Optimal Linear Signal: An Unsupervised Machine Learning Framework to Optimize PnL with Linear Signals.

Pierre RENUCCI

Oct.-Nov. 2023

Abstract

This study presents an unsupervised machine learning approach for optimizing Profit and Loss (PnL) in quantitative finance. Our algorithm, akin to an unsupervised variant of linear regression, maximizes the Sharpe Ratio of PnL generated from signals constructed linearly from exogenous variables.

The methodology employs a linear relationship between exogenous variables and the trading signal, with the objective of maximizing the Sharpe Ratio through parameter optimization. Empirical application on an ETF representing U.S. Treasury bonds demonstrates the model's effectiveness, supported by regularization techniques to mitigate overfitting. The study concludes with potential avenues for further development, including generalized time steps and enhanced corrective terms.

Introduction

In the field of quantitative finance, the creation of signals to generate Profit and Loss (PnL) is a key component. Our initial objective was to find a simple unsupervised machine learning (ML) algorithm capable of exploiting a set of exogenous variables to produce a PnL-effective signal. We observed that the literature in quantitative finance largely favors supervised ML approaches, aiming to predict specific financial magnitudes, such as asset prices, as highlighted in the works of Johnson (2023) [1], Rouf et al. [2], Soni et al. (2022) [3], and Kumar et al. (2022) [4].

There are also unsupervised techniques, as presented by Kelly and Xiu (2023) [5], and Hoang and Wiegratz (2023) [6]. However, these methods are very specific and lack the versatility to be used as general tools, specifically for small dataset. An exception in this category is the generation of signals via Principal Component Analysis (PCA), as explained by Ghorbani and Chong (2020) [7]. Nevertheless, this technique offers little flexibility and does not allow for the explicit optimization of specific criteria within the PnL.

We therefore turned our attention to the literature on optimization in finance, with studies such as those by Jurczenko et al. (2019) [8], Huang et al. (2019) [9], and Cornuéjols et al. (2017, 2018) [10] [11], Reppen et al. (2022) [12], primarily focusing on portfolio optimization. However, these works are more concerned with optimizing portfolios than signal creation.

As a result, we decided to construct our own algorithm, developing an unsupervised counterpart to linear regression aimed at optimizing the Sharpe Ratio of a PnL. The model is based on two hypotheses of linearity: the linearity of the relationship between the exogenous variables and the signal; and the linearity of the relationship between the PnL and the signal. Using these, we can deduce a parametric representation of the PnL. We sought to optimize these parameters to maximize the Sharpe Ratio of the obtained PnL.

The optimal parameters are calculated over a certain training period and are then used to generate the signal subsequently. Other techniques, mainly regularization and correction of this signal, are then addressed and ultimately enable the creation of a very effective strategy on a backtest of about twenty years.

Executive Summary

- We consider an asset with a specific price series price_t and a set of *stationary* and *homoscedastic* variables $X_{1,t},\ X_{2,t},\ \ldots,\ X_{n,t}$.
- The objective is to derive the 'optimal' linear signal extracted from this variable. We work under two linear assumptions:
  - We will consider the signal as a linear combination of the exogenous variables:

$$\mathsf{signal}_t = \alpha_0 + \alpha_1 X_{1,t} + \alpha_2 X_{2,t} + \ldots + \alpha_n X_{n,t}$$

  - At each time step t, the positions taken is proportional to both the signal and the price $\mathsf{pos}_t = \mathsf{price}_t \times \mathsf{signal}_t$ (a negative value would correspond to a short position). This approach establishes a linear relationship between Profit and Loss (PnL) and the trading signal. We get, after computation:

$$\mathsf{PnL}_t = \mathsf{signal}_{t-1} \times (\mathsf{price}_t - \mathsf{price}_{t-1})$$

  This allows us to express PnL as a linear combination of the variables:

$$\mathsf{PnL}_t = \alpha^{\mathrm{T}}[X_{t-1} \times (\mathsf{price}_t - \mathsf{price}_{t-1})]$$

- A parametric expression of the empiric Sharpe Ratio is then derived:

$$\mathcal{L}(\alpha) := \frac{\overline{PnL}}{\sigma(PnL)} = \frac{\alpha^{\mathrm{T}}\mu}{\sqrt{\alpha^{\mathrm{T}}\Sigma\alpha}}$$

  With μ_i and $\Sigma_{i,j}$ the empiric mean and covariance of the variables $(X_{i,t-1} \times (\mathsf{price}_t - \mathsf{price}_{t-1}))_i$

  The optimality criterion is to maximize this empirical Sharpe Ratio. Transforming this into an optimization problem, the alpha that maximizes the objective function $\mathcal{L}(\alpha)$ is:

$$\hat{\alpha} = \frac{\Sigma^{-1}\mu}{\sqrt{\mu^{\mathrm{T}}\Sigma^{-1}\mu}}$$

  These optimal coefficients, upon examination, are seen to create a signal with the highest correlation to the asset's price variations, constrained by a low variance in this correlation. The constraint's trade-off is mecanically set to maximize the sharpe ratio.

- This model can then be utilized as an unsupervised Machine Learning model for finance strategies: trained on a dataset composed of the price and exogenous variables of the last τ days ($t-\tau$, ..., $t-1$), and used to create the present-time signal at t.

  Additional engineering techniques, mainly feature engineering of the exogenous variables, specific regularization techniques, and a corrective factor, are employed to develop and enhance the quantitative finance strategies based on this model.

- This model was tested to create a buy/sell strategy on a specific asset ('IEF', a widely traded Exchange Traded Fund (ETF) that mirrors the performance of U.S. Treasury bonds with maturities of 1-3 years). The strategy yielded qualitative results, exhibiting good metrics in terms of risk-adjusted return: an effective Sharpe ratio of 1.2 when backtested over the period 2000-2023.

  However, a high turnover mitigates the quality of this specific strategy, indicating a need for further improvements.

1 Problem Formulation

Consider an asset whose price evolves according to the time series $(\mathsf{price}_t)_t$. An investment strategy for this asset entails the creation of a time series $(\mathsf{pos}_t)_t$, contingent on time, representing a position on this asset, i.e., the number of shares purchased of this asset multiplied by its price. Once a position pos_t is determined, it is straightforward to compute the PnL: indeed, we have a asset value's variation term $(\mathsf{pos}_t - \mathsf{pos}_{t-1})$ which corresponds to the value variation of the positions and a cashflow term: $\mathsf{price}_t \times (\frac{\mathsf{pos}_{t-1}}{\mathsf{price}_{t-1}} - \frac{\mathsf{pos}_t}{\mathsf{price}_t})$ which corresponds to encashing (resp. disbursing) the value difference of the positions sold (resp. bought) during the period.

Hence, the sum of these two terms is: $\boxed{\mathsf{PnL}_t = \mathsf{pos}_{t-1}\dfrac{\mathsf{price}_t - \mathsf{price}_{t-1}}{\mathsf{price}_{t-1}}}$

Subsequently, this time series pos_t is primarily determined by constructing a principal signal in the form of a time series signal_t. Thus, in the general case, $\mathsf{pos}_t = f_t(\mathsf{price}_t, \mathsf{signal}_t)$. It then becomes essential to ascertain both f_t and signal_t to optimize the PnL. This signal can be constructed in numerous ways, either technically or by utilizing fine economic relations between the asset price and certain exogenous variables. In our case, we will focus on the creation of a signal with a position-taking form as follow :

$$\boxed{\mathsf{pos}_t = \mathsf{price}_t \times \mathsf{signal}_t} \tag{H1}$$

which corresponds to the case where one decides to hold exactly signal_t shares of the asset at any time t. Therefore we can deduce the relation between PnL_t, price_t and signal_t :

$$\boxed{\mathsf{PnL}_t = \mathsf{signal}_{t-1} \times (\mathsf{price}_t - \mathsf{price}_{t-1})}$$

Now let's consider a set of *stationary* and *homoscedastic* variables denoted as $X_{1,t}, X_{2,t}, \ldots, X_{n,t}$ and create a signal as a linear combination of these variables. The sought-after signal is:

$$\boxed{\mathsf{signal}_t = \alpha_0 + \alpha_1 X_{1,t} + \alpha_2 X_{2,t} + \ldots + \alpha_n X_{n,t}} \tag{H2}$$

(That we will note with matrix notation: $\mathsf{signal}_t = \alpha^{\mathrm{T}} X_t$ with: $X_{0,t} = 1$ for the intercept)
Multiplying equation (H2) by $(\mathsf{price}_t - \mathsf{price}_{t-1})$ and then substituting relation between PnL_t, price_t and signal_t, into yields:

$$\boxed{\mathsf{PnL}_t = \alpha^{\mathrm{T}} \tilde{X}_t} \text{ where } \tilde{X}_{i,t} = (\mathsf{price}_t - \mathsf{price}_{t-1}) X_{i,t-1}$$

Now that a parametric expression for our PnL has been established, our goal is to define a metric for optimizing our PnL. The Sharpe Ratio, acts as a gauge for evaluating the risk-adjusted performance of a trading strategy. It is computed as the mean return of the strategy minus the risk-free rate (in practice in quant finance, and in this paper, the risk-free rate is set to 0), divided by the standard deviation of the return, thus rendering a normalization for volatility. This ratio is usefull as it elucidates the return potential but also encapsulates the inherent risk, thereby offering a comprehensive measure of a strategy's efficacy and robustness.

The objective function we aim to maximize is thus defined as the Sharpe Ratio of the PnL, computed over the expectancy and standard deviation:

$$\mathcal{L} := \frac{\mathbb{E}[\mathsf{PnL}]}{\sigma[\mathsf{PnL}]}$$

2 Mathematical Analysis and Optimization

Using the notations $\mu^{\mathrm{T}} = \left[\, \overline{\tilde{X}_{0,t}}, \quad \overline{\tilde{X}_{1,t}}, \quad \ldots, \quad \overline{\tilde{X}_{n,t}} \,\right]$ and $\Sigma = \left[\mathrm{cov}(\tilde{X}_{i,t}, \tilde{X}_{j,t})_{i,j}\right]$ for the empiric mean vector and covariance matrix of $\tilde{X}_t$, we have:

$$\overline{\mathsf{PnL}} = \alpha^{\mathrm{T}} \mu \text{ and } \sigma_{emp.}(\mathsf{PnL}) = \sqrt{\alpha^{\mathrm{T}} \Sigma \alpha}$$

Thus, we can compute the objective function as the empiric sharpe ratio:

$$\mathcal{L}(\alpha) = \frac{\alpha^{\mathrm{T}}\mu}{\sqrt{\alpha^{\mathrm{T}}\Sigma\alpha}}$$

For an invertible Σ, the closed-form solution for the α that maximize $\mathcal{L}$ is[1]:

$$\boxed{\hat{\alpha} = \frac{\Sigma^{-1}\mu}{\sqrt{\mu^{\mathrm{T}}\Sigma^{-1}\mu}}}$$

Finally the Optimal Linear Signal extracted from the exogenous variables can be computed by: $\hat{\alpha}^{\mathrm{T}}X_t$

How to Interpret These Results?

A 'good signal' is defined as a time series signal_t whose 1-day lagged series signal_{t-1} exhibits a high correlation with the asset's price variation $(\mathsf{price}_t - \mathsf{price}_{t-1})$. The methodology we present aims to construct a linear signal that maximally captures the most of this correlation using n exogenous variables, under the constraint of a low variation of this correlation.

In effect, μ_i are estimator of $\mathbb{E}[(\mathsf{price}_t - \mathsf{price}_{t-1}) \times X_{i,t}]$, and actually represents the empirical covariance between the series $(X_{i,t-1})_t$ and the price variations $(\mathsf{price}_t - \mathsf{price}_{t-1})$. The corrective term involving Σ^{-1} is introduced to reduce the variance of the PnL generated by this method.

The trade-off is mecanically set to maximize the sharpe ratio.

Remarks:

- The presence of non-stationarity in the X_t variable introduces bias, predominantly captured by the intercept term α_0. Essentially, the mean contribution of each variable influences the intercept, potentially assigning it an irrelevant value that fails to capture information effectively.
- Heteroscedasticity can introduce bias in the model. This phenomenon occurs because variables with larger amplitude variations are perceived by the model as being more significant compared to others. Such disproportionate emphasis on certain variables due to their variance can skew the model's performance, leading to biased outcomes.

Then, a good practice is to standardize the exogenous variables X_t before using them in the model, it forces them into stationnarity and homoscedasticity.

2.1 Achieving beta neutrality

A limitation of this optimizatio problem is that it is not inherently beta neutral. The signal might correlate with the asset's price or, by extension, the market. However, at the expense of some performance, it is feasible to introduce a constraint to decorrelate the signal from the asset's price.

Incorporating the constraint $\alpha^{\mathrm{T}}\beta = 0$ into the optimization problem alters the vector μ to

$$\tilde{\mu} = \mu - \frac{\mu^{\mathrm{T}}\Sigma^{-1}\beta}{\beta^{\mathrm{T}}\Sigma^{-1}\beta}\beta$$

and then use this altered μ in the optimal beta formula given in 2.

By setting $\beta = \overline{\mathsf{price}_t \times X_t}$, the average of the product of X_t and price_t, we generate a signal uncorrelated with the asset's price, achieving a form of beta neutrality. We can also choose any y_t and $\beta = \overline{y_t \times X_t}$ in order to get a signal uncorrelated to y_t.

[1] We can compute the gradient of $\mathcal{L}$ which is null when $\Sigma\alpha \propto \mu$:

$$\nabla\mathcal{L}(\alpha) = \frac{\mu(\alpha^{\mathrm{T}}\Sigma\alpha) - \Sigma\alpha}{(\alpha^{\mathrm{T}}\Sigma\alpha)^{3/2}}$$

The coefficient of proportionnarity is free since $\mathcal{L}(|cst| \times \alpha) = \mathcal{L}(\alpha)$ we have choosen a positive one that gives: $\alpha^{\mathrm{T}}\Sigma\alpha = 1$ for a unitary standard deviation of the PnL and a positive expectancy.

2.2 Regularization techniques

L1 regularization: When confronting potential redundancy among exogenous variables, L1 regularization emerges as a viable solution. It modifies the objective function to:

$$\tilde{\mathcal{L}}(\alpha) = \mathcal{L}(\alpha) - \lambda_1|\alpha|$$

thereby promoting sparsity in the parameter vector α. No closed-form solution is available in this case, necessitating numerical optimization to find: $\hat{\alpha} = \text{argmax}\,\tilde{\mathcal{L}}$. To maintain parameter comparability and interpretability an alternative formulation can be considered:

$$\tilde{\mathcal{L}}(\alpha) = \mathcal{L}(\alpha) - (\lambda_1 \times \max \mathcal{L}) \times |\alpha|$$

L2 regularization: For non-invertible covariance matrix Σ, L2 regularization can be utilized to ensure invertibility by adding a scaled identity matrix, resulting in:

$$\tilde{\Sigma} = \Sigma + \lambda_2 Id$$

However, to preserve parameter comparability and interpretability, an alternative transformation can be employed:

$$\tilde{\Sigma} = \frac{(\Sigma + \lambda_2 \frac{||\Sigma||}{n} Id)}{(1 + \lambda_2)}$$

PCA regularization: Another approach involves the use of PCA (Principal Component Analysis) before calculating μ and Σ, in order to select only the top k principal components of $\tilde{X}_t$. Caution is required: the PCA should be performed on $\tilde{X}_t$ rather than X_t. This is because the valuable information resides in the covariance of the transformed variables $\tilde{X}_{i,t} = (\mathsf{price}_t - \mathsf{price}_{t-1})X_{i,t-1}$, not in the covariance of the original variables. Therefore, conducting PCA within the OLS model is preferable rather than prior to it.

Denoting μ_i and σ_i, the empiric mean and standard deviation of the i^{th} principal component. Since the principals components are uncorrellated, Σ is diagonal and $\hat{\alpha}$ is computed with :

$$\hat{\alpha_i} = \frac{\mu_i}{\sigma_i^2} / \sqrt{\sum_{j=1}^{k} \frac{\mu_j^2}{\sigma_j^2}}$$

The signal regularized is: $\hat{\alpha}^{\mathrm{T}} \Pi X_t$ with Π the $(k \times n + 1)$ matrix of the projector onto the top k principal components of $\tilde{X}_t$. Note that Π is the projector onto the top k principal components of $\tilde{X}_t$, not of X_t.

Statistical Significance Regularization: For each alpha coefficient, we compute the p-value of α_i and retain only those coefficients $\hat{\alpha}_i$ that meet the criterion:

$$\mathbb{P}[\alpha_i = 0 | \hat{\alpha}_i] < p_{\text{threshold}}$$

The p-value is computed assuming that $\tilde{X}_t$ adheres to a Gaussian distribution. Under null hypothesis subsequent calculations[2] reveal that $\sqrt{\tau}\hat{\alpha}_i \times \hat{\alpha}^{\mathrm{T}}\mu$ follows a Student's t-distribution with $\tau - 1$ degrees of freedom, where τ denotes the length of the time series. This understanding enables the computation of the aforementioned p-value.

Practical use of the ML model

In practical terms, this model is applicable in the development of a trading strategy. This process entails the training of the model over a span of τ days, utilizing datasets $(X_{t-\tau}, ..., X_{t-1})$ and $(\mathsf{price}_{t-\tau}, ..., \mathsf{price}_{t-1})$. From this training dataset, $\hat{\alpha}$ is deduced through the prescribed methodology. Then, a signal for the subsequent day, t, is formulated as $\hat{\alpha}^{\mathrm{T}} X_t$.

This results in a machine learning model able to at generate linear trading signals from exogenous variables. The model is parametric, encompassing parameters such as the training period τ, and a regularization parameter for each regularization technique.

[2]The demonstration is straightforward in the case where Σ is diagonal because we have : $\sqrt{\tau}\hat{\alpha}_i \times \hat{\alpha}^{\mathrm{T}}\mu = \sqrt{\tau}\frac{\mu_i}{\sigma_i}$.

Remark on overfitting: With at most an order of magnitude of 5000 data points, the available dataset is quite limited, inherently posing a constant risk of overfitting. The key to the success of such a machine learning model lies in its ability to effectively implement regularization. This is the rationale behind proposing a variety of complementary regularization methods in the previous section. Beyond feature engineering, a significant portion of the empirical work presented in the next part has been dedicated to constructing and optimizing relevant regularization techniques to mitigate this issue. Still, this problem should be kept in mind as a important risk using this model for a real trading strategy.

3 Empirical Application to a Trading Strategy

The strategies empirically tested in this document is a day-to-day strategies. We consider an asset with an opening price denoted by price_t, reflecting the Open price of the asset in the market. A signal signal_t is generated using the proposed method. At the market opening we buy[3], signal_t shares of the asset, under the condition that the signal is sufficient, establishing a position of:

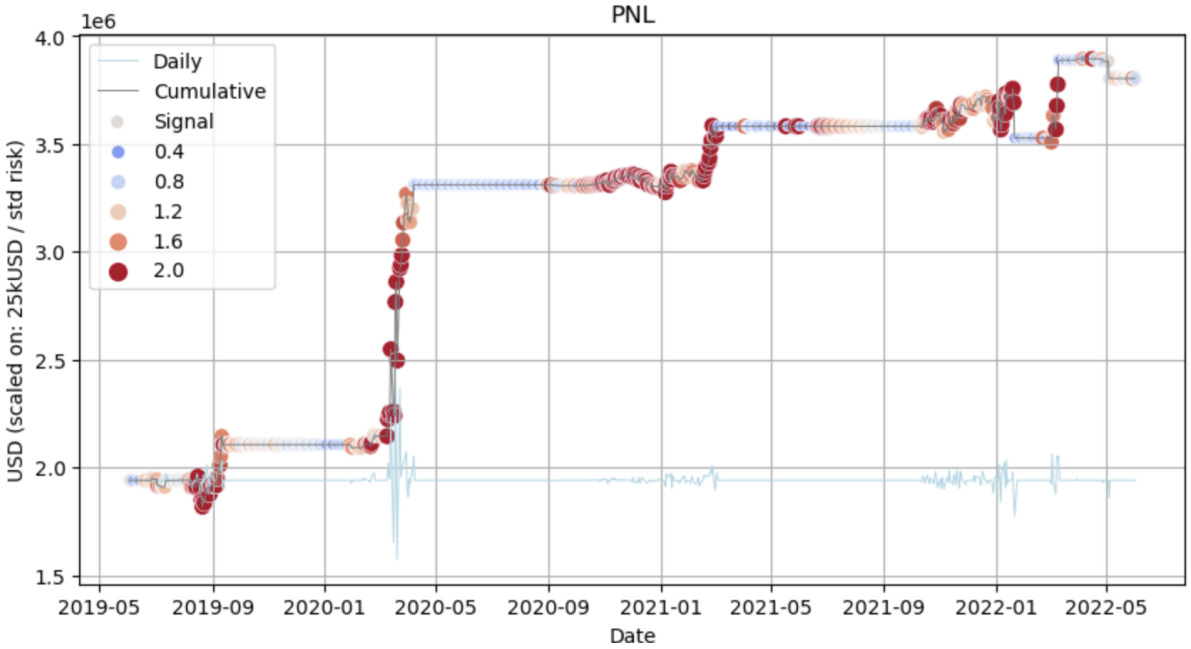
$$\mathsf{pos}_t = \mathsf{signal}_t \times \mathsf{price}_t \text{ if: } Z_{score}(\mathsf{signal}_t) > 1 \text{ else: } 0$$

Traded Products: 'IEF', a widely traded Exchange Traded Fund (ETF) that mirrors the performance of U.S. Treasury bonds of maturities: 1-3 years.

Exogenous variables: Prices of U.S. treasury bonds of differing maturities and various macros, *that have been feature-engineered.*

Data: Open source data from Yahoo finance.

The graph below corroborates the signal's efficacy: intervals of augmented signals (illustrated in red) are congruent with noticeable ascents in the PnL.

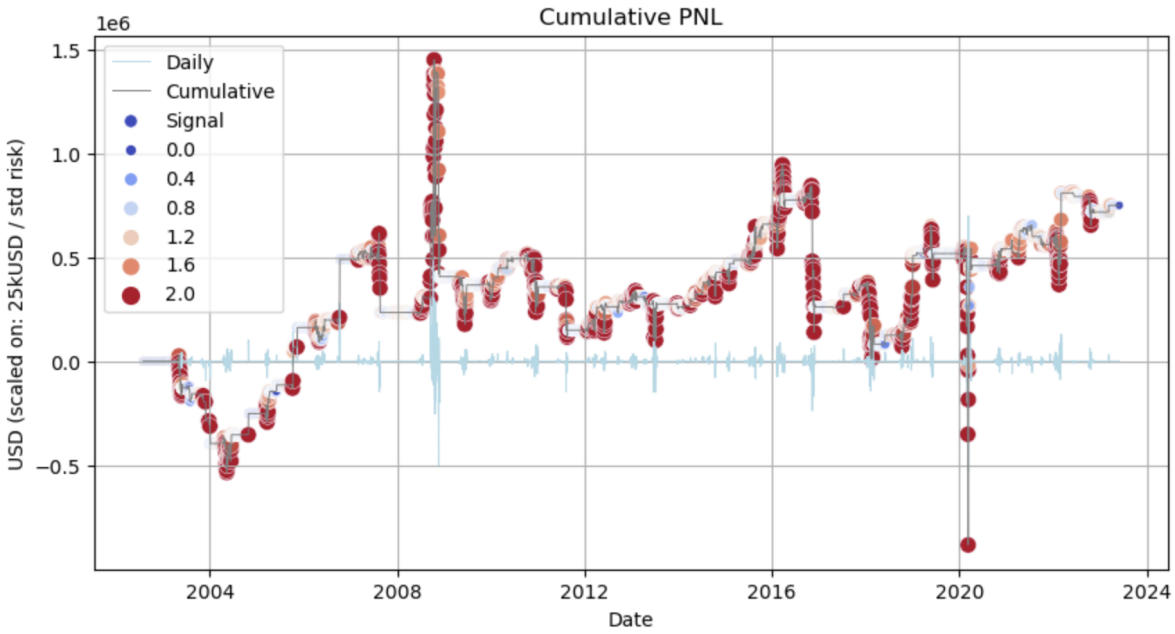


Quantitative Metrics: Sharpe Ratio: 1.25 / effective[4]: 2.1 ;
Turnover: 45.9%; Bips: 18.9

However, it is observed that over a longer period, we are subject to fluctuations: at times, the signal misinterprets the direction. The primary reason is that the signal can detect that the period is conducive to generating PnL, but it remains highly uncertain about which direction to take. Consequently, during each of these periods, the signal can choose the wrong direction, leading to frequent monetary losses in the strategy, which adversely affects the metrics. As a result, we obtain rather mediocre outcomes, as shown in the following figure.

[3]Note: The strategy operates under the assumption that: *(i)* the asset can be acquired at its opening price on day t, *(ii)* without any transaction fees, and that *(iii)* the action of the strategy will not have any effects on the market price.

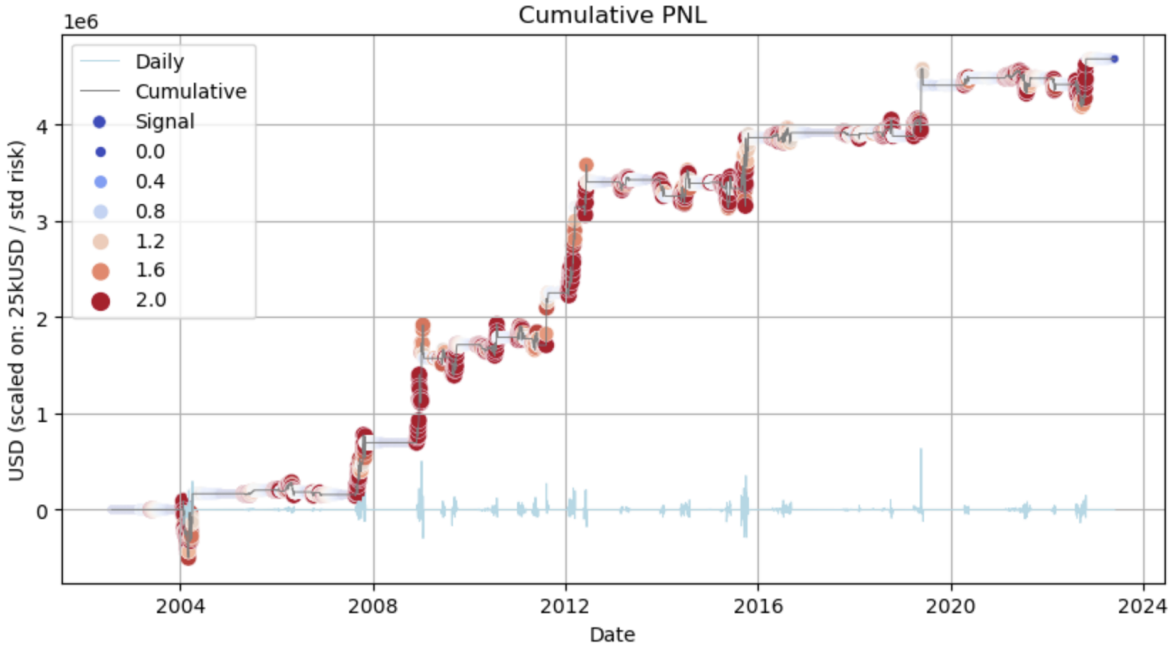
[4]Effective metrics are computed over the days where and effective trading has been made i.e. at days t where $\mathsf{pos}_t \neq 0$



Quantitative Metrics: Sharpe Ratio: 0.09 / effective: 0.46 ;

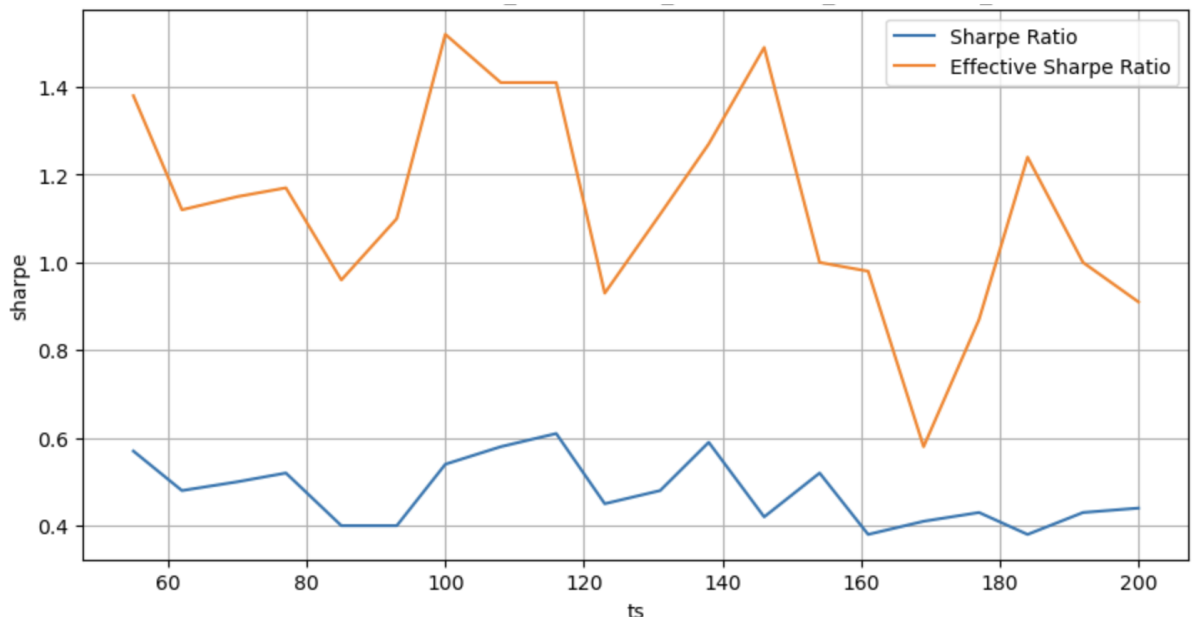
Nevertheless, as depicted in the graph below, two corrective techniques have been identified that drastically improve the results:

- the introduction of a corrective factor, represented as '$\mathsf{signal}_t = \mathrm{sg}(\mathsf{PnL}_t) * \hat{\alpha}^{\mathrm{T}} X_t$' (if the uncorrected model would have lost money on the previous day, the signal is reversed.) and,
- the intensive application of statistical significance regularization,



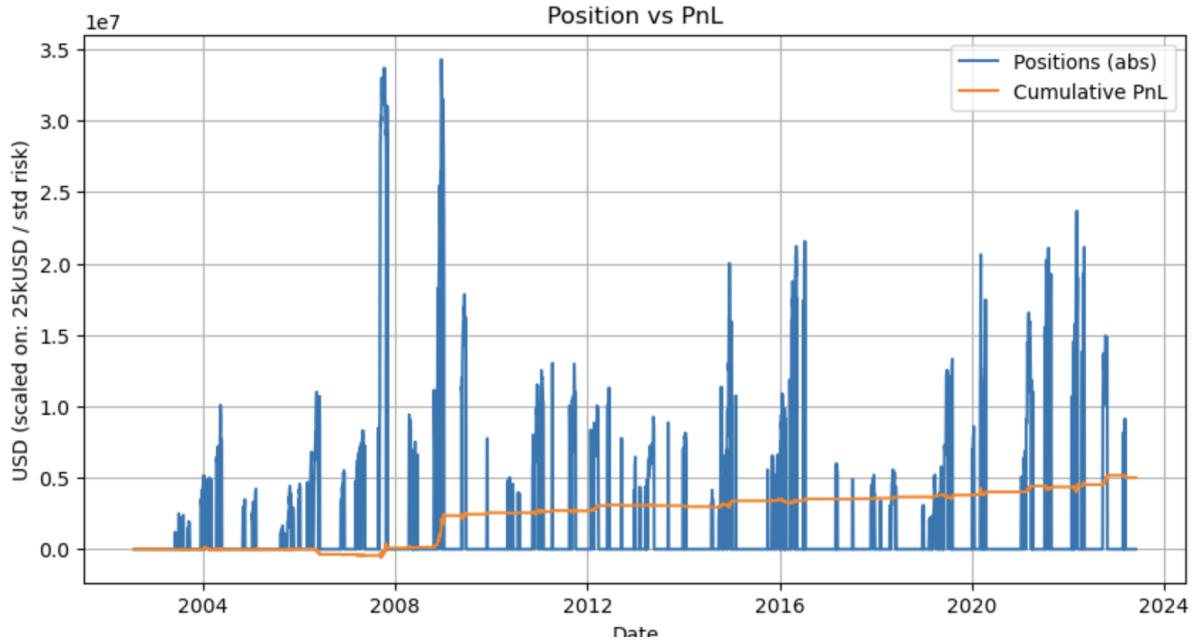
Quantitative Metrics: Sharpe Ratio: 0.55 / effective: 1.19 ;
Turnover: 37.3% / effective: 184.1% ; Bips: 11.9 / effective: 51.1

To verify the functionality of our strategy beyond a specific set of parameters and to identify the optimal range of parameters yielding the best possible results, we conducted a comprehensive analysis by calculating the Sharpe Ratio of the PnL for various training sizes. This exercise is crucial to ensure that the strategy was not just effective for a particular set of parameters (overfitting) but also adaptable and robust across different scenarios.



Sharpe and effective sharpe ratio for different training sizes

Caveat: this strategy is good in terms of Sharpe ratio. Yet, when we look at the effective turnover, we realize that these strategies perform poorly from this indicator's perspective (effective turnover is more than 150%.) Indeed, when plotting the PnL curve versus the curve of positions taken, it becomes apparent that the strategy requires taking large positions to generate significant PnL.



In this specific scenario, to generate a cumulative PnL of 5 million USD, there are brief periods where an investment of up to 35 million USD is required. This observation underscores the strategy's demand for significant capital deployment at sporadic intervals to achieve the desired returns.

Nevertheless, given the high degree of certainty in these strategies (a Sharpe ratio of 1 implies that money is lost in less than 7% of cases), it becomes feasible to obtain such investments through leverage. This approach allows for maximizing potential returns while managing the risk associated with significant capital deployment. Thus, the described strategy either demands substantial sporadic investments, via leverage effect, or it needs to be refined through other quantitative finance methods to reduce this turnover.

Conclusion

This study delineates a comprehensive framework for systematically constructing and optimizing trading strategies, anchored in a solid mathematical approach. The core innovation resides in the formulation of a signal linearly derived from selected exogenous variables. By integrating a linear combination of these variables into a strategic position-taking model, we have demonstrated a method that effectively constructs a signal responsive to market dynamics and predictive of future trends.

The methodology involved formulating the trading signal as a linear combination of a set of exogenous variables, which was subsequently employed to ascertain the position in the asset. The consequent PnL computation, based on the asset's price variations and the held positions over time, underscores the robustness of the proposed framework. This was further validated through empirical application, using publicly accessible market data, to demonstrate the efficacy of the model in constructing a durable trading strategy. Looking forward, the study opens avenues for further technological development, which include:

- **Implement a New Regularization Aimed at Limiting Position Size:** This approach could address the high-turnover issue observed in the empirical tests. By integrating a regularization mechanism that constraints the size of positions taken, the strategy can be fine-tuned to balance PnL generation against the risk of excessive capital allocation. It would result in a sharpe ratio / turnover trade-off. Such regularization would ensure that the strategy remains robust and efficient, without necessitating disproportionately large investments.

- **Generalized Time Steps:** In the 'How to Interpret These Results' section, we have shown that this methods constructs a time series s_t such that its 1-day lagged series s_{t-1} is strongly correlated with the asset price variations ($\mathsf{price}_t - \mathsf{price}_{t-1}$). This concept can be extended to t_s-days lagged time series s_{t-t_s}, by examining their correlation with ($\mathsf{price}_t - \mathsf{price}_{t-t_s}$). The optimal t_s intervals could be determined using methods that study seasonality, such as Fourier analysis.

- **Generalized Linear Signal:** Expanding the signal model to a more generalized form: $\mathsf{signal}_t = f_{\text{act.}}(\alpha^{\mathrm{T}} X_t)$, lead to a new expression for PnL. The samed process can be applied by optimizing the objective function $\mathcal{L}(\alpha)$ accordingly.

  Some activation function $f_{\text{act.}}$ such as a logit activation function inspired by Logistic Regression to create a signal that can take values in [0,1], and can be used as a mitigateur of other signals. This generalization can be explored with various activation functions used as hyperparameters.

- **Continuous Hyperparameter Selection:** In order to mitigate the risk of an overfitting signal on a hyperparameter set, we can implement a dynamic system for adjusting hyperparameters every 'training-size' days, based on the hyperparameters that maximized metrics in the previous period. Note that this bruteforce method will significantly increase the computational cost.

- **Enhance Corrective Terms with Command & Control Theory:** A primary limitation of this model is its relatively low reactivity, as it updates only every 'training-size' days. An approach grounded in command & control theory could significantly enhance the model's responsiveness, enabling real-time adjustments. The 'corrective factor' introduced in the 'Empirical Application to a Trading Strategy' section represents a rudimentary form of this adaptive correction. Investigating the application of command and control theory, especially the use of Kalman filters, to refine the model's corrective terms emerges as a promising avenue for development. Such an advancement would not only add dynamism to the model but also improve its accuracy and adaptability in response to evolving market conditions.

- **Stacking Linear Signal:** Inspired by the concept of stacking regressors used in machine learning, we could create a new, stacked signal by utilizing the Optimal Linear Signal with already efficient signals used as Exogenous Variables, potentially enhancing the signals' strength.

  In particular, by utilizing the 'Achieving Beta Neutrality' section, one can derive a beta-neutral signal from general signals.

- **Boosting Method:** The subsection 'Achieving Beta Neutrality' develops a method to generate a signal uncorrelated to a certain time series. If a signal is generated from n variables and a PnL is deduced from it, it is entirely possible to generate a new residual signal that is uncorrelated to this previous PnL. The sum of these two signals should have been able to capture more relevant

information from the variables, on the condition that some non-linearity has been introduced somewhere, and sufficient regularization avoid overfitting.

On the same principle as Boosting in ML, by iterating each time on the residual signal (orthogonal to the previous one), it is conceivable to construct very powerful signals.

In essence, this study not only lays a foundation for traders and financial analysts seeking to augment their strategies through quantitative methods but also sets the stage for continuous innovation and advancement in systematic trading strategies.

References

[1] Johnson, J. (2023). Machine Learning for Financial Market Forecasting. *Harvard University*.

[2] Rouf, N., Malik, M. B., Arif, T., Sharma, S., Singh, S., Aich, S., & Kim, H.-C. Stock Market Prediction Using Machine Learning Techniques: A Decade Survey on Methodologies, Recent Developments, and Future Directions.

[3] Soni, P., Tewari, Y., & Krishnan, D. (2022). Machine Learning Approaches in Stock Price Prediction: A Systematic Review. *Journal of Physics: Conference Series*.

[4] Kumar, D., Sarangi, P. K., & Verma. A systematic review of stock market prediction using machine learning and statistical techniques. *ScienceDirect*, Volume 49, Part 8, Pages 3187-3191.

[5] Kelly, B. T., & Xiu, D. (2023). Financial Machine Learning. *University of Chicago, Becker Friedman Institute for Economics Working Paper No. 2023-100*.

[6] Hoang, D., & Wiegratz, K. (2023). Machine Learning Methods in Finance: Recent Applications and Prospects. *Social Science Research Network*.

[7] Ghorbani, M., & Chong, E.K.P. (2020). Stock price prediction using principal components. *PLoS ONE*, 15(3): e0230124.

[8] Jurczenko, E., & ale. (2019). Machine Learning Optimization Algorithms & Portfolio Allocation. *arXiv:1909.10233*.

[9] Huang, W., Nakamori, Y., & Wang, S.Y. (2019). Optimization of investment strategies through machine learning. *Procedia Computer Science*, 162, 1035-1042.

[10] Cornuéjols, G., Peña, J., & Tütüncü, R. (2017). Optimization Methods in Finance. *Journal of Computational Finance*, 21(2), 1-12.

[11] Cornuéjols, G., Peña, J., & Tütüncü, R. (2018). Optimization Methods in Finance (2nd Edition). *Cambridge University Press*.

[12] Reppen, A. M., Soner, H. M., & Tissot-Daguette, V. (2022). Deep Stochastic Optimization in Finance.