Data Science in Finance and Economics

DSFE, 4(2): 270–284.
DOI: 10.3934/DSFE.2024011
Received: 14 December 2023
Revised: 12 March 2024
Accepted: 12 March 2024
Published: 21 May 2024

https://www.aimspress.com/journal/dsfe

Research article

The historical lepto-variance of the US stock returns

Vassilis Polimenis

University of Limassol, Cyprus

*** Correspondence:** Email: vasilis.polimenis@uol.ac.cy.

Abstract: Regression trees (RT) involve sorting samples based on a particular feature and identifying the splitting point that yields the highest drop in variance from a parent node to its children. The optimal factor for reducing mean squared error (MSE) is the target variable itself. Consequently, employing the target variable as the basis for splitting sets an upper limit on the reduction of MSE and, equivalently, a lower limit on the residual MSE. Building upon this observation, we define lepto-regression as the process of constructing an RT of a target feature on itself. Lepto-variance pertains to the portion of variance that cannot be mitigated by any regression tree, providing a measure of inherent variance at a specific tree depth. This concept is valuable as it offers insights into the intrinsic structure of the dataset by establishing an upper boundary on the "resolving power" of RTs for a sample. The maximal variance that can be accounted for by RTs with depths up to k is termed the sample k-bit macro-variance. At each depth, the overall variance within a dataset is thus broken into lepto- and macro-variance. We perform 1- and 2-bit lepto-variance analysis for the entire US stock universe for a large historical period since 1926. We find that the optimal 1-bit split is a 30–70 balance. The two children subsets are centered roughly at −1% and 0.5%. The 1-bit macro-variance is almost 42% of the total US stock variability. The other 58% is structure beyond the resolving power of a 1-bit RT. The 2-bit lepto-variance equals 26.3% of the total, with 42% and 47% of the 1-bit lepto-variance of the left and right subtree, respectively.

Keywords: total variance; regression tree; lepto-variance; macro-variance; lepto-ratio; lepto-regression

JEL Codes: C63, C14, G12, G17

1. Introduction

The regression tree (RT) is a machine learning model commonly used for explaining a continuous target variable based on various features. RTs are similar to decision trees, being constructed by recursively partitioning the input space into regions and assigning a constant value to each region. Each internal node in a regression tree represents a decision based on a particular feature. The tree structure is hierarchical, with the first decision at the root node and subsequent split decisions at higher depths creating further branching based on the outcomes of previous decisions. For a brief and concise introduction to RTs see Krzywinski and Altman (2017), Torgo (2011), and Elith et al. (2008).

A great advantage of RTs is their interpretability and simple visual representation, which results in a simple-to-understand decision-making process. Additionally, RTs can capture complex, nonlinear relationships in the data. Using the RT methodology, Polimenis (2022) uniquely defined the sample *lepto-variance* as variance beyond the explanatory power of a RT, and *macro-variance* as the upper bound of sample variability that may be explained.

Unlike the residual mean squared error (MSE) in a regression that depends on the utilized factors, the lepto-variance is a new type of idiosyncratic sample-specific variability that pertains to the portion of variance that cannot be mitigated by any regression tree, thus providing a measure of inherent variance at a specific tree depth. By establishing an upper boundary on the "resolving power" of RTs for a sample, this statistical concept is valuable as it offers insights into the intrinsic structure of the dataset. At each RT depth level, the overall variance within the dataset is broken into lepto- and macro-variance. This is related to the 1-d clustering problem (Grønlund et al., 2017). The k-means problem in higher dimensions is NP-hard (Aloise et al., 2009). Similar techniques have been used by cartographers to produce so-called choropleth or thematic maps, via the so-called natural breaks introduced by Jenks and Caspall (1971). Data in choropleth maps are categorized using a modification of the Jenks natural breaks classification method. These methods cluster data into groups that minimize the within-group variance and maximize the between-group variance.

Following Polimenis (2022), in this paper, the *lepto-regression*, *lepto-variance*, and *lepto-ratio* concepts are defined and then initially explored by providing simple intuitive examples. Then, 1- and 2-bit lepto-regression analysis for the entire US stock universe is performed utilizing historical daily market return data for the previous 96-year period. The sample comprises 25,272 daily returns from July 1, 1926, to June 30, 2022. We find that the optimal 1-bit RT split is roughly a 30–70 balance. The left and right children subsets are centered roughly around −1% and +0.5%. The 1-bit macro-variance is almost 42% of the total US stock variability, while the residual 58% is structured beyond the resolving power of any 1-bit RT. The 2-bit lepto-variance equals 26.3% of the total, with 42% and 47% the 1-bit lepto-variance of the left and right subtree, respectively.

1.1. Motivation

The relationship between the total explanatory power and the number of independent variables is complicated. The total explanatory power of a regression model is often assessed using metrics such as the coefficient of determination (R^2). R-squared represents the dependent variable variability proportion explained by the independent variables and is a measure of how well the independent

variables in a regression model explain dependent variable variation. Adding relevant variables can enhance explanatory power in a linear regression, but careful consideration is needed to avoid overfitting, i.e., a situation of a model fitting the training data closely by capturing noise. Similarly, in a financial regression of stock returns on market-wide factors, the residual (idiosyncratic) variance depends on the factors used in the regression. In general, we can get lower residual variance by adding extra financial factors.

Improving our understanding of the inherent explanatory power of RTs is valuable, as they are a fundamental building block for more advanced ensemble methods, such as random forests and gradient-boosted trees, which combine multiple trees to improve predictive performance and robustness. But, as with linear regression, overfitting the training data is one of the RT challenges, as trees can easily capture noise rather than underlying patterns. Pruning and other regularization techniques are used to address this issue.

1.2. Motivation from the field of financial risk management

Financial risk management is a large field of academic and practical significance for banking and finance. The key starting point of managing risk is to properly quantify it, which effectively means measuring volatility and correlations for the entire investable asset universe. Understanding the sources of volatility is of great interest. Investors place importance on understanding the factors that contribute to investment return volatility, as it directly affects both risk assessment and the overall decision-making process.

The introduction of a model-free method to analyze return variability has always been of great interest to the academic and financial practitioner communities. For example, the volatility index (VIX) introduced by Whaley (1993) is referred to by some as the market "barometer" and is tradable in CBOE. The index was later calculated via a more model-free method developed by Demeterfi et al. (1999). However, the VIX calculation is neither simple nor intuitive.

1.3. The role of idiosyncratic financial risk

When utilizing machine learning techniques in financial analysis, we use financial factors as features, being interested in finding the factors that explain a large fraction of the total stock return variance. Stock return variance that cannot be explained by broad market financial factors is considered idiosyncratic for the specific stock. The total risk of investment results from adding risks determined by exposure to market factors (market risks) and idiosyncratic volatility, which represents the risk component specific to the asset and not determined or related (i.e., orthogonal) to any wider market movements. The pricing implications of idiosyncratic volatility are still not well understood (Ang et al., 2006; Campbell et al., 2001).

In conclusion, as investors aim to diversify portfolios and mitigate risk exposure, quantifying variance is crucial. A novel, model-free, and simple statistical method for analyzing total return variability may enable investors to make better decisions, improve risk assessment, and create portfolios that align with their total risk tolerance and investment objectives.

2. Lepto-regression

In splitting a *nominal* predictor taking q possible labels (*unordered* values), there exist $2^{q-1}-1$ possible binary partitions of these labels. If all these partitions need to be evaluated, the computation becomes prohibitive for any (except for very small) q values. Various theorems related to the concavity of the underlying impurity function allow the problem to be simplified into a linear search of only $q-1$ partitions, where attribute values are sorted based on their strength of correlation with the target. Most notably, Fisher (1958) showed that for a continuous-valued target Y, the least squares partition of a set is contiguous. Breiman et al. (1984) extended this for a decision tree with binary (2-class) target Y (see the discussion in Ripley, 1996 and Hastie et al., 2009).

The process of constructing a regression tree involves recursively partitioning the sample based on the selected features and split thresholds, with a procedure continuing until a stopping criterion (a maximum depth or a minimum number of samples for the node) is met. At each decision node, the algorithm selects a feature and a threshold to split the data into two subsets with the goal of minimizing the residual target variance within each subset. A constant value is assigned to the instances that reach terminal (or leaf) nodes of the tree.

RTs provide a binary splitting of the sample space with minimization criterion the residual sum of squares (i.e., sum of squared error) $RSS=\sum_i\left(y_i-\hat{y}(x_i)\right)^2$, with $\hat{y}(x_i)=\hat{y}_i$ the prediction for y_i given factor values x_i. RTs predict using the average values $\hat{y}_i=\bar{y}_i$ within each subset, as they minimize the residual mean square error $MSE=\frac{1}{n}\sum_i(y_i-\bar{y}_i)^2$.

2.1. Definition of sample lepto-variance

At each internal tree node ($\#j$), the RT performs a sorting of the subsample reaching this node Sj using the chosen split factor xj and finds the split point cj that produces the maximum MSE drop from the node to its children Lj and Rj. Generally, sorting a sample based on a factor xj will not sort the target y. With no loss of generality, assume that the left child L contains the "small" y values (i.e., assume $\overline{y_L}=mean\{y\in L\}<\overline{y_R}=mean\{y\in R\}$).

Definition. A binary split of S into L and R is *sorted* if all target values y in L are smaller than all target values in R.

As an extension of the Fisher (1958) theorem on grouping, the following lemma for RTs is shown in Polimenis (2022)

Lemma 1. In terms of minimizing MSE in a RT, when splitting a sample S, it is always beneficial to utilize a sorted split. Thus, the best factor to use (in terms of MSE drop) is the dependent variable itself.

Proof. Let's assume an unsorted split of a sample S into L and R (with no loss of generality, assume $\overline{y_L}<\overline{y_R}$). Then, the maximum y value of the left subtree u_1 is larger than the minimum value of the right subtree u_0

$$u_1=\max\{y\in L\}>u_0=\min\{y\in R\}$$

But then, we can get a better split by swapping u_0 with u_1 into L and R, respectively. Because moving u_0 into L and u_1 into R will move the center of the L subsample to the left and the center of the right subsample to the right, thus producing a larger separation between the left and right subsamples without changing the relative sample sizes. By the law of total variance, a larger between-group variability means a smaller within-group variability and thus a better split.

Since the best binary split is always a sorted split, and regressing the target on itself allows all sorted splits to be evaluated, using the target as a factor will provide an upper bound on the explained variance (or lower bound on residual MSE).

Definition. We call *lepto-regression* the process of constructing an RT of a target feature on itself, and *sample lepto-variance* as the residual MSE of the lepto-regression.

We use $\mu 1^2$ and $\lambda 1^2$ to denote the 1-bit macro and lepto-variance, respectively (residual MSE for RTs with depth 1). Total variance equals

$$\sigma^2 = \mu 1^2 + \lambda 1^2 \tag{1}$$

3. The $\lambda 1^2$ lepto-variance of simple examples

To get a better understanding of the novel concept of lepto-variance, a few simple example calculations of $\lambda 1^2$ are presented and discussed.

The simplest case is that of the equiprobable 2-member set. Without loss of generality, assume the set $\{-0.5, 0.5\}$. The total variance of this sample is 0.25. In this (degenerate) case, the only split (and thus optimal) is the separation of the two members that produces a variance drop of 0.25 and a residual (lepto) variance remainder equal to zero $\sigma^2 = \mu 1^2$ and $\lambda 1^2 = 0$.

Next, consider the $\{-1, 0, 1\}$ equiprobable set. This will split into $\{-1\}$ and $\{0, 1\}$ with its total variance of $\sigma^2 = \frac{2}{3}$ split into macro-variance of $\mu 1^2 = \frac{1}{2}$ and lepto-variance of $\lambda 1^2 = \frac{1}{6}$

The next case is the equiprobable 4-member set $\{-1.5, -0.5, 0.5, 1.5\}$. With no loss of generality, points are chosen to center at zero to help with mental calculations. In this case, the optimal split is the balanced split producing two equiprobable 2-member sets left and right, with residual variance equal to $\lambda 1^2 = 0.25$. The two clusters $\{-1.5, -0.5\}$ and $\{0.5, 1.5\}$ are centered at -1 and 1, respectively, giving an inter-cluster distance of 2, and a variance drop that equals

$$\mu 1^2 = \text{Variance drop} = 0.5 \times 0.5 \times \textit{distance squared} = 0.25 \times 4 = 1$$

Definition. We define the useful concept of lepto ratio lR^2 as the ratio of lepto-variance (at a specific depth) to total sample variance. For the 1-bit case, this equals $lR1^2 = \lambda 1^2/\sigma^2$.

Using the law of total variance, in the last example, we calculate a total variance of $0.25 + 1 = 1.25$, and a ***lepto ratio*** $lR1^2 = 20\%$ of the total sample variance.

3.1. A first look at lepto-variance and split balance

From the discussion above for the 2- and 4-member sets, it may seem that the optimal split is always a balanced split. It is significant to note that the optimal split is not always balanced. We may understand some of the concepts related to the split balance (i.e., relative cluster weights) using as an

example the 6-member equiprobable sample {−1,0,1,2,3,4}. We may think of the entire sample as comprised of two separate clusters, {−1,0,1} and {2,3,4} centered at 0 and 3, with initial cluster "radius" epsilon equal to 1 and separated by inter-cluster distance delta equal to 3. Total sample variance is thus equal to $\sigma^2 = \frac{\delta^2}{4} + \frac{2}{3}\epsilon^2$. With epsilon = 1 and delta = 3, this equals 2.9167 (see Table 1 below).

Separating the sample in the middle gives a residual variance of the two clusters equal to $\lambda 1^2 = \frac{2}{3}$ and a variance drop of $\frac{\delta^2}{4} = \frac{9}{4} = 2.25$. In this case, the balanced split is optimal. For example, if instead the sample is split by isolating −1, there is a var drop equal to only $\frac{1}{6} \times \frac{5}{6} \times \left(2 - (-1)\right)^2 = 1.25$ and there is $1.67 > \lambda 1^2$ var left unexplained. Splitting at {−1,0} and {1,2,3,4} is better, as it gives a higher var drop equal to $\frac{1}{3} \times \frac{2}{3} \times \left(2.5 - (-0.5)\right)^2 = 2$ and there is 0.9167 var left unexplained. But this is still inferior to the balanced split.

Table 1a. Simple lepto-variance calculation example. Separating the 6-member {−1,0,1,2,3,4} sample (epsilon = 1) in the middle gives the expected variance of the two clusters equal to 0.67 and a variance drop of $\mu 1^2 = 2.25$. In this case, the balanced split is optimal, explaining 77.14% of the total variability of 2.9167. The lepto-variance (highlighted cell) of the sample equals 0.67, giving a 1-bit lepto-variance ratio $lR1^2 = 22.86\%$ of the total variance.

Split point	Sorted sample	Variance drop	Explained var as a fraction	Left child MSE	Right child MSE	Residual MSE	Residual MSE as a fraction of total
1	−1	1.25	42.86%	0.000	2.000	1.6667	57.14%
2	0	2.00	68.57%	0.250	1.250	0.9167	31.43%
3	1	2.25	77.14%	0.667	0.667	0.6667	22.86%
4	2	2.00	68.57%	1.250	0.250	0.9167	31.43%
5	3	1.25	42.86%	2.000	0.000	1.6667	57.14%
6	4		**total var** >>	2.917	0.000	2.9167	

We want to understand what happens if the two clusters get more spread out. Take, for example (by increasing epsilon to 1.2), the set {−1.2, 0, 1.2, 1.8, 3, 4.2}. Again, the two clusters are centered at 0 and 3 (separated by delta = 3) but they are now more spread out (with higher cluster variance). Total variance is 3.21. Separating the set in the middle gives the expected variance of the two clusters equal to $0.67 \times 1.44 = 0.96$ and a variance drop of $0.5 \times 0.5 \times 9 = 2.25$. The best split is still the balanced split, but now the benefit of the balanced split is less pronounced. For example, splitting the sample by isolating −1.2 gives a var drop equal to only $\frac{1}{6} \times \frac{5}{6} \times \left(2.04 - (-1.2)\right)^2 = 1.458$ and there is 1.752 var left unexplained. Splitting at {−1.2, 0} and {1.2, 1.8, 3, 4.2} is even better, as it gives a var drop equal to 2.2050 and there is 1.0050 var left unexplained. But this is still inferior to the balanced split. Observe that isolating the "outlier" −1.2 is now more beneficial than in the previous case, as it explains 45.42% of the total var.

Table 1b. Simple lepto-variance calculation example (cont.). Lepto-regression of a 6-member set with higher within-cluster variance (epsilon = 1.2) at various points. Optimal split is in the middle. It still explains $\mu 1^2 = 2.25$, but this time a smaller fraction of the total variance of $\sigma^2 = 3.21$. The lepto-variance (highlighted cell) of the sample equals $\lambda 1^2 = 0.96$, giving a lepto-variance ratio $lR1^2$ of almost 30%.

Split point	Sorted sample	Variance drop	Explained var as a fraction	Left child MSE	Right child MSE	Residual MSE	Residual MSE as a fraction of total
1	−1.2	1.458	45.42%	0.000	2.102	1.752	54.58%
2	0	2.205	68.69%	0.360	1.328	1.005	31.31%
3	1.2	2.250	70.09%	0.960	0.960	0.960	29.91%
4	1.8	2.205	68.69%	1.328	0.360	1.005	31.31%
5	3	1.458	45.42%	2.102	0.000	1.752	54.58%
6	4.2		**total var** >>	3.210	0.000	3.210	

As the two clusters get more spread out, as in {−1.3, 0, 1.3, 1.7, 3, 4.3}, the balanced split is not optimal any longer. Again, the two internal clusters are centered at 0 and 3, but they are now more spread out (within cluster variance 1.1267). Total sample variance σ^2 is 1.1267 + 2.25 = 3.3767. Separating the sample in the middle again explains $\frac{1}{2} \times \frac{1}{2} \times 9 = 2.25$, but this time represents only 66.63% of the total variance (versus 77.14% in the case of epsilon = 1). The skewed splitting at {−1.3, 0} and {1.3, 1.7, 3, 4.3} is optimal with a var drop equal to $\mu 1^2 = \frac{1}{3} \times \frac{2}{3} \times \left(2.575 - (-0.65)\right)^2 = 2.31125$.

Table 1c. Lepto-regression of a 6-member set with non-balanced optimal split (epsilon = 1.3). The skewed splitting at {−1.3, 0} and {1.3, 1.7, 3, 4.3} is more beneficial (as well as its symmetric split at {−1.3, 0, 1.3, 1.7} and {3, 4.3}). In this case, $\lambda 1^2 = 1.0654$ and $lR1^2 = 31.55\%$.

Split point	Sorted sample	Variance drop	Explained var as a fraction	Left child MSE	Right child MSE	Residual MSE	Residual MSE as a fraction of total
1	−1.3	1.568	46.44%	0.0000	2.1704	1.8087	53.56%
2	0	2.311	68.45%	0.4225	1.3869	1.0654	31.55%
3	1.3	2.250	66.63%	1.1267	1.1267	1.1267	33.37%
4	1.7	2.311	68.45%	1.3869	0.4225	1.0654	31.55%
5	3	1.568	46.44%	2.1704	0.0000	1.8087	53.56%
6	4.3		**total var** >>	3.3767	0.0000	3.3767	

4. Empirical analysis

4.1. Estimation of the historical lepto-variance of US stock returns

Here, the concept of lepto-variance of US stock returns is presented and, to provide some perspective, it is compared with the residual variance when two well-known financial factors, size (SMB) and book-to-value (HML), are used to capture return variability. Specifically, historical daily

percentage return data for US stock returns starting in 1926 are analyzed. High-quality return data from http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html were downloaded on July 30, 2022.[1]

Table 2a. Daily US market percentage return data for a 96-year period are used. The sample comprises 25,272 daily returns for the period July 1, 1926 to June 30, 2022.

	Date	Mkt-RF	SMB	HML	RF	Mkt
0	1926-07-01	0.10	−0.23	−0.28	0.009	0.109
1	1926-07-02	0.45	−0.34	−0.03	0.009	0.459
2	1926-07-06	0.17	0.29	−0.38	0.009	0.179
3	1926-07-07	0.09	−0.59	0.00	0.009	0.099
4	1926-07-08	0.21	−0.38	0.18	0.009	0.219
…	…	…	…	…	…	…
25267	2022-06-24	3.11	−0.36	−0.05	0.003	3.113
25268	2022-06-27	−0.28	0.54	1.24	0.003	−0.277
25269	2022-06-28	−2.10	−0.35	2.36	0.003	−2.097
25270	2022-06-29	−0.20	−0.44	−1.30	0.003	−0.197
25271	2022-06-30	−0.95	0.43	−0.15	0.003	−0.947
25272 rows × 6 columns						

Table 2b presents descriptive return stats for daily returns of the entire US stock market and the two Fama-French factors SMB (size) and HML (value) for a 96-year period (in percentage). The variance for the entire sample is approximately 1.167. The average daily US stock return for this period is 4.2 bp (basis points), of which 3 bp is the risk part and 1.2 bp is the risk-free component.

Table 2b. Using daily market return data for a 96-year period, we see descriptive return stats for the daily US stock market returns and the two Fama-French factors SMB (size) and HML (value) for the 96-year period from July 1, 1926 to June 30, 2022 (in percentage). Values are truncated at two decimal places for better visibility.

	Mkt-Rf	SMB	HML	RF	Mkt
count	25272	25272	25272	25272	25272
mean	0.030	0.0045	0.015	0.012	0.042
std	1.08	0.59	0.62	0.012	1.08
min	−17.44	−11.67	−6.02	−0.00	−17.41
25%	−0.40	−0.25	−0.25	0.00	−0.39
50%	0.06	0.01	0.01	0.01	0.08
75%	0.50	0.27	0.26	0.02	0.51
max	15.76	8.18	9.04	0.06	15.76

In financial asset pricing, a well-known pricing model is the three-factor model of Fama and French (1993). The three-factor model is based on a time-series linear regression of excess portfolio returns of the type

$$R(t) - rf(t) = a + b \cdot [Mkt(t) - rf(t)] + s \cdot SMB(t) + h \cdot HML(t) + e(t) \qquad (2)$$

[1] Copyright 2022 Kenneth R. French.

with $R(t)$ being the return on a security or portfolio for period t, $rf(t)$ the risk-free return, $Mkt(t) - rf(t)$ the excess return on the value-weighted market portfolio above the risk-free asset, $SMB(t)$ the return on a diversified portfolio of small stocks minus the return on a diversified portfolio of big stocks for the period, and $HML(t)$ the difference between the returns on diversified portfolios of high and low book-to-market (B/M) ratio stocks. The three-factor linear model above assumes that the sensitivities b, s, and h in (2) capture the most variation in expected returns, and the true value of the alpha intercept in (2) should be near zero for well-priced securities.

Table 2c. Covariance matrix of the daily US returns and the 2 FF factors (truncated at three decimal places for better visibility).

	Mkt - rf	SMB	HML	rf	Mkt
Mkt - rf	1.167	−0.100	0.110	0	1.167
SMB		0.350	−0.027	0	−0.100
HML			0.387	0	0.110
rf				0	0
Mkt					1.167

Table 2d. Correlation matrix of the daily US returns and the 2 FF factors (truncated at three decimal places for better visibility).

	Mkt-rf	SMB	HML	rf	Mkt
Mkt-rf	1	−0.157	0.163	−0.015	1
SMB		1	−0.073	−0.011	−0.157
HML			1	0.009	0.163
rf				1	−0.004
Mkt					1

In the Figure below, the optimal depth of one RT when the US stock return vector (Mkt) is lepto-regressed is shown. The optimal split is a 30–70 balance, for a Mkt return less than or equal to −0.264. The two children subsets are centered roughly at −1% and 0.5%. Total sample variance is 1.167.

Sample 1-bit Lepto-variance equals

$$\lambda 1^2 = 0.3 \times 0.877 + 0.7 \times 0.593 = 0.678 \tag{3}$$

The *1-bit macro-variance* (max variance drop) thus equals

$$\mu 1^2 = 1.167 - 0.678 = 0.489 \tag{4}$$

This equals almost 42% of the total US stock variability. This implies a 1-bit lepto-ratio $lR1^2 = 58\%$ *comprising structure that cannot be removed by any 1-bit RT.* Observe that the macro-variance could also be computed directly as the variance of a δ-scaled Bernoulli distribution with $p = 0.30$ and $\delta = 1.525 = 0.499 - (-1.026)$ via $\mu 1^2 = 0.3 \times 0.7 \times 1.525^2$.

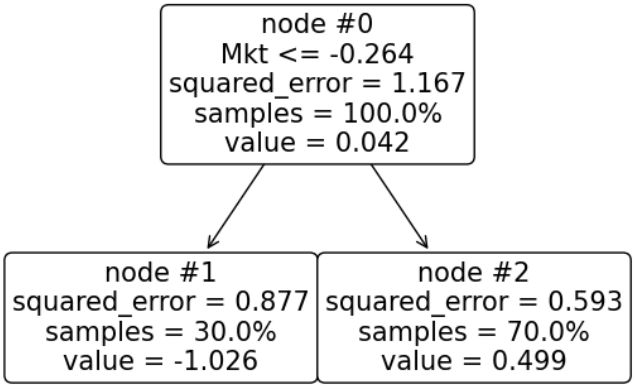

Figure 1a. The optimal 1-bit RT when the historical US stock return vector (Mkt) is lepto-regressed.

To put the historical 1-bit lepto-ratio of 58% in some perspective, the optimal 1-bit RT when US stock returns are regressed on the two Fama-French SMB and HML factors is also estimated and shown in the Figure below. When using the entire historical sample, HML is more efficient than SMB and thus chosen for the 1-bit RT. The tree is highly skewed and can explain very little of the total historical US stock variability. Residual squared error equals 1.1315 (roughly 97% of total MSE).

An interesting new statistic for any feature then is the percentage $mR1^2$ of the sample macro-variance that it can capture with a 1-bit RT. Using 1-bit RT, the Fama-French factors can only explain 0.0355 of the total MSE. This is only $mR1^2 = \frac{0.0355}{0.489} = 7.26\%$ of the 1-bit macro-variance, i.e., the maximum MSE that may be explained by 1-bit RTs. Using SMB explains $mR1^2 = \frac{0.025}{0.489} = 5.11\%$ of the 1-bit macro-variance $\mu 1^2$ (see Table 3).

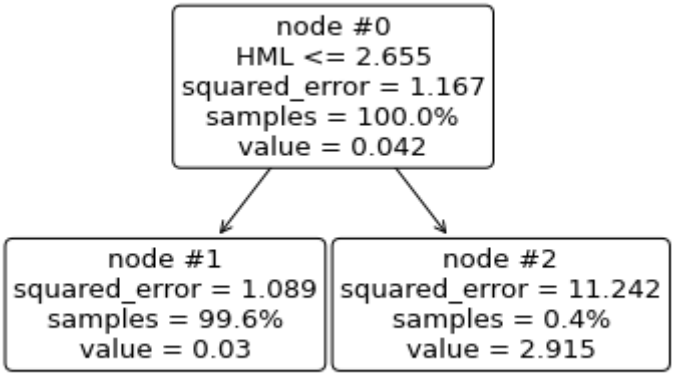

Figure 1b. 1-bit RT for US stock returns regressed on the 2 Fama-French SMB and HML factors (HML is chosen).

In Table 3 below, the summarized depth 1-bit lepto-regression analysis for 96 years of US stock returns and the two Fama-French factors is shown. Using 1-bit RTs, Fama-French factors explain only

a small $mR1^2$ fraction of the total explainable MSE (sample macro-variance). Overall, HML slightly dominates SMB.

Table 3. 1-bit lepto-regression for 96 years data of US stock returns and the two Fama-French factors.

Factor used	Total MSE	Explained MSE		Residual MSE	$mR1^2$
$\boldsymbol{\mu 1^2 + \lambda 1^2 =}$	1.167				
SMB	=	0.025	+	1.142	5.11%
HML	=	0.0355	+	1.1315	7.26%
Mkt	$\boldsymbol{\mu 1^2 =}$	0.489	$\boldsymbol{\lambda 1^2 =}$	0.678	100%

4.2. The 2-bit historical lepto-structure of US returns

The concept of lepto-variance of a sample may also be defined for trees of a maximum depth larger than 1. As we move deeper down on an RT, there will always be less residual variance. The argument of Lemma 1 will *locally* still be valid; at any node, the best split is always achieved by the target itself (via a sorted split). But the greediness of the RT may in rare (degenerate) occasions result in a situation where sorting in a split is sub-optimal (Polimenis, 2022).

For the 4-element set {−1,0,1,2}, the greedy 1-bit split correctly splits to {−1,0} and {1,2} for a final 1-bit residual MSE $\lambda 1^2 = 0.25$, thus explaining 1 out of the 1.25 total variance (i.e., $lR1^2 = 20\%$). But when the greedy 1-bit split is applied on the 4-element set {−1,0,1,4}, it myopically isolates the outlier 4 at the first split, thus explaining $\mu 1^2 = 3$ out of the total $\sigma^2 = 3.5$ (i.e., $lR1^2 = 14.3\%$). This is preferable to the balanced 1-bit split into {−1,0} and {1,4} that would only explain 2.25 out of the total 3.5 (i.e., $mR1^2 = \frac{2.25}{3} = 75\%$). However, the balanced split would allow a better outcome down the tree, as it could capture the entire variation at the 2-bit split. On the contrary, the myopic isolation of the outlier 4 at the first split limits the 2-bit split, thus resulting in a final 2-bit residual MSE equal to $\frac{1}{8} > 0$.

In Polimenis (2022), it is conjectured that the lepto-regression-based split will still achieve the lowest residual squared error at any *average depth*. For example, the greedy 2-bit max depth tree in the example has a lower average depth of 1.75 bits and should not be compared with the balanced split resulting in an average depth of 2 bits. Based on this, the lepto-variance λj^2 of a sample at j-bits is defined as the residual variance when the target is lepto-regressed on itself j times and provides the minimum residual MSE for an *average* depth j. In the {−1,0,1,4} case, the 1-bit lepto-variance equals $\lambda 1^2 = 0.5$, while $\frac{1}{8}$ is the lepto-variance for 1.75 bits. For practical situations, with large sample sizes (> 1K samples) and relatively low-depth trees (less than 3–4 splits), such a situation is highly unlikely to occur, and the distinction between the average and maximum depth of a tree will not matter.

Similarly, μj^2 will denote the j-bit macro variance (RTs with depth j), thus decomposing total variance into $\sigma^2 = \mu j^2 + \lambda j^2$.

4.3. 2-bit lepto-regression analysis for historical US stock returns

Here, the 2-bit lepto-structure analysis for historical US stock return data is performed following the 1-bit analysis of the previous section. In the Figure below, the descriptive statistics and optimal split for the left subtree of the two optimal-depth RT when US returns are lepto-regressed is depicted.

The optimal split point is for returns larger than −1.884, which comprise 88.5% of the total samples reaching the left node. The leftmost child comprises the smallest 3.45% of market returns (11.5% of the initial 30%) with an average −3% return. This is a highly volatile subsection of very negative market returns, with a residual MSE = 1.968. The centermost part of the left child comprises 26.5% (88.5% of the initial 30%) of the total sample (−1.884% < Mkt ≤ −0.264%) and, with a residual MSE = 0.167, it is substantially less volatile. Out of the total MSE of 0.877 that reaches the left subtree, 0.373 is lepto-structure beyond the resolving power of the 2-bit RT. Thus, 42% of the total variability of the left subtree is lepto.

	Mkt
count	7574.000000
mean	-1.026157
std	0.936746
min	-17.413000
25%	-1.240000
50%	-0.747000
75%	-0.459000
max	-0.264000

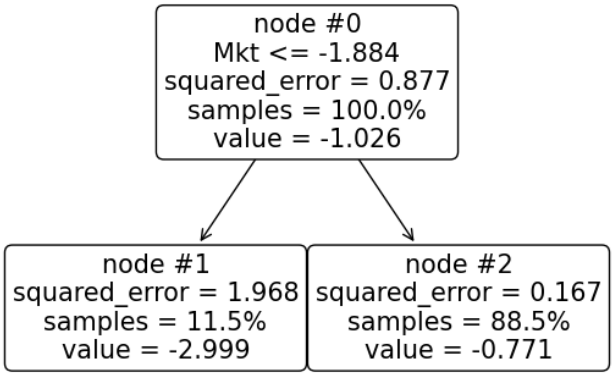

Figure 2a. Descriptive statistics and optimal split for the left subtree of a 2-bit RT when the US return vector is lepto-regressed. *The lepto-variance of the left subtree equals 42% of its total variability.*

In the Figure below, descriptive statistics and optimal split for the right subtree in the optimal 2-bit RT when the US return vector is lepto-regressed is depicted. The optimal split point is for large returns (larger than 1.145%), which comprise 12.2% of the total samples reaching the intermediate right node, or the highest 8.6% of the entire daily return sample (12.2% of the initial 70%) with an average 2% return. This is a highly volatile subsection of very strong market returns, with a residual MSE = 1.393. The centermost part of the right child is the largest subsection, as it comprises 61.5% (87.8% of the initial 70%) of the total sample (−0.264% < Mkt ≤ 1.145) and, with a residual MSE = 0.124, it is substantially less volatile.

	Mkt
count	17698.000000
mean	0.498992
std	0.770257
min	-0.263000
25%	0.038000
50%	0.309000
75%	0.715000
max	15.762000

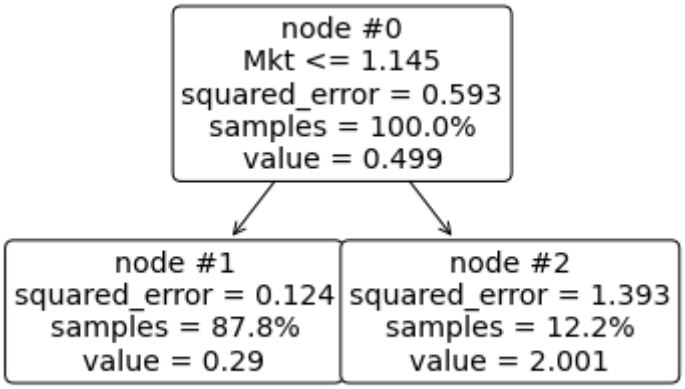

Figure 2b. Descriptive statistics and optimal split for the right subtree in the optimal 2-bit RT when the US return vector is lepto-regressed. For the right subsample (Mkt > − 0.264%), 47% of the total MSE cannot be explained via a RT.

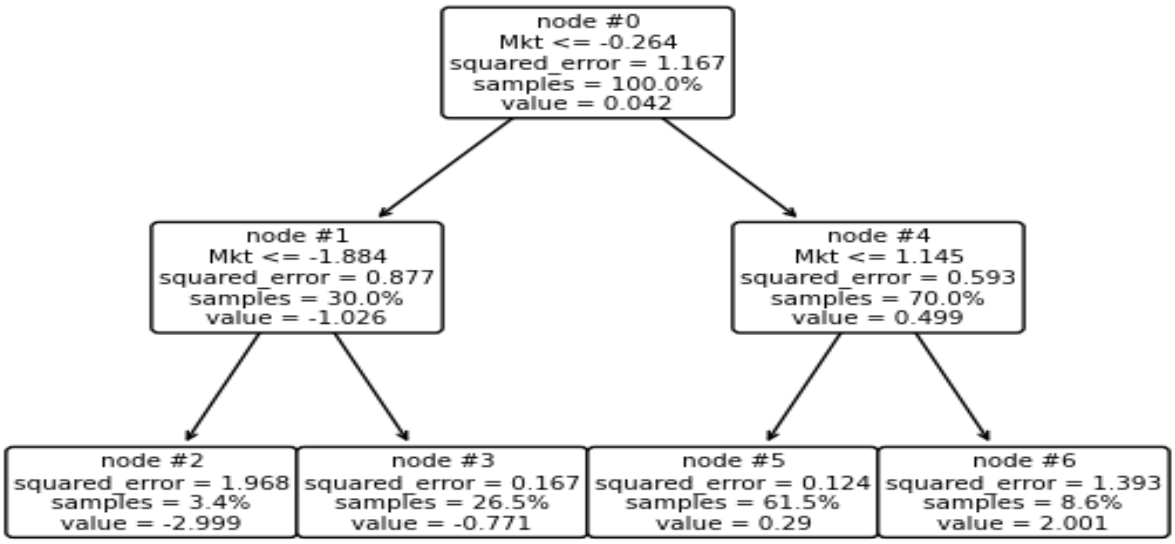

Figure 2c. 2-bit lepto-RT for historical US return vector. From the total historical sample variance of 1.167bp, a 2-bit tree will never be able to explain $\lambda 2^2 = 0.034 \times 1.968 + 0.265 \times 0.167 + 0.615 \times 0.124 + 0.086 \times 1.393 = 0.307$, which implies that the 2-bit lepto-variance explains $\mathrm{lR}2^2 = 26.3\%$ of the total.

5. Conclusions

The lepto-regression of a sample is a novel technique defined as the process of constructing an RT by regressing the target on itself. Due to its simplicity, lepto-regression is an interesting model-free technique and has the potential to reveal important properties of sample structure. It has been

shown in Polimenis (2022) that, since in a regression tree it is always beneficial to generate a sorted split of a sample S, the lepto-regression provides an upper bound in terms of the variability of a target that can be explained. The variance that cannot be explained via the lepto-regression is called sample lepto-variance. The k-bit lepto-variance (λk^2) of a sample is defined as the residual structure after the sample has been lepto-regressed (up to k times) and is the variance that cannot be explained by any set of features. The k-bit macro-variance is the variance captured by the lepto-regression and thus represents the maximum variance that can be captured by any combination of features. The lepto-variance analysis of the entire 96-year period of US stock market daily returns reveals that the 1-bit macro-variance (variance drop) equals 42% of the total US stock variability, while 58% is structure that cannot be explained by any 1-bit RT. The 2-bit lepto-variance equals 26.3% of the total, with 42% and 47% of the 1-bit lepto-variance of the left and right subtree, respectively.

References

Aloise D, Deshpande A, Hansen P, et al. (2009) NP-hardness of Euclidean sum-of-squares clustering. *Mach Learn* 75: 245–248. https://doi.org/10.1007/s10994-009-5103-0

Ang A, Hodrick RJ, Xing Y, et al. (2006) The cross-section of volatility and expected returns. *J Financ* 61: 259–299. https://doi.org/10.1111/j.1540-6261.2006.00836.x

Breiman L, Friedman J, Olshen R, et al. (1984) *Classification and Regression Trees*. Wadsworth, Belmont, CA

Campbell JY, Lettau M, Malkiel BG, et al. (2001) Have Individual Stocks Become More Volatile? An Empirical Exploration of Idiosyncratic Risk. *J Financ* 56: 1–43. https://doi.org/10.1111/0022-1082.00318

Demeterfi K, Derman E, Kamal M, et al. (1999) A Guide to Volatility and Variance Swaps. *J Deriv* 6: 9–32. https://doi.org/10.3905/jod.1999.319129

Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. *J Anim Ecol* 77: 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Fama EF, French KR (1993) Common Risk Factors in the Returns on Stocks and Bonds. *J Financ Econ* 33: 3–56. https://doi.org/10.1016/0304-405X(93)90023-5

Fisher WD (1958) On Grouping for Maximum Homogeneity. *J Am Stat Assoc* 53: 789–798. https://doi.org/10.1080/01621459.1958.10501479

Grønlund A, Larsen KG, Mathiasen A, et al. (2017) Fast exact k-means, k-medians and Bregman divergence clustering in 1D. arXiv:1701.07204.

Hastie T, Tibshirani R, Friedman J (2009) *Elements of Statistical Learning*, Springer. https://doi.org/10.1007/978-0-387-84858-7

Jenks GF, Caspall FC (1971) Error on Choroplethic Maps: Definition, Measurement, Reduction. *Ann Assoc Am Geogr* 61: 217–244. https://doi.org/10.1111/j.1467-8306.1971.tb00779.x

Krzywinski M, Altman N (2017) Classification and regression trees. *Nat Methods* 14: 757–758. https://doi.org/10.1038/nmeth.4370

Polimenis V (2022) The Lepto-Variance of Stock Returns. *Proceedings of the 34th Panhellenic Statistics Conference*, 167–182, Athens. http://dx.doi.org/10.2139/ssrn.4148317

Ripley B (1996) *Pattern Recognition and Neural Networks*. Cambridge University Press. https://doi.org/10.1017/CBO9780511812651

Torgo L (2011) Regression Trees. In: Sammut, C., Webb, G.I. (eds) *Encyclopedia of Machine Learning*. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_711

Whaley Robert E (1993) Derivatives on Market Volatility. *J Deriv* 1: 71–84. https://doi.org/10.3905/jod.1993.407868

AIMS Press