054	+0.922	+0.238	0.123
DPO	**+0.874**	+0.032	**0.037**	**+1.485**	**+0.524**	**0.061**

Table 2: Performance comparison on U.S. market without spectral residual extraction.

2. **Effect of the fractal network**. Introducing the fractal network architecture also improved the performance in multiple evaluation metrics. In both markets, we observed that the fractal network contributed to increasing the AR while keeping the AVOL, which is suggestive of the effectiveness of leveraging the financial inductive bias on the return sequence.
3. **Effect of the normalization**. We also saw the effectiveness of the normalization (5). Comparing `DPO` and `DPO-NV`, the normalization affected both of the AR and the AVOL, resulting in the improvement in the ASR. This may occur because the normalization improves the sample efficiency by reducing the degrees of freedom of the model.

To see the effect of the spectral residuals, we also evaluated our proposed method and the baseline methods on the raw stock returns. Figure 4 and Table 2 show the results. Compared to the corresponding results with spectral residuals, we found that the spectral residuals consistently improved the performance for every method. Some further intriguing observations are summarized as follows.

1. With the spectral residuals, `AR(1)` achieved the best ASR among the baseline methods (Table 1), which has not been observed on the raw return sequence (Table 2). This suggests that the spectral residuals encourage the reversal phenomenon (Poterba and Summers 1988) by suppressing the common market factors. Interestingly, without extracting the spectral residuals, the CWs are crossing during the test period, and no single baseline method consistently beats others (Figure 4). A possible reason is that the strong correlation between the raw stock returns increases the exposure to the common market risks.
2. We found that our network architectures are still effective on the raw sequences. In particular, `DPO` outperformed all the other methods in multiple evaluation metrics.

5 Related Work

5.1 Factor Models and Residual Factors

Trading based on factor models is one of the popular strategies for quantitative portfolio management (e.g., (Nakagawa, Uchida, and Aoshima 2018)). One of the best-known factor models is Fama and French(Fama and French 1992, 1993), and they put forward a model explaining returns in the US equity market with three factors: the market return factor, the size (market capitalization) factor and the value (book-to-market) factor.

Historically, residual factors are treated as errors in factor models(Sharpe 1964). However, (Blitz, Huij, and Martens 2011; Blitz et al. 2013) suggested that there exists predictability in residual factors. In modern portfolio theory, less correlation of investment returns enables to earn larger risk-adjusted returns (Markowitz 1952). Residual factors are less correlated than the raw stock returns by its nature. Consequently, (Blitz, Huij, and Martens 2011; Blitz et al. 2013) demonstrated that residual factors enable to earn larger risk-adjusted returns.

5.2 Asset Price Prediction using Deep Learning

With the recent advance of deep learning, various deep neural networks are applied to stock price prediction (Chen et al. 2019). Some deep neural networks for time series are also applied to stock price prediction (Fischer and Krauss 2018).

Compared to other classical machine learning methods, deep learning enables learning with fewer a priori representational assumptions if provided with sufficient amount of data and computational resources. Even if data is insufficient, introducing inductive biases to a network architecture can still facilitate deep learning (Battaglia et al. 2018). Technical indicators are often used for stock prediction (e.g., (Metghalchi, Marcucci, and Chang 2012; Neely et al. 2014)), and (Li et al. 2019) used technical indicators as inductive biases of a neural network. (Zhang, Aggarwal, and Qi 2017) used a recurrent model that can analyze frequency domains so as to distinguish trading patterns of various frequencies.

6 Conclusions

We proposed a system for constructing portfolios. The key technical ingredients are (i) a spectral decomposition-based method to hedge out common market factors and (ii) a distributional prediction method based on a novel neural network architecture incorporating financial inductive biases. Through empirical evaluations on the real market data, we demonstrated that our proposed method can significantly improve the performance of portfolios on multiple evaluation metrics. Moreover, we verified that each of our proposed techniques is effective on its own, and we believe that our techniques may have wide applications in various financial problems.

Acknowledgments

We thank the anonymous reviewers for their constructive suggestions and comments. We also thank Masaya Abe, Shuhei Noma, Prabhat Nagarajan and Takuya Shimada for helpful discussions.

References

Amihud, Y.; and Mendelson, H. 1987. Trading mechanisms and stock returns: An empirical investigation. *The Journal of Finance* 42(3): 533–553.

Bartholomew, D.; Knott, M.; and Moustaki, I. 2011. *Latent Variable Models and Factor Analysis: A Unified Approach*. Wiley, 3rd edition.

Battaglia, P. W.; Hamrick, J. B.; Bapst, V.; Sanchez-Gonzalez, A.; Zambaldi, V.; Malinowski, M.; Tacchetti, A.; Raposo, D.; Santoro, A.; Faulkner, R.; et al. 2018. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261* .

Biau, G.; and Patra, B. 2011. Sequential Quantile Prediction of Time Series. *IEEE Transactions on Information Theory* 57(3): 1664–1674.

Blitz, D.; Huij, J.; Lansdorp, S.; and Verbeek, M. 2013. Short-term residual reversal. *Journal of Financial Markets* 16(3): 477–504.

Blitz, D.; Huij, J.; and Martens, M. 2011. Residual momentum. *Journal of Empirical Finance* 18(3): 506–521.

Calomiris, C. W.; Love, I.; and Peria, M. S. M. 2010. Crisis" Shock Factors" and the Cross-Section of Global Equity Returns. Technical report, National Bureau of Economic Research.

Cao, G.; Cao, J.; and Xu, L. 2013. Asymmetric multifractal scaling behavior in the Chinese stock market: Based on asymmetric MF-DFA. *Physica A: Statistical Mechanics and its Applications* 392(4): 797–807.

Chen, C.; Zhao, L.; Bian, J.; Xing, C.; and Liu, T.-Y. 2019. Investment behaviors can tell what inside: Exploring stock intrinsic properties for stock trend prediction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2376–2384.

Choudhry, R.; and Garg, K. 2008. A hybrid machine learning system for stock market forecasting. *World Academy of Science, Engineering and Technology* 39(3): 315–318.

Cont, R. 2000. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance* 1: 223–236.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Fama, E. F.; and French, K. R. 1992. The Cross-Section of Expected Stock Returns. *The Journal of Finance* 47(2): 427–465.

Fama, E. F.; and French, K. R. 1993. Common risk factors in the returns on stocks and bonds. *Journal of* .

Fama, E. F.; and French, K. R. 2015. A five-factor asset pricing model. *Journal of Financial Economics* 116(1): 1–22.

Fischer, T.; and Krauss, C. 2018. Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research* 270(2): 654–669.

Graves, A.; Mohamed, A.; and Hinton, G. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 6645–6649. ISSN 2379-190X.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation* 9(8): 1735–1780.

Imajo, K.; Minami, K.; Ito, K.; and Nakagawa, K. 2020. Deep Portfolio Optimization via Distributional Prediction of Residual Factors. *arXiv preprint arXiv:2012.07245* .

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* .

Jegadeesh, N.; and Titman, S. 1993. Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *The Journal of Finance* 48(1): 65–91.

Kan, R.; and Zhou, G. 2007. Optimal portfolio choice with parameter uncertainty. *Journal of Financial and Quantitative Analysis* 42(3): 621–656.

Kintzel, D. 2007. Portfolio theory, life-cycle investing, and retirement income. *Social Security Administration Policy Brief* 2007-02.

Koenker, R. 2005. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press. doi:10.1017/CBO9780511754098.

Krauss, C.; Do, X. A.; and Huck, N. 2017. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research* 259(2): 689–702.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In Pereira, F.; Burges, C. J. C.; Bottou, L.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 25*, 1097–1105. Curran Associates, Inc. URL http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

LeCun, Y.; Haffner, P.; Bottou, L.; and Bengio, Y. 1999. Object Recognition with Gradient-Based Learning. In *Shape, Contour and Grouping in Computer Vision*, 319–345. Springer Berlin Heidelberg.

Lee, M.; Song, J.; Kim, S.; and Chang, W. 2018. Asymmetric market efficiency using the index-based asymmetric-MFDFA. *Physica A: Statistical Mechanics and its Applications* 512(15): 1278–1294.

Li, Z.; Yang, D.; Zhao, L.; Bian, J.; Qin, T.; and Liu, T.-Y. 2019. Individualized indicator for all: Stock-wise technical indicator optimization with stock embedding. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 894–902.

Lin, T.-C.; and Liu, X. 2018. Skewness, individual investor preference, and the cross-section of stock returns. *Review of Finance* 22(5): 1841–1876.

Lux, T.; and Marchesi, M. 2000. Volatility clustering in financial markets: a microsimulation of interacting agents. *International journal of theoretical and applied finance* 3(04): 675–702.

Malkiel, B. G.; and Fama, E. F. 1970. Efficient capital markets: A review of theory and empirical work. *The journal of Finance* 25(2): 383–417.

Mandelbrot, B. B. 1997. The variation of certain speculative prices. In *Fractals and scaling in finance*, 371–418. Springer.

Markowitz, H. 1952. Portfolio selection. *The journal of finance* 7(1): 77–91.

Mensi, W.; Hamdi, A.; Shahzad, S. J. H.; Shafiullah, M.; and Al-Yahyaee, K. H. 2018. Modeling cross-correlations and efficiency of Islamic and conventional banks from Saudi Arabia: Evidence from MF-DFA and MF-DXA approaches. *Physica A: Statistical Mechanics and its Applications* 502(15): 576–589.

Metghalchi, M.; Marcucci, J.; and Chang, Y.-H. 2012. Are moving average trading rules profitable? Evidence from the European stock markets. *Applied Economics* 44(12): 1539–1559.

Meucci, A. 2009. Managing Diversification. *Risk* 22(5): 74–79.

Mitra, S. K. 2009. Optimal combination of trading rules using neural networks. *International business research* 2(1): 86–99.

Nakagawa, K.; Uchida, T.; and Aoshima, T. 2018. Deep factor model. In *ECML PKDD 2018 Workshops*, 37–50. Springer.

Neely, C. J.; Rapach, D. E.; Tu, J.; and Zhou, G. 2014. Forecasting the equity risk premium: the role of technical indicators. *Management science* 60(7): 1772–1791.

Pal, S. K.; and Mitra, S. 1992. Multilayer perceptron, fuzzy sets, classifiaction. *IEEE Transactions on Neural Networks* 3(5): 683–697.

Partovi, M. H.; and Caputo, M. 2004. Principal portfolios: recasting the efficient frontier. *Economics Bulletin* 7(3): 1–10.

Pasini, G. 2017. Principal component analysis for stock portfolio management. *International Journal of Pure and Applied Mathematics* 115(1): 153–167.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.

Peters, E. E. 1994. *Fractal market analysis: applying chaos theory to investment and economics*, volume 24. John Wiley & Sons.

Poterba, J. M.; and Summers, L. H. 1988. Mean reversion in stock prices: Evidence and Implications. *Journal of Financial Economics* 22(1): 27–59.

Shah, V. H. 2007. Machine learning techniques for stock prediction. *Foundations of Machine Learning— Spring* 1(1): 6–12.

Sharpe, W. F. 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance* 19(3): 425–442.

Sharpe, W. F. 1994. The sharpe ratio. *Journal of portfolio management* 21(1): 49–58.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15(56): 1929–1958.

Szado, E. 2009. VIX futures and options: A case study of portfolio diversification during the 2008 financial crisis. *The Journal of Alternative Investments* 12(2): 68–85.

Wang, J.; Zhang, Y.; Tang, K.; Wu, J.; and Xiong, Z. 2019. AlphaStock: A Buying-Winners-and-Selling-Losers Investment Strategy using Interpretable Deep Reinforcement Attention Networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1900–1908.

Zhang, L.; Aggarwal, C.; and Qi, G.-J. 2017. Stock price prediction via discovering multi-frequency trading patterns. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2141–2149.