






Modelling Financial Markets

Chapter 13: Machine Learning in Finance



Daniel Traian PELE

Bucharest University of Economic Studies
IDA Institute Digital Assets
Blockchain Research Center
AI4EFin Artificial Intelligence for Energy Finance
Romanian Academy, Institute for Economic Forecasting
MSCA Digital Finance

What Can Machine Learning Predict in Finance?



Campbell R. Harvey.
Photo: Xuthoria (2026); CC BY-SA 4.0; Wikimedia Commons

"Hundreds of papers and factors attempt to explain the cross-section of expected returns. Given this extensive data mining, it does not make sense to use the usual criteria for establishing significance."

— *Harvey, Liu & Zhu (2016)*

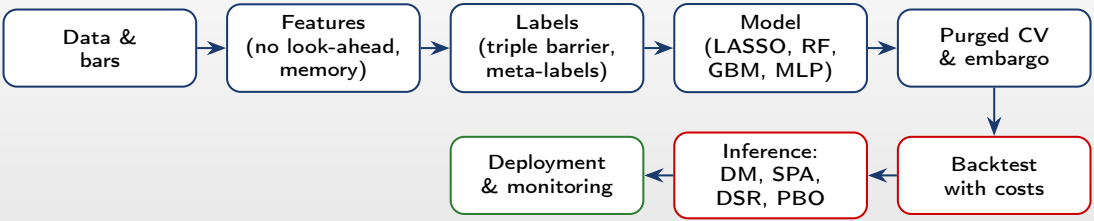
- After hundreds of factors tried on the same data, a new factor needs $t > 3$, not $t > 2$
- 5-day direction of the S&P 500 (Standard & Poor's), $n = 6{,}269$: no model beats the "always up" benchmark
  - the benchmark has 58.1% accuracy; the logit has 57.7% and predicts "up" on 98.3% of days
- Walk-forward backtest 2010–2026: the Sharpe ratio of the GBM (Gradient Boosting Machine) strategy is 0.33
  - the Sharpe ratio of buy-and-hold is 0.66
- On a target with nothing to predict, shuffled K-fold reports 69.4% accuracy
  - purged K-fold with embargo reports 49.6%

▸ Central chart: Model Comparison: Fold Means

What connects the steps of the analysis

- Part I: the prediction problem and the basic methods
  - features and fractional differencing, labels, validation without information leakage
- Part II: an S&P 500 case study with formal tests, selection and backtest overfitting
  - machine learning (ML) for inference, a research case study
  - artificial intelligence (AI) for scientific discovery (proofs and estimators: Appendix)
- Abbreviations in the diagram
  - LASSO (L_1-penalised regression), RF (random forest), GBM (gradient boosting machine)
  - MLP (multilayer perceptron), CV (cross-validation)
- What you should be able to do after this chapter
  - compute features and labels without look-ahead
  - compute the score of a model validated with purging and embargo
  - derive why truncated FFD (Fixed-width window Fractional Differentiation) of an $I(1)$ price stays $I(1)$
  - interpret, against a benchmark, the Pesaran–Timmermann (PT) and Diebold–Mariano (DM) tests, the deflated Sharpe ratio (DSR), the superior predictive ability (SPA) test and the probability of backtest overfitting (PBO)
  - Further Topics and Appendix: reading for familiarity
- Detailed examples and advanced topics are in Further Topics, after the Conclusions
  - the footer button leads there; the Back button returns to the starting slide
- Most failures come from the process, not the algorithm (López de Prado (2018)); two failures dominate
  - **look-ahead bias**: using information not available at decision time
  - **selection bias**: reporting the best of many trials

What We Will Explain: from Data to a Tested Claim



- Features use only the information available at the decision date
- Validation measures what still holds on observations the model has not seen
- A predicted direction justifies a trade only after uncertainty and costs are assessed

Why Machine Learning in Finance, and Why It Is Hard

The promise

- Flexible, non-linear functional forms
  - interactions found automatically
- Hundreds of candidate predictors (the "factor zoo", Chapter 3)
- Alternative data: text, images, order books, on-chain data

The constraints

- Very low **signal-to-noise ratio**
  - one serially dependent history that cannot be re-run
- **Non-stationarity**: regimes, breaks
  - markets react to their users
- Labels must be constructed
  - errors cost real money
- Massive risk of **overfitting** (fitting noise)

Adaptive Markets Hypothesis: Lo (2004)

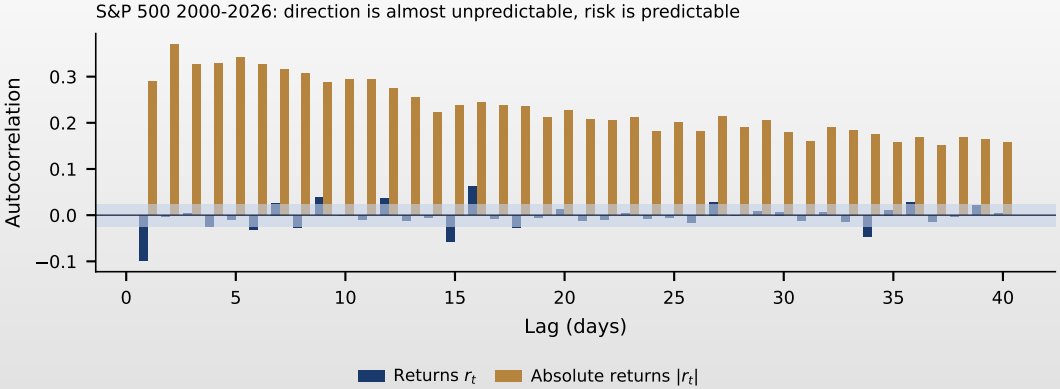
- Predictability appears, is arbitraged away and re-appears
  - a model that worked yesterday is a **hypothesis**, not a law (market efficiency: Chapter 2)
- The hard part is not fitting a model but **not fooling yourself**
  - correct labels, correct validation, honest backtests

What Is Predictable? Direction vs. Risk (1/2)

- The question: which part of a financial series is predictable at all?
- Autocorrelation at lag k: $\rho_k = \mathrm{Corr}(x_t, x_{t-k})$
  - for the daily log return $r_t = \log P_t - \log P_{t-1}$ and for $|r_t|$
- Next chart: lag k on the horizontal axis, $\hat{\rho}_k$ on the vertical axis
  - blue bars for r_t, brown bars for $|r_t|$
- The light-blue band, $\pm 1.96/\sqrt{T}$, is the 95% band under no autocorrelation
  - T: number of observations
  - for $|r_t|$, the band requires a finite fourth moment, which Chapter 1 questions
- Autocorrelations of returns r_t are smaller than those of $|r_t|$, yet 13 of the 40 lie outside the band
  - at lag 1, $\hat{\rho}_1 = -0.099$
  - the band is neither simultaneous over 40 lags nor robust to heteroskedasticity
  - leaving the band does not show profitability after costs
- Autocorrelations of $|r_t|$ are large and persistent
- ML has the best chance where there is structure
  - **volatility, risk, liquidity, cross-sectional ranking**

MFM_ch13_model_comparison

What Is Predictable? Direction vs. Risk (2/2)



MFM_ch13_model_comparison ▶ Interactive

Evidence, and the Economic Value of a Small R^2_{OOS}

- Gu, Kelly & Xiu (2020): monthly returns of individual US (United States) stocks, 1957–2016
  - they compare linear models, penalised regressions, principal component regression (PCR), partial least squares (PLS)
  - random forests, gradient boosting and neural nets
- Their fit measure: the out-of-sample (OOS) R^2; economic value of a small R^2: Further Topics
  - how much of the squared return a forecast removes, relative to forecasting zero

$$R^2_{OOS} = 1 - \frac{\sum_t (r_t - \hat{r}_t)^2}{\sum_t r_t^2}$$

  - r_t: realised return in month t of the test period; $\hat{r}_t$: forecast made with information up to $t-1$
  - the denominator is the error of the zero forecast $\hat{r}_t = 0$: $R^2_{OOS} > 0$ if the model beats it, $R^2_{OOS} < 0$ if it does worse
- Results (Table 1, monthly R^2_{OOS})
  - three-layer network 0.40%, gradient boosting 0.34%, random forest 0.33%
  - OLS (Ordinary Least Squares) with 3 predictors 0.16%, OLS with all 900+ predictors −3.46%
- Trees and nets win through **non-linear interactions**
  - momentum, liquidity and volatility dominate
- The benchmark matters
  - Welch & Goyal (2008) show that most published predictors lose to the **historical mean**
  - Gu, Kelly & Xiu (2020) and Chapter 14 use a zero forecast

Loss, Penalty and Five Model Families (1/5)

- The question of this section
  - why, with a weak signal, does variance control matter more than flexibility?
- Supervised learning picks, within a model family, the function that fits the training data best
  - after a penalty for complexity

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L\big(y_i, f(\mathrm{x}_i)\big) + \lambda\, \Omega(f)$$

  - (y_i, x_i), $i = 1, \ldots, n$: target and feature vector of observation i; $\mathcal{F}$: the model family (e.g. all trees of depth 3)
  - L: loss, the squared error $(y - f)^2$ or, for a probability f of $y = 1$, the log-loss $-[y \log f + (1 - y) \log(1 - f)]$
  - $\Omega(f)$: complexity penalty with weight $\lambda \geq 0$; $\lambda = 0$: no penalty, large λ: a simpler model
- λ is chosen by K-fold cross-validation (CV) on folds $\mathcal{P}_1, \ldots, \mathcal{P}_K$:

$$CV(\lambda) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i \in \mathcal{P}_k} L\big(y_i, \hat{f}_\lambda^{(-k)}(\mathrm{x}_i)\big), \qquad \hat{\lambda} = \arg\min_{\lambda} CV(\lambda)$$

  - n_k: number of observations in fold k; $\hat{f}_\lambda^{(-k)}$: the model fitted on the other $K - 1$ folds

Loss, Penalty and Five Model Families (2/5)

- Why we penalise: the expected test error at a point x_0 (Hastie, Tibshirani & Friedman (2009), Ch. 7)
  - for $y = f(x) + \varepsilon$ with $\mathbb{E}[\varepsilon] = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$:

$$\mathbb{E}\big[(y_0 - \hat{f}(x_0))^2\big] = \underbrace{\sigma^2}_{\text{noise}} + \underbrace{\big(\mathbb{E}[\hat{f}(x_0)] - f(x_0)\big)^2}_{\text{bias}^2} + \underbrace{\mathrm{Var}\big(\hat{f}(x_0)\big)}_{\text{variance}}$$

  - $\mathbb{E}$: expectation over possible training samples and over the noise of the new observation $y_0 = f(x_0) + \varepsilon_0$
  - bias: the systematic error of a too-simple family
  - variance: the sensitivity of $\hat{f}$ to the training sample; it grows with model flexibility
  - σ^2: variance of the unpredictable part of y, which no model can reduce
- Derivation: add and subtract $\mathbb{E}[\hat{f}(x_0)]$ in $y_0 - \hat{f}(x_0)$
  - the cross terms have mean zero because ε_0 is independent of the training sample
- In finance, return forecasts reduce the error very little
  - an out-of-sample R^2 below 1%: the model removes less than 1% of the benchmark's squared loss
  - this does not identify σ^2: the remaining error also holds misspecification, omitted variables and estimation error
  - so penalties and tree averaging cut variance at the price of a small bias

Loss, Penalty and Five Model Families (3/5)

- Penalised regressions, with p standardised features and coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \mathrm{x}_i^\top \boldsymbol{\beta}\right)^2 + \lambda \left[\alpha \|\boldsymbol{\beta}\|_1 + \tfrac{1-\alpha}{2} \|\boldsymbol{\beta}\|_2^2\right]$$

  - $\|\boldsymbol{\beta}\|_1 = \sum_j |\beta_j|$, $\|\boldsymbol{\beta}\|_2^2 = \sum_j \beta_j^2$; $\alpha \in [0, 1]$: share of the absolute-value penalty
- Ridge, $\alpha = 0$ (Hoerl & Kennard (1970)): $\hat{\boldsymbol{\beta}} = \left(\mathrm{X}^\top \mathrm{X} + \frac{n\lambda}{2} \mathrm{I}\right)^{-1} \mathrm{X}^\top \mathrm{y}$
  - shrinks all coefficients, none exactly to zero
  - stabilises the estimate when predictors are correlated
- LASSO (least absolute shrinkage and selection operator), $\alpha = 1$ (Tibshirani (1996))
  - the penalty $|\beta_j|$ has a kink at zero, so some coefficients become exactly zero
  - LASSO selects features
- Elastic net, $0 < \alpha < 1$ (Zou & Hastie (2005)): keeps groups of correlated predictors together
  - LASSO, by contrast, arbitrarily keeps one of them
- The chapter's linear benchmark is the penalised logit
  - $p = P(y = 1 \mid \mathrm{x}) = 1/\left(1 + e^{-\mathrm{x}^\top \boldsymbol{\beta}}\right)$, with the log-loss in place of the squared loss

Loss, Penalty and Five Model Families (4/5)

- Trees: at each node v choose the feature j and threshold s that reduce the Gini impurity the most
  - Gini impurity: $G = 1 - \sum_c \hat{\pi}_c^2$

$$(j^*, s^*) = \arg\min_{j,\, s} \left[\frac{n_L}{n_v} G(L) + \frac{n_R}{n_v} G(R) \right], \qquad L = \{x_j \le s\},\ R = \{x_j > s\}$$

  - $\hat{\pi}_c$: share of class c in the node; n_v, n_L, n_R: observations in the node and in its two children
  - $G = 0$ for a pure node; with two classes $G = 2\hat{\pi}(1 - \hat{\pi})$, at most 0.5 at 50/50; a leaf predicts the share of class 1
- Random forest (RF, Breiman (2001)): $\hat{f}_{RF}(\mathrm{x}) = \frac{1}{B} \sum_{b=1}^{B} T_b(\mathrm{x})$
  - each tree T_b grown on a bootstrap sample with a random subset of features at each split

$$\mathrm{Var}\big(\hat{f}_{RF}(\mathrm{x})\big) = \rho\, \sigma_T^2 + \frac{1 - \rho}{B}\, \sigma_T^2$$

  - σ_T^2: variance of one tree; ρ: correlation between two trees (Hastie, Tibshirani & Friedman (2009), Ch. 15)
  - random feature subsets lower ρ, the only term that more trees cannot remove
  - the out-of-bag error is **optimistic** when labels overlap

Random Forests

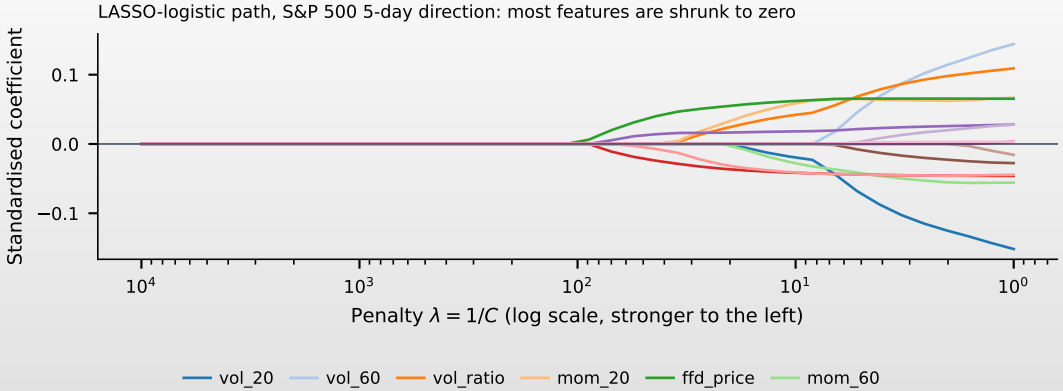
Loss, Penalty and Five Model Families (5/5)

- Gradient boosting machine (GBM, Friedman (2001))
  - add, step by step, a small tree that corrects the errors of the current model

$$r_{im} = -\frac{\partial L(y_i, F)}{\partial F}\bigg|_{F=F_{m-\mathbf{1}}(\mathbf{x}_i)}, \qquad F_m(\mathbf{x}) = F_{m-\mathbf{1}}(\mathbf{x}) + \eta\, h_m(\mathbf{x})$$

  - F_m: the model after m steps; h_m: the tree fitted to the pseudo-residuals r_{im}; for the log-loss on the logit score, $r_{im} = y_i - \hat{p}_{m-\mathbf{1}}(\mathbf{x}_i)$
  - $\eta \in (0, 1]$: learning rate; a small η learns slowly and overfits less
- Multilayer perceptron (MLP): hidden layers $\mathbf{h}^{(\ell)} = g\big(\mathbf{W}^{(\ell)}\mathbf{h}^{(\ell-\mathbf{1})} + \mathbf{b}^{(\ell)}\big)$, $\mathbf{h}^{(\mathbf{0})} = \mathbf{x}$
  - the output: $\hat{p} = 1/\big(1 + e^{-(\mathbf{w}^\top \mathbf{h}^{(D)} + b)}\big)$
  - $\mathbf{W}^{(\ell)}$, $\mathbf{b}^{(\ell)}$: weights and biases of layer ℓ; D: number of hidden layers; $g(z) = \max(0, z)$: the ReLU (rectified linear unit), element by element
  - the weights minimise the log-loss by gradient descent
  - early stopping: a chronological validation set (the last 10% of the training data, purged); training stops when the validation log-loss stops falling
  - richer architectures are in Chapter 14

LASSO Regularisation Path (1/2)



MFM_ch13_regularization ▶ Interactive

LASSO Regularisation Path (2/2)

- L_1-penalised logistic regression for the 5-day direction of the S&P 500, 14 standardised features
- Chart labels
  - vol_20, vol_60, vol_ratio: the 20- and 60-day volatilities and their ratio
  - mom_20, mom_60: the log returns over the last 20 and 60 days
  - ffd_price: the FFD log price
- Horizontal axis: the penalty $\lambda = 1/C$, on a log scale, stronger to the left
  - vertical axis: the standardised coefficient
- Strong penalty (left of the chart): every coefficient is zero
  - the first feature to enter is the FFD (fixed-width window fractional differentiation) log price
  - FFD log price: a scaled price level (next section)
- Weak penalty: the volatility features get the largest coefficients
- All coefficients are small in absolute value
  - weak signal, consistent with near-efficient markets

MFM_ch13_regularization

Stationarity, Memory and the Fractional Difference (1/2)

- The question of this section
  - how do we turn a price into a stationary feature without erasing its memory?
- Log prices $X_t = \log P_t$ behave as $I(1)$
  - P_t: closing price on day t
- Returns $\Delta \log P_t = \log P_t - \log P_{t-1}$ are $I(0)$ but **erase the memory** of the price level
- Integration order: a series X_t is $I(\delta)$ if $(1-B)^\delta X_t$ is $I(0)$
  - B: backshift operator, $BX_t = X_{t-1}$
  - $I(0)$: stationary with short memory; $I(1)$: a random walk type series, stationary after one difference
  - δ may be fractional: the series is stationary only for $\delta < \frac{1}{2}$, with long memory when $0 < \delta < \frac{1}{2}$ (Hosking (1981))
- Idea of fractional differencing: difference by less than 1
  - removing the unit root while keeping part of the memory

Stationarity, Memory and the Fractional Difference (2/2)

Definition 1 (Fractional difference of order d)

- $\tilde{X}_t = (1-B)^d X_t = \sum_{k=0}^{\infty} w_k X_{t-k}$, with $w_0 = 1$ and $w_k = -w_{k-1}\frac{d-k+1}{k}$

- Notation
  - d: differencing order chosen by the analyst; k: lag in days
  - w_k: weight of the value k days back, from the binomial expansion of $(1-B)^d$
- $d = 1$: $w = (1, -1, 0, \dots)$, ordinary returns
- $0 < d < 1$: weights decay hyperbolically, so memory is kept
- An $I(1)$ log price differenced with d is $I(1-d)$
  - non-stationary for $d \leq \frac{1}{2}$
  - stationary with long memory for $\frac{1}{2} < d < 1$
- The infinite weights sum to $(1-1)^d = 0$
  - the level cancels only in the limit (asymptotics in the Appendix)
- López de Prado (2018) proposes the minimum d at which an ADF (Augmented Dickey–Fuller) unit-root test rejects
  - test equation and critical value on the ADF slide; this section shows why that rule fails

The Fixed-Width Window (FFD)

Definition 2 (Fixed-width-window fractional differentiation (FFD))

- Truncate the weights: keep $w_0, \ldots, w_{l^*}$ with $|w_k| \geq \tau$ (e.g. $\tau = 10^{-4}$), drop the rest (López de Prado (2018))
- $\tilde{X}_t = \sum_{k=0}^{l^*} w_k X_{t-k}$ for every $t > l^*$

- τ: smallest weight kept; l^*: the last lag kept, i.e. the window length
  - the first l^* days have too little history and are dropped
- Every $\tilde{X}_t$ uses the same number of lags
  - only past values enter $\Rightarrow$ no look-ahead
- The smaller d, the more lags are needed
  - $d = 0.25$, $\tau = 10^{-4} \Rightarrow l^* = 444$ lags
- The truncated weights no longer sum to zero
  - $\sum_{k=0}^{l^*} w_k = 0.178$ at $d = 0.25$ (0.040 at $d = 0.5$)

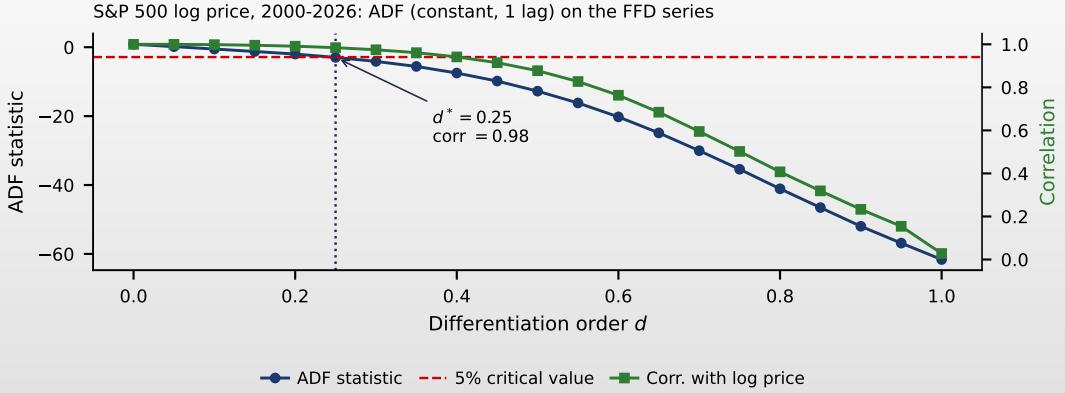
The ADF Statistic Along d: S&P 500 (1/3)

- Augmented Dickey–Fuller (ADF) unit-root test: the t-statistic of γ in

$$\Delta \tilde{X}_t = c + \gamma \, \tilde{X}_{t-1} + \sum_{i=1}^{p} \phi_i \, \Delta \tilde{X}_{t-i} + e_t$$

  - H_0: $\gamma = 0$, a unit root; p: number of lagged differences, which absorb the autocorrelation of e_t
  - with a constant, reject H_0 at 5% below ≈ -2.86
- The rule of López de Prado (2018)
  - compute the FFD series on a grid of d between 0 and 1
  - pick the smallest d at which the ADF rejects
- Next chart: d on the horizontal axis
  - the ADF statistic (blue) on the left axis
  - the correlation of the FFD series with the log price (green) on the right axis
  - red dashed line: the 5% critical value; vertical line: $d^* = 0.25$

The ADF Statistic Along d: S&P 500 (2/3)



MFM_ch13_ffd ▶ Interactive

The ADF Statistic Along d: S&P 500 (3/3)

- FFD log price, S&P 500 2000–2026, constant and 1 lag: ADF $= -2.96$
  - so the rule rejects from $d = 0.25$
- At $d = 0.25$, $\mathrm{Corr}(\tilde{X}_t, \log P_t) = 0.98$
  - the FFD series is almost the log price
- Same series, with 34 lags chosen by the Akaike information criterion (AIC)
  - ADF $= -0.72$, no rejection
- The verdict depends on the number of lags
- A correlation of 0.98 with an $I(1)$ price is a warning, not a virtue
- Type of evidence: comparison of procedures (it shows which method works better on the same data, not why)
  - the same log price, two lag choices, two verdicts

MFM_ch13_ffd

Truncated FFD of an $I(1)$ Price Is Still $I(1)$

- Decompose the FFD series:

$$\tilde{X}_t = \sum_{k=0}^{l^*} w_k X_{t-k} = \Big(\sum_{k=0}^{l^*} w_k\Big) X_t - \sum_{k=1}^{l^*} w_k \left(X_t - X_{t-k}\right)$$

  - first term: today's level times the sum of the truncated weights; second term: a weighted sum of the k-day changes $X_t - X_{t-k}$
- If X_t is $I(1)$, each $X_t - X_{t-k}$ is $I(0)$
  - the first term is $0.178\,X_t$ at $d = 0.25$: a **scaled random walk**
- The unit-root component has small variance next to the stationary part
  - so ADF with few lags confuses the two
  - Monte Carlo: pure random walk, $T = 6{,}274$ days (the length of the S&P 500 FFD series), FFD with $d = 0.25$: ADF (1 lag) rejects in **97.8%** of 500 samples, at a nominal 5%
- A finite filter whose weights do not sum to zero keeps a level component, whatever the ADF p-value
- Consequence: the FFD log-price in the case study is a scaled price level, not a stationary feature
  - the memory is measured directly in Further Topics

MFM_ch13_inference

From Fixed-Horizon to Triple-Barrier Labels (1/2)

- The question of this section
  - which label reproduces a trader's real decision, with take-profit, stop-loss and a maximum holding period?
- **Fixed horizon**: the label is the sign of the h-day log return
  - with a dead zone around zero

$$r_{t,t+h} = \log P_{t+h} - \log P_t, \qquad y_t = \begin{cases} +1 & \text{if } r_{t,t+h} > \kappa \\ -1 & \text{if } r_{t,t+h} < -\kappa \\ 0 & \text{otherwise} \end{cases}$$

  - P_t: closing price on day t; h: horizon in days; $\kappa \geq 0$: threshold, moves smaller than κ get the label 0
- Three problems
  - a constant κ ignores time-varying volatility
  - the path is ignored: a stop-loss could have been hit before $t + h$
  - the classes are unbalanced: S&P 500, $h = 10$, $\kappa = 0$: 60.5% up vs. 39.5% down

From Fixed-Horizon to Triple-Barrier Labels (2/2)

Definition 3 (Triple-barrier label, López de Prado (2018))

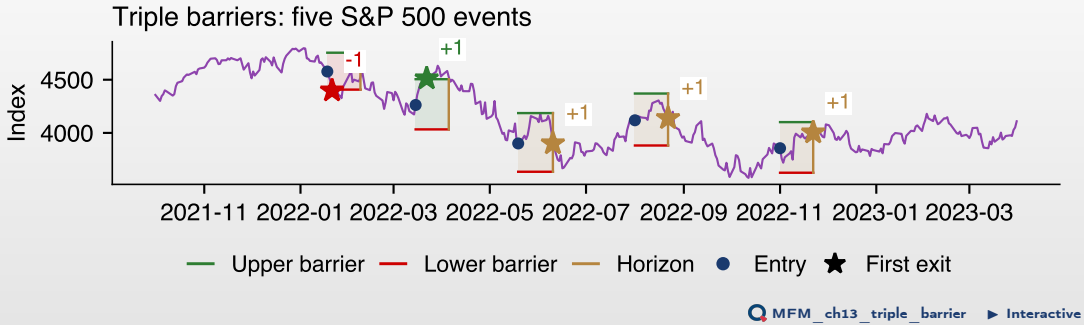
- At event time t_0, with daily volatility σ_{t_0} and horizon h
  - upper barrier $\log P_t - \log P_{t_0} \geq \pi_u \sigma_{t_0} \sqrt{h}$ (profit-taking)
  - lower barrier $\leq -\pi_l \sigma_{t_0} \sqrt{h}$ (stop-loss), vertical barrier $t = t_0 + h$
- Label = **first barrier touched** at t_1
  - $+1$ upper, -1 lower; at the vertical barrier $\text{sign}(r_{t_0,t_1})$ (or 0)

- $\sigma_{t_0} \sqrt{h}$: the typical h-day move; $\pi_u, \pi_l > 0$: barrier widths in units of it
- Each label has a **span** $[t_0, t_1]$: the input of purging in validation
- σ_{t_0}: exponentially weighted moving average (EWMA) volatility, span 50
  - square root of the EWMA of squared deviations of daily log returns

$$\sigma_t^2 \approx \alpha \sum_{k \geq 0} (1-\alpha)^k \left(r_{t-k} - \bar{r}_t\right)^2, \qquad \alpha = \frac{2}{\text{span} + 1} = \frac{2}{51} \approx 0.039$$

  - $\bar{r}_t$: the exponentially weighted mean return; span 50: an effective memory of about two trading months
- Why volatility-scaled barriers
  - a fixed threshold would be hit almost daily in 2008 or March 2020 and rarely in calm years

Triple Barrier: S&P 500 in 2022 (1/2)

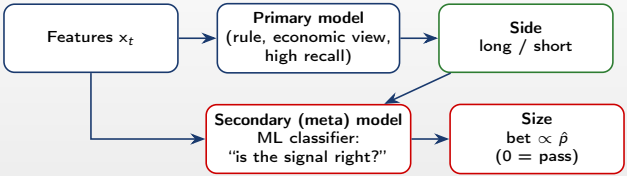


Triple Barrier: S&P 500 in 2022 (2/2)

- Five events during the 2022 bear market, $\pi_u = \pi_l = 1$, $h = 15$ days
  - the colour of each star is the barrier touched first
- Chart: the index level
  - green lines: upper barrier, red: lower barrier, brown: horizon
  - dots: entry, stars: first barrier touched
- January: stop-loss hit (-1); March: profit-taking ($+1$)
  - the other three end at the vertical barrier (amber stars) above their entry level
  - so the sign rule labels them $+1$
- A label $+1$ therefore does not mean that the profit barrier was touched

MFM_ch13_triple_barrier

Meta-Labelling: Deciding Whether to Act (1/2)



- Meta-label $y_t^{meta} = 1$ if the primary signal was profitable, 0 otherwise
  - profitable: barrier touched in its direction
- The secondary model estimates $\hat{p} = P(y_t^{meta} = 1 \mid x_t)$
  - the probability that the signal of day t is right given the features x_t
- ML learns **when to trust** an interpretable primary rule, not the direction itself

Meta-Labelling: Deciding Whether to Act (2/2)

- A filter decides which signals to act on; three scores describe it

$$\text{precision} = \frac{TP}{TP + FP}, \qquad \text{recall} = \frac{TP}{TP + FN}, \qquad F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

  - TP: profitable signals taken; FP: unprofitable signals taken; FN (False Negative): profitable signals skipped
  - acting on every signal gives recall 1; F_1: their harmonic mean, between 0 and 1
- Example (Seminar 13, exercise B11): moving-average crossover on Bitcoin
  - the 20- and 50-day moving-average crossover (MA20/MA50), 1,660 test trades
  - with the meta filter, precision rises from 0.475 to 0.517, recall falls from 1 to 0.541, and F_1 falls from 0.644 to 0.529
  - the mean gross log return per trade, before costs, moves from -0.27% to $+0.19\%$
- Meta-labelling raises precision at the cost of recall
  - with costs per trade, precision and the return per trade matter, not F_1

Leakage: What Crosses the Train/Test Boundary

- The question of this section
  - why does ordinary K-fold overstate a model's skill on financial series?

Definition 4 (Leakage)

- The training set contains information that is also present in the test set
  - so the test score no longer measures performance on genuinely new data

- Serially correlated features ($x_t \approx x_{t+1}$) and overlapping labels ($y_t \approx y_{t+1}$)
  - a shuffled K-fold puts **near-identical** days on both sides
- A flexible model (RF, GBM) memorises the neighbour $\Rightarrow$ excellent "out-of-sample" score
  - tuning on a leaky CV selects the **most overfitted** model
- Other sources of leakage
  - transformations fitted on the full sample (scaling, principal component analysis, the choice of d for FFD)
  - survivorship bias, data revisions, wrong timestamps

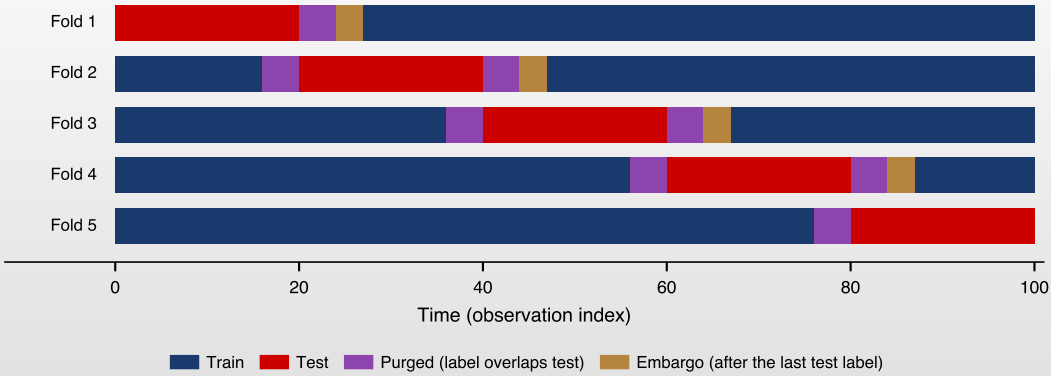
Purged K-Fold with Embargo (1/2)

- Folds are **contiguous** blocks in time (López de Prado (2018))
  - test fold k holds the observations $t_k^a, \ldots, t_k^b$
  - their labels end at the latest at $T_k = \max_{i \in \text{test}_k} t_{1,i}$
- The training set keeps only the observations that do not touch the test:

$$\mathcal{A}_k = \left\{ i \notin \text{test}_k : \; [t_{0,i}, t_{1,i}] \cap [t_k^a, T_k] = \emptyset \;\; \text{and} \;\; t_{0,i} \notin (T_k, \; T_k + e] \right\}$$

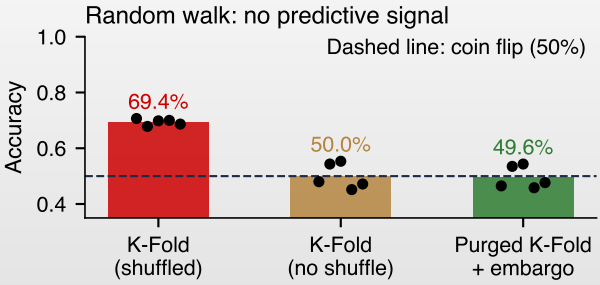
  - the first condition is **purging**: it removes labels that overlap the span of the test labels
  - the second is the **embargo**: e observations after the end of the last test label, T_k (e.g. 1%), needed because features look back in time
- The purged CV score is the mean of K scores
  - each from a model fitted on $\mathcal{A}_k$ and evaluated on fold k
- With labels of length h, a boundary removes h observations before and $h + e$ after the test block
  - an interior fold loses $2h + e$ observations
  - a 10-observation example: Appendix
- Next chart: five folds in rows, observation index on the horizontal axis
  - blue train, red test, purple purged, brown embargo

Purged K-Fold with Embargo (2/2)



MFM_ch13_purged_cv ▶ Interactive

Leakage Experiment with an Unpredictable Target



- Price = pure random walk
  - label = sign of the 20-day forward return
  - 5 features = 20-day rolling means of independent noise
- True accuracy is 50% by construction
- Random forest, 5 folds
- **Shuffled K-fold: 69.4%**
  - contiguous K-fold 50.0%
  - purged K-fold + embargo 49.6%
- Shuffling fails because labels overlap *and* features persist
  - here contiguous folds already remove most of the leak: only the fold boundaries share returns, and purging cleans them
  - a result that looks too good is usually leakage
- Type of evidence: comparison of procedures

MFM_ch13_purged_cv ▶ Interactive

Model Selection versus Deployment

- Ordinary K-fold is valid for a purely autoregressive model with serially uncorrelated errors (Bergmeir, Hyndman & Koo (2018))
  - when it fails: Appendix
- **Model selection**: purged K-fold, or combinatorial purged CV (CPCV)
  - CPCV tests on many combinations of groups and returns a **distribution** of Sharpe ratios instead of one number
- **Deployment**: walk-forward (WF)
  - train on $[0, T_k]$, test on $(T_k, T_{k+1}]$, roll forward (T_k: end of the k-th training window)
  - realistic, but a single path with high variance: it serves the final evaluation, not the choice (WF and CPCV: Appendix)

Case Study: Can the 5-Day Direction of the S&P 500 Be Predicted?

- S&P 500 index and VIX, daily closes of the course data, 2001–2026: $n = 6{,}269$ observations
  - the sample starts in October 2001: the first 444 days of 2000–2001 feed the FFD window
  - VIX: the CBOE (Chicago Board Options Exchange) volatility index, the 30-day implied volatility of S&P 500 options
- Why the S&P 500: the most liquid equity index, with a long VIX history
  - if its direction were predictable, arbitrage would be cheap
- Target: $y_t = 1\{\log P_{t+5} - \log P_t > 0\}$
  - equal to 1 if the index is higher 5 days later
  - the fixed-horizon label with $h = 5$, $\kappa = 0$
- Why 5 days: a weekly trading horizon
  - about $6{,}269/5 \approx 1{,}250$ non-overlapping labels
- The label span $[t, t + 5]$ is used for purging
- Validation: purged 5-fold CV, 1% embargo, the same folds for every model
  - each fold covers about 5 years, hence at least one crisis episode

Case Study: Metrics, against the "Always Up" Baseline

- Accuracy: share of days whose direction is predicted correctly
- Receiver operating characteristic (ROC) area under the curve (AUC)
  - the probability that a randomly chosen up day gets a higher predicted probability than a randomly chosen down day
- AUC $= 0.5$: no skill; AUC $= 1$: perfect ranking
- Example: up days with probabilities 0.7 and 0.4, down days with 0.5 and 0.2
  - 3 of the 4 pairs are ranked correctly, so AUC $= 0.75$
- Why both metrics
  - accuracy depends on the 0.5 threshold
  - AUC compares only the ranking of the probabilities
- **Fold mean**: the metric computed in each of the 5 test folds, then averaged
- **Pooled**: the metric computed once on all out-of-fold predictions together
  - used by the formal tests

Case Study: 14 Features and Four Models

Features, information up to day t

- **Momentum**: $\log P_t - \log P_{t-h}$
  - the log return over the last h days, $h \in \{1, 5, 20, 60, 120\}$
- **Realised volatility**
  - annualised standard deviation (SD) of daily log returns over 20 and 60 days, and their ratio
- **Distance to moving average**
  - $\log P_t$ minus the mean of $\log P$ over 50 or 200 days
- **RSI(14)**: $100 - 100/(1 + RS)$
  - RS = smoothed average gain over average loss, 14 days
  - between 0 and 100, high after rises
- **VIX**: log level and 5-day change
- **FFD** log price, $d = 0.25$
  - chosen on the full sample (ablation after the model comparison)

Models, deliberately conservative

- **Logit**: L2-penalised (Ridge), standardised features
- **RF**: 400 trees, minimum leaf 100
- **GBM**: histogram-based
  - depth 3, learning rate 0.03, 200 trees
- **MLP**: hidden layers of 32 and 16 units, early stopping on a purged chronological validation set
- Large leaves and shallow trees limit variance

A Classifier That Predicts Only the Majority Class (1/2)

	Pred. up	**Pred. down**
Actual up	TP = 3,575	FN = 65
Actual down	FP = 2,589	TN = 40

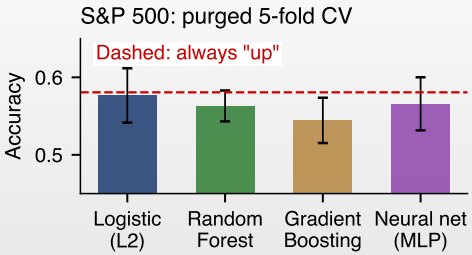
- L2-penalised logit, $n = 6{,}269$
  - pooled out-of-fold predictions of purged 5-fold CV
- TP, FN, FP, TN: true positives, false negatives, false positives, true negatives

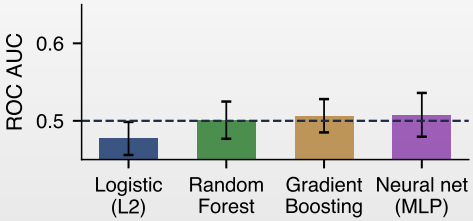
- Accuracy $= (TP + TN)/n = 57.7\%$
- Precision $= TP/(TP + FP) = 0.580$
- Recall $= TP/(TP + FN) = 0.982$
- $F_1 = 0.729$ (formula on the meta-labelling slide)
- The model predicts "up" on 98.3% of days
  - it has essentially learned "always up"

A Classifier That Predicts Only the Majority Class (2/2)

- **Majority baseline** from the same table
  - actual up days $(3{,}575 + 65)/6{,}269 = 3{,}640/6{,}269 = 58.1\%$
  - the accuracy of predicting "up" every day, with no model at all
- A 57.7%-accurate model is *worse* than doing nothing
- The cause is class imbalance
  - with weak features the predicted probabilities stay near the base rate 0.58, above the 0.5 threshold
  - so almost every day is classified "up"
  - the case study keeps the fixed-horizon label on purpose, so that the baseline is transparent
  - triple-barrier labels change the class balance the classifier sees (Seminar 13, exercises B1 and B8)
- Other baselines: yesterday's direction, a moving-average rule, buy-and-hold for trading
- A model has value only if it beats the baseline **out of sample**
  - after costs, by more than sampling error

Model Comparison: Fold Means





Fold mean	Logit	RF	GBM	MLP
Accuracy	57.7%	56.3%	54.4%	56.6%
AUC	0.477	0.501	0.507	0.508

- Bars: mean over the 5 purged folds; whiskers: ± 1 SD across folds
- Baseline "always up": **58.1%**; no model beats it

MFM_ch13_model_comparison ▶ Interactive

▸ Further Topics: FFD ablation, Newey–West

Are the Differences Significant? Pesaran–Timmermann and AUC Intervals (1/2)

- **Pesaran–Timmermann** (PT, Pesaran & Timmermann (1992)): are predicted and realised directions related?

$$PT = \frac{\hat{P} - \hat{P}^*}{\widehat{\mathrm{SE}}} \xrightarrow{d} N(0,1), \qquad \hat{P}^* = \hat{p}_y \hat{p}_x + (1-\hat{p}_y)(1-\hat{p}_x)$$

  - $\hat{P}$: share of days with the correct sign; $\hat{p}_y$, $\hat{p}_x$: shares of realised and of predicted "up"; $\hat{P}^*$: share correct expected if predictions were independent of outcomes
  - H_0: independence; $\widehat{\mathrm{SE}}$: standard error of $\hat{P} - \hat{P}^*$ under H_0 (Appendix)
  - $PT > 1.645$: skill at 5% (one-sided); $PT < 0$: fewer hits than chance
- PT assumes independent days
  - with overlapping labels we also use the slope t in $y_t = a + b\,x_t + u_t$
  - Newey–West errors with $q = 4$ (Newey–West slide)
  - x_t the predicted and y_t the realised direction, $b > 0$ means skill
- **AUC**: the DeLong standard error (SE) assumes independent days (Appendix)
- Hence a circular block bootstrap on blocks of $L = 20$ consecutive days (Künsch (1989))
  - four times the label length, which keeps the serial dependence

Are the Differences Significant? Pesaran–Timmermann and AUC Intervals (2/2)

Pooled out-of-fold	Logit	RF	GBM	MLP
$\hat{P}$ / $\hat{P}^*$	0.577 / 0.578	0.563 / 0.570	0.544 / 0.554	0.566 / 0.569
PT / HAC t (lag 4)	-0.80 / -0.62	-2.27 / -1.67	-1.97 / -1.49	-0.83 / -0.60
AUC	0.480	0.495	0.501	0.497
block-bootstrap 95% CI	[0.456; 0.505]	[0.470; 0.520]	[0.478; 0.525]	[0.474; 0.520]

- No significant positive directional skill: every block-bootstrap AUC interval contains 0.5
- With HAC (Heteroskedasticity and Autocorrelation Consistent) standard errors at lag 4, no model differs significantly from chance ($|t| \leq 1.67$)
  - PT, which treats the days as independent, gives -2.27 for RF: overlapping labels inflate significance
- Logit, for contrast: the DeLong 95% CI (Confidence Interval) is [0.466; 0.494] and excludes 0.5
  - but the DeLong CI treats days as independent; the block-bootstrap interval contains 0.5
- Consistent with weak-form efficiency (Chapter 2)
- Type of evidence: comparison of procedures
- Look for signal in cross-sections (Gu, Kelly & Xiu (2020))
  - in risk targets such as volatility (Chapters 5 and 9)
  - in crypto delisting (Będowska-Sójka, Wójcik & Pele (2026)) and in text (Chapter 15)

MFM_ch13_inference

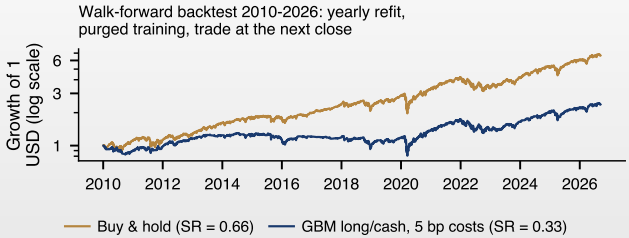
Walk-Forward Backtest: Strategy and Design

- The question
  - would a trader following the model have earned more, after costs, than by simply holding the index?
- Position $w_t \in \{0, 1\}$: long the S&P 500 if $\hat{p}_t > 0.5$, otherwise cash
  - decided with information up to t
  - $\hat{p}_t$: the model's probability of "up"
- Strategy return: the index return when invested
  - minus a cost each time the position changes
  - the decision at the close of day t is traded at the close of day $t + 1$

$$r^s_{t+\mathbf{2}} = w_t \, r_{t+\mathbf{2}} - c \, |w_t - w_{t-\mathbf{1}}|$$

  - $r_{t+\mathbf{2}}$: log return of the index from the close of $t + 1$ to that of $t + 2$; turnover $|w_t - w_{t-\mathbf{1}}|$: 1 on the days the position changes
  - $c = 5$ basis points (bp), i.e. 0.05% per unit traded: a prudent cost for an exchange-traded fund on the S&P 500
  - $r - c$ approximates $\log\left((1 - c)\, e^r\right) = r + \log(1 - c)$, the log return after paying the cost
- Exposure $\bar{w} = \frac{\mathbf{1}}{T} \sum_t w_t$: share of the T days spent invested
- Cash earns 0; interest on cash days is ignored
  - both Sharpe ratios use daily log returns with risk-free rate 0, annualised with $\sqrt{252}$
- Model: GBM refitted every January, 2010–2026, on all past data
  - purging labels that overlap the test year; FFD log price with $d = 0.15$, chosen on data up to 2009
- Benchmark: buy-and-hold ($w_t = 1$: exposure 100%, no turnover)

Walk-Forward Backtest: Results (1/2)



- 2010–2026: strategy Sharpe **0.33** vs. buy-and-hold **0.66** (annualised)
  - average exposure 76%
- The model stayed mostly long
  - its exits from the market (24% of the days) lowered the Sharpe ratio
- Chart: growth of one dollar invested, log scale
  - brown buy-and-hold, blue the GBM strategy

MFM_ch13_walk_forward ▶ Interactive

Walk-Forward Backtest: Results (2/2)

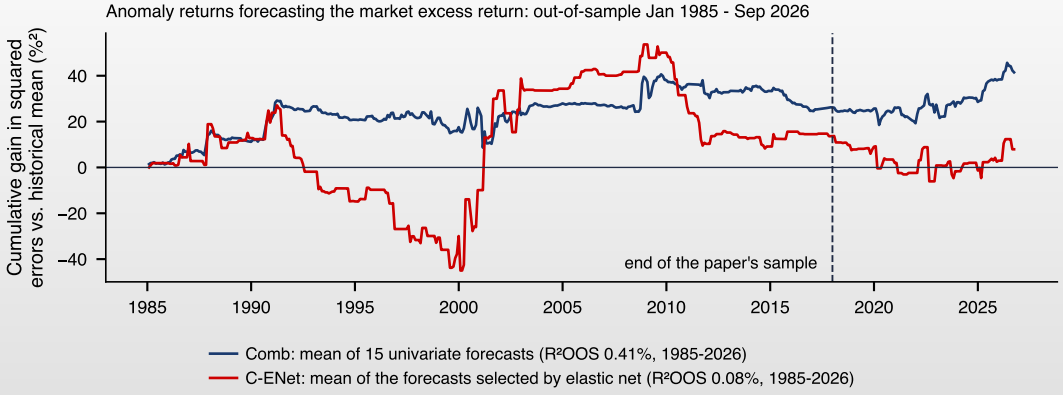
- The Sharpe-ratio difference, -0.33, has $p = 0.001$
  - Ledoit–Wolf HAC test, in Further Topics
  - requires a finite fourth moment, which Chapter 1 questions
- The strategy is significantly worse than buy-and-hold; there is no evidence of skill
- Type of evidence: comparison of procedures (the GBM strategy against buy-and-hold)
- Sensitivity of the strategy SR (Sharpe Ratio)
  - 0.31 with $d = 0.25$, chosen on the full sample, hence with information from the test period
  - 0.48 without the FFD log price
  - 0.43 when trading at the same close, an idealisation with no execution delay

MFM_ch13_walk_forward

Case Study: Do Cross-Sectional Anomalies Predict the Market?

- Dong, Li, Rapach & Zhou (2022): do long-short anomaly returns predict next month's market excess return?
  - anomaly: a portfolio long the stocks with high average returns on a characteristic and short those with low average returns
  - examples: value (the book-to-market ratio), momentum, accruals, net share issuance
- The cross-sectional literature and market-return predictability had been studied separately
- The benchmark is hard to beat: Welch & Goyal (2008) show that most published predictors do not beat the historical mean
- The paper's data: 100 US anomalies, monthly returns, January 1970 – December 2017
  - the first out-of-sample forecast has 180 months for 100 predictors
  - the regression on all predictors learns noise
- Method: shrinkage, i.e. elastic net, forecast combination and dimension reduction
- Main claim: anomalies have statistically and economically significant out-of-sample predictive power for the market
- Proposed explanation: asymmetric limits of arbitrage and persistent correction of overpricing
  - overpricing is harder to arbitrage than underpricing, because short selling is costly
  - hence the short leg of the anomalies carries most of the information

Anomalies and the Market: Out-of-Sample Gain over the Historical Mean



MFM_ch13_anomaly_market ▶ Interactive

Anomalies and the Market: Method and What the Replication Does Differently

- Predictive regression of anomaly i, estimated on an expanding window and re-estimated every month

$$r_{t+\mathbf{1}} = a_i + b_i\, x_{i,t} + \varepsilon_{i,t+\mathbf{1}}, \qquad \hat{r}^{(i)}_{t+\mathbf{1}} = \hat{a}_i + \hat{b}_i\, x_{i,t}$$

  - $r_{t+\mathbf{1}}$: market excess return (Mkt-RF) in month $t+1$
  - $x_{i,t}$: long-short return of anomaly i in month t
- Comb (simple combination): the mean of the N forecasts $\hat{r}^{(i)}_{t+\mathbf{1}}$
- C-ENet (combination elastic net): elastic net in the regression of r_t on the N forecasts, with non-negative coefficients
  - the forecast is the mean of the selected forecasts
  - ENet: elastic net on all predictors; OLS: the regression on all predictors
- The paper's periods: training 1970–1979, validation 1980–1984, out-of-sample 1985–2017
- The replication, on the Kenneth French Data Library
  - 15 anomalies instead of 100: SMB (Small Minus Big, size factor), HML (High Minus Low, value factor), RMW (Robust Minus Weak, profitability factor), CMA (Fama & French (2015)), momentum, short- and long-term reversal
  - and the extreme-decile spreads for accruals, net issuance, beta, variance, residual variance, E/P, CF/P (cash flow, CF, to price), D/P
  - elastic-net penalty ($\ell_\mathbf{1}$ and $\ell_\mathbf{2}$ in equal proportions) chosen by the small-sample corrected AIC
  - out-of-sample period extended to September 2026

Anomalies and the Market: Result and Limits

- Measure: R^2_{OOS} against the historical mean $\bar{r}_t$, computed up to month $t-1$ (Campbell & Thompson (2008))

$$R^2_{OOS} = 1 - \frac{\sum_t (r_t - \hat{r}_t)^2}{\sum_t (r_t - \bar{r}_t)^2}$$

  - the chart cumulates the squared errors of the historical mean minus those of the model
  - the curve rises in the months in which the model beats the historical mean
- The paper's figures, published version (Table II), 1985–2017: C-ENet 2.81%, ENet 2.03%, Comb 0.89%, OLS −2,513.86%
- The replication, 1985–2017: Comb 0.35%, C-ENet 0.18%, ENet −4.75%, OLS −6.84%
- The replication, 1985–2026: Comb 0.41% (one-sided Clark–West p 0.045), C-ENet 0.08% ($p = 0.15$)
  - Clark–West: the adjusted difference regressed on a constant, Newey–West errors with $\lfloor 4(T/100)^{2/9} \rfloor = 5$ lags
  - Comb p with 3, 6 and 12 lags: 0.063, 0.044 and 0.032; without the HAC correction: 0.068
  - the lag count is fixed by the rule, not chosen by significance
  - the Clark & West (2007) test requires a finite fourth moment, which Chapter 1 questions
- After the paper's sample, 2018–2026: Comb 0.61% ($p = 0.12$), C-ENet −0.23%
- Type of evidence: the ranking of procedures (forecast combination ahead of the regression on all predictors)
  - it does not support the size of the paper's gains
- Limits: 15 anomalies do not reproduce the 100
  - with 15 forecasts, C-ENet selects no forecast in 57% of the months and then uses the historical mean

MFM_ch13_anomaly_market

Selection Bias and the False Strategy Theorem

- The question of this section
  - how large is the best Sharpe ratio among N trials when none has skill?
- A researcher tries N strategies and reports the best: **selection bias**
  - even if all have zero true Sharpe ratio, $\max_n \widehat{SR}_n$ grows with N
- Harvey, Liu & Zhu (2016): given hundreds of published factors, a new factor needs $t > 3$, not 2

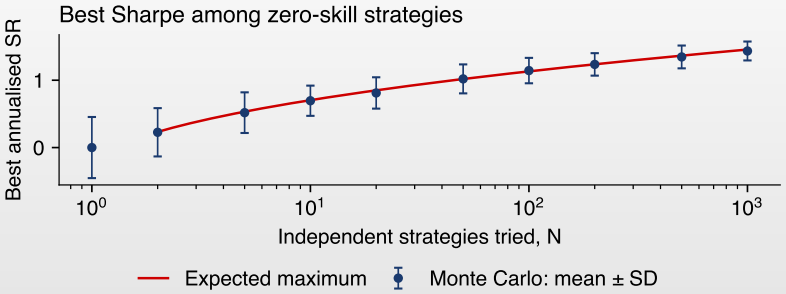
Theorem 1 (False Strategy Theorem, Bailey & López de Prado (2014))

- The **expected maximum Sharpe ratio** of N strategies with no skill
  - if $\widehat{SR}_n$, $n = 1, \ldots, N$, are independent and approximately $N(0, \mathrm{Var}[\widehat{SR}])$, then

$$SR_{\mathbf{0}} \equiv \mathbb{E}\Big[\max_n \widehat{SR}_n\Big] \approx \sqrt{\mathrm{Var}[\widehat{SR}]}\left[(1-\gamma_E)\,\Phi^{-\mathbf{1}}\left(1-\tfrac{\mathbf{1}}{N}\right) + \gamma_E\,\Phi^{-\mathbf{1}}\left(1-\tfrac{\mathbf{1}}{Ne}\right)\right]$$

- Notation
  - $\widehat{SR}_n$: estimated Sharpe ratio of strategy n; Φ: standard Normal distribution function; $\Phi^{-\mathbf{1}}$: its quantile function
  - $\mathrm{Var}[\widehat{SR}]$: sampling variance of one estimated SR, $\approx 1/T$ per period for i.i.d. returns, T the number of return periods
  - $\gamma_E \approx 0.5772$: Euler–Mascheroni constant
- $SR_{\mathbf{0}}$ is the Sharpe ratio the luckiest of N worthless strategies is expected to show
  - it rises with N and with the sampling noise $\sqrt{\mathrm{Var}[\widehat{SR}]}$

The Best of N Zero-Skill Strategies



- Strategies with i.i.d. zero-mean returns, 5 years of daily data (annualised $\mathrm{Var}[\widehat{SR}] \approx 1/5$)
- Trying $N = 1000$ variants yields a best Sharpe ratio of ≈ 1.45 — **with no skill at all**
- Horizontal axis: number of strategies N (log scale); vertical axis: best annualised SR
  - red line: the theorem; dots: Monte Carlo mean $\pm$ one SD, in line with the theorem

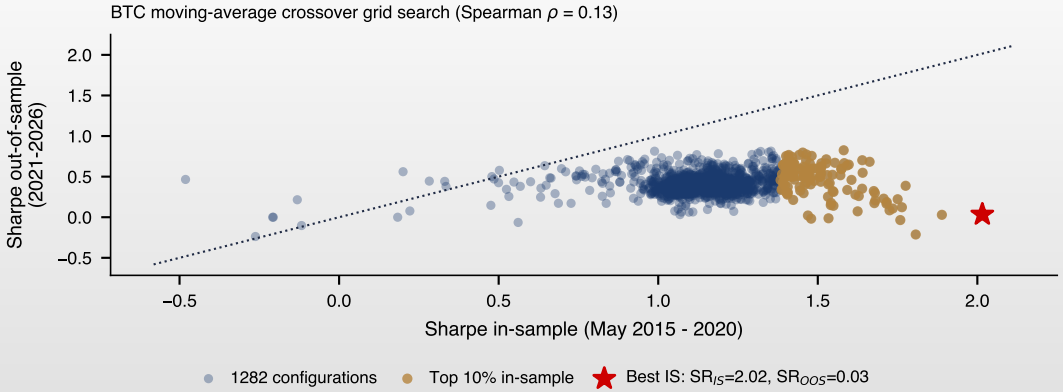
MFM_ch13_backtest_overfitting ▶ Interactive

In-Sample (IS) vs. Out-of-Sample (OOS): Bitcoin Grid Search (1/2)

- Moving-average (MA) crossover on Bitcoin
  - long when the fast MA of the log price exceeds the slow MA by more than a band
- 1,284 configurations (fast, slow, band)
  - in sample (IS) 16 May 2015 – 2020, out of sample (OOS) 2021–2026
  - the tested family, fixed in sample only: the 1,282 rules with non-zero in-sample variance
  - 3 of them stay in cash throughout the OOS period: OOS SR undefined, counted as 0
  - the decision at the close of day t is traded at the close of $t+1$
- Next chart: IS Sharpe ratio on the horizontal axis, OOS Sharpe ratio on the vertical axis
  - brown points: the top 10% in sample; red star: the selected rule; dotted line: equal SRs
- Selected rule (11/20 days): SR_{IS} **2.02** $\rightarrow$ SR_{OOS} **0.03** (annualised with $\sqrt{365}$)
- Spearman rank correlation between IS and OOS SR: $\rho = 0.13$
- Median OOS SR 0.50 for the top 10% in sample vs. 0.41 for all rules
- The result is consistent with overfitting of the selected rule
- A positive family SR does not show incremental predictive information
  - skill against buy-and-hold needs a separate test (SPA, in Further Topics)
- Type of evidence: comparison of procedures (in-sample choice, judged out of sample)

MFM_ch13_backtest_overfitting

In-Sample (IS) vs. Out-of-Sample (OOS): Bitcoin Grid Search (2/2)



MFM_ch13_backtest_overfitting ▶ Interactive

Sharpe Ratio Inference: Standard Error and PSR (1/2)

- The question
  - how precise is an estimated Sharpe ratio?
  - how do we compare it with a benchmark?
- Per-period Sharpe ratio of daily log returns, risk-free rate 0: $\widehat{SR} = \bar{r}/\hat{\sigma}$
- Shape of the distribution, with μ, σ the mean and SD of r
  - skewness $\gamma_3 = \mathbb{E}[(r-\mu)^3]/\sigma^3$
  - ordinary kurtosis $\gamma_4 = \mathbb{E}[(r-\mu)^4]/\sigma^4$ (3 for the Normal distribution)
- Sampling variance of $\widehat{SR}$, i.i.d. returns, delta method (Lo (2002); Opdyke (2007)):

$$\mathrm{Var}(\widehat{SR}) \approx \frac{1}{T}\left(1 - \gamma_3\, SR + \frac{\gamma_4 - 1}{4}\, SR^2\right)$$

  - T: number of returns; SR: true per-period Sharpe ratio
  - negative skewness and fat tails widen the uncertainty; Normal distribution ($\gamma_3 = 0$, $\gamma_4 = 3$): $(1 + \frac{1}{2}SR^2)/T$
  - the formula requires a finite fourth moment, which Chapter 1 questions

Sharpe Ratio Inference: Standard Error and PSR (2/2)

Definition 5 (Probabilistic Sharpe Ratio (PSR), Bailey & López de Prado (2012))

- Called by its authors the probability that the true SR exceeds a benchmark SR^*
  - formally Φ of a z-statistic, with the variance above in the denominator:

$$\widehat{PSR}(SR^*) = \Phi\left(\frac{(\widehat{SR} - SR^*)\sqrt{T-1}}{\sqrt{1 - \hat{\gamma}_3\,\widehat{SR} + \frac{\hat{\gamma}_4 - 1}{4}\,\widehat{SR}^2}}\right)$$

- SR^*: the benchmark, e.g. 0 or the expected maximum SR_0
  - $\hat{\gamma}_3$, $\hat{\gamma}_4$: sample skewness and kurtosis
- Reading: $\widehat{PSR} = 1 - p$, with p the one-sided p-value of $H_0 : SR \leq SR^*$
- $\widehat{PSR} > 0.95$ means rejecting H_0 at 5%, not a 95% probability for the true SR
  - under i.i.d. returns and the delta-method approximation

Deflated Sharpe Ratio: the Best Bitcoin Rule

- **Deflated Sharpe Ratio** (DSR, Bailey & López de Prado (2014)): $DSR = \widehat{PSR}(SR_{\mathbf{0}})$
  - the PSR with the benchmark set to $SR_{\mathbf{0}}$, the expected maximum of the False Strategy Theorem
  - $SR_{\mathbf{0}}$ computed with $\text{Var}[\widehat{SR}_n]$ of the trials
  - corrects for selection (N trials), non-Normality and sample length; always report N
  - $DSR \leq 0.95$: insufficient evidence at the chosen threshold, not proof that the strategy has no skill
- Worked example, best in-sample rule: $T = 2{,}057$ days
  - $\widehat{SR} = 0.1055$ a day (2.02 annualised), $\hat{\gamma}_{\mathbf{3}} = 0.16$, $\hat{\gamma}_{\mathbf{4}} = 12.23$
  - $\hat{\gamma}_{\mathbf{4}}$ and the DSR require a finite fourth moment, which Chapter 1 questions
  - denominator $\sqrt{1 - 0.16 \cdot 0.1055 + \frac{\mathbf{12.23}-\mathbf{1}}{\mathbf{4}} \cdot 0.1055^{\mathbf{2}}} = 1.007$; $N = 1{,}282$ trials
  - $z = (\widehat{SR} - SR_{\mathbf{0}})\sqrt{T-1}/1.007$, the argument of Φ; $DSR = \Phi(z)$; SD $= \sqrt{\text{Var}[\widehat{SR}_n]}$

$\text{Var}[\widehat{SR}_n]$ from	SD	$SR_{\mathbf{0}}$ (per day)	z / DSR
dispersion of the 1,282 trial SRs	0.0109	0.036	3.11 / 0.999
independent trials, $1/T$	0.0220	0.073	1.45 / 0.926

- Row 1: the observed dispersion of the 1,282 SRs
  - it includes the correlation between rules
- Row 2: rules treated as independent, Var $= 1/T$, the conservative variant
- The verdict depends on this choice: the OOS SR of 0.03 sides with the conservative value (DSR $= 0.926$, below 0.95)

Inference after Variable Selection: Post-Double Selection (1/2)

- The question of this section
  - how do we get a valid confidence interval for the effect of one variable after ML has chosen the controls?
- The LASSO path answers "which feature enters first", not "which feature matters"
  - OLS t-statistics after LASSO ignore the selection step
- A control correlated with the regressor of interest d_t but only weakly with y_t gets dropped
  - its effect moves into the coefficient of d_t (omitted-variable bias)
- Partially linear model: the effect of d_t is linear
  - the controls enter through unknown functions

$$y_t = \theta\, d_t + g(\mathrm{x}_t) + \varepsilon_t, \qquad d_t = m(\mathrm{x}_t) + v_t$$

  - y_t: outcome; d_t: regressor of interest (not the FFD order d); θ: its effect, the parameter to estimate
  - x_t: the p controls; g: their effect on y_t; m: how they predict d_t; ε_t, v_t: errors
  - approximately sparse: a few controls explain most of g and m

Inference after Variable Selection: Post-Double Selection (2/2)

- Post-double selection (Belloni, Chernozhukov & Hansen (2014)):
  1. LASSO of y_t on $x_t \Rightarrow$ selected set $\hat{S}_y$
  2. LASSO of d_t on $x_t \Rightarrow$ selected set $\hat{S}_d$
  3. OLS of y_t on d_t and $x_{\hat{S}_y \cup \hat{S}_d}$, robust standard error
- $\hat{S}_y$, $\hat{S}_d$: indices of the controls kept by each LASSO
  - $x_{\hat{S}_y \cup \hat{S}_d}$: the controls kept by either
- A control is missed only if it matters for *neither* equation
  - so the omitted bias is second order
- Valid under independent sampling and sparsity
  - with HAC errors an approximation for time series
  - penalty and conditions in the Appendix

Double/Debiased Machine Learning (DML): Orthogonal Scores and Cross-Fitting

- Chernozhukov et al. (2018): residualise both variables with *any* learner (LASSO, RF, boosting)
  - then regress residual on residual

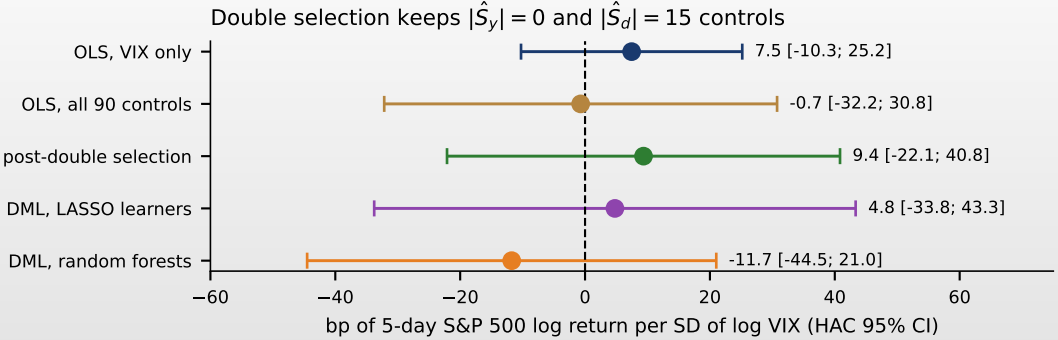
$$\tilde{y}_t = y_t - \hat{\ell}(\mathrm{x}_t), \quad \tilde{v}_t = d_t - \hat{m}(\mathrm{x}_t), \qquad \hat{\theta} = \frac{\sum_t \tilde{v}_t \tilde{y}_t}{\sum_t \tilde{v}_t^2}$$

  - $\hat{\ell}(\mathrm{x}) = \hat{\mathbb{E}}[y \mid \mathrm{x}]$: ML prediction of y_t from the controls; $\hat{m}(\mathrm{x})$: ML prediction of d_t
  - $\tilde{y}_t$, $\tilde{v}_t$: the parts of y_t and d_t the controls do not explain; $\hat{\theta}$: OLS slope of $\tilde{y}_t$ on $\tilde{v}_t$
- Neyman-orthogonal score $\psi = (\tilde{y} - \theta\tilde{v})\,\tilde{v}$
  - its mean is zero at the true θ
  - first-order errors in $\hat{\ell}$ and $\hat{m}$ do not bias $\hat{\theta}$ (rate conditions in the Appendix)
- Cross-fitting: fit $\hat{\ell}$, $\hat{m}$ on $K-1$ folds, residualise the held-out fold, repeat
  - removes the own-observation overfitting bias
- The published theory assumes independent sampling
  - with overlapping 5-day labels we use contiguous folds purged by 5 days around the test fold and HAC standard errors
  - an adaptation to serial dependence outside the published theory

Worked Example: Does the VIX Predict S&P 500 Returns? (1/3)

- The question
  - does the VIX level predict the S&P 500 return over the next 5 days, given a large set of controls?
- y_t: 5-day forward log return, in basis points (bp)
  - d_t: log VIX, standardised (1 SD $= 0.356$)
- 90 controls: 12 technical features, their squares and pairwise products
  - $n = 6{,}269$ days; HAC errors with 4 lags
- The single LASSO drops the VIX and every control
  - "no effect", with no standard error
- Double selection keeps $|\hat{S}_y| = 0$ controls in the y_t equation and $|\hat{S}_d| = 15$ in the d_t equation
  - $|\cdot|$: number of elements
- Next chart: one estimator per row
  - the dot is the estimate of θ (bp per week per SD of log VIX)
  - the bar its 95% HAC CI, the dashed line no effect

Worked Example: Does the VIX Predict S&P 500 Returns? (2/3)



MFM_ch13_inference ▶ Interactive

Worked Example: Does the VIX Predict S&P 500 Returns? (3/3)

- Double selection: 95% CI for θ $[-22.1;\ 40.8]$
  - in bp per week per SD of log VIX
- OLS with log VIX only: 7.5 $[-10.3;\ 25.2]$
  - OLS with all 90 controls, double selection and DML (Double/Debiased Machine Learning) give -0.7, 9.4, 4.8 (LASSO) and -11.7 (random forests)
  - all with intervals about twice as wide
- Controls that predict d_t absorb part of the variation of log VIX
  - less own variation in d_t means a wider interval
  - the price of protection against omitted-variable bias
- The sign change under DML-RF (DML with random forests) is noise
  - every point estimate lies inside the double-selection and DML intervals
- Honest intervals contain zero and large effects of both signs
  - the data cannot decide
- Type of evidence: association (the variables move together; this does not prove causation)
  - a predictive regression: even a non-zero θ would not show that the VIX causes returns

MFM_ch13_inference

The Virtue of Complexity, and a Critical Reading [deeper]

- Kelly, Malamud & Zhou (2024): ridge regressions on P random Fourier features of a few predictors
  - rolling windows as short as $T = 12$ months, P in the thousands
  - random Fourier features: sin and cos of random linear combinations of the predictors
  - theory: with $P > T$ the minimum-norm solution (the exact fit with the smallest $\|\boldsymbol{\beta}\|_2$) shrinks implicitly ("double descent")
  - R^2_{OOS} may be negative while the Sharpe ratio of market timing **rises** with complexity; complexity is not automatically overfitting
- Nagel (2025): for $P \gg T$ the forecast is a similarity-weighted average of the T training returns
  - in short windows similarity is mostly recency, so the method becomes a **volatility-timed momentum** strategy, which loses on artificial data with reversals
- Avramov, Cheng & Metzker (2023): deep-learning signal profits come from hard-to-arbitrage stocks
  - excluding microcaps, distressed firms or high-volatility episodes attenuates them, and costs erode them further
- Chen, Pelger & Zhu (2024): impose economic structure instead
  - the no-arbitrage moment conditions become the loss of a deep network (next slides)
- Survey of the field: Kelly & Xiu (2023); textbook: Dixon, Halperin & Bilokon (2020)

Case Study: Deep Learning in Asset Pricing, Question and Method [deeper]

- Chen, Pelger & Zhu (2024): estimate the stochastic discount factor (SDF) M_{t+1} that prices **all individual US stocks**
  - conditioning on 46 firm characteristics and on the history of 178 macroeconomic series
- No arbitrage: the SDF prices every excess return (Section I.A, eq. 1)

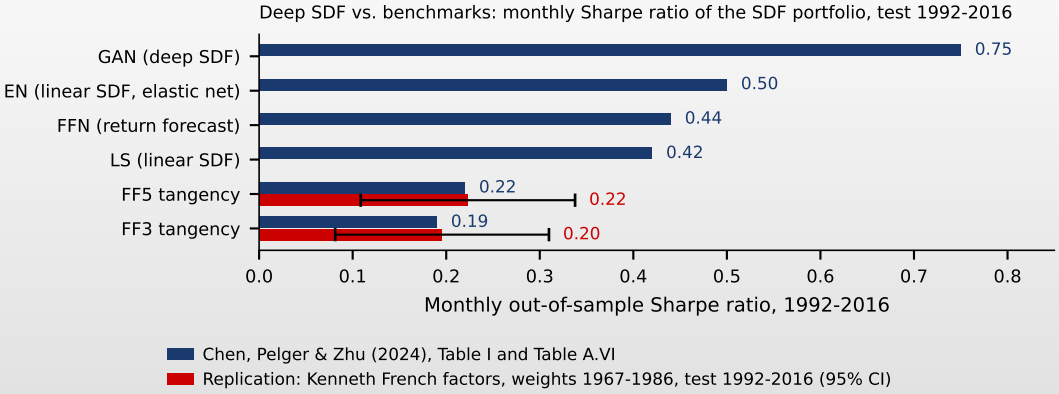
$$E_t[M_{t+1}R^e_{t+1,i}] = 0, \qquad M_{t+1} = 1 - \omega_t^\top R^e_{t+1}$$

  - $R^e_{t+1,i}$: excess return of stock i over the Treasury bill from month t to $t+1$; R^e_{t+1}: the vector of all stocks
  - E_t: expectation given the information at t; ω_t: portfolio weights, a function of characteristics and macro states
- $F_{t+1} = \omega_t^\top R^e_{t+1}$, the SDF portfolio, is the conditional tangency portfolio
  - tangency portfolio: highest Sharpe ratio; and

$$E_t[R^e_{t+1,i}] = \beta_{t,i}\, E_t[F_{t+1}]$$

  - $\beta_{t,i}$: loading of stock i on F; a stock's expected excess return is its loading times the premium of F (derivation in Further Topics)

Deep Learning in Asset Pricing: Out-of-Sample Sharpe Ratios (1/2) [deeper]



MFM_ch13_sdf_replication ▶ Interactive

Deep Learning in Asset Pricing: Out-of-Sample Sharpe Ratios (2/2) [deeper]

- Blue: the paper's figures (published version), monthly out-of-sample SR 1992–2016
  - GAN (the paper's generative adversarial network) 0.75, EN (linear SDF with elastic net) 0.50, FFN (the feedforward forecasting network) 0.44, LS (unregularised linear SDF) 0.42 (Table I)
  - FF5 (Fama–French five-factor model) 0.22, FF3 0.19 (Table A.VI)
  - the regularised no-arbitrage model already beats the forecasting network
- Red: FF3 and FF5 tangency portfolios $w \propto \hat{\Sigma}^{-1}\hat{\mu}$ on the Kenneth French factors
  - weights fixed on 1967–1986
  - FF5 0.22 [0.11; 0.34], FF3 0.20 [0.08; 0.31] (SE 0.058, Lo (2002))
  - the SE requires a finite fourth moment, which Chapter 1 questions
  - $\hat{\mu}$, $\hat{\Sigma}$: sample mean vector and covariance matrix of the factor returns
- Type of evidence: comparison of procedures (out-of-sample SRs of the models)
  - after 2016 the same fixed weights give a monthly SR of 0.06 (FF5) and 0.03 (FF3): Further Topics

MFM_ch13_sdf_replication

An Open Question: Discovery or Data Mining at Scale?

- Chen, Lopez-Lira & Zimmermann (2022): mining 29,000 accounting ratios for $t > 2$
  - gives cross-sectional predictability similar to peer-reviewed predictors
  - for both, about half remains after the original samples
- Novy-Marx & Velikov (2025): an LLM (large language model) pipeline mines over 30,000 signals
  - keeps 96 that pass a fixed protocol
  - writes hundreds of complete papers, with invented motivations and some imagined citations
  - HARKing (hypothesising after the results are known) at scale
- Open: when a machine proposes features, model and story
  - what is the number of trials?
  - what separates a discovery from the best of many draws?
- The discovery loop
  - literature → competing hypotheses → data and code → robustness fixed in advance
  - → referee-style critique → report every test
  - reusable prompts in the Appendix
- What the human checks
  - every reference resolves (DOI, arXiv)
  - every trial is logged, so M in the correction counts them all
  - correlated tests get a common yardstick (a placebo or a joint bootstrap of the scores)
  - transformations live inside the training fold

Mini-Case: 102 Tests, One Old Effect Passes the Placebo (1/2)

- A typical assistant proposal: "add more indicators"
  - "momentum, moving-average distances, volatility ratios, RSI and VIX changes over several windows"
- 34 features $\times$ 3 horizons (1, 5, 20 days) = 102 predictive regressions
  - on 2001–2015, Newey–West t; only targets whose window ends by 31 December 2015 (1, 5 and 20 days dropped at the boundary)
  - regression j: $r_{t,t+h} = \alpha_j + \beta_j x_{j,t} + e_t$, with $x_{j,t}$ feature j and $r_{t,t+h}$ the next h-day return; $H_{0,j} : \beta_j = 0$
- Corrections for $M = 102$ tests, $p_{(1)} \leq \cdots \leq p_{(M)}$ the ordered p-values
  - Holm: reject while $p_{(j)} \leq 0.05/(M - j + 1)$; controls the FWER (Family-Wise Error Rate)
  - Benjamini–Hochberg (BH, Benjamini & Hochberg (1995)): reject the j smallest, for the largest j with $p_{(j)} \leq j\, q/M$
  - controls the false discovery rate (FDR) at q, the expected share of false rejections among all rejections
  - shift placebo: the 95% quantile of $\max_j |t_j|$ when the target is shifted circularly
  - the shift keeps the correlation between the features but also changes their link with conditional volatility: a heuristic yardstick, not an exact test

Mini-Case: 102 Tests, One Old Effect Passes the Placebo (2/2)

- $|t| > 1.96$: **22** of 102, against 5.1 expected under the global null at 5%
- Holm (FWER 5%): **0**; BH (FDR 10%): **13**; $|t| > 3$ (Harvey, Liu & Zhu (2016)): **4**
- Shift placebo: 95% quantile of $\max_j |t_j| = 3.32$
  - 500 circular shifts of the target
  - one predictor exceeds it, the 1-day return on the next day's return, $t = -3.38$: short-term reversal, a known effect
- Holdout 2016–2026: 3 of the 22 keep their sign with $|t| > 1.96$
  - nominal replications, all three reversal variants; the VIX-based "finds" fail
  - with Holm at 5% over the 22 checks none passes (smallest p 0.019, threshold $0.05/22 = 0.0023$)
- BH needs independence or positive dependence of the true-null p-values
  - correlated predictors do not ensure it
- Type of evidence: association
  - short-term reversal is a predictive regularity, not an identified mechanism

MFM_ch13_ai_discovery

Your Turn: a Project Seed

- **Questions**
  - can an AI-assisted feature search find a predictor of index returns (S&P 500, BET – the Bucharest Stock Exchange benchmark index, Bitcoin)?
  - does it survive a pre-registered correction and a holdout?
  - does it go beyond short-term reversal?
- **Data**: S&P 500 and Bitcoin are in the course data; BET is not
  - a team that chooses BET names the source (e.g. the Bucharest Stock Exchange) and freezes the sample
  - it logs the download date, n and the first and last close
- **Deliverables**
  - a feature list fixed before the data are touched
  - all tests with Holm, BH and the max-$|t|$ null
  - a separate holdout
  - DSR and PBO for any strategy built on the survivors
- **Log** (templates)
  - each prompt that shaped a decision, in `AI_USE.md`
- Each invented reference, leak or undercounted trial in `AI_ERRORS.md`, with the test that caught it

Key Takeaways

Four empirical lessons

- An ADF rejection is not stationarity
  - truncated FFD of the S&P 500 log price stays $I(1)$
  - estimate memory directly with local Whittle
- Shuffled K-fold "finds" skill in pure noise
- On the S&P 500 no model beats "always up" (PT, DM, GW)
- The best of many backtests must be deflated or bootstrapped
  - on the Bitcoin grid the best in-sample rule keeps nothing out of sample

Next

- Chapter 14
  - deep learning and time-series foundation models
- Chapter 15
  - large language models and sentiment

A Checklist Before Any Claim

- ☐ Point-in-time data; transformations fitted on training folds only
- ☐ Volatility-scaled labels with recorded spans
- ☐ Purged K-fold or CPCV with embargo for selection; walk-forward for the final test
- ☐ A naive baseline; PT, DM / CW (Clark–West, test for nested models) / GW and block-bootstrap intervals
- ☐ Costs included and every trial logged
- ☐ DSR, SPA and PBO for the selected strategy
- ☐ "Selected" or "important" is not "significant": double selection or DML for inference

Check Your Understanding

1. Why does the logit with 57.7% accuracy show no skill when 58.1% of the days are "up"?
2. What does a correlation of 0.98 between the FFD series ($d = 0.25$) and the log price say about an ADF rejection?
3. Why does shuffled K-fold report 69.4% on a target with nothing to predict, while purged K-fold with embargo reports 49.6%?
4. How does SR_0, the expected maximum of N zero-skill strategies, change when N rises from 10 to 1,000?
5. Why does the best Bitcoin rule, with an in-sample SR of 2.02, get a DSR of 0.999 or 0.926 depending on the variance used?
6. Why do the DeLong interval for the AUC, [0.466; 0.494], and the block-bootstrap interval, [0.456; 0.505], lead to different conclusions?
7. Which question about the VIX does double selection answer that the LASSO path does not?

The Practical Decision

- **What I choose**
  - for the 5-day direction of the S&P 500: buy-and-hold, not the GBM strategy
  - in the 2010–2026 walk-forward, the Sharpe ratio is 0.66 for buy-and-hold and 0.33 for GBM
  - for any model: purged CV with embargo for selection, walk-forward for the final evaluation
- **Under which conditions**
  - the directional signal stays weak: no model beats the "always up" benchmark of 58.1%
  - each change of position costs 5 bp
  - every trial is logged and the number of trials N is reported
- **What result would make me drop it**
  - a model that beats buy-and-hold out of sample, after costs
  - with a PT statistic above 1.645
  - with a block-bootstrap AUC interval entirely above 0.5
  - with a DSR above 0.95 under the conservative variance $1/T$ (the best Bitcoin rule has 0.926)

Evidence, and the Economic Value of a Small R^2_{OOS} (1/2)

- Campbell & Thompson (2008): what a small R^2 is worth to an investor
  - returns have a predictable and an unpredictable part

$$r_{t+1} = \mu + x_t + \varepsilon_{t+1}, \qquad R^2 = \frac{\mathrm{Var}(x_t)}{\mathrm{Var}(r_{t+1})}$$

  - μ: mean return; x_t: predictable component, known at t, mean zero; ε_{t+1}: unpredictable shock
  - R^2: share of the return variance explained by the predictor
- An unconstrained mean-variance investor who knows x_t and re-optimises every period
  - no costs, no estimation error; SR_t: the conditional Sharpe ratio at t

$$\mathbb{E}[SR_t^2] = \frac{SR_{BH}^2 + R^2}{1 - R^2}$$

  - $\mathbb{E}[SR_t^2]$ is not, in general, the square of the unconditional Sharpe ratio of the dynamic strategy
  - the optimal mean-variance utility at t is $SR_t^2/(2\gamma)$, γ the risk aversion, so the gain reads as mean utility
  - SR_{BH}: SR of buy-and-hold, which ignores x_t; $R^2 = 0$ gives $\mathbb{E}[SR_t^2] = SR_{BH}^2$

MFM_ch13_inference

Evidence, and the Economic Value of a Small R^2_{OOS} (2/2)

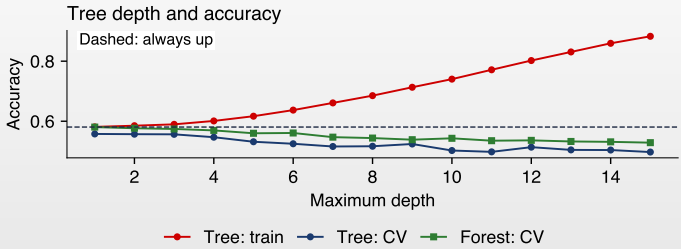
- S&P 500, monthly log returns 2001–2026
  - $SR_{BH} = 0.128$ a month (0.44 annualised with $\sqrt{12}$)
  - $R^2 = 0.5\%$ a month $\Rightarrow \sqrt{\mathbb{E}[SR_t^2]} = 0.51$ annualised ($+15\%$ over SR_{BH})
  - $R^2 = 1\% \Rightarrow 0.57$ (about $+27\%$)
- For a **single index** the effective sample is far smaller than for thousands of stocks
  - the situation of the S&P 500 case study

MFM_ch13_inference

Where the Campbell–Thompson Formula Comes From

- Same model: $r_{t+1} = \mu + x_t + \varepsilon_{t+1}$
  - $\mathbb{E}[x_t] = 0$, ε_{t+1} independent of x_t
  - write $\sigma_r^2 = \mathrm{Var}(r_{t+1})$
- Step 1, the risk the informed investor still bears
  - $\mathrm{Var}(r_{t+1}) = \mathrm{Var}(x_t) + \sigma_\varepsilon^2$ and $R^2 = \mathrm{Var}(x_t)/\sigma_r^2$, so
$$\sigma_\varepsilon^2 = (1 - R^2)\,\sigma_r^2$$
- Step 2, the investor's conditional Sharpe ratio at t
  - mean $\mu + x_t$, variance σ_ε^2, squared SR $(\mu + x_t)^2/\sigma_\varepsilon^2$
- Step 3, average over t: $\mathbb{E}[(\mu + x_t)^2] = \mu^2 + \mathrm{Var}(x_t) = \mu^2 + R^2\sigma_r^2$
- Step 4, divide and use $SR_{BH} = \mu/\sigma_r$:
$$\mathbb{E}[SR_t^2] = \frac{\mu^2 + R^2\sigma_r^2}{(1 - R^2)\,\sigma_r^2} = \frac{SR_{BH}^2 + R^2}{1 - R^2}$$
- So the formula gives the average squared *conditional* SR, not the unconditional SR of the strategy
  - for an investor who re-optimises each month, in monthly units
- Numbers, $R^2 = 0.005$
  - $\mathbb{E}[SR_t^2] = (0.128^2 + 0.005)/0.995 = 0.02149$, $\sqrt{\mathbb{E}[SR_t^2]} = 0.1466$ a month
  - $\times\sqrt{12} = 0.51$ a year; $0.1466/0.128 = 1.15$

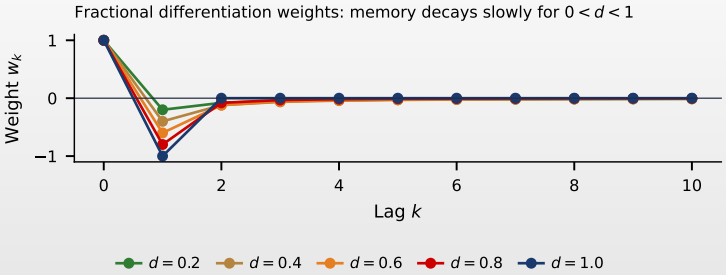
Bias–Variance in Practice: Tree Depth



- S&P 500, 5-day direction: training accuracy rises with depth to $\approx 88\%$ at depth 15
  - purged-CV accuracy **falls** towards 50%
  - purged CV drops from training the observations whose labels overlap the test set (Section 5)
- Horizontal axis: maximum tree depth; vertical axis: accuracy
  - red training, blue CV of one tree, green CV of the forest
- Random forests degrade more slowly, but never beat the "always up" baseline (dashed line)

MFM_ch13_regularization

Fractional Differentiation Weights



- Horizontal axis: lag k; vertical axis: weight w_k, for five values of d
- For $d = 1$ only the first lag matters
- For small d the weights of distant lags remain non-zero
- All weights after w_0 are negative
  - $\tilde{X}_t$ compares today's level with a weighted history

MFM_ch13_ffd

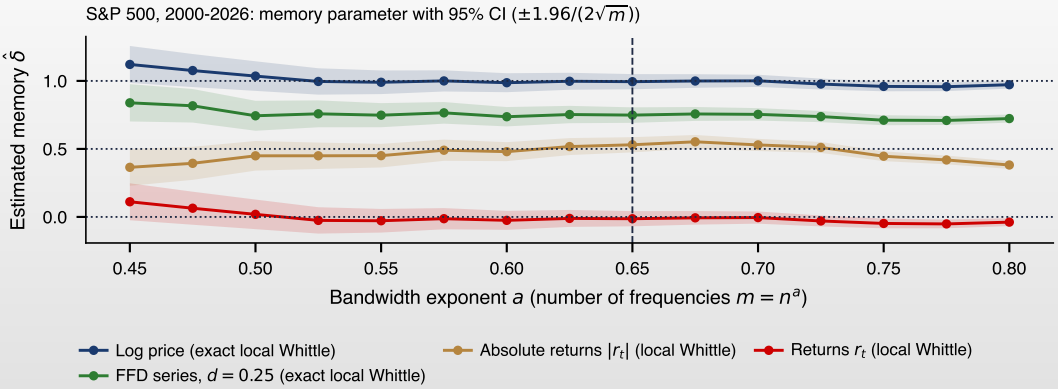
Memory Estimated Directly: Local Whittle (1/3)

- Idea: a long-memory series has most of its variance in slow cycles
  - the slope of its spectrum near frequency 0 reveals δ (here λ is the frequency, not the penalty)

$$f(\lambda) \sim G\,\lambda^{-2\delta} \quad \text{as } \lambda \to 0$$

  - $f(\lambda)$: spectral density, the variance contributed by cycles of frequency λ (period $2\pi/\lambda$ days)
  - $G > 0$: scale constant; $\sim$: the ratio of the two sides tends to 1
  - $\delta > 0$: f explodes near 0 (long memory); $\delta = 0$: flat (short memory); $\delta < 0$: f vanishes near 0
- Local Whittle (LW, Robinson (1995)): fits this law only on the first $m = n^a$ Fourier frequencies
  - $\lambda_j = 2\pi j/n$, $j = 1, \ldots, m$
  - n: sample size; $a \in (0, 1)$: bandwidth exponent; a larger a uses more frequencies: less variance, more bias from short-run dynamics
- Exact local Whittle (ELW, Shimotsu & Phillips (2005)): also valid for non-stationary $\delta \geq \frac{1}{2}$
  - both estimators in the Appendix
- 95% confidence interval (CI): $\hat{\delta} \pm 1.96/(2\sqrt{m})$
  - the standard error $1/(2\sqrt{m})$ depends only on m

Memory Estimated Directly: Local Whittle (2/3)



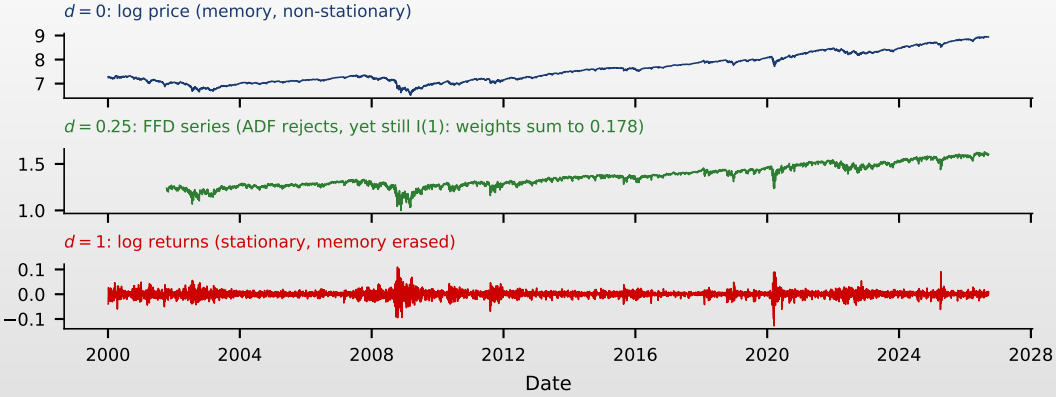
MFM_ch13_ffd

Memory Estimated Directly: Local Whittle (3/3)

- Chart: the exponent a (with $m = n^a$) on the horizontal axis, $\hat{\delta}$ on the vertical axis
  - with its 95% CI band
  - the vertical line marks $a = 0.65$, the value used below (S&P 500, 2000–2026)
- Log price (ELW): $\hat{\delta} = 0.99$, CI [0.94; 1.05]
- FFD series (ELW): $\hat{\delta} = 0.75$, CI [0.69; 0.81]
- Returns: $\hat{\delta} = -0.01$, CI [−0.07; 0.04]
- Absolute returns $|r_t|$: $\hat{\delta} = 0.53$, CI [0.47; 0.59]
- The FFD series stays near $1 - d$ over the frequency band used, although its integration order is still 1
- The long memory is in **volatility** (Chapters 5 and 9)
  - so fractional filters make sense for $|r_t|$, not for the price

MFM_ch13_ffd

Prices, FFD Series and Returns (1/2)



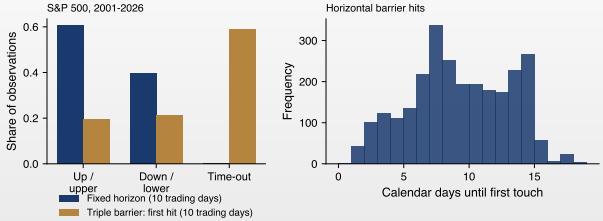
MFM_ch13_ffd

Prices, FFD Series and Returns (2/2)

- The top panel shows the S&P 500 log price, an $I(1)$ series
- The middle panel shows the FFD series with $d = 0.25$: the price trend is still visible
  - because the series is $0.178 \log P_t$ plus a stationary part
- The bottom panel shows log returns, with no trend
  - only volatility clustering is visible

MFM_ch13_ffd

Label Distribution: Fixed Horizon vs. Triple Barrier



- S&P 500, 2001–2026, $h = 10$, barriers at $\pm 1\,\sigma\sqrt{h}$
  - 19.6% up, 21.3% down, 59.0% time-out
  - labels after the sign rule at the vertical barrier: 59.1% $+1$ and 40.9% -1 (fixed horizon: 60.5% up)
- The chart counts **exit types**: time-out = exit at the vertical barrier
  - under the sign rule these 59.0% still get ± 1 from the sign of their return
  - under the "0" convention they form a third class
- Right panel: histogram of calendar days until the first horizontal barrier is touched
  - typically after 1–2 weeks

MFM_ch13_triple_barrier

Overlapping Labels and Sample Uniqueness (1/2)

- Label i spans $[t_{0,i}, t_{1,i}]$
  - consecutive labels share returns $\Rightarrow$ samples are **not i.i.d.**
- Concurrency: the number of labels whose span contains day t

$$c_t = \sum_i 1\{t_{0,i} \leq t \leq t_{1,i}\}$$

- Average uniqueness of label i: the average share of each day of its span that belongs to it alone

$$\bar{u}_i = \frac{1}{t_{1,i} - t_{0,i} + 1} \sum_{t=t_{0,i}}^{t_{1,i}} \frac{1}{c_t}$$

  - $\bar{u}_i = 1$: no shared return; $\bar{u}_i = 1/5$: the harmonic mean of c_t over the span of label i is five

Overlapping Labels and Sample Uniqueness (2/2)

Consequences

- Effective sample size $\approx \sum_i \bar{u}_i \ll n$
  - n: the number of labels
  - an overlap heuristic, not a variance-based measure
- Bagging draws near-duplicates
  - CV scores too optimistic

Remedies

- Weight samples by $\bar{u}_i$ (and by $|r|$)
- Sequential bootstrap
  - in RF, the fraction of observations drawn for each tree $\approx \bar{u}$, the mean uniqueness

Transformations on Training Data Only: the FFD Ablation

- The "minimum d" was chosen on 2000–2026 and then used inside purged CV
  - the test folds helped choose a feature
  - the same ADF rule on data up to 2009 gives $d = 0.15$; on the full sample $d = 0.25$
- Case study without the FFD log-price (purged 5-fold, mean over folds)
  - accuracy: logit 57.5%, RF 56.4%, GBM 55.6%, MLP 56.5%, all below 58.1%
  - AUC: logit 0.483, RF 0.520, GBM 0.514, MLP 0.526
- The negative conclusion does not depend on this feature
- Correct protocol
  - estimate d (local Whittle, with its CI) inside each training fold only
  - prefer features with an economic meaning: distances to moving averages, fractional filters of $|r_t|$

MFM_ch13_inference

Overlapping Labels and the Newey–West Long-Run Variance

- The significance tests of the case study (Pesaran–Timmermann) average a daily series z_t
  - z_t: a hit, a loss differential d_t, a regression score
  - with 5-day labels, neighbouring z_t share returns, so $\mathrm{Var}(\bar{z}) \neq \mathrm{Var}(z_t)/n$
- Newey–West (Newey & West (1987)) heteroskedasticity and autocorrelation consistent (HAC) long-run variance
  - with Bartlett weights:

$$\hat{\Omega}_{NW} = \hat{\Gamma}_0 + 2\sum_{j=1}^{q}\Big(1 - \frac{j}{q+1}\Big)\hat{\Gamma}_j, \qquad \hat{\Gamma}_j = \frac{1}{n}\sum_{t=j+1}^{n}(z_t - \bar{z})(z_{t-j} - \bar{z})$$

  - $\hat{\Gamma}_j$: sample autocovariance of z_t at lag j; q: number of lags ("lag 4" in the tables); $\widehat{\mathrm{SE}}(\bar{z}) = \sqrt{\hat{\Omega}_{NW}/n}$
  - lag rule: h-day labels share returns up to $h-1$ days apart, so $q \geq h-1 = 4$ is the minimum here
- Toy example: z_t = sum of the next 5 i.i.d. daily returns of variance σ^2
  - $\Gamma_j = (5-j)\sigma^2$ for $j \leq 4$, 0 beyond
  - i.i.d. formula: $\Gamma_0 = 5\sigma^2$; true long-run variance $\Gamma_0 + 2(4+3+2+1)\sigma^2 = 25\sigma^2$: an i.i.d. t is $\sqrt{5} \approx 2.2$ times too large
  - Bartlett with $q = 4$: $5 + 2(0.8 \cdot 4 + 0.6 \cdot 3 + 0.4 \cdot 2 + 0.2 \cdot 1) = 17\sigma^2$; the weights damp far lags, so $q = h-1$ is a floor, not a generous choice

Comparing Forecasts: Diebold–Mariano, Clark–West, Giacomini–White (1/2)

- Two probability forecasts of $y_t \in \{0, 1\}$
  - model 1 gives $\hat{p}_{1,t}$
  - the benchmark (model 0, e.g. a constant) gives $\hat{p}_{0,t}$
- Loss of a forecast $\hat{p}_t$
  - the log-loss of the loss-and-penalty slide, with $f = \hat{p}_t$, or the Brier loss $(y_t - \hat{p}_t)^2$
- Loss differential (a different d_t from FFD and DML): $d_t = L(y_t, \hat{p}_{1,t}) - L(y_t, \hat{p}_{0,t})$
  - $d_t > 0$: model 1 worse on day t
- **Diebold–Mariano** (DM, Diebold & Mariano (1995)): is the mean loss differential zero?

$$DM = \frac{\bar{d}}{\sqrt{\hat{\Omega}_{NW}/n}} \xrightarrow{d} N(0,1) \quad \text{under } H_0 : \mathbb{E}[d_t] = 0$$

  - $\bar{d}$: mean of d_t over the n test days; $\hat{\Omega}_{NW}$: Newey–West long-run variance of d_t (Newey–West formula, with $z_t = d_t$)
  - $DM > 1.96$: model 1 significantly worse at 5%; $DM < -1.96$: significantly better

Comparing Forecasts: Diebold–Mariano, Clark–West, Giacomini–White (2/2)

- **Nested models**: the logit with 14 features nests the constant, $\beta = 0$
  - under H_0 the large model estimates zeros with noise, $\mathbb{E}[d_t] > 0$
  - DM therefore rarely favours it
- **Clark–West** (CW, Clark & West (2007)): one-sided HAC t-test of $\mathbb{E}[f_t] = 0$ against $\mathbb{E}[f_t] > 0$

$$f_t = (y_t - \hat{p}_{0,t})^2 - \left[(y_t - \hat{p}_{1,t})^2 - (\hat{p}_{0,t} - \hat{p}_{1,t})^2\right]$$

  - the Brier-loss gain of model 1, plus the term $(\hat{p}_{0,t} - \hat{p}_{1,t})^2$ that removes the noise cost of estimating zero coefficients; a large positive t: the features add information
  - derived for nested linear models estimated by ordinary least squares (OLS); for penalised-logit probabilities an approximation, not a theorem
- **Giacomini–White** (GW, Giacomini & White (2006)): tests the forecasting *method* as deployed
  - with a bounded estimation window
- By contrast, CW asks whether the population model adds information
  - the GW statistic in the Appendix

Feature Importance: MDI and MDA

- The question
  - which features does the model rely on?
  - which of them help prediction on new data?
- Mean decrease impurity (MDI), trees only, computed in sample:

$$\mathrm{MDI}_j = \frac{1}{B}\sum_{b=\mathbf{1}}^{B} \sum_{v \in T_b:\ j(v)=j} \frac{n_v}{n}\Big[G_v - \frac{n_L}{n_v}\,G_L - \frac{n_R}{n_v}\,G_R\Big]$$

  - the sum of the Gini reductions of the nodes v that split on feature j, weighted by the share n_v/n of observations reaching the node
  - favours continuous features with many possible thresholds
- Mean decrease accuracy (MDA), any model, on test data:

$$\mathrm{MDA}_j = S\big(\hat{f};\, \mathcal{D}_{test}\big) - \frac{1}{R}\sum_{r=\mathbf{1}}^{R} S\big(\hat{f};\, \mathcal{D}_{test}^{\pi_r(j)}\big)$$

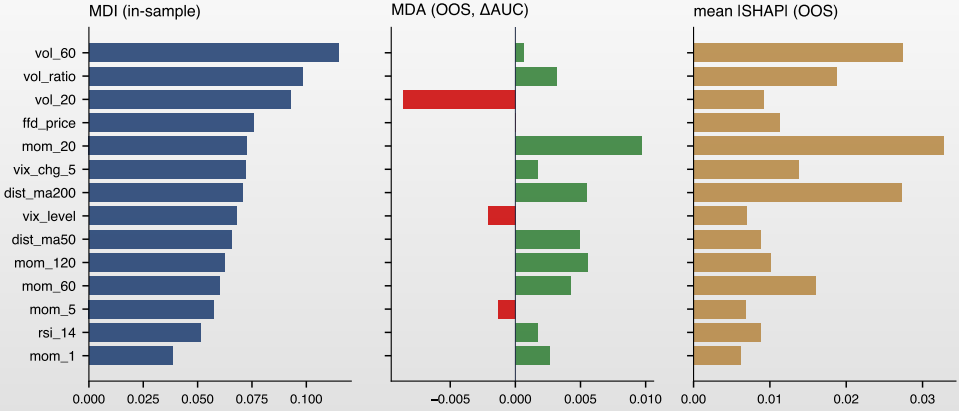
  - S: the score (here AUC); $\mathcal{D}_{test}^{\pi_r(j)}$: the test data with column j randomly permuted, repetition r
  - $\mathrm{MDA}_j < 0$: permuting raised the score; read it with the spread over the R permutations and the estimation uncertainty
  - a small negative MDA alone does not show that a feature hurts

Feature Importance: SHAP Values

- SHAP (Lundberg & Lee (2017)): the Shapley value $\phi_j(\mathbf{x})$
  - the average marginal contribution of feature j to the prediction at $\mathbf{x}$
  - averaged over all coalitions of other features (coalition formula: Appendix)
- Additive: $\hat{f}(\mathbf{x}) = \phi_0 + \sum_j \phi_j(\mathbf{x})$, ϕ_0 the average prediction
- Global importance of feature j: $\frac{1}{n}\sum_i |\phi_j(\mathbf{x}_i)|$, computed on test data
- MDI answers "what did the model use?"
  - SHAP attributes the model's prediction across features and does not use the realised outcome
  - MDA measures the change in score on the evaluated set when a feature is permuted
- Correlated features share and dilute importance (substitution)
  - cluster them first (López de Prado (2020))

Shapley Values

Feature Importance in the Case Study (1/2)



MFM_ch13_feature_importance

Feature Importance in the Case Study (2/2)

- Fit on the first 70%, evaluated on the last 30% after a purge
  - **RF** gives MDI and MDA, **GBM** gives SHAP
  - so the differences are not purely methodological
- Panels: MDI (in sample), MDA and mean $|\phi_j|$ (out of sample)
  - MDA: out-of-sample AUC drop; red bars are negative
- Labels beyond the LASSO path
  - mom_*h*, *h*-day momentum; dist_ma50, dist_ma200, the distance to the moving average
  - rsi_14, the RSI(14); vix_level, vix_chg_5, the VIX level and its 5-day change
- MDI ranks volatility first
  - MDA and SHAP point to 20-day momentum and the distance from the 200-day moving average
- 20-day volatility: high MDI, **negative** MDA
  - a sign of overfitting, not a proof
- All importances are tiny ($\Delta\text{AUC} < 0.01$)
  - how to read them: Appendix
- Correlated features: MDA and SHAP **extrapolate** (Hooker, Mentch & Zhou (2021))
  - permutation creates off-manifold points

MFM_ch13_feature_importance

Walk-Forward: Do the Forecasts Beat a Constant?

- Walk-forward 2010–2026, $n = 4{,}198$ days, 60.2% "up"
  - benchmark: the constant forecast = training share of "up"
  - DM, Clark–West and GW use Newey–West with 4 lags
- Logit (nests the constant): log-loss 0.6824 vs. 0.6756
  - DM $t = 2.50$ ($p = 0.012$): the estimated logit is significantly *worse*
  - Clark–West $t = -0.36$ (one-sided $p = 0.64$): no evidence that the features carry information
  - Giacomini–White $= 6.29$ (χ^2_2, $p = 0.043$): the logit is worse at 5% as well; approximate, since the window expands (rolling-window result in the Appendix)
- GBM: log-loss 0.7027, worse than a constant
  - DM $t = 5.81$, GW $= 33.6$ ($p < 0.001$)
  - Pesaran–Timmermann -2.74
- Conclusion: no model beats "always up" or a constant forecast

MFM_ch13_inference

Annualising the Sharpe Ratio and Comparing Two Strategies

- Annualising with $\sqrt{A}$ needs zero autocorrelation; otherwise $SR_A = \eta(A)\, SR$ (Lo (2002))
  - A: return periods per year (252 trading days; 365 for Bitcoin)
  - $\eta(A) = A/\sqrt{A + 2\sum_{k=1}^{A-1}(A-k)\rho_k}$, ρ_k: autocorrelation of returns at lag k
  - positive autocorrelations, $\rho_k > 0$, make $\eta(A) < \sqrt{A}$
- Example: monthly returns, $A = 12$, with $\rho_1 = 0.2$ and all other autocorrelations zero
  - $\eta = 12/\sqrt{12 + 2 \cdot 11 \cdot 0.2} = 12/\sqrt{16.4} = 2.96$, against $\sqrt{12} = 3.46$
  - so the annual SR is about 15% lower
- Two strategies on the same days: test $SR_a - SR_b$
  - with the HAC delta method or a studentised block bootstrap (Ledoit & Wolf (2008))
  - not with two separate intervals

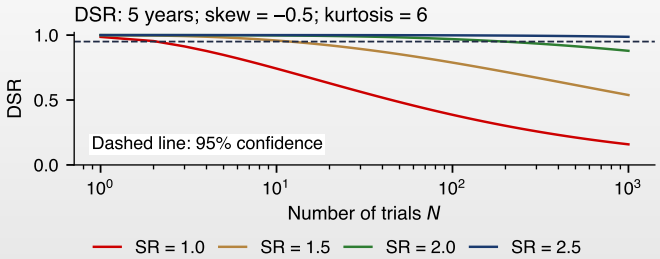
Correlated Trials in the DSR: Alternative Approximations

Two approximations with different assumptions

- (a) raw N, with $\mathrm{Var}[\widehat{SR}_n]$ = dispersion of the trial SRs: row 1 of the Bitcoin DSR table
- (b) $\mathrm{Var}[\widehat{SR}_n] = 1/T$, with an effective number of independent trials $\hat{N} < N$
  - Seminar 13, exercise B6
- Used together, dispersion *and* $\hat{N}$ can count the same dependence effect twice (toy model below)

- Why (a) works, toy model: N zero-skill trials $\widehat{SR}_n = \sigma(\sqrt{\rho}\, Z + \sqrt{1-\rho}\, e_n)$
  - each of variance $\sigma^2 = 1/T$, independent standard Normal $Z, e_1, \ldots, e_N$, pairwise correlation ρ
  - the common shock moves all trials together, so only $\sqrt{1-\rho}\,\sigma\, e_n$ creates selection: $\mathbb{E}[\max_n \widehat{SR}_n] = \sqrt{1-\rho} \times$ the independent-trial value
  - the dispersion across trials is $(1-\rho)\sigma^2$, so (a) captures this scaling
  - the two rows have SD ratio $0.0109/0.0220 = 0.50$, which this toy model reads as $\rho \approx 1 - 0.50^2 \approx 0.75$
- The dispersion also contains true differences between rules
  - with real rules (a) is an approximation
- Recommendation: report the conservative row and either (a) or (b), with N
  - call a rule a discovery only if it also passes the conservative row

Deflated Sharpe Ratio in Practice



- 5 years of daily returns, skewness -0.5, kurtosis 6
  - horizontal axis: number of trials N (log scale), vertical axis: DSR, dashed line: the 0.95 threshold
- $N = 1$: no selection, $SR_0 = 0$, so $DSR = \widehat{PSR}(0)$; for an annualised SR of 1.0, DSR $= 0.986$
  - SR $= 1.0$ stays above 0.95 at $N = 2$ (0.954) and falls below from $N = 3$ (0.913)
- SR $= 1.5$ survives ≈ 10 trials, SR $= 2.0 \approx 200$
  - only SR $= 2.5$ survives 1000

MFM_ch13_backtest_overfitting

Data-Snooping Tests on the Bitcoin Grid (1/3)

- The question
  - does any of the N rules beat the benchmark, once we account for having tried them all?
- $r_{k,t}$: return of rule k on day t; $r_{0,t}$: benchmark return
  - $d_{k,t} = r_{k,t} - r_{0,t}$, with mean $\bar{d}_k$ over the n days
- $H_0 : \max_k \mathbb{E}[d_{k,t}] \leq 0$: no rule beats the benchmark
- Null distribution from a **stationary bootstrap** (Politis & Romano (1994))
  - resample blocks of whole days, of random length
  - $\bar{d}_k^*$: the mean in a bootstrap replication
- The same days for all rules keep the correlation between them
  - no "effective" N needed
- Reality Check (RC, White (2000)): statistic $T^{RC} = \max_k \sqrt{n}\,\bar{d}_k$
  - its p-value is the share of replications with $\max_k \sqrt{n}\,(\bar{d}_k^* - \bar{d}_k) \geq T^{RC}$

Data-Snooping Tests on the Bitcoin Grid (2/3)

- Superior Predictive Ability (SPA, Hansen (2005)): studentised, $T^{SPA} = \max\left(\max_k \sqrt{n}\,\bar{d}_k/\hat{\omega}_k,\ 0\right)$
  - $\hat{\omega}_k$: the bootstrap SD of $\sqrt{n}\,\bar{d}_k$
  - replications are recentred with $\bar{d}_k\,1\{\sqrt{n}\,\bar{d}_k/\hat{\omega}_k \geq -\sqrt{2\log\log n}\}$: clearly bad rules are not recentred, so adding junk rules does not destroy power
- Romano–Wolf StepM (Romano & Wolf (2005)): names *which* rules beat the benchmark
  - with family-wise error rate (FWER) control, the probability of at least one false rejection
  - step 1: reject $H_{0,k}$ for every k with $\sqrt{n}\,\bar{d}_k/\hat{\omega}_k > \hat{c}_1$, where $\hat{c}_1$ is the $1-\alpha$ quantile of $\max_k \sqrt{n}\,(\bar{d}_k^{*} - \bar{d}_k)/\hat{\omega}_k$
  - step j: recompute the quantile on the rules not yet rejected; stop when no further rule is rejected

Data-Snooping Tests on the Bitcoin Grid (3/3)

- Setup: 1,282 rules, in-sample $n = 2{,}057$ days, mean block 20 days, 2,000 replications

Benchmark	largest studentised mean	RC p	SPA p	StepM rules kept
cash ($r_0 = 0$)	4.20	0.001	0.002	117
buy-and-hold	1.36 (the 11/20 rule)	0.60	0.50	0

- Against cash, some trend rule earns a positive mean: Bitcoin's own trend
  - no rule beats holding Bitcoin, even in sample
  - consistent with the OOS SR of 0.03
- DSR needs N and $\mathrm{Var}[\widehat{SR}]$, hence its two verdicts
  - SPA takes both from the data

MFM_ch13_inference

Probability of Backtest Overfitting (PBO)

Combinatorially symmetric cross-validation (CSCV): Bailey, Borwein, López de Prado & Zhu (2017)

1. Matrix of returns: T periods $\times$ N configurations; split the rows into S contiguous blocks
2. For each of the $\binom{S}{S/2}$ choices, half of the blocks are in sample (IS), the rest out of sample (OOS)
3. Pick the best IS configuration n^*
   - rank all N configurations OOS in ascending order
   - set $\bar{\omega} = \text{rank}(n^*)/(N+1) \in (0,1)$
4. Logit $\lambda = \log \frac{\bar{\omega}}{1-\bar{\omega}}$ (a different λ from the penalty); $PBO = \Pr[\lambda \leq 0]$
   - the share of splits in which the IS winner ranks at or below the OOS median

- $\bar{\omega}$: relative OOS rank of the IS winner, near 1 if it stays the best
  - $\lambda > 0$: above the OOS median
- $PBO \approx 0$: the IS winner reliably stays above the OOS median
  - $PBO \approx 0.5$ or more: selecting the IS winner is no better than picking at random
- PBO evaluates the **research procedure**, not one strategy
  - it is not the probability of losing money
  - like DSR it needs **every** trial logged (Seminar 13, exercise B12)

From the Pricing Condition to the Beta Representation

- Three steps (written unconditionally; conditionally, replace E by E_t)
  1. With $M = 1 - F$: $E[(1 - F)R_i^e] = 0$, so $E[R_i^e] = E[R_i^e F]$ for every stock
  2. F is itself an excess return; apply step 1 to $R_i^e = F$: $E[F] = E[F^2]$
  3. Define $\beta_i = E[R_i^e F]/E[F^2]$; then $E[R_i^e] = E[R_i^e F] = \beta_i E[F^2] = \beta_i E[F]$
- This β_i is defined with second moments
  - when step 1 holds it equals the regression beta, since $\mathrm{Cov}(R_i^e, F) = \beta_i E[F^2] - \beta_i E[F]^2 = \beta_i \mathrm{Var}(F)$
  - estimated on data with pricing errors, the two can differ
  - so compute the second-moment form, as the paper's loading network does (Appendix)
- Example: $E[F] = E[F^2] = 0.01$ (1% a month), $E[R_i^e F] = 0.005$
  - $\Rightarrow \beta_i = 0.5$, $E[R_i^e] = 0.5\%$ a month
  - the regression beta gives the same value, $(0.005 - 0.005 \cdot 0.01)/(0.01 - 0.01^2) = 0.5$
- Why the loss changes: any ω that prices all assets must satisfy step 1 for every i
  - step 1 is a first-moment condition
  - this is what the adversarial GMM (Generalized Method of Moments) loss on the next slide enforces

Case Study: Deep Learning in Asset Pricing, Question and Method (1/2)

- Why: forecasting fits a second moment where signal-to-noise is low
  - forecasting: Gu, Kelly & Xiu (2020), the benchmark feedforward network (FFN)
- The no-arbitrage condition targets a first moment
  - averaging over time removes much of the noise
- Method: adversarial generalized method of moments (GMM, eq. 3, after Hansen & Jagannathan (1997))

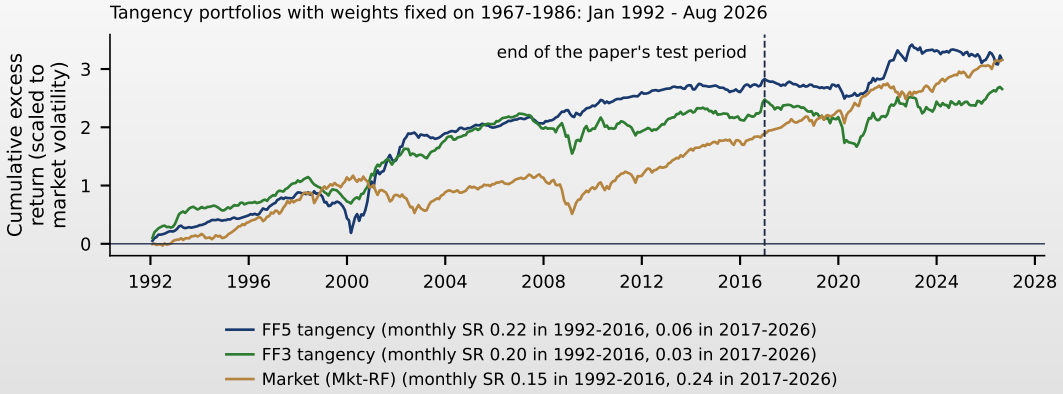
$$\min_{\omega} \max_{g} \frac{1}{N} \sum_{j=1}^{N} \left(E\left[M_{t+1} R^{e}_{t+1,j} \, g(I_t, I_{t,j}) \right] \right)^2$$

  - I_t: macroeconomic information; $I_{t,j}$: characteristics of stock j; N: number of stocks
  - g: instrument function chosen by an adversary to make the pricing errors as large as possible
  - ω minimises this worst case

Case Study: Deep Learning in Asset Pricing, Question and Method (2/2)

- Networks
  - an FFN for the weights ω and an adversary g that builds the hardest test assets
  - a long short-term memory network (LSTM) that turns the macro history into 4 states
  - the full model is labelled GAN (Generative Adversarial Network)
- Chronological split: training 1967–1986, validation 1987–1991, test 1992–2016
- Benchmarks
  - the unregularised (LS) or elastic net (EN) linear SDF
  - the tangency portfolios of the Fama–French three- and five-factor models (FF3, FF5, Fama & French (2015))
- Configurations ranked by the validation Sharpe ratio
  - data and tuning in the Appendix

Deep Learning in Asset Pricing: the Benchmarks after 2016 (1/2)



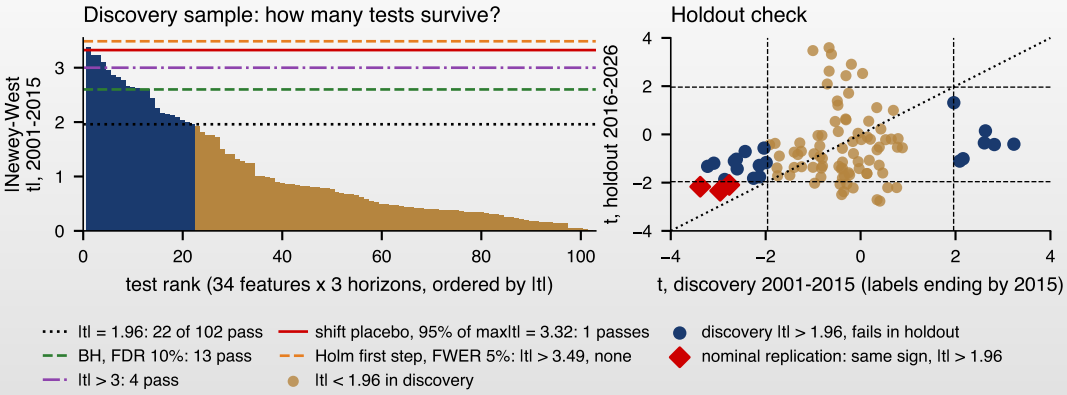
MFM_ch13_sdf_replication

Deep Learning in Asset Pricing: the Benchmarks after 2016 (2/2)

- Same fixed weights, January 1992 – August 2026
  - each series scaled to the volatility of the market excess return on 1992–2016
  - monthly SR 2017–2026 ($T = 116$, SE ≈ 0.09): FF5 0.06, FF3 0.03, market 0.24
  - Mkt-RF in the chart: the market excess return; on 1992–2016: 0.22, 0.20, 0.15
- Replicable with public data: the Fama–French benchmarks
  - not replicable: the stock-level GAN, LS and EN (the panel of the Center for Research in Security Prices, CRSP, is not in the course data)
  - the post-2016 fall is a warning, not yet a test, and the same question applies to a GAN frozen in 1991 (details in the Appendix)

MFM_ch13_sdf_replication

Mini-Case: Thresholds and the Holdout (1/2)



MFM_ch13_ai_discovery

Mini-Case: Thresholds and the Holdout (2/2)

- Left: the 102 discovery $|t|$ in decreasing order
- Right: discovery vs. holdout t
  - the three nominal replications (red) are all short-term reversal; none passes Holm at 5%
  - the significant VIX features collapse towards zero

MFM_ch13_ai_discovery

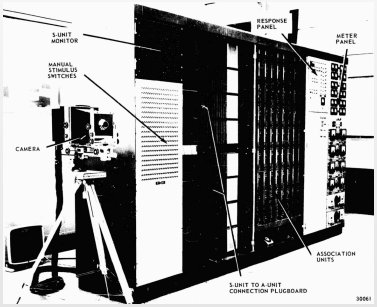
Further Study on Quantinar

Advanced courses on quantinar.com

- Random Forests
- Shapley Values
- Shapley Curves: A Smoothing Perspective
- Machine Learning in Financial Risk
- Introduction to Reinforcement Learning
- Generative Adversarial Networks
- MSCA (Marie Skłodowska-Curie Actions) Digital: The Need for Explainable AI
- Business Analytics and Data Science
- Model Selection Criteria
- PAC (Probably Approximately Correct) Learning

Back

From the Perceptron to Deep Blue



Mark I Perceptron, operator's manual (1960)

Figure: J.C. Hay, A.E. Murray (1960); public domain; Wikimedia Commons



IBM Deep Blue, Computer History Museum

Photo: Anton Chiang (2011); CC BY 2.0; Wikimedia Commons

- **Perceptron** Rosenblatt (1958)
  - a learning rule for a single artificial neuron
  - the Mark I: hardware built to recognise simple images
- 1997: Deep Blue, by IBM (International Business Machines), beats world chess champion Garry Kasparov (Campbell, Hoane & Hsu (2002))
- Deep Blue searched with hand-built rules
  - modern ML (this chapter) *learns* the rules from data

Back

The In-Sample R^2 of a Predictor without Skill

- A monthly R^2 of 0.5% is economically large
  - but sampling noise is of the same order
- In-sample OLS R^2 of one predictor with no skill: $nR^2 \approx \chi^2_1$
  - intercept, independent errors, $n \approx 300$ months
  - n: number of months; χ^2_1: chi-squared distribution with 1 degree of freedom, mean 1 and variance 2
  - so R^2 has mean $1/n \approx 0.3\%$ and standard deviation $\sqrt{2}/n \approx 0.5\%$
- R^2_{OOS} can be negative and has no such null distribution
  - it needs its own test (Clark–West)

Back

Worked Example: How a Forest Averages Its Trees

- Random forest of the case study (400 trees, minimum leaf size 100)
  - trained on the first four purged folds
- Prediction for the last observation (11 September 2026)
  - P(up in 5 days) from each tree
  - individual trees range from 0.31 to 0.88 (standard deviation 0.11)
  - 81.5% of the trees vote "up" ($p > 0.5$)
  - forest probability = average of the 400 trees = 0.608
- Single trees disagree strongly
  - the average is far more stable
- Stability is not accuracy
  - the forest still does not beat "always up" (case study)

Back

Sampling Financial Data: Bars

Time and tick bars

- A **bar** = one row summarising a sampling interval
  - open, high, low, close, volume
- Time bars: every fixed interval (1 minute, 1 day)
  - oversample quiet periods
  - fat-tailed returns
- Tick bars: one bar every N transactions

Volume, dollar and information bars

- One bar every V shares or every D dollars traded
  - sampling by **information arrival**
- Returns closer to i.i.d. draws from the Normal distribution
  - Clark (1973); Ané & Geman (2000)
- Imbalance / runs bars
  - sample when order flow becomes one-sided

- This chapter uses daily time bars
  - tick data (crypto exchanges) in Chapters 10 and 16

Back

Fractional Weights and Their Asymptotics

Definition 6 (Fractional difference operator)

- For a real order d:

$$(1-B)^d = \sum_{k=\mathbf{0}}^{\infty} \binom{d}{k} (-B)^k, \qquad \tilde{X}_t = \sum_{k=\mathbf{0}}^{\infty} w_k X_{t-k}$$

- Weights: $w_{\mathbf{0}} = 1$, $w_k = -w_{k-\mathbf{1}} \frac{d-k+\mathbf{1}}{k}$
  - for $d \notin \{0, 1, 2, \dots\}$: $w_k = \frac{\Gamma(k-d)}{\Gamma(-d)\,\Gamma(k+\mathbf{1})}$ and $w_k \sim \frac{k^{-(\mathbf{1}+d)}}{\Gamma(-d)}$ as $k \to \infty$

- $\binom{d}{k} = d(d-1)\cdots(d-k+1)/k!$: the binomial coefficient, defined for any real d
  - Γ: the gamma function, $\Gamma(k+1) = k!$
- The weights decay like $k^{-(\mathbf{1}+d)}$, more slowly for smaller d
- For $0 < d < 1$: $\sum_{k=\mathbf{0}}^{\infty} w_k = (1-1)^d = 0$
  - it follows from the binomial series at $B = 1$, which converges for $d > 0$
  - $d = 0$: $w_{\mathbf{0}} = 1$ and the rest zero, so the sum is 1; $d = 1$: $w = (1, -1)$, sum 0
  - for $d < 0$ the series of weights does not converge

Back

Local Whittle and Exact Local Whittle (1/2)

- Periodogram: the sample estimate of the spectral density
  - at the Fourier frequency $\lambda_j = 2\pi j/n$

$$I(\lambda_j) = \frac{1}{2\pi n}\Big|\sum_{t=1}^{n} X_t\, e^{-\mathrm{i}\lambda_j t}\Big|^2$$

- Local Whittle (Robinson (1995)): choose δ to fit $f(\lambda) \approx G\lambda^{-2\delta}$
  - on the first m frequencies

$$\hat{\delta} = \arg\min_{\delta}\ \log\Big(\frac{1}{m}\sum_{j=1}^{m}\lambda_j^{2\delta} I(\lambda_j)\Big) - \frac{2\delta}{m}\sum_{j=1}^{m}\log\lambda_j, \qquad \sqrt{m}\,(\hat{\delta}-\delta) \xrightarrow{d} N\big(0, \tfrac{1}{4}\big)$$

  - n: sample size; m: number of frequencies used; i: the imaginary unit; G has been profiled out
  - $\xrightarrow{d}$: convergence in distribution; hence the standard error $1/(2\sqrt{m})$

Back

Local Whittle and Exact Local Whittle (2/2)

- Exact local Whittle (ELW, Shimotsu & Phillips (2005); unknown mean: Shimotsu (2010))
  - periodogram of $(1-B)^{\delta} X_t$, valid also for non-stationary $\delta \geq \frac{1}{2}$
- Why the FFD series looks like $\hat{\delta} \approx 0.75$
  - above frequency $1/l^*$ the truncated filter acts as $(1-B)^{0.25}$
  - only $n/(2\pi l^*) \approx 2$ of the $m = 294$ frequencies lie below $1/l^*$, where the unit root returns
  - l^*: the FFD window length (444 lags)
  - the estimate is the apparent memory over the band, not the integration order of the series

MFM_ch13_ffd

Back

Worked Example: Purging with 10 Observations

- Observations $t = 0, \ldots, 9$
  - label of t ends at $t + 2$ (capped at 9)
- 5 contiguous folds of 2; embargo = 1 observation

Test	Removed	Train
0, 1	2, 3, 4	5–9
2, 3	0, 1, 4, 5, 6	7–9
4, 5	2, 3, 6, 7, 8	0, 1, 9
6, 7	4, 5, 8, 9	0–3
8, 9	6, 7	0–5

- Test $\{4, 5\}$
  - labels of 2 and 3 end at 4 and 5 $\Rightarrow$ purged
  - 6 and 7 start on or before day 7, when the last test label ends $\Rightarrow$ purged
  - the embargo removes 8, the first observation after day 7
- With overlapping labels and the embargo, an interior fold loses $2h + e = 5$ of 10 observations

Back

When Is K-Fold Valid for Time Series?

- Bergmeir, Hyndman & Koo (2018): standard K-fold CV is valid for a **purely autoregressive** model
  - $y_t = g(y_{t-1}, \ldots, y_{t-p}) + \varepsilon_t$ with serially uncorrelated errors
  - g: any, possibly non-linear, function of the last p values; ε_t: error
  - the errors in the test fold are uncorrelated with the training errors, so the CV error is unbiased for the OOS error
- It fails when the **errors** are serially correlated
  - overlapping labels: h-day forward returns share $h-1$ daily returns
  - persistent features that are not in the model's lag set (rolling means, levels such as the FFD log-price)
  - under-specified dynamics (too few lags)
- The leakage experiment isolates exactly this
  - 20-day labels and 20-day rolling features
- Check before choosing: the autocorrelation of the CV residuals
  - if it is significant, purge and embargo

Back

Walk-Forward and Combinatorial Purged CV

Walk-forward (WF)

- Mimics real-time deployment
  - no future data in training
- But a **single path**
  - high variance, easy to overfit by repeated trials

Combinatorial purged CV (CPCV)

- N contiguous groups
  - test on all $\binom{N}{k}$ combinations of k groups, with purging and embargo
- $\varphi = \frac{k}{N}\binom{N}{k}$: the number of complete backtest paths
  - e.g. $N = 6$, $k = 2$: 15 splits, 5 paths
- A **distribution** of Sharpe ratios instead of one number

- Each group is tested φ times, so the test predictions form φ full histories

Back

Testing Directional Accuracy: Pesaran–Timmermann

- H_0: the predicted sign x_t is independent of the realised sign y_t (Pesaran & Timmermann (1992))
  - $\hat{P}$, $\hat{P}^*$, $\hat{p}_y$, $\hat{p}_x$ as on the main slide, n: number of forecasts

$$PT = \frac{\hat{P} - \hat{P}^*}{\sqrt{\widehat{V}(\hat{P}) - \widehat{V}(\hat{P}^*)}} \xrightarrow{d} N(0,1), \qquad \widehat{V}(\hat{P}) = \frac{\hat{P}^*(1-\hat{P}^*)}{n}$$

  - $\widehat{V}(\hat{P}^*) = \left[(2\hat{p}_y - 1)^2 \hat{p}_x(1-\hat{p}_x) + (2\hat{p}_x - 1)^2 \hat{p}_y(1-\hat{p}_y)\right]/n + 4\hat{p}_y\hat{p}_x(1-\hat{p}_y)(1-\hat{p}_x)/n^2$
  - $\widehat{V}(\cdot)$: estimated variance under H_0; $\hat{P}^*$ is itself estimated, hence the second term
- PT conditions on the marginal frequencies automatically
  - no need to guess the baseline
- Logit of the case study: $\hat{p}_y = 0.581$, $\hat{p}_x = 0.983$
  - one-sided $p = 0.79$ for $PT = -0.80$

MFM_ch13_inference

Back

ROC AUC and Its Sampling Uncertainty

- Receiver operating characteristic (ROC) curve, over all thresholds
  - true positive rate $TP/(TP+FN)$ against false positive rate $FP/(FP+TN)$
- Area under the curve (AUC)
  - $\text{AUC} = P(\hat{p}_{\mathbf{pos}} > \hat{p}_{\mathbf{neg}}) + \frac{1}{2}P(\hat{p}_{\mathbf{pos}} = \hat{p}_{\mathbf{neg}}) = U/(n_{\mathbf{1}}n_{\mathbf{0}})$
  - $\hat{p}_{\mathbf{pos}}$, $\hat{p}_{\mathbf{neg}}$: predicted probabilities of a randomly drawn positive (up) and negative (down) observation
  - $n_{\mathbf{1}}$, $n_{\mathbf{0}}$: numbers of positives and negatives; U: Mann–Whitney statistic, the number of (positive, negative) pairs ranked correctly, ties counted $\frac{1}{2}$
- DeLong standard error (DeLong, DeLong & Clarke-Pearson (1988)): from the placement values
  - placement value: the share of the other class that each observation outranks
  - assumes independent observations
- Logit, pooled out-of-fold AUC $= 0.480$
  - DeLong 95% CI [0.466; 0.494]: "significantly worse than random"
  - block bootstrap 95% CI [0.456; 0.505]: noise around 0.5
- An AUC below 0.5 is anti-ranking only if it survives a test that respects the dependence

MFM_ch13_inference

Back

Giacomini–White, Statistic and Rolling Window

- Giacomini & White (2006): the null concerns the forecasting **method**, not the population model
  - method: model + estimation window + refitting rule

$$H_0 : \mathbb{E}\big[d_{t+\tau} \mid \mathcal{F}_t\big] = 0 \;\Rightarrow\; \mathbb{E}\big[\mathrm{h}_t\, d_{t+\tau}\big] = 0, \qquad GW = n\,\bar{\mathrm{Z}}^{\top}\hat{\Omega}^{-1}\bar{\mathrm{Z}} \xrightarrow{d} \chi^2_q$$

- Notation
  - $d_{t+\tau}$: loss differential of the two forecasts made at t for horizon $\tau = 5$ days (here τ is the horizon, not a threshold); $\mathcal{F}_t$: information available at t
  - h_t: test functions known at t (here 1 and the lagged loss differential, $q = 2$); $\mathrm{Z}_t = \mathrm{h}_t d_{t+\tau}$
  - $\bar{\mathrm{Z}}$: mean of Z_t over the n forecasts; $\hat{\Omega}$: HAC covariance matrix of Z_t
- Reading: under H_0 no information at t predicts which method will do better
- The theory requires a bounded (rolling) estimation window
  - so that estimation noise never vanishes
  - an expanding window lies outside the theorem
- S&P 500 logit, rolling 8-year window: GW $= 26.7$ ($p < 0.001$)
  - the logit forecasts worse than the constant

MFM_ch13_inference

Back

Shapley Values: the Coalition Formula

Definition 7 (Shapley value of feature j (SHAP, Lundberg & Lee (2017)))

- Average marginal contribution of j over all coalitions $S \subseteq \{1, \ldots, p\} \setminus \{j\}$:

$$\phi_j = \sum_S \frac{|S|!\,(p - |S| - 1)!}{p!} \big[v(S \cup \{j\}) - v(S)\big], \qquad \hat{f}(\mathbf{x}) = \phi_{\mathbf{0}} + \sum_j \phi_j$$

- Interventional SHAP: $v(S) = \mathbb{E}[\hat{f}(\mathbf{x}_S, \mathbf{X}_{\bar{S}})]$
  - averaging $\hat{f}$ with the features outside S drawn from a background sample
  - $\phi_{\mathbf{0}} = v(\emptyset)$ = average prediction
- Notation
  - p: number of features; S: a coalition, a subset of the other features, $|S|$ its size, $\bar{S}$ the features outside it
  - $v(S)$: prediction when only the features in S are known; $\hat{f}(\mathbf{x})$: model prediction for $\mathbf{x}$
- The weight of S is a probability
  - the probability that exactly the features in S come before j in a random ordering of the p features
- Conditional SHAP uses $v(S) = \mathbb{E}[\hat{f}(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S]$
  - the two differ when features are dependent

Shapley Values

Back

Reading Feature Importances

- Remedies for correlated features
  - permute a cluster's columns together, with the same row permutation, or refit without the feature
  - conditional and time-block permutations are separate variants
- Importances are estimates and need their own uncertainty analysis
  - honest trees give asymptotically Normal forest *predictions* and CIs (Wager & Athey (2018))
  - they do not give CIs for importances
- **Backtesting is not a research tool**, importance analysis is: López de Prado (2018)

Back

Double Selection and DML, Penalty and Conditions (1/2)

- Partially linear model: $\mathbb{E}[\varepsilon_t \mid d_t, \mathrm{x}_t] = 0$, $\mathbb{E}[v_t \mid \mathrm{x}_t] = 0$, $\mathrm{Var}(v_t) > 0$
  - g, m approximately sparse
  - the errors are unpredictable from the regressors; d_t has variation of its own beyond the controls
- Post-double selection: the heteroskedasticity-robust standard error is valid (Belloni, Chernozhukov & Hansen (2014))
  - under independent sampling, regularity conditions and sparsity $s^2 \log^2(p \vee n)/n \to 0$
  - s: number of controls that matter; p: number of controls; $p \vee n = \max(p, n)$
- LASSO objective, with feature-specific penalty loadings

$$\frac{1}{n}\sum_t (y_t - \mathrm{x}_t^\top \beta)^2 + \frac{\lambda}{n}\sum_j \hat{\psi}_j |\beta_j|, \qquad \lambda = 2c\sqrt{n}\,\Phi^{-1}(1 - \gamma/(2p))$$

  - plug-in constants $c = 1.1$, $\gamma = 0.1/\log n$; loadings $\hat{\psi}_j^2 = \frac{1}{n}\sum_t x_{tj}^2 \hat{e}_t^2$, $\hat{e}_t$ the residuals of the previous step, iterated

Back

Double Selection and DML, Penalty and Conditions (2/2)

- DML (Chernozhukov et al. (2018)): with $\ell(x) = \mathbb{E}[y \mid x]$, it suffices that
  - with consistency and regularity:

$$\|\hat{m} - m\| \left(\|\hat{\ell} - \ell\| + \|\hat{m} - m\| \right) = o_P(n^{-1/2})$$

  - $\|\hat{m} - m\|$: root mean squared estimation error of the learner for m; same for ℓ
  - $o_P(n^{-1/2})$: tends to zero in probability faster than $1/\sqrt{n}$
  - e.g. both errors $o_P(n^{-1/4})$: each learner may converge more slowly than a parametric model, because only the product of the errors matters

Back

Machine Learning for Inference in Asset Pricing

- Feng, Giglio & Xiu (2020): does a new factor add pricing information given hundreds of existing ones?
  - double selection of the controlling factors from the zoo, then a valid t-test of the new factor's risk price
  - the inferential answer to Harvey, Liu & Zhu (2016), who raise the hurdle to $t > 3$
- Heterogeneous effects: honest causal forests (Wager & Athey (2018))
  - they give CIs for effects that vary with covariates
- Three rules
  - prediction and inference are different targets: a good predictor can give a biased $\hat{\theta}$
  - orthogonalise before you test, and cross-fit with folds that respect time
  - report CIs; "selected" or "important" is not "significant"

Back

Deep Learning for Sequences

Recurrent networks (LSTM; GRU, gated recurrent unit)

- Hidden state carries memory: $h_t = g(h_{t-1}, x_t)$
  - g: a non-linear function with trained weights, x_t: input at t
- Gates control what is kept or forgotten
  - mitigating vanishing gradients
- Used for volatility, limit-order-book and intraday forecasting

Transformers

- Self-attention: $\text{softmax}\left(QK^\top/\sqrt{d_k}\right)V$
  - query, key, value matrices Q, K, V (linear maps of the inputs)
  - d_k: key dimension
  - softmax turns each row of scores into weights that sum to 1
- Long-range dependence, parallel training
- Backbone of large language models (Chapter 15)

Back

Time-Series Foundation Models

Pre-trained forecasters

- Pre-trained on large collections of series, used **zero-shot** (without re-training)
  - Chronos (Ansari et al., 2024), TimesFM (Das et al., 2024), Moirai (Woo et al., 2024), Lag-Llama (Rasul et al., 2023)
- Useful baselines for volatility and volume forecasting
- For returns they inherit the same signal-to-noise limits
- Pre-training data may overlap your test period
  - a new form of leakage

- Covered in depth in Chapter 14
  - with the risk-forecast evaluation tools of Chapter 8

Back

Reinforcement Learning, Generative Models, Explainable AI, Risk

Reinforcement learning (RL)

- A policy $\pi(a \mid s)$ chooses actions a in states s to maximise cumulative reward
- Optimal execution, market making, deep hedging: Buehler et al. (2019)

Explainable AI

- Required by regulators (EU AI Act, credit scoring, model risk)
- SHAP, LIME (local interpretable model-agnostic explanations), partial dependence

Generative models

- GANs, variational autoencoders (VAEs), diffusion models
  - synthetic scenarios for stress tests

ML for risk

- Value-at-Risk (VaR): conditional quantiles (quantile forests, quantile-loss boosting)
- Expected Shortfall (ES): average of fitted tail quantiles or a joint VaR–ES loss
- Volatility: a squared-return target (Chapters 5 and 9)

- **Quantinar:** Reinforcement Learning · GANs · Explainable AI · ML in Financial Risk

Back

Deep Learning in Asset Pricing, Data (Section III.A)

- Monthly returns of all CRSP stocks, January 1967 – December 2016
  - excess over the one-month Treasury-bill rate
- 46 firm characteristics in six groups (Table A.II)
  - past returns, investment, profitability, intangibles, value, trading frictions
  - Only stock-months with all 46 available: about 10,000 of 31,000 stocks
  - ranked each month into cross-sectional quantiles
- 178 macroeconomic series, transformed to stationarity
  - 124 of FRED-MD (Federal Reserve Economic Data, Monthly Database, McCracken–Ng), the St. Louis Fed's monthly macroeconomic database (McCracken & Ng (2016))
  - the 46 monthly medians of the characteristics, 8 predictors of Welch & Goyal (2008)

Back

Deep Learning in Asset Pricing, Method (Sections I–II, Appendix C) (1/2)

- Objective (eq. 3) and networks: main slide
  - the adversary g uses 8 instruments
  - three steps: unconditional SDF, maximise over g, minimise again (Section II.D)
- Tuning: 384 configurations ranked by the validation Sharpe ratio (SR)
  - chosen: 2 layers of 64 units, dropout, Adam; ensemble of 9 fits (Table A.VIII)
  - loadings: a second FFN fits $E_t[R^e_{t+\mathbf{1},i} F_{t+\mathbf{1}}] \propto \beta_{t,i}$
- The linear LS/EN benchmark: SDF $\omega_{t,i} = \theta^\top I_{t,i}$
  - on long and short characteristic legs
  - $\omega_{t,i}$: weight of stock i; $I_{t,i}$: its characteristics; θ: one coefficient per characteristic, the same for all stocks and months

Back

Deep Learning in Asset Pricing, Method (Sections I–II, Appendix C) (2/2)

- Metrics (II.F): the SR of F and two fit measures
  - $\hat{\epsilon}_{t+1,i}$: pricing error, the part of $R^e_{t+1,i}$ not explained by $\hat{\beta}_{t,i}F_{t+1}$
- Explained variation (EV)
  - one minus the average squared pricing error over the average squared excess return

$$\text{EV} = 1 - \frac{\overline{\hat{\epsilon}^2}}{\overline{(R^e)^2}}$$

  - overline: time average of the monthly cross-sectional means; EV $= 1$: the factor explains all variation
- Cross-sectional R^2: how well the model explains average returns

$$\text{XS-}R^2 = 1 - \frac{\frac{1}{N}\sum_i \frac{T_i}{T}\left(\bar{\hat{\epsilon}}_i\right)^2}{\frac{1}{N}\sum_i \frac{T_i}{T}\left(\bar{R}^e_i\right)^2}$$

  - $\bar{\hat{\epsilon}}_i$, $\bar{R}^e_i$: time averages for stock i; T_i: months of data on stock i, T: months in the sample

Back

Deep Learning in Asset Pricing, Results (1/2)

- Source: Chen, Pelger & Zhu (2024), *Management Science*
- Out-of-sample SR of the SDF (Table I), annualised: GAN 2.6, EN 1.7, FFN 1.5
  - monthly values on the main slide
- Individual stocks (Table I)
  - EV 8% (GAN) vs. 3–4%; cross-sectional R^2 23% vs. 14–19%
  - Reading: the no-arbitrage objective improves both fits, although it maximises neither
- Deciles sorted on the loadings β (Table II, test period, annualised)
  - top minus bottom 0.39; FF5 alpha 0.39, $t = 9.96$
  - the GRS test rejects the CAPM (capital asset pricing model), FF3 and FF5 ($p = 0.00$)
  - GRS (Gibbons–Ross–Shanken) test: H_0 all alphas of the decile portfolios are zero
  - Reading: loadings predict returns, linearly ($R^2 = 0.97$ of mean return on β, Figure 8), and the factor models do not price them

Back

Deep Learning in Asset Pricing, Results (2/2)

- 46 sets of characteristic-sorted deciles (Table III)
  - GAN cross-sectional R^2 above 90% for all
  - EV short-term reversal 0.70 vs. 0.43 (EN), momentum 0.54 vs. 0.26
  - Reading: gains come from extreme deciles and interactions
- Macro states (Figure 6): the 178 raw series make the out-of-sample SR collapse
  - GAN loses about 10% without macro states and 20% with unconditional moments only
  - Reading: the dynamics of macro data must be modelled (LSTM), not fed as the last increment
- Largest stocks (Table IV, monthly SR)
  - models fitted on all stocks, evaluated on the subset
  - about 1,500 stocks GAN 0.41, FFN 0.30, EN 0.15
  - about 550 stocks 0.26, 0.24, 0.23
  - re-estimated on the 1,500: 0.42, 0.28, 0.42
  - Reading: the edge narrows sharply among the largest stocks; EN catches up when re-estimated on them

Back

Deep Learning in Asset Pricing, Replication and Extension

- With the Kenneth French Data Library and FRED-MD
  - on the published split
  - Replicable: FF3 and FF5 tangency weights estimated on 1967–1986 (footnote 23), monthly SR on 1992–2016 next to Table A.VI
  - Not replicable with the course data: the published LS and EN (Tables I and III) use rank-weighted legs of 46 characteristics and stock-level betas
  - Analogue on portfolio data, an extension rather than a replication
  - the linear SDF ($\hat{\theta}_{LS}$ in closed form, $\hat{\theta}_{EN}$ with the objective of Section II.F) on halves (deciles 1–5, 6–10) of the 15 characteristics that the library sorts into deciles
  - its levels are not comparable with Tables I and III

Back

Deep Learning in Asset Pricing, Limits and an Open Extension

- Limitations stated by the authors
  - The positivity of the SDF is not imposed (footnote 9)
  - only stock-months with all characteristics, which mainly drops very small stocks (Section III.A)
  - Frictions (Section III.I, Figure 17): annual SR 1.73 without the smallest 40% of stocks
  - 2.07 without the widest 40% bid–ask spreads
  - 1.87 without the lowest 40% turnover; see also Avramov, Cheng & Metzker (2023)
  - In-sample metrics exceed the out-of-sample ones, a sign of some overfitting (Section III.C)
  - the hidden states are not comparable across fits, so they cannot be averaged (Section II.E)
- Open extension: freeze the 1967–1991 design
  - evaluate 2017–2026, after the paper's sample
  - does the linear no-arbitrage SDF still beat FF5?
  - do LSTM states of FRED-MD add to it?

Back

Deep Learning in Asset Pricing, Reading the Two Charts

- GAN 0.75 vs. FF5 0.22
  - with Lo SEs 0.065 and 0.058, the SE of the difference is at most 0.12 whatever their correlation
  - the gap exceeds 4 SEs (i.i.d. Normal returns)
  - a paired test on aligned returns (Ledoit & Wolf (2008)) is sharper
- The fixed FF5 weights load mainly on profitability and investment
  - RMW 0.42, CMA 0.30 (weights scaled to $\sum |w| = 1$)
  - FF3 puts 0.61 on HML
  - both combinations were chosen on 1967–1986

MFM_ch13_sdf_replication

Back

The Discovery Loop, Step by Step (1/2)

- **Literature**
  - `"Summarise the evidence on how much published return predictability survives out of sample and after publication: DOI or arXiv id, sample, headline decay estimate."`
  - Resolve every identifier; Semantic Scholar, Google Scholar, Elicit for the search
- **Hypotheses**
  - `"For a feature that predicts S&P 500 returns in 2001-2015, give competing explanations (risk premium, reversal, microstructure, data mining) and what each predicts for 2016-2026 and for other indices."`
- **Data and code**
  - `"Write Python that builds 34 technical features from the course's daily S&P 500 and VIX closes with information up to day t only, and reports Newey-West t-statistics for 1-, 5- and 20-day returns in 2001-2015."`
  - Any assistant (Claude, ChatGPT, Gemini, GitHub Copilot)

Back

The Discovery Loop, Step by Step (2/2)

- **Robustness**, fixed in advance
  - `"Write the pre-analysis plan: the full list of features and horizons, the holdout, the corrections (Holm, BH, max-t) and the result that would count as a failure."`
  - Placebo: shifting the target circularly (250–3,000 days) breaks its alignment with the features but keeps each series' own dependence
  - a stationary target does not give the joint null of all zero slopes: the shift also changes the features' link with conditional volatility
  - hence a heuristic yardstick; an alternative: a joint bootstrap of the scores, centred under the null, with explicit assumptions
- **Critique**
  - `"Act as a referee: which of these significant features are one effect counted several times, and which are known anomalies rather than new ones?"`
- **Report**: all M tests, all trials including the discarded ones
  - DSR and PBO for any strategy built on them

Back

Acronyms used in this material (1/3)

- **ADF**: Augmented Dickey–Fuller (test)
- **AI**: Artificial Intelligence
- **AIC**: Akaike Information Criterion
- **AUC**: Area Under the (ROC) Curve
- **BET**: Bucharest Exchange Trading (index)
- **BH**: Benjamini–Hochberg (procedure)
- **C-ENet**: Combination Elastic Net
- **CAPM**: Capital Asset Pricing Model
- **CBOE**: Chicago Board Options Exchange
- **CF**: Cash Flow (in CF/P: cash flow to price)
- **CI**: Confidence Interval
- **CMA**: Conservative Minus Aggressive (investment factor)
- **CPCV**: Combinatorial Purged Cross-Validation
- **CRSP**: Center for Research in Security Prices
- **CSCV**: Combinatorially Symmetric Cross-Validation
- **CV**: Cross-Validation
- **CW**: Clark–West (test for nested models)
- **DM**: Diebold–Mariano (test)
- **DML**: Double/Debiased Machine Learning
- **DSR**: Deflated Sharpe Ratio
- **ELW**: Exact Local Whittle (estimator)
- **ENet**: Elastic Net
- **ES**: Expected Shortfall
- **EU**: European Union
- **EV**: Explained Variation
- **EWMA**: Exponentially Weighted Moving Average
- **FDR**: False Discovery Rate
- **FF3**: Fama–French three-factor model
- **FF5**: Fama–French five-factor model
- **FFD**: Fixed-width window Fractional Differentiation
- **FFN**: FeedForward Network
- **FN**: False Negative

Acronyms used in this material (2/3)

- **FP**: False Positive
- **FRED-MD**: Federal Reserve Economic Data, Monthly Database (McCracken–Ng)
- **FWER**: Family-Wise Error Rate
- **GAN**: Generative Adversarial Network
- **GB**: Gradient Boosting (the seminar's name for GBM)
- **GBM**: Gradient Boosting Machine (here not geometric Brownian motion)
- **GMM**: Generalized Method of Moments
- **GRS**: Gibbons–Ross–Shanken (test)
- **GRU**: Gated Recurrent Unit
- **GW**: Giacomini–White (test)
- **HAC**: Heteroskedasticity and Autocorrelation Consistent (standard errors)
- **HARKing**: Hypothesising After the Results are Known
- **HML**: High Minus Low (value factor)
- **IBM**: International Business Machines (Corporation)
- **IS**: In-Sample
- **LASSO**: Least Absolute Shrinkage and Selection Operator
- **LIME**: Local Interpretable Model-agnostic Explanations
- **LLM**: Large Language Model
- **LS**: Linear SDF estimated by unregularised Least Squares
- **LSTM**: Long Short-Term Memory (network)
- **LW**: Local Whittle (estimator)
- **MA**: Moving Average
- **MA20**: 20-day Moving Average
- **MDA**: Mean Decrease Accuracy
- **MDI**: Mean Decrease Impurity
- **ML**: Machine Learning
- **MLP**: MultiLayer Perceptron
- **MSCA**: Marie Skłodowska-Curie Actions

Acronyms used in this material (3/3)

- **OLS**: Ordinary Least Squares
- **OOS**: Out-Of-Sample
- **PAC**: Probably Approximately Correct (learning)
- **PBO**: Probability of Backtest Overfitting
- **PCR**: Principal Component Regression
- **PLS**: Partial Least Squares
- **PSR**: Probabilistic Sharpe Ratio
- **PT**: Pesaran–Timmermann (test of directional accuracy)
- **RC**: Reality Check (White's test for data snooping)
- **ReLU**: Rectified Linear Unit, $\max(0, z)$
- **RF**: Random Forest
- **RL**: Reinforcement Learning
- **RMW**: Robust Minus Weak (profitability factor)
- **ROC**: Receiver Operating Characteristic (curve)
- **RSI**: Relative Strength Index
- **SD**: Standard Deviation
- **SDF**: Stochastic Discount Factor
- **SE**: Standard Error
- **SHAP**: SHapley Additive exPlanations
- **SMB**: Small Minus Big (size factor)
- **SPA**: Superior Predictive Ability (test)
- **SR**: Sharpe Ratio
- **TN**: True Negative
- **TP**: True Positive
- **US**: United States
- **VAE**: Variational AutoEncoder
- **VaR**: Value-at-Risk
- **VIX**: CBOE Volatility IndeX
- **WF**: Walk-Forward (validation)

References (1/9)

- Ané, T., Geman, H. (2000). Order flow, transaction clock, and normality of asset returns. *Journal of Finance*, 55(5), 2259–2284.
- Ansari, A.F. et al. (2024). Chronos: learning the language of time series. *arXiv:2403.07815*.
- Avramov, D., Cheng, S., Metzker, L. (2023). Machine learning vs. economic restrictions: evidence from stock return predictability. *Management Science*, 69(5), 2587–2619.
- Bailey, D.H., Borwein, J., López de Prado, M., Zhu, Q.J. (2017). The probability of backtest overfitting. *Journal of Computational Finance*, 20(4), 39–69.
- Bailey, D.H., López de Prado, M. (2012). The Sharpe ratio efficient frontier. *Journal of Risk*, 15(2), 3–44.
- Bailey, D.H., López de Prado, M. (2014). The deflated Sharpe ratio: correcting for selection bias, backtest overfitting and non-normality. *Journal of Portfolio Management*, 40(5), 94–107.
- Będowska-Sójka, B., Wójcik, P., Pele, D.T. (2026). Early warning systems for cryptocurrency markets: predicting 'zombie' assets using machine learning. *North American Journal of Economics and Finance*, 81, 102543.

References (2/9)

- Belloni, A., Chernozhukov, V., Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2), 608–650.
- Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1), 289–300.
- Bergmeir, C., Hyndman, R.J., Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120, 70–83.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Buehler, H., Gonon, L., Teichmann, J., Wood, B. (2019). Deep hedging. *Quantitative Finance*, 19(8), 1271–1291.
- Campbell, J.Y., Thompson, S.B. (2008). Predicting excess stock returns out of sample: can anything beat the historical average? *Review of Financial Studies*, 21(4), 1509–1531.
- Campbell, M., Hoane, A.J., Hsu, F. (2002). Deep Blue. *Artificial Intelligence*, 134(1–2), 57–83.

References (3/9)

- Chen, A.Y., Lopez-Lira, A., Zimmermann, T. (2022). Does peer-reviewed research help predict stock returns? *arXiv:2212.10317*.
- Chen, L., Pelger, M., Zhu, J. (2024). Deep learning in asset pricing. *Management Science*, 70(2), 714–750.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1), C1–C68.
- Clark, P.K. (1973). A subordinated stochastic process model with finite variance for speculative prices. *Econometrica*, 41(1), 135–155.
- Clark, T.E., West, K.D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1), 291–311.
- Das, A., Kong, W., Sen, R., Zhou, Y. (2024). A decoder-only foundation model for time-series forecasting. *Proceedings of the 41st International Conference on Machine Learning (ICML)*, PMLR 235.
- DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3), 837–845.

References (4/9)

- Diebold, F.X., Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.
- Dixon, M.F., Halperin, I., Bilokon, P. (2020). *Machine Learning in Finance: From Theory to Practice*. Springer.
- Dong, X., Li, Y., Rapach, D.E., Zhou, G. (2022). Anomalies and the expected market return. *Journal of Finance*, 77(1), 639–681.
- Fama, E.F., French, K.R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1), 1–22.
- Feng, G., Giglio, S., Xiu, D. (2020). Taming the factor zoo: a test of new factors. *Journal of Finance*, 75(3), 1327–1370.
- Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.
- Giacomini, R., White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74(6), 1545–1578.
- Gu, S., Kelly, B., Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5), 2223–2273.

References (5/9)

- Hansen, L.P., Jagannathan, R. (1997). Assessing specification errors in stochastic discount factor models. *Journal of Finance*, 52(2), 557–590.
- Hansen, P.R. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4), 365–380.
- Harvey, C.R., Liu, Y., Zhu, H. (2016). . . . and the cross-section of expected returns. *Review of Financial Studies*, 29(1), 5–68.
- Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning*. 2nd ed. Springer.
- Hoerl, A.E., Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67.
- Hooker, G., Mentch, L., Zhou, S. (2021). Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31(6), 82.
- Hosking, J.R.M. (1981). Fractional differencing. *Biometrika*, 68(1), 165–176.
- Kelly, B., Malamud, S., Zhou, K. (2024). The virtue of complexity in return prediction. *Journal of Finance*, 79(1), 459–503.

References (6/9)

- Kelly, B., Xiu, D. (2023). Financial machine learning. *Foundations and Trends in Finance*, 13(3–4), 205–363.
- Künsch, H.R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17(3), 1217–1241.
- Ledoit, O., Wolf, M. (2008). Robust performance hypothesis testing with the Sharpe ratio. *Journal of Empirical Finance*, 15(5), 850–859.
- Lo, A.W. (2002). The statistics of Sharpe ratios. *Financial Analysts Journal*, 58(4), 36–52.
- Lo, A.W. (2004). The adaptive markets hypothesis. *Journal of Portfolio Management*, 30(5), 15–29.
- López de Prado, M. (2018). *Advances in Financial Machine Learning*. Wiley.
- López de Prado, M. (2020). *Machine Learning for Asset Managers*. Cambridge University Press.

References (7/9)

- Lundberg, S.M., Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
- McCracken, M.W., Ng, S. (2016). FRED-MD: a monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4), 574–589.
- Nagel, S. (2025). Seemingly virtuous complexity in return prediction. NBER Working Paper 34104.
- Newey, W.K., West, K.D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703–708.
- Novy-Marx, R., Velikov, M. (2025). AI-powered (finance) scholarship. NBER Working Paper 33363.
- Opdyke, J.D. (2007). Comparing Sharpe ratios: so where are the p-values? *Journal of Asset Management*, 8(5), 308–336.
- Pesaran, M.H., Timmermann, A. (1992). A simple nonparametric test of predictive performance. *Journal of Business & Economic Statistics*, 10(4), 461–465.
- Politis, D.N., Romano, J.P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89(428), 1303–1313.

References (8/9)

- Rasul, K. et al. (2023). Lag-Llama: towards foundation models for probabilistic time series forecasting. *arXiv:2310.08278*.
- Robinson, P.M. (1995). Gaussian semiparametric estimation of long range dependence. *Annals of Statistics*, 23(5), 1630–1661.
- Romano, J.P., Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4), 1237–1282.
- Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386–408.
- Shimotsu, K. (2010). Exact local Whittle estimation of fractional integration with unknown mean and time trend. *Econometric Theory*, 26(2), 501–540.
- Shimotsu, K., Phillips, P.C.B. (2005). Exact local Whittle estimation of fractional integration. *Annals of Statistics*, 33(4), 1890–1933.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58(1), 267–288.

References (9/9)

- Wager, S., Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228–1242.
- Welch, I., Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21(4), 1455–1508.
- White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5), 1097–1126.
- Woo, G. et al. (2024). Unified training of universal time series forecasting transformers. *arXiv:2402.02592*.
- Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67(2), 301–320.
- **Code and data:** github.com/danpele/MFM/tree/main/Quantlets/Ch_13