







Modelarea piețelor financiare

Seminarul 13: Machine Learning în finanțe — aplicații



Daniel Traian PELE
Antoaneta Amza

Academia de Studii Economice din București
IDA Institute Digital Assets
Blockchain Research Center
AI4EFin Artificial Intelligence for Energy Finance
Academia Română, Institutul de Prognoză Economică
MSCA Digital Finance

Are un semnal Bitcoin putere predictivă la validarea în afara eșantionului?

- Întrebarea seminarului
  - **rămîne valid vreun semnal de machine learning pe Bitcoin după etichetare, validare și selecție corecte?**
- Livrabile
  - refaceți exemplele A rezolvate și rezolvați independent A7
  - un grafic și un tabel pentru fiecare exercițiu B
  - codul corectat din C1
  - o propoziție de interpretare pentru fiecare exercițiu
- Notebook-ul de azi: deschideți notebook-ul seminarului în Colab
  - fiecare cerință indică secțiunea din notebook
- Seminar: prof. Daniel Traian Pele, danpele@ase.ro; drd. Antoaneta Amza, antoaneta.amza@csie.ase.ro

Randamente și convenții între curs și seminar

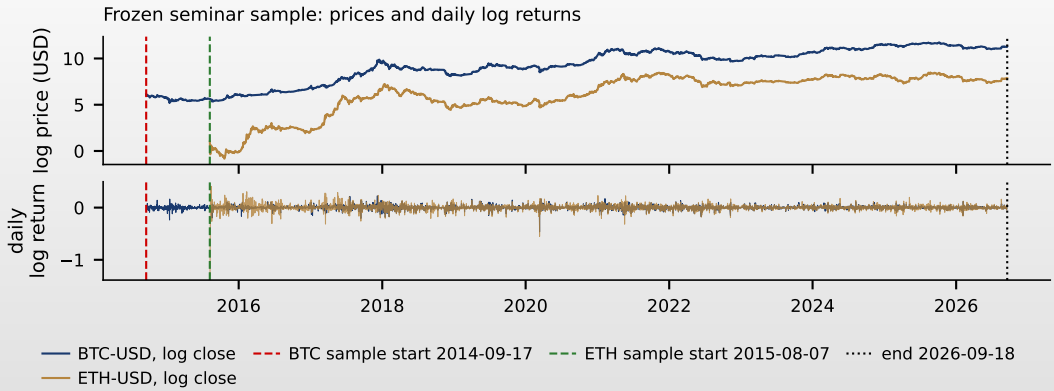
- Randamentele logaritmice (slide-ul „Notație și funcțiile date") sînt în unități zecimale. Pentru un preț care crește de la 100 la 110, randamentul simplu este 10%, iar cel log aproximativ 9,53%. Procentele se obțin înmulțind valorile zecimale cu 100.
- În EWMA (Exponentially Weighted Moving Average, media mobilă ponderată exponențial), α din seminar este ponderea informației noi, iar λ din curs este ponderea trecutului: $\alpha + \lambda = 1$.
- La metoda celor trei bariere, *pt* și *sl* sînt multiplicatorii barierei de profit și ai celei de limitare a pierderii (*stop-loss*), notați π_u și π_l în curs.
- În Logit-L1, penalizarea λ este inversul lui C („Modelul logistic cu penalizare L1"). În figura cursului axa este $1/C$, inversată; în figura seminarului este C: o penalizare mai mare înseamnă un C mai mic.

Datele folosite

Serie	Cîmp	Eșantion fixat	Folosit în
Bitcoin	close	2014-09-17 – 2026-09-18, 4 385 de închideri	B1, B3–B6, B8–B14, C1, C2
Ether	close	2015-08-07 – 2026-09-18, 4 061 de închideri	B7
VIX	close	zilele de tranzacționare din SUA, preluate pe calendarul BTC	B14
Factorii și decilele Kenneth French	Data Library, lunar	1967 – ultima lună	C3

- Închiderile zilnice din datele cursului, cîte o serie pe activ (cîmpul folosit: close); criptomonedele se tranzacționează în fiecare zi, deci un an are $A = 365$ de perioade
- Randamente logaritmice zilnice (Notație) pe calendarul propriu al fiecărui activ; nicio închidere lipsă în eșantioanele fixate; evenimentele fără o fereastră completă de h zile la final nu se etichetează

Pregătire: rulați și verificați (1/2)



Pregătire: rulați și verificați (2/2)

```
# notebook section 0 defines load_data
btc = load_data('btc')['Close']
eth = load_data('eth')['Close']
print(len(btc), btc.index[0], btc.index[-1])
print(len(eth), eth.index[0], eth.index[-1])
r = np.log(btc).diff().dropna()
print(round(r.std() * np.sqrt(365), 3))
```

- Rezultat așteptat pentru BTC (Bitcoin): 4 385 de închideri, de la 17 sep. 2014 (457,33) la 18 sep. 2026 (80 901,46); volatilitatea anualizată 0,666
- Pentru ETH (Ether, criptomoneda rețelei Ethereum): 4 061 de închideri, din 7 aug. 2015
- Dacă un număr diferă, opriți-vă și verificați seria, cîmpul și datele

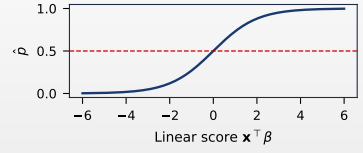
MFM_ch13_seminar

Notație și funcțiile date

- P_t: închiderea din ziua t; $x_t = \ln P_t$; $r_t = x_t - x_{t-\mathbf{1}}$; B: operatorul de întîrziere (lag), $Bx_t = x_{t-\mathbf{1}}$
- Momentul deciziei: închiderea zilei t; în grila de reguli (B6, B12, B13) tranzacția este idealizată la aceeași închidere, deci poziția cîștigă $r_{t+\mathbf{1}}$; cu execuție la închiderea următoare ar cîștiga $r_{t+\mathbf{2}}$
- d: ordinul unui filtru fracționar, o alegere („Diferențierea fracționară"); δ: memoria unei serii, $I(\delta)$, un parametru necunoscut pe care îl estimăm („Memoria, testul ADF și local Whittle")
- $\sigma_{t_\mathbf{0}}$: volatilitatea EWMA a randamentelor logaritmice zilnice, span 50 („Etichete cu orizont fix și volatilitatea EWMA"); h: orizontul etichetei; $[t_\mathbf{0}, t_\mathbf{1}]$: intervalul unei etichete
- SR: raportul Sharpe zilnic al randamentelor logaritmice zilnice, rata fără risc 0 („Etichete cu orizont fix și volatilitatea EWMA"); $\gamma_\mathbf{3}$, $\gamma_\mathbf{4}$: asimetria și kurtosis obișnuit („Raportul Sharpe probabilistic")

Funcția dată (secțiunea 0 a notebook-ului)	Rezultatul returnat
`ffd_weights(d, threshold)`, `frac_diff_ffd`	ponderile $w_\mathbf{0}, \ldots, w_K$; seria filtrată de la ziua K
`local_whittle`, `exact_local_whittle`	$(\hat{\delta}, \mathrm{SE})$ cu $m = \lfloor n^{\mathbf{0,65}} \rfloor$ frecvențe
`get_daily_vol`, `triple_barrier`	volatilitatea EWMA; un rînd pe eveniment: $t_\mathbf{1}$, randament, etichetă, tipul ieșirii
`PurgedKFold`, `build_features`	pliuri purjate cu embargo; cele 11 caracteristici din B3
`probabilistic_sharpe_ratio`, `deflated_sharpe_ratio`	PSR (raportul Sharpe probabilistic) și DSR (raportul Sharpe deflatat) pentru un SR zilnic

Modelul logistic cu penalizare L1



- Cadrul: caracteristicile x_t cunoscute la închiderea zilei t, eticheta $y_t \in \{0, 1\}$ („sus" $= 1$)
  - un clasificator furnizează $\hat{p}_t = \hat{P}(y_t = 1 \mid x_t)$ și prezice „sus" cînd $\hat{p}_t > 0{,}5$
- **Logit-L1** (Tibshirani (1996)): legătură logistică, coeficienți micșorați printr-o penalizare în valoare absolută

$$\hat{p} = \frac{1}{1 + e^{-\mathbf{x}^\top \boldsymbol{\beta}}}, \qquad \min_{\boldsymbol{\beta}} \sum_t \ell_t(\boldsymbol{\beta}) + \lambda \sum_j |\beta_j|$$

  - $\mathbf{x}^\top \boldsymbol{\beta} = \sum_j \beta_j x_j$: scorul liniar; $\beta_j > 0$: o valoare mai mare a lui x_j crește $\hat{p}$
  - ℓ_t: log loss („Log loss"); $\lambda \geq 0$: ponderea penalizării; în scikit-learn $C = 1/\lambda$
  - panoul de jos: cînd C scade, coeficienții slabi devin exact 0

Arbori, random forest și gradient boosting

- **Arborele de clasificare**: împarte spațiul caracteristicilor prin reguli $x_j \leq c$; în frunza R prezice proporția etichetelor „sus"

$$\hat{p}_R = \frac{1}{n_R} \sum_{t \in R} y_t, \qquad G(R) = 2\hat{p}_R(1 - \hat{p}_R)$$

  - n_R: observațiile din R; G: impuritatea Gini, 0 pentru o frunză pură; fiecare ramificare este aleasă astfel încît să reducă cel mai mult $\sum n_R G(R)$
  - mărimea minimă a frunzei (de exemplu 50): o valoare mai mare dă arbori mai netezi, cu mai puțină supraajustare (*overfitting*)

- **Random forest** (RF, Breiman (2001)): media a B arbori, fiecare crescut pe un eșantion bootstrap, cu o submulțime aleatoare de caracteristici la fiecare ramificare

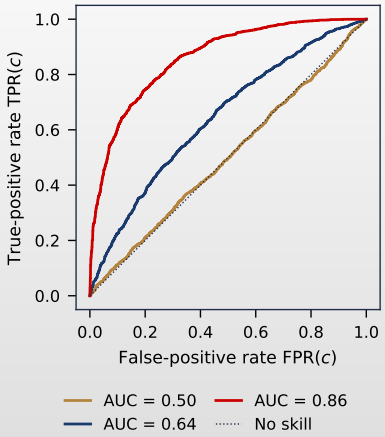
$$\hat{p}_{RF}(\mathrm{x}) = \frac{1}{B} \sum_{b=\mathbf{1}}^{B} \hat{p}_b(\mathrm{x})$$

- **Gradient boosting** (GB, Friedman (2001)): arbori adăugați unul cîte unul, pe scala log-șanselor, $\ln[\hat{p}/(1-\hat{p})]$

$$F_m(\mathrm{x}) = F_{m-\mathbf{1}}(\mathrm{x}) + \eta\, h_m(\mathrm{x}), \qquad \hat{p} = \frac{1}{1 + e^{-F_M(\mathbf{x})}}$$

  - h_m: un arbore mic estimat pe reziduurile $y_t - \hat{p}_{m-\mathbf{1},t}$ (gradientul negativ al log loss)
  - $\eta \in (0, 1]$: rata de învățare; M: numărul de arbori; un η mic învață lent și supraajustează mai puțin

Acuratețea, curba ROC și AUC



- **Acuratețea** $= \frac{1}{n}\sum_t 1\{\hat{y}_t = y_t\}$, $\hat{y}_t = 1\{\hat{p}_t > 0{,}5\}$
  - reperul: „mereu în sus", a cărui acuratețe este proporția etichetelor „sus", nu 50%
- **Curba ROC** (receiver operating characteristic): pentru fiecare prag c, rata adevărat pozitivelor (TPR, True Positive Rate) și rata fals pozitivelor (FPR, False Positive Rate)

$$\mathrm{TPR}(c) = P(\hat{p}_t > c \mid y_t = 1), \quad \mathrm{FPR}(c) = P(\hat{p}_t > c \mid y_t = 0)$$

- **AUC** (Hanley & McNeil (1982)): aria de sub curba ROC

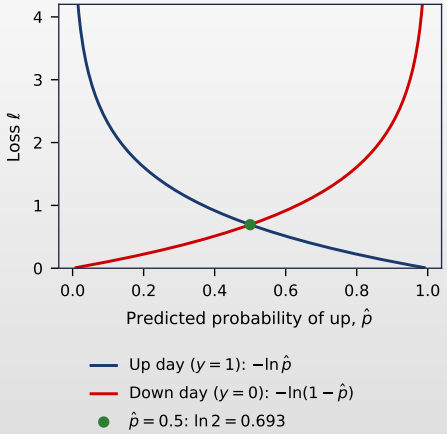
$$\mathrm{AUC} = P(\hat{p}_+ > \hat{p}_-) + \tfrac{1}{2}P(\hat{p}_+ = \hat{p}_-)$$

  - $\hat{p}_+$, $\hat{p}_-$: scorurile unei zile „sus" și ale unei zile „jos" alese la întîmplare
  - 0,5: nicio capacitate de ordonare (diagonala); 1: ordonare perfectă; sub 0,5: ordonare inversă
  - Exemplu: zilele „sus" (0,7; 0,4), zilele „jos" (0,5; 0,2): 3 din 4 perechi ordonate corect, AUC $= 0{,}75$

Acuratețea față de reper

- Întrebare pentru sală: o acuratețe de 55% la 5 zile pe Bitcoin dovedește abilitate?
- Răspuns: nu; 54,8% dintre ferestrele de 5 zile sînt „sus” (B3a)
- Regula „mereu în sus” are deci deja o acuratețe de 54,8%; un model se compară cu acest reper, nu cu 50%

Log loss



- **Log loss** (pierderea logaritmică): penalizează probabilitățile greșite date cu încredere

$$\ell_t = -\big[y_t \ln \hat{p}_t + (1 - y_t) \ln(1 - \hat{p}_t)\big]$$

  - $\ell_t \geq 0$: $-\ln \hat{p}_t$ într-o zi „sus", $-\ln(1 - \hat{p}_t)$ într-o zi „jos"
  - constanta $\hat{p}_t = 0{,}5$ are pierderea $\ln 2 = 0{,}693$ în fiecare zi
  - Exemplu: $y_t = 1$; $\hat{p}_t = 0{,}6$ are pierderea 0,51, $\hat{p}_t = 0{,}2$ are pierderea 1,61

- **Scorul Brier**: eroarea pătratică a unei probabilități, $b_t = (y_t - \hat{p}_t)^2 \in [0, 1]$

- AUC evaluează ordonarea; log loss și Brier evaluează calitatea probabilităților, inclusiv calibrarea: un model poate ordona bine și totuși avea un log loss mai mare (B4)

Compararea a două prognoze

- Predicție în afara pliului: fiecare zi este prognozată de un model care nu a fost antrenat pe ea („Validarea fără informație din viitor" arată ce zile mai trebuie excluse)
- Compararea a două prognoze, A și B: statistica Diebold–Mariano (DM) (Diebold & Mariano (1995))

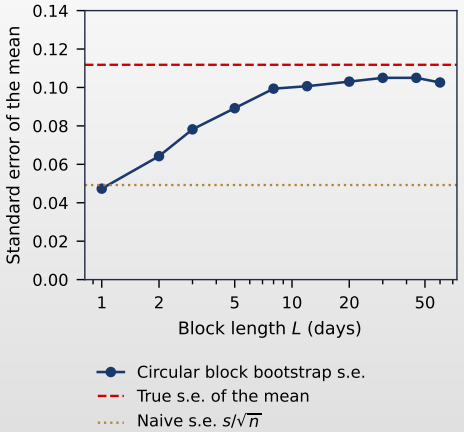
$$d_t = \ell_t^A - \ell_t^B, \qquad DM = \frac{\bar{d}}{\widehat{\mathrm{SE}}(\bar{d})}$$

  - ℓ_t^A, ℓ_t^B: pierderile lui A și B în ziua t; $d_t > 0$: A este mai slabă în acea zi; $\bar{d}$: media lui d_t
  - H_0: pierderi așteptate egale; DM are aproximativ distribuția Normală standard; $|DM| > 1{,}96$: diferență semnificativă la 5%
- $\widehat{\mathrm{SE}}(\bar{d})$: eroarea standard (SE, standard error) Newey–West (Newey & West (1987)), pentru că valorile d_t vecine sînt dependente

$$\widehat{\mathrm{SE}}(\bar{d}) = \sqrt{\frac{1}{n}\Big[\hat{\Gamma}_0 + 2\sum_{j=1}^{q}\Big(1 - \frac{j}{q+1}\Big)\hat{\Gamma}_j\Big]}$$

  - $\hat{\Gamma}_j$: autocovarianța de selecție a lui d_t la lagul j; q: numărul de laguri (cel puțin suprapunerea etichetelor)
  - echivalent: regresăm d_t pe o constantă cu erori HAC (robuste la heteroscedasticitate și autocorelație); valoarea t este DM
  - statistica t naivă (i.i.d.) folosește doar $\hat{\Gamma}_0$ și exagerează semnificația cînd $\hat{\Gamma}_j > 0$

Incertitudinea unui AUC și bootstrap-ul pe blocuri

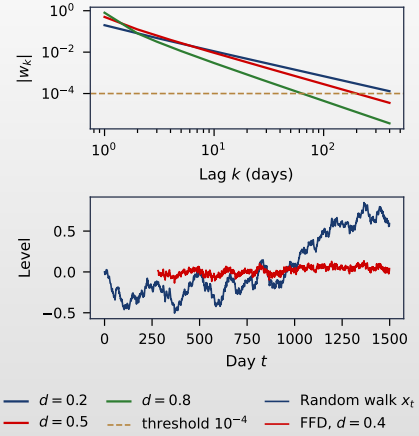


- **DeLong** (DeLong, DeLong & Clarke-Pearson (1988)): o eroare standard în formă închisă a AUC, care tratează zilele ca independente
  - $z = (\mathrm{AUC} - 0{,}5)/\mathrm{SE} \approx N(0,1)$ sub H_0: AUC $= 0{,}5$
- **Bootstrap circular pe blocuri** (Künsch (1989))
  - alegem începuturi aleatoare și luăm blocuri de L zile consecutive (cu reluare de la început după ultima zi) pînă la n zile
  - recalculăm statistica $\hat{\theta}_b^*$ în fiecare dintre cele B replici

$$\mathrm{SE}^* = \mathrm{sd}(\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*), \quad CI_{95} = [\hat{\theta}^*_{(2{,}5\%)},\ \hat{\theta}^*_{(97{,}5\%)}]$$

- CI_{95}: intervalul de încredere (CI, confidence interval) percentil de 95%
- $L = 1$ este bootstrap-ul obișnuit; blocurile păstrează dependența din interiorul lor
- Graficul (sume suprapuse pe 5 zile, simulate): $L = 1$ dă eroarea naivă; de la $L \approx 10$, SE este aproape de cea adevărată

Diferențierea fracționară



- Diferențierea de ordin d (Hosking (1981)), cu operatorul lag $Bx_t = x_{t-1}$ și log-prețul x_t

$$(1-B)^d x_t = \sum_{k \geq 0} w_k x_{t-k}, \quad w_0 = 1, \; w_k = -w_{k-1} \frac{d-k+1}{k}$$

  - k: lagul, în zile; w_k: ponderea valorii de acum k zile
  - $d = 0$ păstrează nivelul, $d = 1$ dă randamentul $x_t - x_{t-1}$
  - $d = 0{,}5$: $w = (1; -0{,}5; -0{,}125; -0{,}0625; \ldots)$; ponderile care scad încet păstrează o parte din nivelurile vechi: „memorie"

- **FFD** (diferențiere fracționară cu fereastră fixă, *Fixed-width window Fractional Differentiation*; López de Prado (2018)): păstrăm ponderile pînă la primul $|w_k| < \tau$ (de exemplu 10^{-4}) și aplicăm aceleași $K+1$ ponderi în fiecare zi
  - K: ultimul lag păstrat; un d mai mic cere o fereastră mai lungă (panoul de sus)
  - panoul de jos: FFD cu $d = 0{,}4$ al unui mers aleator pare să revină la medie

Memoria, testul ADF și local Whittle

- **Memoria** δ: x_t este $I(\delta)$ cînd $(1-B)^{\delta} x_t$ are memorie scurtă
  - $\delta < \frac{1}{2}$: staționară; $\frac{1}{2} \leq \delta < 1$: nestaționară, cu revenire la medie; $\delta = 1$: rădăcină unitară (mers aleator)
  - un filtru netrunchiat de ordin d aplicat unei serii $I(1)$ lasă memoria $1-d$
- **Testul ADF (Augmented Dickey–Fuller)** (Dickey & Fuller (1979)): H_0: rădăcină unitară ($\gamma = 0$) în regresia

$$\Delta x_t = c + \gamma x_{t-1} + \sum_{j=1}^{p} \phi_j \Delta x_{t-j} + \varepsilon_t$$

  - statistica: valoarea t a lui $\hat{\gamma}$; distribuția Dickey–Fuller, nu t; respingem dacă este sub $-2{,}86$ (5%, cu constantă)
  - regula d minim (sau regula ADF minimă, López de Prado (2018)): cel mai mic d la care ADF respinge; A1 întreabă ce dovedește aceasta
- **Local Whittle** (Robinson (1995)): $\hat{\delta}$ minimizează, pe cele mai mici m frecvențe Fourier $\lambda_j = 2\pi j/n$

$$R(\delta) = \ln\Big(\frac{1}{m}\sum_{j=1}^{m} \lambda_j^{2\delta} I(\lambda_j)\Big) - \frac{2\delta}{m}\sum_{j=1}^{m} \ln \lambda_j, \qquad \mathrm{SE}(\hat{\delta}) = \frac{1}{2\sqrt{m}}$$

  - $I(\lambda_j)$: periodograma; n: lungimea eșantionului; $m = \lfloor n^{0{,}65} \rfloor$, $\lfloor \cdot \rfloor$: partea întreagă
  - local Whittle exact (Shimotsu & Phillips (2005)) este valid și pentru $\delta \geq \frac{1}{2}$

Etichete cu orizont fix și volatilitatea EWMA

- **Eticheta cu orizont fix** pe h zile

$$y_t = 1\{\ln P_{t+h} - \ln P_t > 0\}$$

  - $1\{A\} = 1$ dacă A are loc și 0 altfel; h: orizontul, în zile
  - eticheta zilei t folosește randamentele din $[t, t+h]$: etichetele vecine au randamente comune

- **Volatilitatea EWMA** a randamentelor logaritmice zilnice r_t

$$\sigma_t^2 = (1-\alpha)\,\sigma_{t-1}^2 + \alpha\, r_t^2, \qquad \alpha = \frac{2}{\text{span} + 1}$$

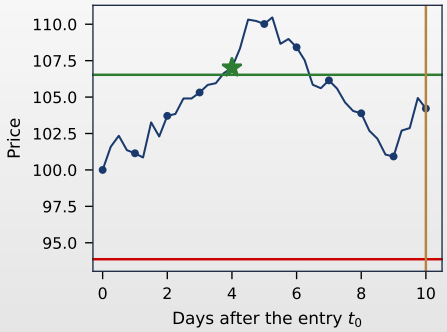
  - span 50: $\alpha = 2/51 \approx 0{,}039$ pentru ziua cea mai recentă; ponderea zilei $t-k$ este $\alpha(1-\alpha)^k$
  - `get_daily_vol` folosește abaterea standard EWMA (cu media scăzută); σ_{t_0} scalează barierele de pe slide-ul următor

- **Raportul Sharpe** (SR) al randamentelor zilnice, rata fără risc 0

$$\widehat{SR} = \frac{\bar{r}}{\hat{\sigma}}, \qquad \widehat{SR}_{ann} = \widehat{SR}\sqrt{A}$$

  - $\bar{r}$, $\hat{\sigma}$: media și abaterea standard a randamentelor zilnice; A: zile pe an (365 pentru Bitcoin, 252 pentru acțiuni)

Metoda celor trei bariere

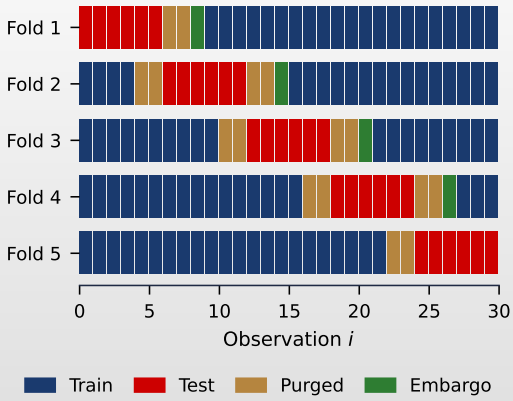


- **Trei bariere** (López de Prado (2018)): de la intrarea t_0, două bariere orizontale și una verticală

$$P_{t_0}e^{+pt\,\sigma_{t_0}\sqrt{h}}, \qquad P_{t_0}e^{-sl\,\sigma_{t_0}\sqrt{h}}, \qquad t_0 + h$$

  - $\sigma_{t_0}\sqrt{h}$: variația logaritmică tipică pe h zile; pt, sl: lățimile de profit și de stop-loss, în unități ale acesteia
  - eticheta $+1$ / -1 la prima atingere orizontală; altfel semnul randamentului la $t_0 + h$; t_1: ziua ieșirii
- Graficul (simulat): $\sigma = 2\%$, $h = 10$, $pt = sl = 1$: barierele 106,53 și 93,87; închiderea din ziua 4 este 107,02, deci $t_1 = 4$ și eticheta este $+1$, deși traiectoria scade ulterior
- Se verifică doar închiderile zilnice: o atingere în timpul zilei este ratată

Validarea fără informație din viitor



- Validarea încrucișată (CV, cross-validation) în K pliuri: antrenare pe $K-1$ pliuri, test pe cel rămas
  - pliurile **amestecate** combină zilele; cele **contigue** păstrează blocuri de timp
- **Purjarea** (López de Prado (2018)): eliminăm din antrenare orice observație al cărei interval de etichetă se suprapune cu intervalul unei etichete de test
- **Embargoul**: eliminăm și cele e observații de după sfîrșitul ultimei etichete de test, pentru că caracteristicile privesc înapoi
- Exemplu: $n = 10$, intervalele $[i, i+2]$, pliul de test $\{4, 5\}$ (etichete pînă în ziua 7), $e = 1$
  - eliminăm $\{2, 3\}$ înainte, $\{6, 7\}$ după (se suprapun cu intervalul închis de test pînă în ziua 7) și $\{8\}$ sub embargo; antrenăm pe $\{0, 1, 9\}$
- Graficul: $n = 30$, 5 pliuri, intervalele $[i, i+2]$, $e = 1$
- Fără purjare, o zi de antrenare are randamente comune cu o zi de test, iar un model flexibil memorează observația vecină (B2)

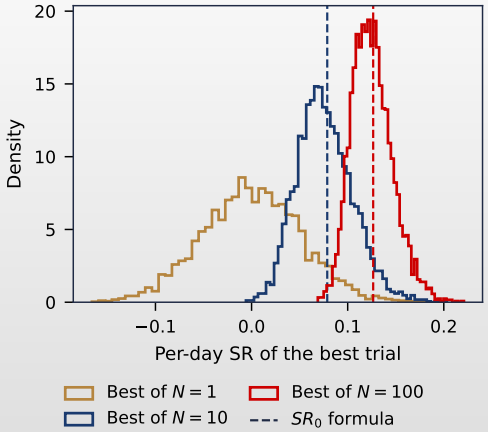
Raportul Sharpe probabilistic

- Raportul Sharpe probabilistic (PSR, Bailey & López de Prado (2012)): aproximația Normală a lui unu minus valoarea p (*p-value*) unilaterală față de un reper zilnic SR^*

$$PSR(SR^*) = \Phi\Big(\frac{(\widehat{SR} - SR^*)\sqrt{T-1}}{D}\Big), \qquad D = \sqrt{1 - \gamma_3 \widehat{SR} + \frac{\gamma_4 - 1}{4}\widehat{SR}^2}$$

  - $\widehat{SR}$: SR zilnic estimat; T: numărul de randamente; Φ: funcția de repartiție a distribuției Normale standard
  - $\gamma_3 = \frac{1}{T}\sum_t (r_t - \bar{r})^3 / \hat{\sigma}^3$: asimetria; $\gamma_4 = \frac{1}{T}\sum_t (r_t - \bar{r})^4 / \hat{\sigma}^4$: kurtosis obișnuit (3 pentru distribuția Normală)
  - $D/\sqrt{T-1}$: eroarea standard a lui $\widehat{SR}$; asimetria negativă și cozile groase măresc D și scad PSR
  - $D = 1$ în prima aproximare pentru randamente cu distribuția Normală ($\gamma_3 = 0$, $\gamma_4 = 3$, $\widehat{SR}$ mic)
- Interpretare: $PSR(0) = 0{,}95$ corespunde unei valori p unilaterale de 5% față de un SR nepozitiv, sub randamente i.i.d. și metoda delta, pentru **o singură** strategie stabilită dinainte; nu este o probabilitate posterioară a parametrului

Selecția și raportul Sharpe deflatat



- Selecția: cea mai bună dintre N strategii fără abilitate reală are totuși un SR pozitiv; valoarea lui așteptată este

$$SR_0 = \sqrt{V}\left[(1-\gamma_E)\Phi^{-1}(1-\tfrac{1}{N}) + \gamma_E\Phi^{-1}(1-\tfrac{1}{Ne})\right]$$

  - V: scala variației rapoartelor Sharpe ale încercărilor; ilustrarea sub ipoteza nulă folosește $V = 1/T$, iar dispersia empirică între încercări este o aproximație diferită; Φ^{-1}: funcția cuantilă a distribuției Normale standard; $e \approx 2{,}718$; $\gamma_E = 0{,}5772$: constanta lui Euler

- **Raportul Sharpe deflatat** (DSR, Bailey & López de Prado (2014))

$$DSR = PSR(SR_0)$$

  - PSR față de maximul așteptat sub noroc; peste 0,95 dă o valoare p unilaterală sub 5% față de reperul ajustat, sub aproximația precizată

- Graficul (simulat, $T = 400$, $V = 1/T$): SR-ul maxim se deplasează spre dreapta cînd N crește; liniile întrerupte: $SR_0 = 0{,}079$ ($N = 10$) și $0{,}127$ ($N = 100$)

Raportul Sharpe după alegerea celei mai bune strategii

- SR zilnic $\widehat{SR} = 0{,}10$, $T = 400$, $D = 1$: $PSR(0) = \Phi(0{,}10\sqrt{399}) = \Phi(2{,}00) = 0{,}977$
- Dacă este cea mai bună dintre $N = 10$ încercări, cu $V = 1/T$
  - $\Phi^{-1}(0{,}9) = 1{,}282$, $\Phi^{-1}(1 - \frac{1}{10e}) = 1{,}789$, $\sqrt{V} = 1/20$
  - $SR_0 = \frac{1}{20}[0{,}4228 \cdot 1{,}282 + 0{,}5772 \cdot 1{,}789] = 0{,}079$
  - $DSR = \Phi\big((0{,}10 - 0{,}079)\sqrt{399}\big) = \Phi(0{,}43) = 0{,}66$
- Interpretare: luat singur, SR-ul este convingător (0,977); ca cel mai bun din zece încercări, nu mai este ($0{,}66 < 0{,}95$)
- Mai mult în Capitolul 13: teorema strategiei false, importanța caracteristicilor, backtest-urile cu evaluare pe ferestre succesive (*walk-forward*) și studiul de caz S&P 500 (*Standard & Poor's*)

Instrumente pentru pachetele de extindere

- **Precizia, recall-ul (sensibilitatea), F1** pentru decizia „luăm tranzacția" (B11); TP (True Positive), FP (False Positive), FN (False Negative): adevărat pozitive, fals pozitive, fals negative

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

- **Importanța caracteristicilor** (B3c)
  - MDI (*mean decrease in impurity*, scăderea medie a impurității): reducerea lui $\sum n_R G(R)$ datorată ramificărilor după caracteristica j, mediată pe arbori; în eșantion
  - MDA (*mean decrease in accuracy*, scăderea medie a scorului): $\mathrm{AUC} - \mathrm{AUC}_{(j)}$, cu $\mathrm{AUC}_{(j)}$ AUC-ul de test după permutarea aleatoare a caracteristicii j pe setul evaluat; o valoare negativă se citește cu dispersia permutărilor și cu incertitudinea estimării
  - SHAP (SHapley Additive exPlanations, Lundberg & Lee (2017)): $\hat{p}(\mathrm{x}) = \phi_{\mathbf{0}} + \sum_j \phi_j(\mathrm{x})$, o împărțire echitabilă a unei predicții între caracteristici; atribuie predicția, față de rezultatul realizat; importanța: media lui $|\phi_j|$
- **Corelația de rang Spearman**: corelația Pearson a rangurilor; 1 = ordonări identice

Harta exercițiilor

Exercițiu	Întrebare	Tip	Model
A1	ce păstrează un filtru fracționar trunchiat din nivelul prețului?	Rezolvat	–
A2, B1	cum se comportă etichetele cu trei bariere pe traiectorii simple și pe Bitcoin?	Rezolvat	–
A3, B2	ce elimină purjarea și ce simulează amestecarea?	Rezolvat	–
A7	cum schimbă o etichetă mai scurtă numărul observațiilor pierdute?	Propus	A3
B3a, B3b	depășesc trei modele reperul „mereu în sus”? Cît de incert este AUC?	Rezolvat	–
B4	prognozează un model mai bine decît o constantă?	Propus	B3a, B3b
C1	ce este greșit într-un răspuns AI?	Propus	B2, B3a
A4	depășește cel mai bun dintre 100 de backtest-uri norocul?	Rezolvat	–
După curs	B5: face diferențierea fracționară prețul Bitcoin staționar? B6: cea mai bună dintre 120 de reguli față de noroc	Rezolvat	A1, A4
Pachete de extindere	A2 extins, A5, A6, A8–A10, B3c, B7–B14, C2, C3	ambele	la fiecare cerință

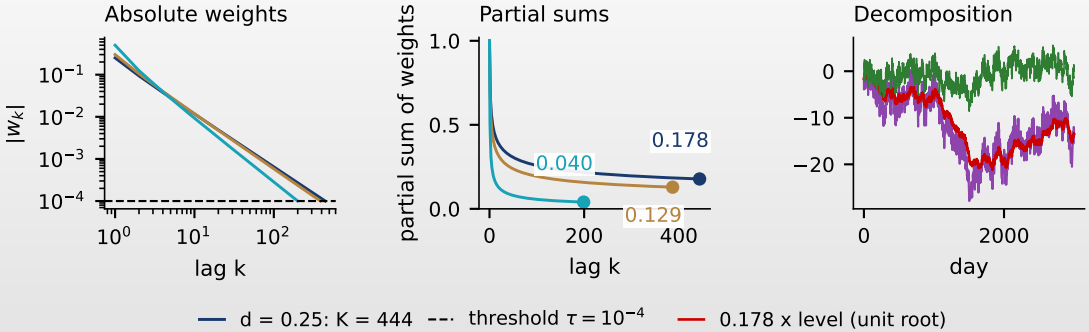
- **[Rezolvat]**: rezolvare completă în slide-uri și în notebook, model de urmat; **[Propus]**: îl rezolvați, apoi se discută la seminar
- A: derivări și exemple numerice mici, pe hîrtie; B: inferență pe date cripto, în notebook, cu o întrebare de interpretare; C: critica unui răspuns AI (Artificial Intelligence) și puncte de plecare pentru proiect
- Partea I: de la A1 la B2; Partea a II-a: de la B3a la A4; B5, B6 folosesc instrumente din curs (local Whittle exact, numărul efectiv de încercări)

A1 [Rezolvat]: ponderile FFD și fereastra păstrată — cerința

- **Întrebarea**: ce păstrează un filtru fracționar trunchiat din nivelul prețului?
- **Date de intrare**: ponderile FFD w_k și regula lor de trunchiere („Diferențierea fracționară"), cu $\tau = 10^{-4}$; $K =$ ultimul lag păstrat, deci fereastra folosește $K+1$ observații; fără fișier de date
- **Sarcini**
  1. Calculați $w_0, \ldots, w_4$ pentru $d = 0{,}25$ cu relația de recurență.
  2. Cu `ffd_weights(0.25)`, raportați K, lungimea ferestrei $K+1$ și $\sum_{k \le K} w_k$; repetați pentru $d = 0{,}3$ și $d = 0{,}5$.
  3. Scrieți $\tilde{X}_t = \sum_{k=0}^{K} w_k X_{t-k}$ ca $(\sum_k w_k) X_t$ minus o combinație a diferențelor $X_t - X_{t-k}$.
  4. Decideți dacă $\tilde{X}_t$ al unui log-preț $I(1)$ este staționar.
- **Raportați**: cele cinci ponderi, un tabel (d, K, $K+1$, suma ponderilor), descompunerea și decizia
- **Interpretare**: regula d minim alege cel mai mic d la care un test ADF respinge: ce dovedește o astfel de respingere?
- Notebook: secțiunea A1

Înapoi Lista exercițiilor

A1 [Rezolvat]: rezolvare (1/2) — ponderile FFD și fereastra păstrată



A1 [Rezolvat]: rezolvare (2/2) — ponderile FFD și fereastra păstrată

1. $w_1 = -0{,}25$, $w_2 = -0{,}25 \cdot 0{,}75/2 = -0{,}09375$, $w_3 = -0{,}0547$, $w_4 = -0{,}0376$: toate ponderile după w_0 sînt negative
2. Tabel:

d	K	$K+1$	$\sum_{k \le K} w_k$
0,25	444	445	0,178
0,30	387	388	0,129
0,50	199	200	0,040

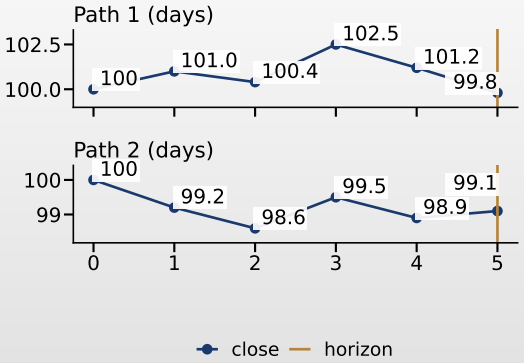
3. $\tilde{X}_t = \left(\sum_{k \le K} w_k\right) X_t - \sum_{k=1}^{K} w_k (X_t - X_{t-k})$
4. Fiecare $X_t - X_{t-k}$ este $I(0)$, dar primul termen este $0{,}178\, X_t$ la $d = 0{,}25$: $\tilde{X}_t$ rămîne $I(1)$

- Panoul din dreapta: un mers aleator simulat; seria FFD urmărește nivelul scalat (roșu), care poartă rădăcina unitară (corelație 0,93)
- Exemplu simplu: ponderile $(1; -0{,}5)$ dau $X_t - 0{,}5X_{t-1} = 0{,}5X_t + 0{,}5(X_t - X_{t-1})$; un termen păstrează nivelul nestaționar, celălalt este un increment staționar
- Interpretare: în acest model filtrat $I(1)$ cunoscut, respingerea ADF induce în eroare; descompunerea algebrică, nu respingerea, stabilește nestaționaritatea

MFM_ch13_seminar

A2 [Rezolvat]: etichete cu trei bariere pe două traiectorii — cerința

- **Întrebările**
  - ce barieră este atinsă prima?
  - ce etichetă rezultă?
- **Date de intrare**: intrare $P_{t_0} = 100$, volatilitate zilnică $\sigma_{t_0} = 1\%$, bariera verticală $h = 5$ zile, $pt = sl = 1$; barierele din secțiunea „Metoda celor trei bariere"; doar închideri zilnice
- **Sarcini**
  1. Calculați bariera superioară și cea inferioară.
  2. Desenați-le pe cele două traiectorii și marcați prima atingere.
  3. Etichetați fiecare traiectorie: $+1$ sus, -1 jos, iar la bariera verticală semnul lui $\ln(P_{t_0+h}/P_{t_0})$.
- **Raportați**: cele două bariere, ziua și tipul ieșirii, cele două etichete
- **Interpretare**: de ce poate primi $+1$ o traiectorie care se încheie sub 100? Notebook: secțiunea A2



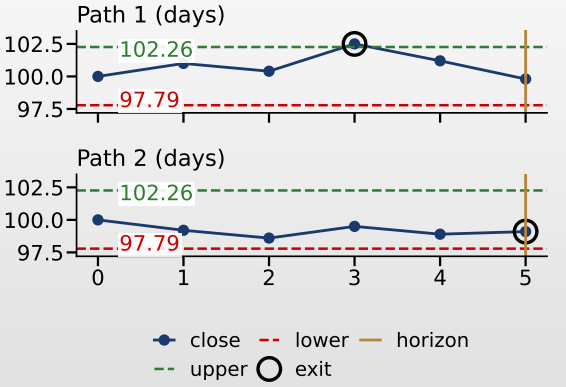
MFM_ch13_seminar

înapoi Lista exercițiilor

A2 [Rezolvat]: rezolvare — etichete cu trei bariere pe două traiectorii

1. Lățimea $0{,}01\sqrt{5} = 0{,}02236$: bariera superioară $100e^{\mathbf{0{,}02236}} = 102{,}26$, cea inferioară $100e^{\mathbf{-0{,}02236}} = 97{,}79$
2. Traiectoria 1: ziua 3; $102{,}5 \geq 102{,}26$, atingere sus; traiectoria 2: nicio atingere orizontală în 5 zile
3. Traiectoria 1: $+1$ (scăderea ulterioară la 99,8 nu contează); traiectoria 2: $\ln(99{,}1/100) = -0{,}90\%$ la bariera verticală: -1

⊡ Interpretare: eticheta reține prima ieșire pe care ar fi luat-o un trader, nu prețul de la sfîrșitul ferestrei; o ieșire verticală păstrează semnul randamentului final, deci „vertical” și „-1” sînt noțiuni diferite

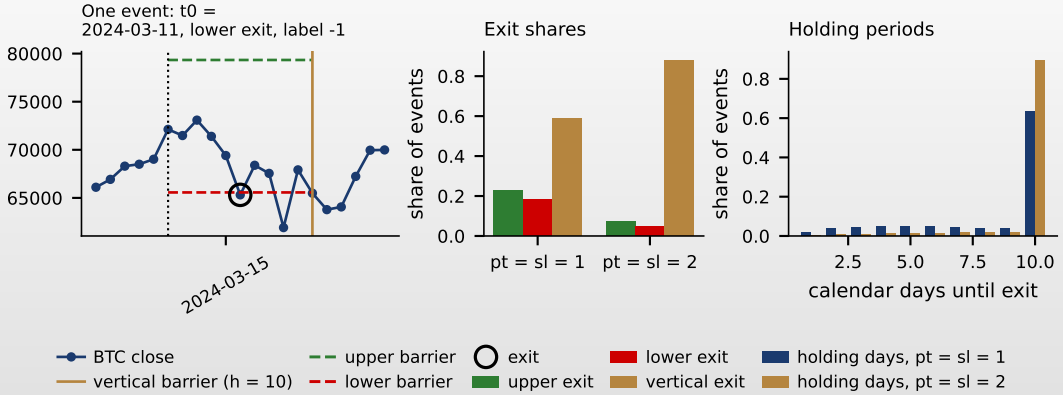


B1 [Rezolvat]: etichete cu trei bariere pe Bitcoin — cerința

- **Întrebările**
  - cît de des atinge Bitcoin o barieră orizontală în 10 zile?
  - cît de echilibrate sînt etichetele?
- **Date și setări**: închideri zilnice BTC; fiecare zi este un eveniment t_0; σ_{t_0} = volatilitatea EWMA („Etichete cu orizont fix și volatilitatea EWMA"; `get_daily_vol`); $h = 10$ zile; monitorizare pe închideri zilnice; evenimentele fără 10 închideri ulterioare sînt eliminate (4 373 de evenimente, 19 sep. 2014 – 8 sep. 2026)
- **Sarcini**
  1. Cu `triple_barrier`, etichetați fiecare eveniment pentru $pt = sl = 1$ și $pt = sl = 2$; afișați înregistrarea unui eveniment (t_0, σ_{t_0}, prețurile barierelor, t_1, tipul ieșirii, randamentul, eticheta).
  2. Calculați ponderile ieșirilor sus, jos și verticale, durata medie de deținere și ponderea etichetelor $+1$.
  3. Faceți tabelul încrucișat tip de ieșire – etichetă finală și comparați cu eticheta cu orizont fix de 10 zile, pe aceleași evenimente.
- **Raportați**: înregistrarea evenimentului, un tabel pe variante, tabelul încrucișat, graficul ponderilor și al duratelor
- **Interpretare**
  - de ce circa 56% dintre etichete sînt $+1$?
  - ce acuratețe trebuie atunci depășită?
- Notebook: secțiunea B1

Înapoi Lista exercițiilor

B1 [Rezolvat]: rezolvare (1/2) — etichete cu trei bariere pe Bitcoin



B1 [Rezolvat]: rezolvare (2/2) — etichete cu trei bariere pe Bitcoin

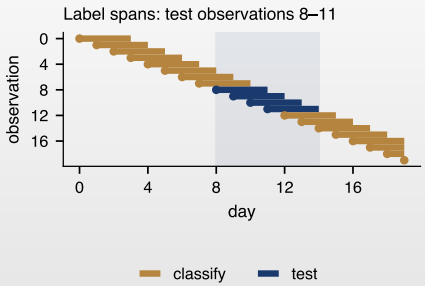
	Sus	Jos	Vertical	Zile	$+1$
$pt = sl = 1$	22,8%	18,2%	59,0%	8,25	55,7%
$pt = sl = 2$	7,2%	4,9%	87,9%	9,58	55,6%
orizont fix 10 zile	–	–	100%	10	55,5%

- Evenimentul din 11 mar. 2024: închidere 72 124, $\sigma = 3{,}0\%$, bariere 79 336 și 65 568; ieșire jos pe 16 mar., randament $-9{,}9\%$, eticheta -1
- $pt = sl = 1$: ieșirile verticale se împart în 1 442 de etichete $+1$ și 1 139 de etichete -1; cele orizontale dau 995 de $+1$ și 797 de -1
- Interpretare: BTC a crescut în eșantion, deci „mereu în sus" (circa 56%) este reperul; barierele mai largi apropie etichetele de cele cu orizont fix

MFM_ch13_seminar

A3 [Rezolvat]: costul purjării — cerința

- **Întrebarea**: ce observații de antrenare trebuie eliminate cînd etichetele de test se suprapun cu ale lor?
- **Date de intrare**: $n = 20$ de observații $i = 0, \ldots, 19$ (de la zero) în $K = 5$ pliuri contigue de $L = 4$; eticheta lui i acoperă intervalul închis $[i, i + h]$, $h = 3$; embargo $e = 1$ observație după sfîrșitul ultimei etichete de test; pliul de test $\{8, \ldots, 11\}$, ale cărui etichete ajung la ziua 14
- **Sarcini**
  1. Enumerați observațiile dinaintea pliului de test al căror interval ajunge la ziua 8 (purjate înainte).
  2. Enumerați observațiile de după pliu care încep cel tîrziu în ziua 14 (purjate după) și observația sub embargo, prima după ziua 14.
  3. Numărați reuniunea eliminărilor și proporția pierdută din cele 16 observații din afara testului; pentru un pliu interior, cu destule observații de ambele părți, arătați că proporția este $\frac{2h+e}{n-L}$.
- **Raportați**: mulțimile eliminate, proporția pierdută, formula
- **Interpretare**: de ce supraestimează K-fold cu amestecare acuratețea în acest cadru? Notebook: secțiunea A3

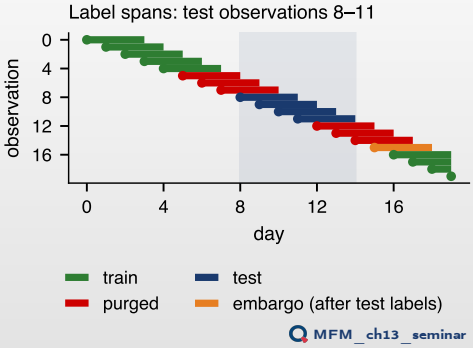


MFM ch13_seminar
◂ Înapoi ▸ Lista exercițiilor

A3 [Rezolvat]: rezolvare — costul purjării

1. Purjate înainte: $\{5, 6, 7\}$, pentru că $i + 3 \geq 8$
2. Purjate după: $\{12, 13, 14\}$, pentru că încep pînă în ziua 14; embargo: $\{15\}$, prima observație după ziua 14
3. Reuniunea: $\{5, 6, 7, 12, 13, 14, 15\}$, 7 din 16 pierdute (43,75%); setul de antrenare $\{0, \ldots, 4\} \cup \{16, \ldots, 19\}$

- Pliu interior general: h observații înainte; după, h etichete care încep pînă la sfîrșitul ultimei etichete de test, apoi e sub embargo; proporția $\frac{2h+e}{n-L} = \frac{3+3+1}{16}$
- Interpretare: amestecarea pune vecini cu randamente comune de ambele părți ale împărțirii; un model flexibil îi memorează (B2)



A7 [Propus]: purjare cu etichete mai scurte — cerința

- **Model**: A3 [Rezolvat]
- **Date de intrare**: ca în A3, cu intervalele închise $[i, \min(i+2, 19)]$ ($h = 2$), embargo $e = 1$, pliul de test $\{12, \ldots, 15\}$
- **Sarcini**
  1. Pe o axă a indicilor 0–19, completați un tabel cu coloanele purjate înainte, purjate după, embargo și reuniunea.
  2. Verificați numărătoarea cu formula din A3.
  3. Pliu interior cu $n = 250$, $K = 5$ ($L = 50$), $h = 20$, $e = 1$: calculați proporția pierdută; repetați cu $K = 10$ ($L = 25$).
- **Raportați**: tabelul pe axa 0–19, verificarea cu formula și proporțiile pierdute pentru $K = 5$ și $K = 10$
- **Interpretare**
  - ce proporție din eșantion elimină purjarea cînd etichetele sînt lungi și pliurile scurte?
  - ce soluții ați propune?
- Notebook: secțiunea A7

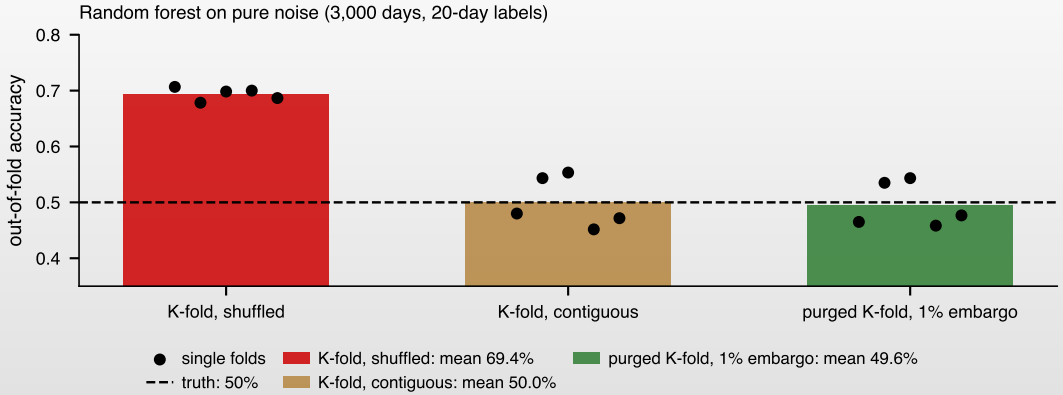
Înapoi Lista exercițiilor

B2 [Rezolvat]: K-fold cu purjare față de K-fold amestecat — cerința

- **Întrebarea**: cît de mult supraestimează amestecarea acuratețea cînd ținta nu este predictibilă? Un diagnostic sintetic, fără date de piață
- **Date de intrare** (notebook, secțiunea B2, seed 42): un mers aleator de 3 000 de zile; eticheta = semnul randamentului pe 20 de zile înainte, intervale $[i, i + 20]$; 5 caracteristici = medii pe 20 de zile ale unui zgomot independent; acuratețea adevărată 50%
- **Sarcini**
  1. Verificați clasa dată `PurgedKFold`: găsiți linia care purjează înainte de pliul de test, linia care aplică embargoul și pe cea care purjează după el; rulați-o pe exemplul din A3.
  2. Estimați un random forest (Breiman (2001); 200 de arbori, frunza minimă 5, seed 42) cu 5-fold amestecat, 5-fold contiguu și 5-fold purjat cu embargo 1% (30 de observații).
  3. Raportați cele cinci acurateți pe pliuri și media lor pentru fiecare schemă.
- **Raportați**: verificarea pe A3, un grafic și un tabel cu acuratețile pe pliuri
- **Interpretare**
  - în acest cadru, ce elimină scurgerea de informație (*leakage*): purjarea sau renunțarea la amestecare?
  - s-ar reduce diferența cu etichete pe 1 zi?

Înapoi Lista exercițiilor

B2 [Rezolvat]: rezolvare (1/2) — K-fold cu purjare față de K-fold amestecat



B2 [Rezolvat]: rezolvare (2/2) — K-fold cu purjare față de K-fold amestecat

- Liniile clasei: `before = idx[t1 < test_t0]` purjează înainte; `later = idx[(idx > stop) & (t0 > test_t1)]` găsește prima observație care începe după ultima etichetă de test; `after_start = min(later[0] + embargo, len(X))` aplică embargoul după ea
- Pe exemplul din A3 clasa elimină $\{5, 6, 7, 12, 13, 14, 15\}$ pentru pliul de test $\{8, \ldots, 11\}$, ca la calculul pas cu pas

Schema	Acurateți pe pliuri; media
amestecat	0,707 0,678 0,698 0,700 0,687; **0,694**
contiguu	0,480 0,543 0,553 0,452 0,472; 0,500
purjat	0,465 0,535 0,543 0,458 0,477; **0,496**

- Seturile de antrenare purjate păstrează 2 330–2 380 din 2 400 de observații
- Interpretare: aici pliurile contigue elimină deja cea mai mare parte a scurgerii de informație
  - doar granițele pliurilor au randamente comune; purjarea elimină observațiile de la granițe
  - scurgerea de informație apare cînd există etichete suprapuse *și* caracteristici persistente (Bergmeir, Hyndman & Koo (2018))
  - acest seed ilustrează rezultatul doar pentru acest cadru
- Cu etichete pe 1 zi (aceeași simulare), acuratețile medii sînt: amestecat 0,496; contiguu 0,489; purjat 0,496
  - etichetele nu se mai suprapun, deci amestecarea nu mai are ce memora și diferența dispare (Bergmeir, Hyndman & Koo (2018))

MFM_ch13_seminar

Cadrul de modelare pe Bitcoin (B3, B4, B9, B10, B14)

- Ținta: $y_t = 1\{\ln P_{t+\mathbf{5}} - \ln P_t > 0\}$, 5 zile calendaristice (criptoactivele se tranzacționează zilnic); B3–B4 folosesc această etichetă cu orizont fix, iar B1 ieșiri prin trei bariere; intervalul etichetei $[t, t+5]$
- Rînduri: fiecare zi cu toate caracteristicile și o fereastră completă de 5 zile: $n = 4\,181$, 4 apr. 2015 – 13 sep. 2026; 54,8% „sus"
- Validare: `PurgedKFold(5, t1, pct_embargo=0.01)`: 5 pliuri contigue de 836–837 de zile, embargo 41 de zile; aceleași pliuri pentru toate modelele

11 caracteristici (`build_features`)	Definiție, informație pînă în ziua t
`mom_1, mom_5, mom_20, mom_60, mom_120`	$x_t - x_{t-h}$
`vol_20, vol_60, vol_ratio`	abaterea standard a lui r_t pe 20 și 60 de zile $\times\sqrt{365}$; raportul lor
`dist_ma50, dist_ma200`	x_t minus media lui x pe ultimele 50 sau 200 de zile
`rsi_14`	indicele de forță relativă, 14 zile, împărțit la 100

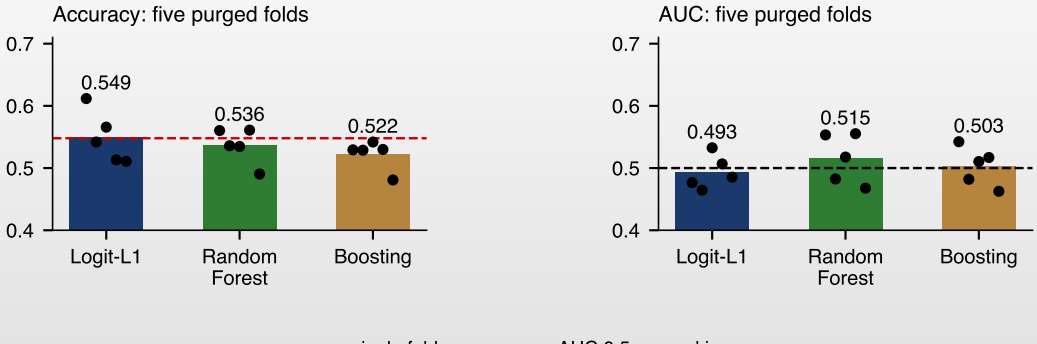
- Modele: Logit-L1 (standardizat, $C = 0{,}05$, Tibshirani (1996)); RF (400 de arbori, frunza minimă 50, Breiman (2001)); GB (pe histograme, adîncime 3, rată de învățare 0,03; 200 de arbori, Friedman (2001)); seed 42
- Se salvează pentru exercițiile următoare: un rînd pe dată cu eticheta, pliul, cele trei probabilități în afara pliului și prognoza constantă (proporția „sus" din antrenare în acel pliu)

B3a [Rezolvat]: trei modele față de „mereu în sus” — cerința

- **Întrebarea**: prezice vreunul dintre cele trei modele direcția Bitcoin pe 5 zile mai bine decît „mereu în sus”?
- **Date**: cadrul de modelare pe Bitcoin (slide-ul anterior)
- **Sarcini**
  1. Estimați Logit-L1, RF și GB pe fiecare pliu de antrenare purjat și prognozați pliul de test; salvați probabilitățile în afara pliului.
  2. Pentru fiecare model, calculați acuratețea ($\hat{p} > 0{,}5$) și ROC AUC în fiecare pliu de test și mediile lor (**medii pe pliuri**).
  3. Calculați acuratețea „mereu în sus” a fiecărui pliu de test, adică proporția etichetelor „sus”, și media pe pliuri.
  4. Calculați AUC **agregat**: un singur AUC pe toate cele 4 181 de probabilități în afara pliurilor.
- **Raportați**: un tabel (acuratețe, AUC, abaterea între pliuri), reperul, graficul cu punctele pe pliuri, proporția prognozelor „sus”
- **Interpretare**: de ce o acuratețe de 54,8% nu este aici o dovadă de abilitate? Notebook: secțiunea B3

Înapoi Lista exercițiilor

B3a [Rezolvat]: rezolvare (1/2) — trei modele față de „mereu în sus”



B3a [Rezolvat]: rezolvare (2/2) — trei modele față de „mereu în sus”

	Logit-L1	RF	GB
Acurateţe, media pe pliuri	54,9%	53,6%	52,2%
AUC, media pe pliuri	0,493	0,515	0,503
AUC agregat	0,468	0,512	0,497
Prezis „sus”	99,6%	69,5%	68,4%

- „Mereu în sus”: proporțiile pe pliuri 61,3%; 53,5%; 56,6%; 51,3%; 51,3%, media 54,8%; RF și GB rămîn sub el; Logit-L1 (54,9%) se reduce practic la „mereu în sus”
- Media pe pliuri și AUC agregat diferă: cinci ordonări mediate față de o singură ordonare pe toate zilele, care amestecă și calibrarea pliurilor
- Interpretare: 54,8% este proporția ferestrelor de 5 zile „sus”, nu o abilitate; întrebarea următoare este cît de incert este 0,512 al RF (B3b)

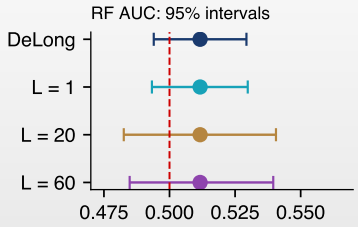
MFM_ch13_seminar

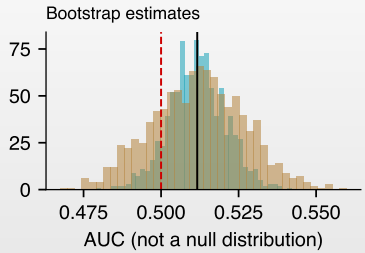
B3b [Rezolvat]: un CI pentru AUC robust la dependență — cerința

- **Întrebarea**: diferă semnificativ AUC agregat de 0,512 al RF de 0,5?
- **Date**: probabilitățile RF în afara pliurilor și etichetele salvate în B3a ($n = 4\,181$); AUC ca în „Acuratețea, curba ROC și AUC", 0,5 = nicio ordonare
- **Sarcini**
  1. Calculați eroarea standard DeLong (DeLong, DeLong & Clarke-Pearson (1988)), CI 95%, $z = (\mathrm{AUC} - 0{,}5)/\mathrm{SE}$ și valoarea p bilaterală.
  2. Reeșantionați rîndurile cu un bootstrap circular pe blocuri (Künsch (1989)): blocuri de $L \in \{1, 20, 60\}$ zile consecutive, 1 000 de replici, seed 42; recalculați de fiecare dată AUC.
  3. Raportați SE bootstrap, CI 95% percentil și proporția replicilor egale cu sau sub 0,5, pentru fiecare L.
- **Raportați**: un grafic cu intervalele și distribuțiile bootstrap pentru $L = 1$ și $L = 20$
- **Interpretare**
  - de ce este intervalul bootstrap pe blocuri mai larg?
  - ce lungime de bloc ați justifica?
- Notebook: secțiunea B3

◂ Înapoi ▸ Lista exercițiilor

B3b [Rezolvat]: rezolvare — un CI pentru AUC robust la dependență





AUC 0.5 bootstrap L = 20
bootstrap L = 1 observed AUC

- DeLong: SE 0,0090, CI [0,494; 0,529], $z = 1{,}29$, $p = 0{,}20$; $L = 1$: SE 0,0092, CI [0,493; 0,530], ca DeLong
- $L = 20$: SE 0,0150, CI [0,483; 0,541]; 21,9% dintre replici $\leq 0{,}5$; $L = 60$: SE 0,0141, CI [0,485; 0,540]
- Interpretare: etichetele suprapuse pe 5 zile și gruparea volatilității (*volatility clustering*) fac zilele vecine dependente; SE robustă la dependență este de 1,7 ori SE DeLong. Un bloc cel puțin egal cu orizontul etichetei, stabil între 20 și 60 de zile: $L = 20$ se poate justifica. Histogramele sînt distribuții ale estimatorului, nu distribuții sub ipoteza nulă, cu modelele estimate ținute fixe

B4 [Propus]: Diebold–Mariano pentru prognoze de probabilitate — cerința

- **Model**: B3a și B3b [Rezolvat]
- **Întrebarea**: prognozează vreun model probabilitatea unei săptămîni „sus" mai bine decît o constantă?
- **Date de intrare**: probabilitățile salvate în B3a, limitate la $[10^{-6}, 1-10^{-6}]$; log loss ℓ_t din secțiunea „Log loss"; rînd rezolvat, 13 iul. 2015, $y_t = 0$: constantă $\hat{p} = 0{,}529$, $\ell = -\ln(1-0{,}529) = 0{,}753$; RF $\hat{p} = 0{,}4953$, $\ell = 0{,}684$
- **Sarcini**
  1. Calculați log loss mediu pentru Logit-L1, RF, GB și constantă.
  2. Pentru RF față de Logit-L1, RF față de constantă și GB față de constantă, formați $d_t = \ell_t^A - \ell_t^B$ ($d_t > 0$: modelul A mai slab) și regresați d_t pe o constantă cu erori Newey–West (Newey & West (1987)), laguri 4 și 20: statistica DM este valoarea t (Diebold & Mariano (1995)), bilateral.
  3. Calculați statistica t naivă i.i.d. a lui $\bar{d}$ și comparați.
- **Raportați**: un tabel cu log loss, $\bar{d}$, DM (laguri 4, 20) cu valorile p, t naiv; un grafic cu d_t cumulat
- **Interpretare**: poate un model cu AUC peste 0,5 să aibă un log loss mai mare decît o constantă? Probabilitățile din CV agregat sînt descriptive: o afirmație despre utilizare cere prognoze evaluate pe ferestre succesive. Notebook: secțiunea B4

◂ Înapoi ▸ Lista exercițiilor

C1 [Propus]: critica unui răspuns AI — răspunsul

- **Promptul** trimis unui asistent AI: „Prognozează direcția Bitcoin de mîine din randamentele trecute și evaluează modelul în afara eșantionului, cu validare încrucișată în 5 pliuri”
- **Răspunsul** (`btc`: închideri zilnice 2014-09-17 – 2026-09-18; fără importuri):

```
r = np.log(btc).diff()
X = pd.DataFrame({'r1': r, 'r5': r.rolling(5).sum(), 'vol20': r.rolling(20).std(),
                  'trend10': r.rolling(10, center=True).mean()})
data = pd.concat([X, r.shift(-1).rename('fwd')], axis=1).dropna()
y = (data['fwd'] > 0).astype(int)                # 1 if tomorrow is up
Xs = StandardScaler().fit_transform(data[X.columns])
cv = KFold(n_splits=5, shuffle=True, random_state=0)
model = LogisticRegression(penalty='l1', C=0.1, solver='liblinear')
auc = cross_val_score(model, Xs, y, cv=cv, scoring='roc_auc')
print(round(auc.mean(), 3))
```

- „Rezultat: 0,697. Un AUC de 0,70 în afara eșantionului, pe 4 361 de zile, este un semnal puternic, tranzacționabil; validarea încrucișată este în afara eșantionului, deci nu există supraajustare”

◂ Înapoi ▸ Lista exercițiilor

C1 [Propus]: critica unui răspuns AI — cerința

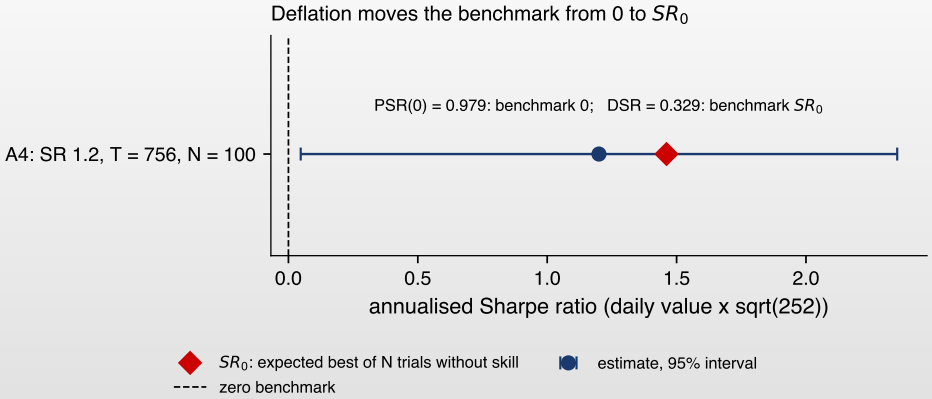
- **Model**: B2 și B3a [Rezolvat]; B9 (extindere) este aceeași capcană într-un model mai mare
- **Sarcini**
  1. Dați cinci constatări adnotate, fiecare cu numărul liniei: două surse de scurgere de informație din test, o schemă de validare care nu corespunde utilizării reale, două afirmații nesusținute.
  2. Pentru fiecare, spuneți cum ați detecta-o: o verificare a rezultatului, un rezultat din B2 sau B3a, o regulă din secțiunea „Noțiuni și notație".
  3. Scrieți un singur pipeline corectat: caracteristici calculate doar din trecut, scalarea în fiecare pliu de antrenare (`make_pipeline`), `TimeSeriesSplit(n_splits=5, gap=20)`, AUC mediu pe cele 5 pliuri de test.
  4. Raportați un tabel înainte/după: codul AI, apoi fiecare corecție adăugată pe rînd.
  5. Treceți erorile în formatul `AI_ERRORS.md`: cererea, răspunsul, eroarea, cum ați găsit-o, corecția.
- **Interpretare**
  - ce zile intră în `trend10` în ziua t? Verificați alinierea pe seria simplă $0, 1, \ldots, 30$, cu același apel `rolling`.
  - care dintre ele este ținta?
- Un AUC neobișnuit de mare este un motiv să verificați codul, nu o dovadă de scurgere de informație

A4 [Rezolvat]: PSR și DSR, fișă de calcul — cerința

- **Întrebarea**: este convingător un raport Sharpe anualizat de 1,2 pe trei ani, dacă este cel mai bun din 100 de încercări?
- **Date de intrare** (strategie pe acțiuni, $A = 252$ de zile de tranzacționare pe an): SR anualizat $\widehat{SR}_{an} = 1{,}2$, $T = 756$ randamente zilnice, $\hat{\gamma}_3 = -0{,}4$, kurtosis obișnuit $\hat{\gamma}_4 = 5$; $N = 100$ încercări independente, $V = \mathrm{Var}[\widehat{SR}_n] = 1/T$ (Bailey & López de Prado (2012); Bailey & López de Prado (2014))
- **Sarcini**
  1. Calculați SR zilnic $\widehat{SR} = 1{,}2/\sqrt{252}$ și numitorul D din secțiunea „Raportul Sharpe probabilistic".
  2. Calculați $z = \widehat{SR}\sqrt{T-1}/D$ și $PSR(0) = \Phi(z)$.
  3. Calculați maximul așteptat SR_0 din secțiunea „Selecția și raportul Sharpe deflatat", zilnic și anualizat.
  4. Calculați $DSR = PSR(SR_0) = \Phi\big((\widehat{SR} - SR_0)\sqrt{T-1}/D\big)$.
- **Raportați**: fiecare număr intermediar, PSR(0) și DSR; derivarea lui D este în extindere (A4 extins)
- **Interpretare**
  - pentru randamente i.i.d., reper zero și 100 de încercări, ce spune DSR?
  - ce N l-ar schimba?
- Notebook: secțiunea A4

◂ Înapoi ▸ Lista exercițiilor

A4 [Rezolvat]: rezolvare (1/2) — PSR și DSR, fișă de calcul



A4 [Rezolvat]: rezolvare (2/2) — PSR și DSR, fișă de calcul

1. $\widehat{SR} = 1{,}2/\sqrt{252} = 0{,}07559$; $D = \sqrt{1 + 0{,}4 \cdot 0{,}07559 + 1 \cdot 0{,}07559^2} = 1{,}01782$
2. $z = 0{,}07559 \cdot \sqrt{755}/1{,}01782 = 2{,}041$; $PSR(0) = \Phi(2{,}041) = 0{,}979$
3. $\sqrt{V} = 1/\sqrt{756} = 0{,}03637$; $\Phi^{-1}(0{,}99) = 2{,}326$, $\Phi^{-1}(1 - \frac{1}{100e}) = 2{,}680$;
   $SR_0 = 0{,}03637\,[0{,}4228 \cdot 2{,}326 + 0{,}5772 \cdot 2{,}680] = 0{,}0920$ pe zi, 1,46 anualizat
4. $z = (0{,}07559 - 0{,}0920)\sqrt{755}/1{,}01782 = -0{,}444$; $DSR = \Phi(-0{,}444) = 0{,}329$

⊡ Interpretare: numai prin noroc, maximul așteptat al SR anualizat din 100 de încercări i.i.d. fără abilitate este 1,46

- peste valoarea observată de 1,2: dovada nu atinge pragul ilustrativ de 0,95
- cu $N = 1$, luați direct $SR_0 = 0$ (nu evaluați aproximația cu cuantile pentru N mare): DSR este egal cu $PSR(0) = 0{,}979$

MFM_ch13_seminar

Proiectul de echipă: alegeți un traseu

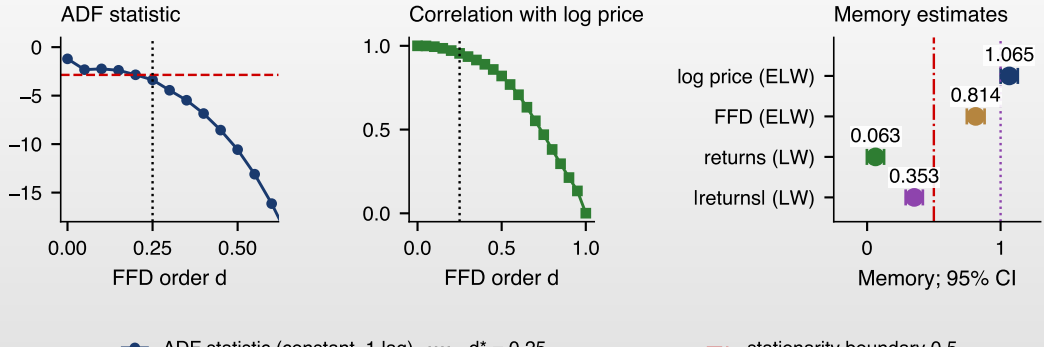
- **Standard**: un activ la alegere, prin B1, B3a, B3b, B5 și B6, cu matricea randamentelor tuturor regulilor încercate salvată
- **Avansat**: C2 (model fundațional față de boosting) sau C3 (un factor stocastic de actualizare, SDF, fără arbitraj, pe decile), în pachetele de extindere
- Dovezi minime: codul ca Quantlet, un grafic, testul față de un reper, numărul de încercări; folosirea AI în `AI_USE.md`, erorile depistate în `AI_ERRORS.md`
- Precizați sursa datelor și data descărcării; fixați eșantionul, raportați numărul de observații și prima/ultima închidere cu datele lor. Un rezultat negativ documentat riguros este valid.
- **Criterii de evaluare**: etichete și validare fără scurgere de informație; incertitudinea fiecărui rezultat (CI, DSR); numărul de încercări raportat; codul reproductibil; principala limită precizată

B5 [Rezolvat]: memoria log-prețului Bitcoin — cerința

- **Întrebările**
  - ce d alege regula ADF minimă pentru Bitcoin?
  - care este memoria δ a log-prețului?
- **Date**: $x_t = \ln P_t$, BTC 2014-09-17 – 2026-09-18 ($n = 4\,385$); funcții disponibile: `frac_diff_ffd`, `exact_local_whittle`, `local_whittle`
- **Sarcini**
  1. Pentru $d \in \{0; 0{,}05; \ldots; 1\}$, filtrați x_t ($\tau = 10^{-4}$), rulați `adfuller(x, maxlag=1, regression='c', autolag=None)` (Dickey & Fuller (1979)) și notați statistica și $\mathrm{Corr}(\tilde{x}_t, x_t)$ pe datele păstrate de filtru.
  2. Raportați d^*, cel mai mic d cu statistica sub valoarea critică la 5%.
  3. Estimați δ cu CI 95% ($m = \lfloor n^{0{,}65} \rfloor$ frecvențe Fourier pentru fiecare serie, ca în „Memoria, testul ADF și local Whittle"): local Whittle exact (ELW, *Exact Local Whittle*; Shimotsu & Phillips (2005)) pentru x_t și seria FFD la d^*, local Whittle (LW, Robinson (1995)) pentru r_t și $|r_t|$.
  4. Testați $H_0 : \delta = 1$ pentru x_t cu $z = (\hat{\delta} - 1)/\mathrm{SE}$, din estimarea nerotunjită.
- **Raportați**: graficul ADF–corelație, d^*, un tabel cu cele patru estimări și CI, z și p
- **Interpretare**
  - este seria FFD la d^* staționară?
  - unde se află memoria Bitcoin?
- Notebook: secțiunea B5

◂ Înapoi ▸ Lista exercițiilor

B5 [Rezolvat]: rezolvare (1/2) — memoria log-prețului Bitcoin



B5 [Rezolvat]: rezolvare (2/2) — memoria log-prețului Bitcoin

- ADF: $-1{,}21$ ($d = 0$); $-2{,}32$ (0,05); $-2{,}24$ (0,10); $-2{,}39$ (0,15); $-2{,}85$ (0,20); $-3{,}38$ (0,25); valoarea de 5% $-2{,}86$: $d^* = 0{,}25$, corelația 0,955; traiectoria nu este monotonă în d
- ELW log-preț 1,065 [1,000; 1,129] ($m = 232$); FFD la d^* 0,814 [0,747; 0,880]; LW randamente 0,063 [$-0{,}002$; 0,127]; $|r_t|$ 0,353 [0,289; 0,417]
- $z = (1{,}0647 - 1)/0{,}0328 = 1{,}97$, p bilateral $= 0{,}049$: o respingere la limită spre $\delta > 1$ (memorie peste rădăcina unitară), nu spre staționaritate
- Interpretare: seria FFD are $\hat{\delta} \approx 1 - d$ în banda de frecvențe, iar CI-ul ei este mult peste $\frac{1}{2}$: nu este staționară; memoria lungă a Bitcoin este în volatilitate, $|r_t|$

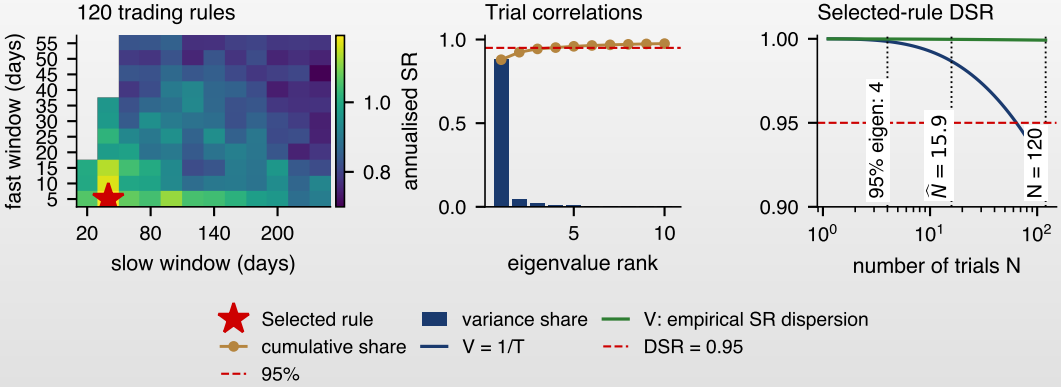
MFM_ch13_seminar

B6 [Rezolvat]: căutare în grilă, DSR și numărul efectiv de încercări — cerința

- **Întrebarea**: după căutarea printre 120 de reguli de medii mobile pe Bitcoin, depășește cea mai bună dintre ele norocul?
- **Reguli**: fereastra rapidă $f \in \{5, 10, \ldots, 55\}$, lentă $s \in \{20, 40, \ldots, 240\}$, $s > f$: $N = 120$; semnalul din ziua t: $1\{\overline{x}_{f,t} > \overline{x}_{s,t}\}$ ($\overline{x}_{f,t}$: media log-prețului pe ultimele f zile), deținut în ziua $t+1$: $r^{(f,s)}_{t+\mathbf{1}} = \text{semnal}_t \cdot r_{t+\mathbf{1}}$; long/numerar, fără costuri, rata fără risc 0
- **Date**: start comun în prima zi în care fiecare regulă are semnal, 15 mai 2015, pînă la 18 sep. 2026: $T = 4\,145$ zile, $A = 365$
- **Sarcini**
  1. Construiți matricea de randamente $T \times 120$; alegeți regula cu cel mai bun SR; raportați SR, asimetria, kurtosis și PSR(0) (Bailey & López de Prado (2012)).
  2. Din matricea de corelație 120×120: valorile proprii, numărul care atinge 95% din varianță și $\hat{N} = \bar{\rho} + (1 - \bar{\rho})N$, cu $\bar{\rho}$ corelația medie din afara diagonalei (euristici, nu teoreme).
  3. Calculați $SR_{\mathbf{0}}$ și DSR (Bailey & López de Prado (2014)) pentru $N = 120$, $\hat{N}$ și numărul de la 95%, cu $V = 1/T$.
- **Raportați**: o hartă a rapoartelor Sharpe cu regula aleasă, graficul valorilor proprii, un tabel DSR
- **Interpretare**: ce N și ce V ați raporta unui comitet de investiții? Notebook: secțiunea B6

◂ Înapoi ▸ Lista exercițiilor

B6 [Rezolvat]: rezolvare (1/2) — căutare în grilă, DSR și numărul efectiv de încercări



B6 [Rezolvat]: rezolvare (2/2) — căutare în grilă, DSR și numărul efectiv de încercări

$V = 1/T$	$N = 120$	$\hat{N} = 15{,}9$	$N = 4$
SR_0 (an.)	0,77	0,53	0,31
DSR	0,922	0,987	0,998

- Cea mai bună regulă $(5, 40)$: SR 1,19 anualizat (buy-and-hold 0,78); asimetrie 0,13, kurtosis 12,8; $\mathrm{PSR}(0) = 1{,}000$
- Execuție idealizată la aceeași închidere; cu tranzacția la închiderea următoare, SR anualizat al regulii $(5, 40)$ este 1,18
- Cea mai mare valoare proprie, $\lambda_1 = 105{,}5$, explică 87,9%; 4 valori proprii ating 95%; $\bar{\rho} = 0{,}875$, $\hat{N} = 15{,}9$
- Interpretare: cu $V = 1/T$, tratarea celor 120 de încercări corelate ca independente deflatează excesiv. Curba cu dispersia estimată se citește la $N = 120$; reducerea suplimentară a lui N poate număra corelația de două ori
  - DSR testează „mai bine decît norocul", nu „mai bine decît buy-and-hold" (B13)
  - regula cîștigătoare a fost aleasă pe tot eșantionul: o afirmație cere o regulă fixată și un eșantion de verificare neatins (Capitolul 13, grila Bitcoin)

MFM_ch13_seminar

Pachetele de extindere: condițiile fiecăruia

Pachet	Exerciții	Condiții
1. Memorie	A5, B7	A1, B5
2. Etichete	A2 extins [Rezolvat], A6, B8	A2, B1
3. Validare și scurgere de informație	A7, B9	A3, B2, cadrul de modelare
4. Inferență predictivă	A8 [Rezolvat], A9, B3c [Rezolvat], B10 [Rezolvat]	B3a, B3b, B4
5. Strategii și selecție	A4 extins [Rezolvat], A10, B11 [Rezolvat], B12, B13	A4, B6
6. Inferență cu multe controale	B14	exemplul din curs (VIX)
7. Proiecte	C2, C3	Capitolul 14 (C2); studiul de caz din curs (C3)

- Fiecare exercițiu are secțiunea lui în notebook, cu aceeași etichetă; rezolvați-le în ordine, în cadrul unui pachet
- C2 și C3 sînt puncte de plecare posibile pentru proiectul de echipă

A5 [Propus]: lungimea ferestrei și respingerile ADF prea dese — cerința

- **Model**: A1 [Rezolvat]
- **Întrebarea**: cît de des respinge ADF rădăcina unitară în seria FFD a unui mers aleator pur?
- **Simulare** (`rng = default_rng(42)`): pentru $d \in \{0{,}3; 0{,}5\}$ și fiecare dintre 500 de traiectorii, calculați K, extrageți $T + K$ inovații $\varepsilon_s \sim N(0,1)$ cu $T = 2\,500$, puneți $X_t = \sum_{s \le t} \varepsilon_s$ (X pornește de la prima extragere), aplicați filtrul finit și păstrați exact T valori; apoi repetați cu $d = 0{,}25$ și fără filtru (control)
- **Sarcini**
  1. Rulați `adfuller(f, maxlag=1, regression='c', autolag=None)` pe fiecare traiectorie și notați respingerile la 5%.
  2. Pe aceleași traiectorii, rulați testul cu laguri alese prin criteriul informațional Akaike (AIC, `autolag='AIC'`), cu maximul implicit statsmodels: rotunjirea în sus a lui $12(T/100)^{1/4}$, unde T este lungimea seriei filtrate.
  3. Raportați fiecare frecvență de respingere cu eroarea ei Monte Carlo $\sqrt{\hat{p}(1-\hat{p})/500}$.
- **Interpretare**: fiecare traiectorie este $I(1)$
  - de ce nu distinge testul ADF cu un lag un mers aleator scalat de o serie staționară?
  - ce d ar face staționar un preț $I(1)$?
- Notebook: secțiunea A5

◂ Înapoi ▸ Lista exercițiilor

B7 [Propus]: robustețea regulii ADF minime pentru Ether — cerința

- **Model**: B5 [Rezolvat]
- **Întrebarea**: cît de robustă este regula ADF minimă pe un activ care a crescut de circa 1 000 de ori?
- **Date**: log-închideri ETH-USD; trei specificații, fiecare cu scanarea completă a grilei $d \in \{0; 0{,}05; \ldots; 1\}$ și un lag ADF

Rînd	Eșantion	Termeni deterministici	Valoarea critică 5%
1	2015-08-07 – 2026-09-18 ($n = 4\,061$)	constantă	$-2{,}86$
2	2015-08-07 – 2026-09-18	constantă + trend	$-3{,}41$
3	2018-01-01 – 2026-09-18 ($n = 3\,183$)	constantă	$-2{,}86$

- **Sarcini**
  1. Pentru fiecare rînd, găsiți d^* și notați valorile ADF la $d = 0$ și $d = 0{,}05$.
  2. Estimați ELW cu CI 95% pentru log-preț pe rîndurile 1 și 3.
  3. Estimați LW cu CI 95% pentru $|r_t|$ pe tot eșantionul.
- **Raportați**: un tabel cu d^* și cele două valori ADF pe fiecare rînd; cele două estimări ELW și estimarea LW, cu CI
- **Interpretare**
  - ce d ați folosi într-un model?
  - de ce acel d?
- Notebook: secțiunea B7

înapoi Lista exercițiilor

A2 extins [Rezolvat]: probabilități de atingere — cerința

- **Întrebarea**: ce proporție de evenimente ar trebui să se încheie la bariera verticală dacă prețurile ar fi o mișcare browniană?
- **Cadru**: log-prețul σW_t, W_t o mișcare browniană standard (fără tendință), barierele $+a = pt\,\sigma\sqrt{h}$ și $-b = -sl\,\sigma\sqrt{h}$; τ = primul moment în care una este atinsă; martingalele W_t și $W_t^2 - t$ (procese a căror valoare viitoare așteptată este cea curentă); moment de oprire τ cu $\mathbb{E}[\tau] < \infty$; teorema de oprire opțională: atunci $\mathbb{E}[W_\tau] = 0$ și $\mathbb{E}[W_\tau^2] = \mathbb{E}[\tau]$ (Karatzas & Shreve (1998))
- **Sarcini**
  1. Fără barieră verticală, derivați P(sus prima) și $\mathbb{E}[\tau]$ prin oprire opțională.
  2. Pentru $pt = sl = 1$ și $h = 10$, calculați $P(\tau > h)$ cu monitorizare continuă, cu $P(\tau > h) = \frac{4}{\pi}\sum_{n\geq 0}\frac{(-1)^n}{2n+1}e^{-(2n+1)^2\pi^2/8}$.
  3. Aproximați monitorizarea zilnică mutînd fiecare barieră în afară cu $0{,}5826\,\sigma$ (Broadie, Glasserman & Kou (1997)); apoi simulați 400 000 de traiectorii de 10 pași zilnici gaussieni (seed 42) și comparați.
- **Raportați**: P(sus prima) și $\mathbb{E}[\tau]$; $P(\tau > h)$ cu monitorizare continuă, aproximat și simulat, cu SE simulării
- **Interpretare**: B1 găsește 59,0% ieșiri verticale pe Bitcoin: ce parte din diferența față de valoarea browniană ține de monitorizare și ce parte de date?

Înapoi Lista exercițiilor

A2 extins [Rezolvat]: rezolvare — probabilități de atingere

1. W_t este martingal: $0 = p\,a - (1-p)\,b \Rightarrow p = \frac{b}{a+b} = \frac{sl}{pt+sl}$; $\sigma^2 W_t^2 - \sigma^2 t$ este martingal: $\sigma^2 \mathbb{E}[\tau] = p\,a^2 + (1-p)\,b^2 = ab \Rightarrow \mathbb{E}[\tau] = pt \cdot sl \cdot h$ zile
2. Monitorizare continuă, exact: $P(\tau > h) = 1{,}2732 \cdot 0{,}2912 - \cdots = 0{,}371$
3. Monitorizare zilnică, aproximativ: $c = 1 + 0{,}5826/\sqrt{10} = 1{,}184$, înlocuind $\pi^2/8$ cu $\pi^2/(8c^2)$ obținem 0,528; simulat cu pași zilnici: 0,517 (SE 0,001)

⊡ Interpretare: monitorizarea discretă explică cea mai mare parte a diferenței dintre 37% și 59%; cele 7 puncte rămase pot proveni din cozile groase, gruparea volatilității și eroarea volatilității EWMA, pe care această comparație nu le separă

A6 [Propus]: bariere asimetrice — cerința

- **Model**: A2 și A2 extins [Rezolvat]
- **Date de intrare**: $P_{t_0} = 50$, $\sigma_{t_0} = 2\%$, $h = 4$, $pt = 2$, $sl = 1$; traiectoriile (i) 50; 51,5; 53,0; 49,5; 47,9 și (ii) 50; 52,0; 54,5; 53,0; 51,0
- **Sarcini**
  1. Calculați ambele bariere, desenați-le pe cele două traiectorii și etichetați fiecare traiectorie. O atingere orizontală în ultima zi are prioritate față de ieșirea verticală.
  2. Opțional (tendință): pentru $X_t = \mu t + \sigma W_t$ (μ: tendința pe zi) fără barieră verticală, funcția de scală $s(x) = e^{-2\mu x/\sigma^2}$ rezolvă $\frac{\sigma^2}{2}s'' + \mu s' = 0$, deci $s(X_t)$ este martingal; derivați $P(\text{sus prima}) = \frac{e^{2\mu b/\sigma^2}-1}{e^{2\mu b/\sigma^2}-e^{-2\mu a/\sigma^2}}$ și evaluați-o cu $pt = 2$, $sl = 1$, schimbînd acum volatilitatea zilnică la $\sigma = 3\%$, $\mu = 0{,}1\%$ pe zi, $h = 10$.
- **Raportați**: barierele, etichetele, formula și valoarea ei
- **Interpretare**: formula nu are barieră verticală: ce traiectorii ar cenzura o barieră verticală la h? Notebook: secțiunea A6

Înapoi Lista exercițiilor

B8 [Propus]: bariere asimetrice pe Bitcoin — cerința

- **Model**: B1 [Rezolvat] și A2 extins
- **Date**: evenimentele, volatilitatea EWMA și etichetarea din B1, $h = 10$; variantele $(pt, sl) = (2, 1)$ și $(1, 2)$
- **Sarcini**
  1. Raportați ponderile ieșirilor, ponderea etichetelor $+1$ și durata medie de deținere pentru ambele variante.
  2. Calculați ponderea ieșirilor sus între ieșirile orizontale, sus/(sus + jos).
  3. Comparați-o cu $sl/(pt + sl)$ (fără barieră verticală) și cu o simulare: 400 000 de traiectorii de 10 pași zilnici gaussieni, în unități de σ, bariere la $pt\sqrt{10}$ și $-sl\sqrt{10}$, seed 42, tendință zero și tendința BTC $\bar{r}/\hat{\sigma} = 0{,}034$ pe zi; dați SE a simulării.
- **Raportați**: un tabel cu ponderile ieșirilor, ponderea $+1$ și durata medie pentru ambele variante; ponderea ieșirilor sus, cu valoarea teoretică și cele simulate, cu SE
- **Interpretare**: cum schimbă alegerea echilibrul claselor pe care îl vede un clasificator? Asimetria este o ipoteză de modelare, nu o regulă optimă de tranzacționare. Notebook: secțiunea B8

Înapoi Lista exercițiilor

B9 [Propus]: identificarea scurgerii de informație — cerința

- **Model**: B2 [Rezolvat] și C1
- **Date**: cadrul de modelare plus o a 12-a caracteristică, construită în notebook ca `ma11 = r.rolling(11, center=True).mean()`; RF cu 300 de arbori (nu 400 ca în B3a), frunza minimă 50, seed 42, 5-fold purjat; 4 181 de rînduri
- **Sarcini**
  1. Scrieți datele $t-5,\dots,t+5$ care intră în `ma11` în ziua t și marcați-le pe cele din fereastra țintei $t+1,\dots,t+5$; ordonați cele 12 caracteristici după MDI dintr-un random forest ajustat pe toate rîndurile (diagnostic descriptiv; scorurile predicției rămîn în afara pliului).
  2. Calculați corelația lui `ma11` și a mediei retrospective `r.rolling(11).mean()` cu randamentul pe 5 zile înainte.
  3. Raportați acuratețea și AUC pentru: 11 caracteristici + `ma11` centrată; 11 + retrospectivă; cele 11 de bază, pe aceleași rînduri, pliuri și setări.
- **Interpretare**: de ce nu poate purjarea să detecteze această scurgere de informație? Dați un al doilea exemplu de informație din viitor într-un flux de caracteristici. Notebook: secțiunea B9

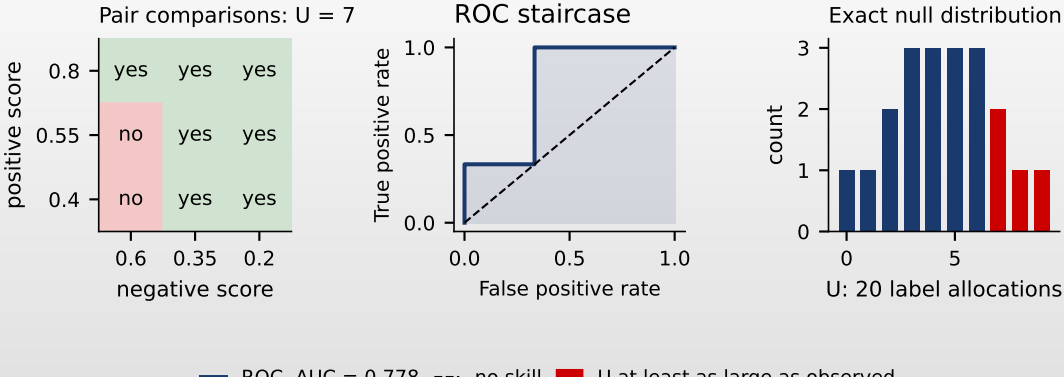
◂ Înapoi ▸ Lista exercițiilor

A8 [Rezolvat]: AUC pas cu pas — cerința

- **Întrebările**
  - ce măsoară AUC?
  - cît de incert este cu șase observații?
- **Date de intrare**: scorurile a 3 pozitive (0,80; 0,55; 0,40) și 3 negative (0,60; 0,35; 0,20); $n_1 = n_0 = 3$ pozitive și negative; $U =$ numărul perechilor (pozitiv, negativ) ordonate corect, $A = U/(n_1 n_0) = U/9$; SE Hanley–McNeil (Hanley & McNeil (1982)): $\mathrm{SE}^2 = \frac{A(1-A)+(n_1-1)(Q_1-A^2)+(n_0-1)(Q_2-A^2)}{n_1 n_0}$, $Q_1 = \frac{A}{2-A}$, $Q_2 = \frac{2A^2}{1+A}$
- **Sarcini**
  1. Completați grila de comparație 3×3 (este pozitivul peste negativ?) și calculați U și A.
  2. Desenați curba ROC în trepte și verificați că aria ei este A.
  3. Calculați SE și statistica Wald $z = (A - 0{,}5)/\mathrm{SE}$, bilateral.
  4. Enumerați cele 20 de moduri de a eticheta 3 dintre cele 6 scoruri ca pozitive, calculați U pentru fiecare și $P(U \geq 7)$ exact.
- **Raportați**: grila, U, A, curba ROC, SE, z, p și $P(U \geq 7)$ exact
- **Interpretare**: ce ipoteză a SE nu mai este îndeplinită pentru etichete suprapuse pe 5 zile? Notebook: secțiunea A8

Înapoi Lista exercițiilor

A8 [Rezolvat]: rezolvare (1/2) — AUC pas cu pas



A8 [Rezolvat]: rezolvare (2/2) — AUC pas cu pas

1. 0,80 este peste 3 negative, 0,55 peste 2; 0,40 peste 2: $U = 7$, $A = 0{,}778$; (2) aria curbei în trepte este $\frac{1}{3} \cdot \frac{1}{3} + \frac{2}{3} \cdot 1 = \frac{7}{9}$
3. $Q_1 = 0{,}636$, $Q_2 = 0{,}681$, $\mathrm{SE} = 0{,}207$, $z = 1{,}34$, $p = 0{,}18$ (asimptotic, nesigur cu șase puncte)
4. Exact: $P(U \geq 7) = 4/20 = 0{,}20$, p bilateral $= 0{,}40$ cu etichete interschimbabile: șase observații nu pot separa abilitatea de noroc

⊡ Ambele SE presupun observații independente; etichetele suprapuse au randamente comune, deci eșantionul efectiv este mai mic: B3b folosește în schimb un bootstrap pe blocuri

MFM_ch13_seminar

A9 [Propus]: AUC ca statistică U și varianța DeLong — cerința

- **Model**: A8 [Rezolvat]
- **Date de intrare**: 4 pozitive (0,70; 0,62; 0,48; 0,45); 4 negative (0,66; 0,50; 0,40; 0,30); nucleul $\psi(s_i, s_j) = 1\{s_i > s_j\} + \frac{1}{2}1\{s_i = s_j\}$, cu s_i scorul pozitivei i, s_j al negativei j, $1\{\cdot\}$ funcția indicator; $n_1 = n_0 = 4$
- **Sarcini**
  1. Completați grila 4×4 a lui ψ; calculați U, AUC și SE Hanley–McNeil.
  2. Mediile pe rînduri $V_{10}(i)$ și pe coloane $V_{01}(j)$ ale grilei (valori de plasare); variațele lor de selecție S_{10}, S_{01}, cu numitorii $n_1 - 1$ și $n_0 - 1$.
  3. SE DeLong $= \sqrt{S_{10}/n_1 + S_{01}/n_0}$ (DeLong, DeLong & Clarke-Pearson (1988)).
  4. Demonstrație opțională: prin descompunerea Hoeffding, AUC este asimptotic suma a două medii independente ale valorilor de plasare.
- **Raportați**: grila, U, AUC, valorile de plasare, S_{10}, S_{01} și cele două SE
- **Interpretare**: argumentul este asimptotic, nu o garanție exactă pentru eșantioane mici: ce înseamnă asta cu 8 scoruri? Notebook: secțiunea A9

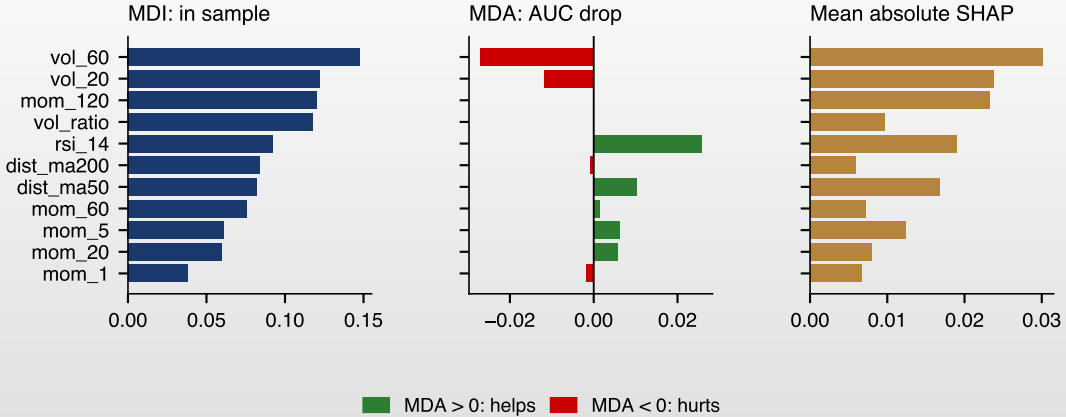
◂ Înapoi ▸ Lista exercițiilor

B3c [Rezolvat]: concordanța măsurilor de importanță — cerința

- **Întrebarea**: ordonează măsurile de importanță în eșantion și în afara lui cele 11 caracteristici la fel?
- **Date**: cadrul de modelare, cu o împărțire cronologică separată: RF (400 de arbori, frunza minimă 50) estimat pe primele 70% (pînă la 8 apr. 2023), evaluat după o pauză de 5 + 41 de zile, din 24 mai 2023; o singură împărțire, pentru că importanțele cer un singur model estimat
- **Sarcini**
  1. MDI din random forest-ul estimat; MDA ca scăderea AUC de test cînd o caracteristică este permutată (10 repetări); media |SHAP| pe 300 de zile de test (Lundberg & Lee (2017)).
  2. Corelațiile de rang Spearman între cele trei măsuri.
- **Raportați**: un grafic cu cele trei măsuri pe caracteristici și tabelul corelațiilor Spearman
- **Interpretare**
  - ce înseamnă un MDA negativ?
  - de ce estompează caracteristicile corelate toate cele trei măsuri?
- Notebook: secțiunea B3

◂ Înapoi ▸ Lista exercițiilor

B3c [Rezolvat]: rezolvare (1/2) — concordanța măsurilor de importanță



B3c [Rezolvat]: rezolvare (2/2) — concordanța măsurilor de importanță

- MDI pune primele `vol_60`, `vol_20`, `mom_120`; MDA este negativ pentru ambele caracteristici de volatilitate și cel mai mare pentru `rsi_14` (0,026)
- Spearman: MDI–SHAP 0,75, MDI–MDA −0,42, MDA–SHAP −0,10: ordonările în eșantion și în afara lui nu concordă
- Interpretare: un MDA negativ înseamnă că, pe date noi, caracteristica permutată dă un scor mai bun decît cea reală: un semn de supraajustare, în limita zgomotului de eșantionare; caracteristicile corelate își împart și își diluează importanța, iar permutarea extrapolează în afara datelor

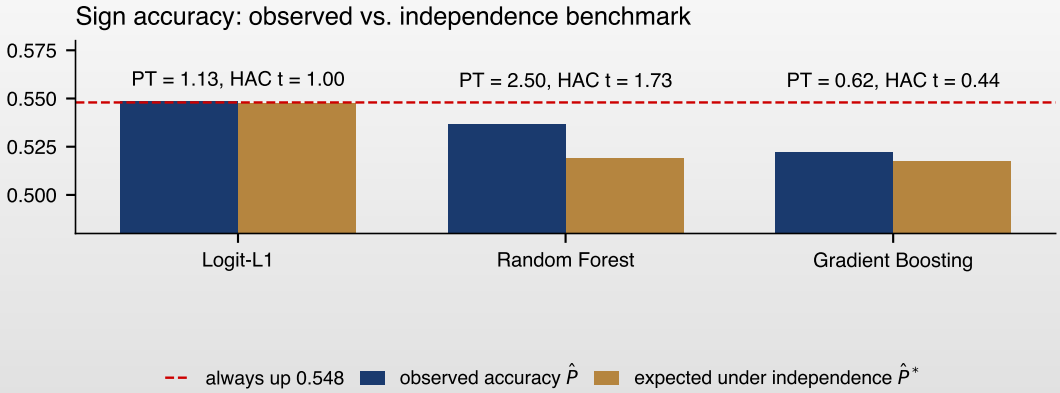
MFM_ch13_seminar

B10 [Rezolvat]: Pesaran–Timmermann și Clark–West — cerința

- **Întrebările**
  - conțin semnele prezise informație?
  - adaugă caracteristicile informație unei prognoze constante?
- **Date**: probabilitățile salvate în B3a și constanta din B4; $x_t = 1\{\hat{p}_t > 0{,}5\}$. Precizare înainte de a începe: Clark–West (CW, Clark & West (2007)) este derivat pentru modele liniare imbricate și prognoze reale în afara eșantionului; pe probabilitățile unui logit penalizat din CV purjat agregat este o corecție aproximativă, descriptivă
- **Sarcini**
  1. Pentru fiecare model: $\hat{P}$ = proporția semnelor corecte, $\hat{P}^* = \hat{p}_y\hat{p}_x + (1-\hat{p}_y)(1-\hat{p}_x)$ și statistica Pesaran–Timmermann (PT), $PT = (\hat{P} - \hat{P}^*)/\sqrt{\hat{V}(\hat{P}) - \hat{V}(\hat{P}^*)}$ (Pesaran & Timmermann (1992)); $\hat{p}_y$, $\hat{p}_x$: proporțiile de „sus" realizate și prezise; n: numărul de zile; $\hat{V}(\hat{P}) = \hat{P}^*(1-\hat{P}^*)/n$,
     $\hat{V}(\hat{P}^*) = [(2\hat{p}_y - 1)^2\hat{p}_x(1-\hat{p}_x) + (2\hat{p}_x - 1)^2\hat{p}_y(1-\hat{p}_y)]/n + 4\hat{p}_y\hat{p}_x(1-\hat{p}_y)(1-\hat{p}_x)/n^2$.
  2. Statistica t a pantei în $y_t = a + b\,x_t + u_t$ cu erori Newey–West (4 laguri), bilateral.
  3. Logit-L1 ($\hat{p}_{\mathbf{1},t}$) față de constantă ($\hat{p}_{\mathbf{0},t}$): $f_t = (y_t - \hat{p}_{\mathbf{0},t})^2 - [(y_t - \hat{p}_{\mathbf{1},t})^2 - (\hat{p}_{\mathbf{0},t} - \hat{p}_{\mathbf{1},t})^2]$; CW = t HAC pentru $\bar{f}$, unilateral (pozitiv = caracteristicile ajută); DM pe pierderea Brier, bilateral.
- **Raportați**: un tabel cu $\hat{P}$, $\hat{P}^*$, PT și t_{HAC} pe model; CW și DM, cu valorile lor p
- **Interpretare**: de ce pot DM și Clark–West să nu fie de acord la modele imbricate? Notebook: secțiunea B10

◂ Înapoi ▸ Lista exercițiilor

B10 [Rezolvat]: rezolvare (1/2) — Pesaran–Timmermann și Clark–West



B10 [Rezolvat]: rezolvare (2/2) — Pesaran–Timmermann și Clark–West

	Prezis sus	$\hat{P}$	$\hat{P}^*$	PT	t_{HAC}
Logit-L1	99,6%	0,549	0,548	1,13	1,00
RF	69,5%	0,536	0,519	2,50	1,73
GB	68,4%	0,522	0,518	0,62	0,44

- RF: acuratețea de 53,6% este sub „mereu în sus” (54,8%); $PT = 2{,}50$ respinge independența semnelor doar dacă zilele sînt independente; cu erori HAC $t = 1{,}73$, $p = 0{,}08$: nesemnificativ la 5%
- Clark–West, Logit-L1 față de constantă: $t = -0{,}44$, p unilateral $= 0{,}67$; DM pe Brier: $t = 1{,}27$ (pozitiv = logit mai slab), $p = 0{,}20$
- DM întreabă dacă logit-ul *estimat* prognozează mai bine, iar zgomotul de estimare îl dezavantajează sub ipoteza nulă imbricată; CW adaugă înapoi $(\hat{p}_0 - \hat{p}_1)^2$ și întreabă dacă coeficienții populației sînt zero; aici niciunul nu găsește semnal
- Dependența semnelor (PT) nu înseamnă depășirea reperului clasei majoritare și nici un log loss mai bun (B4)

MFM_ch13_seminar

A4 extins [Rezolvat]: varianța raportului Sharpe — cerința

- **Întrebarea**: de unde vine numitorul D din A4?
- **Cadru**: T randamente i.i.d. cu media μ, varianța σ^2, asimetria γ_3, kurtosis obișnuit γ_4; $SR = \mu/\sigma = g(m_1, m_2) = m_1/\sqrt{m_2 - m_1^2}$, cu $m_j = \mathbb{E}[r^j]$; metoda delta: $T\,\mathrm{Var}(\widehat{SR}) \approx \nabla g^\top \Sigma \nabla g$, cu ∇g gradientul lui g în raport cu (m_1, m_2) și Σ matricea de covarianță a lui (r, r^2)
- **Sarcini**
  1. Calculați gradientul ∇g.
  2. Calculați $\mathrm{Var}(r)$, $\mathrm{Cov}(r, r^2)$ și $\mathrm{Var}(r^2)$ în funcție de $\mu, \sigma, \gamma_3, \gamma_4$.
  3. Arătați că $T\,\mathrm{Var}(\widehat{SR}) \approx \nabla g^\top \Sigma \nabla g = 1 - \gamma_3 SR + \frac{\gamma_4 - 1}{4} SR^2$ (Lo (2002)).
- **Raportați**: gradientul, cele trei momente și derivarea pas cu pas
- **Interpretare**: ce moment face ca o strategie cu asimetrie negativă să pară mai precisă decît este?

Înapoi Lista exercițiilor

A4 extins [Rezolvat]: rezolvare — varianța raportului Sharpe

1. $\nabla g = \left(\frac{1}{\sigma} + \frac{\mu^2}{\sigma^3},\ -\frac{\mu}{2\sigma^3}\right)$
2. $\mathrm{Var}(r) = \sigma^2$, $\mathrm{Cov}(r, r^2) = \sigma^3\gamma_3 + 2\mu\sigma^2$, $\mathrm{Var}(r^2) = \sigma^4(\gamma_4 - 1) + 4\mu\sigma^3\gamma_3 + 4\mu^2\sigma^2$
3. Cu $S = \mu/\sigma$ și $(a, b) = \nabla g$, cele trei produse sînt

$$a^2\mathrm{Var}(r) = 1 + 2S^2 + S^4,$$
$$2ab\mathrm{Cov}(r, r^2) = -S\gamma_3 - 2S^2 - S^3\gamma_3 - 2S^4,$$
$$b^2\mathrm{Var}(r^2) = S^2(\gamma_4 - 1)/4 + S^3\gamma_3 + S^4.$$

4. În $\nabla g^\top \Sigma \nabla g$ termenii în μ^2/σ^2 se reduc împreună cu termenii cubici și de gradul patru și rămîne $1 - \gamma_3 SR + \frac{\gamma_4 - 1}{4} SR^2$; pentru distribuția Normală ($\gamma_3 = 0$, $\gamma_4 = 3$) $1 + \frac{1}{2} SR^2$

⊡ Interpretare: cu $SR > 0$, asimetria negativă adaugă $|\gamma_3| SR$ la varianță, iar cozile groase adaugă $\frac{\gamma_4 - 3}{4} SR^2$: ignorarea lor subestimează SE; A4 a folosit $\gamma_3 = -0{,}4$, $\gamma_4 = 5$

A10 [Propus]: DSR pe un eșantion mai scurt și constanta Gumbel — cerința

- **Model**: A4 și A4 extins [Rezolvat]
- **Date de intrare**: $\widehat{SR}_{an} = 1{,}5$, $T = 504$ randamente zilnice ($A = 252$ pe an), $\hat{\gamma}_3 = -0{,}2$, $\hat{\gamma}_4 = 6$, $N = 20$ încercări independente, $V = 1/T$; constanta lui Euler $\gamma_E = 0{,}5772$ (nu asimetria)
- **Sarcini**
  1. Completați fișa din A4: $\widehat{SR}$, D, $PSR(0)$, SR_0 (zilnic și anualizat), DSR.
  2. Opțional: cu parametrul de poziție $b_N = \Phi^{-1}(1 - \frac{1}{N})$ și scala $a_N = \Phi^{-1}(1 - \frac{1}{Ne}) - b_N$, maximul $\max_n Z_n$ a N extrageri independente Z_n din distribuția Normală standard are aproximativ o distribuție Gumbel. Folosiți $\mathbb{E}[M] = b + \gamma_E a$ pentru $M \sim \mathrm{Gumbel}(b, a)$; arătați că media lui este $(1 - \gamma_E) b_N + \gamma_E \Phi^{-1}(1 - \frac{1}{Ne})$.
  3. Opțional: verificați prin simulare pentru $N = 100$, 20 000 de extrageri, seed 42, și desenați distribuția.
- **Raportați**: fișa completată, media Gumbel și media simulată, cu SE
- **Interpretare**
  - este dovada mai puternică decît în A4?
  - trece de 0,95?
- Notebook: secțiunea A10

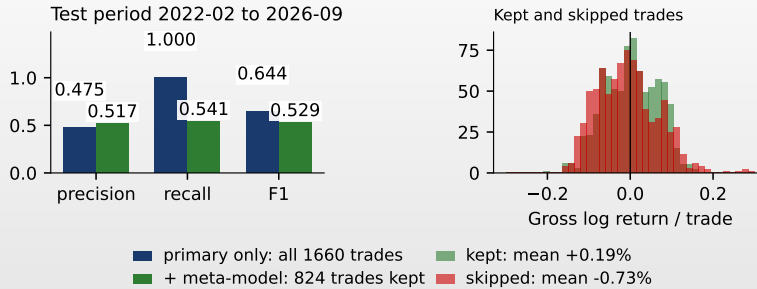
Înapoi Lista exercițiilor

B11 [Rezolvat]: meta-etichetarea unei reguli de medii mobile — cerința

- **Întrebarea**: poate un clasificator să indice cînd merită urmată o regulă simplă de trend?
- **Direcția primară**: $s_t = \text{sign}(MA_{20}(P)_t - MA_{50}(P)_t)$ pe prețuri, o tranzacție candidată în fiecare zi eligibilă (o direcție zilnică permanentă, nu doar la încrucișări); ieșire prin cele trei bariere $pt = sl = 1$, $h = 10$; meta-eticheta brută $1\{s_t \cdot r_{t_0,t_1} > 0\}$ (López de Prado (2018)); $MA_k(P)_t$: media închiderilor pe ultimele k zile; r_{t_0,t_1}: randamentul logaritmic de la intrare la ieșire
- **Modelul secundar**: RF (400 de arbori, frunza minimă 50, ponderi echilibrate ale claselor) pe cele 11 caracteristici plus s_t; antrenare pe primele 60% din 4 174 de evenimente (pînă la 9 feb. 2022), pauză 10 zile, test 20 feb. 2022 – 6 sep. 2026 (1 660 de tranzacții)
- **Sarcini**
  1. Calculați matricea de confuzie a regulii „intrăm în tranzacție dacă $\hat{p} > 0{,}5$".
  2. Calculați precizia, recall-ul și F1 înainte de filtrare (toate tranzacțiile) și după.
  3. Calculați randamentul brut pe tranzacție pentru tranzacțiile păstrate și pentru cele omise.
- **Raportați**: matricea de confuzie; un tabel cu precizia, recall-ul și F1 înainte și după; randamentul pe tranzacție al celor două grupuri
- **Interpretare**: tranzacțiile se suprapun: acesta nu este un backtest de portofoliu; ce ar adăuga o regulă de alocare a capitalului și costurile? Notebook: secțiunea B11

◂ Înapoi ▸ Lista exercițiilor

B11 [Rezolvat]: rezolvare — meta-etichetarea unei reguli de medii mobile



1. Executate: 426 cîștigătoare, 398 pierzătoare; omise: 362 cîștigătoare, 474 pierzătoare (824 din 1 660 păstrate, 49,6%)
2. Înainte: precizie 0,475, recall 1 (fiecare cîștigătoare este executată), F1 0,644; după: 0,517; 0,541; 0,529
3. Randamentul logaritmic brut pe tranzacție: toate $-0{,}27\%$, păstrate $+0{,}19\%$, omise $-0{,}73\%$

⊡ Interpretare: filtrarea crește precizia și randamentul pe tranzacție, iar F1 scade; pașii următori sînt un test bootstrap pe blocuri al acestui randament, costurile, o regulă de alocare pentru tranzacțiile suprapuse și un DSR pe toate variantele de meta-model încercate

MFM_ch13_seminar

B12 [Propus]: probabilitatea de supraajustare a backtest-ului — cerința

- **Model**: B6 [Rezolvat]; PBO (probabilitatea de supraajustare a backtest-ului, *Probability of Backtest Overfitting*) este definită la sarcina 3
- **Date**: matricea de randamente $4\,145 \times 120$ din B6, împărțită în $S = 16$ blocuri contigue cu `np.array_split` (un bloc de 260 de zile, cincisprezece de 259) (Bailey, Borwein, López de Prado & Zhu (2017))
- **Exercițiu pregătitor**: folosiți primele patru coloane de strategii și primele patru blocuri consecutive ale matricei date; pentru fiecare dintre cele șase alegeri a două blocuri de antrenare, raportați cîștigătorul, rangul lui pe celelalte două blocuri și PBO rezultat; calculați raportul Sharpe din randamentele zilnice selectate
- **Sarcini**
  1. Pentru fiecare dintre cele $\binom{16}{8} = 12\,870$ alegeri de 8 blocuri în eșantion, alegeți regula cu cel mai bun SR în eșantion.
  2. Ordonați crescător toate cele 120 de reguli în afara eșantionului (ranguri medii la egalitate), $\omega = \text{rang}/(N+1)$, $\lambda = \ln \frac{\omega}{1-\omega}$.
  3. Raportați $\text{PBO} = P(\lambda \leq 0)$, λ median, SR median în eșantion (IS, in-sample) și în afara eșantionului (OOS, out-of-sample) al cîștigătorului, proporția cu SR OOS < 0 și histograma lui λ.
- **Interpretare**: PBO este proporția cîștigătorilor selectați aflați la sau sub mediana OOS, nu probabilitatea de a pierde bani; este grila afectată de supraajustare? Notebook: secțiunea B12

◂ Înapoi ▸ Lista exercițiilor

B13 [Propus]: Reality Check, SPA și StepM pe cele 120 de reguli — cerința

- **Model**: B6 [Rezolvat]; funcții disponibile: `stationary_bootstrap_index` și `snooping_tests` (notebook, secțiunea B13)
- **Ipoteza nulă**: $H_0 : \max_k \mathbb{E}[d_{k,t}] \le 0$ cu $d_{k,t} = r_{k,t} - r_{\mathbf{0},t}$; repere numerar ($r_{\mathbf{0}} = 0$) și buy-and-hold ($r_{\mathbf{0}} = r_t$); aceste teste compară randamentele *medii*, DSR privește raportarele Sharpe
- **Bootstrap** (Politis & Romano (1994)): un singur șir de indici pe replică, aplicat tuturor coloanelor (zile întregi), bloc mediu de 20 de zile (în fiecare zi un bloc se încheie cu probabilitatea $1/20$), 2 000 de replici, seed 42; $\hat{\omega}_k = \sqrt{n}\cdot$ abaterea standard bootstrap a lui $\bar{d}_k$
- **Sarcini**
  1. Reality Check (RC, White (2000)), cu $\bar{d}_k^*$ media regulii k într-o replică și n numărul de zile: $p =$ proporția replicilor cu $\max_k \sqrt{n}(\bar{d}_k^* - \bar{d}_k) \ge \max_k \sqrt{n}\,\bar{d}_k$.
  2. SPA (*Superior Predictive Ability*, Hansen (2005)): studentizat, $t_k = \sqrt{n}\,\bar{d}_k/\hat{\omega}_k$, recentrînd doar regulile cu $t_k \ge -\sqrt{2\ln\ln n}$; StepM (Romano & Wolf (2005)) la 5%: respingem fiecare regulă peste cuantila de 95% a maximului pe regulile active, repetăm pînă nu mai respingem.
  3. Raportați pe fiecare reper: cea mai mare medie studentizată, ambele valori p, numărul ipotezelor nule respinse; desenați distribuția bootstrap a maximului.
- **Interpretare**
  - ce metodă tratează încercările corelate fără să aleagă N?
  - depășește vreo regulă deținerea Bitcoin?

◂ Înapoi ▸ Lista exercițiilor

B14 [Propus]: selecție dublă, puterea predictivă a VIX pentru Bitcoin — cerința

- **Model**: B10 [Rezolvat] (regresie cu erori HAC); penalizarea și selecția dublă sînt definite mai jos
- **Precizare**: penalizarea plug-in presupune scoruri independente; cu rezultate suprapuse pe 5 zile, intervalul de mai jos este o aproximare ilustrativă, nu un CI de 95% validat
- **Date**: cadrul de modelare ($n = 4\,181$); y_t = randamentul logaritmic BTC pe 5 zile înainte, în puncte de bază (bp, basis points; 1 bp = 0,01%); d_t = log VIX (*CBOE Volatility IndeX*, indicele de volatilitate CBOE) la închiderea din aceeași dată calendaristică, reportat peste weekenduri și sărbătorile din SUA (ultima închidere cunoscută la închiderea BTC), standardizat; controale: cele 11 caracteristici, pătratele și produsele lor două cîte două ($p = 77$), standardizate
- **Penalizarea** (Belloni, Chernozhukov & Hansen (2014)): $\min_\beta \frac{1}{n}\sum_t (y_t - x_t^\top \beta)^2 + \frac{\lambda}{n}\sum_j \hat{\psi}_j |\beta_j|$, $\lambda = 2c\sqrt{n}\,\Phi^{-1}(1-\gamma/(2p))$, $c = 1{,}1$, $\gamma = 0{,}1/\ln n$, ponderile $\hat{\psi}_j$ (scala zgomotului pentru controlul j) iterate; dată ca `rlasso`
- **Sarcini**
  1. Estimați un LASSO (Least Absolute Shrinkage and Selection Operator) pentru y_t pe (d_t, x_t): este păstrat VIX?
  2. Găsiți $\hat{S}_y$ și $\hat{S}_d$, controalele păstrate de un LASSO al lui y_t, respectiv al lui d_t, pe x_t.
  3. Estimați OLS (Ordinary Least Squares) pentru y_t pe d_t și $x_{\hat{S}_y \cup \hat{S}_d}$ cu erori HAC (4 laguri): $\hat{\theta}$ (coeficientul lui d_t), SE și intervalul; repetați OLS doar cu VIX și cu toate cele 77 de controale.
- **Raportați**: controalele selectate; un tabel cu $\hat{\theta}$, SE și intervalul celor trei regresii
- **Interpretare**: de ce nu dă LASSO-ul simplu o inferență validă? Citiți $\hat{\theta}$ ca un coeficient predictiv, nu ca o primă de risc cauzală asociată fricii. Notebook: secțiunea B14

◂ Înapoi ▸ Lista exercițiilor

C2 [Propus]: un model fundațional față de boosting — descrierea proiectului

- **Întrebarea**: prognozează un model fundațional pentru serii de timp din Capitolul 14, folosit fără reantrenare (*zero-shot*), direcția Bitcoin pe 5 zile mai bine decît gradient boosting și o constantă?
- **Protocol** (proiect de echipă, notat după protocol și dovezi, nu după semnul rezultatului)
  1. Numiți modelul și versiunea, data-limită documentată a preantrenării și testați doar pe date ulterioare; altfel declarați contaminarea ca limită.
  2. Prognozați direct distribuția randamentului pe 5 zile (nu cuantile zilnice însumate); $P(r_{t,t+\mathbf{5}} > 0)$ prin interpolarea liniară a funcției de repartiție predictive în 0, cu tratarea cozilor declarată.
  3. Reestimați gradient boosting și constanta la fiecare moment pe o fereastră mobilă de lungime fixă (de ex. 3 ani); log loss pe aceleași date; testul Giacomini–White (Giacomini & White (2006)) cu erori HAC; un grafic comparativ.
  4. Opțional: o regulă long/numerar pentru $P(\text{sus}) > 0{,}5$ cu costuri și DSR peste toate configurațiile și prompturile încercate.
- **Livrabile**: un folder Quantlet (cod, `Metainfo.txt`, un grafic) și un slide cu concluzia și principala limită

◂ Înapoi ▸ Lista exercițiilor

C3 [Propus]: un SDF fără arbitraj pe decile — descrierea proiectului (1/2)

- **Întrebarea**: întrece un factor stocastic de actualizare (SDF) liniar fără arbitraj, pe portofolii sortate după caracteristici, portofoliile tangente Fama–French (Fama & French (2015)) în afara eșantionului, ca în studiul de caz din curs (Chen, Pelger & Zhu (2024))?
- **Date** (inventarul arhivelor și coloanelor): portofoliile decile ponderate după valoare ale celor 15 sortări univariate din Kenneth French Data Library (mărime, valoare contabilă/de piață, profitabilitate operațională, investiții, cîștig/preț, flux de numerar/preț, randamentul dividendului, momentum, reversal pe termen scurt și lung, accruals, beta de piață, emisiuni nete de acțiuni, varianță, varianță reziduală), lunar, minus dobînda pe o lună; fișierele de factori FF3 (modelul Fama–French cu trei factori) și FF5 (cu cinci factori)
- **Componente**: pentru fiecare caracteristică, media decilelor 1–5 și a decilelor 6–10: 30 de factori $\tilde{F}_{t+1}$; doar lunile cu toate cele 30 de componente și FF5, fără imputare; salvați panelul fixat, cu data descărcării
- **Limită**: jumătățile decilelor ponderate după valoare înlocuiesc componentele ponderate după rang ale celor 46 de caracteristici pe acțiuni individuale: o extindere pe date de portofoliu, nu o replicare a modelului GAN (rețea generativă adversarială) la nivel de acțiune
- **Împărțire**: estimare pe 1967–1986, reglare pe 1987–1991, test pe 1992–2016 și din 2017 pînă la ultima lună
- **Atenție**: condițiile de evaluare și un SDF liniar neconstrîns nu garantează singure absența arbitrajului; pozitivitatea lui m_{t+1} nu este impusă

◂ Înapoi ▸ Lista exercițiilor

C3 [Propus]: un SDF fără arbitraj pe decile — descrierea proiectului (2/2)

- **Model**: A4 (inferență pentru raportul Sharpe) și B10 (teste predictive) [Rezolvat]
- **Etape**
  1. Reproduceți reperele FF3 și FF5: ponderile tangente $w \propto \hat{\Sigma}^{-1}\hat{\mu}$ pe 1967–1986, SR lunar pe 1992–2016 cu CI 95% (Lo (2002)).
  2. Cu μ = media lui $\tilde{F}$ și Q = media lui $\tilde{F}\tilde{F}^\top$ (un moment de ordinul doi, nu o covarianță) pe 1967–1986: $\hat{\theta}_{LS} = Q^{-1}\mu$ (LS: cele mai mici pătrate; rezolvați sistemul, nu inversați), $\hat{\theta}_{EN}$ (EN: elastic net) $= \arg\min_\theta \|\mu - Q\theta\|_2^2 + \lambda_2\|\theta\|_2^2 + \lambda_1\|\theta\|_1$ pe o grilă raportată de ponderi de penalizare $\lambda_1, \lambda_2 \geq 0$, aleasă după SR pe 1987–1991.
  3. Cu ponderile fixate: SR al lui $F = \hat{\theta}^\top\tilde{F}$ pe ambele perioade de test; diferențe de SR pe perechi, pe lunile aliniate (metoda delta HAC sau bootstrap pe blocuri); DSR peste grilă.
  4. Coeficienții $\hat{\beta}_i = \overline{R_i^e F}$ pe 1967–1986; proiectați lunar randamentele de test pe $\hat{\beta}$; variația explicată $1 - \overline{\hat{\epsilon}^2}/\overline{(R^e)^2}$, cu $\hat{F}_t = \hat{\beta}^\top R_t^e/(\hat{\beta}^\top\hat{\beta})$ factorul proiectat lunar și $\hat{\epsilon}_{t,i} = R^e_{t,i} - \hat{\beta}_i\hat{F}_t$ reziduul; bara este o medie în timp, și R^2 transversal pe caracteristică; un grafic cu randamentele medii realizate față de cele implicate.
- **Interpretare**
  - de ce poate un SDF liniar fără arbitraj să întreacă o rețea de prognoză (Chen, Pelger & Zhu (2024), Tabelul I)?
  - se menține rezultatul de după 2016 după corecția pentru configurațiile încercate?
- **Livrabile**: un folder Quantlet (cod, `Metainfo.txt`, un grafic) și un slide cu concluzia și limita ei

C3: inventarul decilelor și factorilor Kenneth French

- Descărcați factorii lunari din Kenneth French Data Library: `F-F_Research_Data_Factors_CSV.zip` și `F-F_Research_Data_5_Factors_2x3_CSV.zip`.
- Pentru cele 15 sortări după caracteristici, salvați numele exacte ale arhivelor, coloanele de decile, ponderarea, codurile lipsă, unitățile și lunile comune într-un inventar. Fixați fișierele descărcate înainte de estimare.
- $\hat{\beta}_i$ (etapa 4) este definit pînă la o constantă; ea se simplifică în $\hat{F}_t = \hat{\beta}^\top R_t^e / (\hat{\beta}^\top \hat{\beta})$. Calculați reziduurile cu $\hat{F}_t$, nu cu F_t.

Acronime folosite în acest material (1/3)

- **ADF**: Augmented Dickey–Fuller (test) — *testul Dickey–Fuller augmentat*
- **AI**: Artificial Intelligence — *inteligență artificială*
- **AIC**: Akaike Information Criterion — *criteriul informațional Akaike*
- **AUC**: Area Under the (ROC) Curve — *aria de sub curba ROC*
- **CI**: Confidence Interval — *interval de încredere*
- **CV**: Cross-Validation — *validare încrucișată*
- **CW**: Clark–West (test for nested models) — *testul Clark–West pentru modele imbricate*
- **DM**: Diebold–Mariano (test) — *testul Diebold–Mariano de comparare a prognozelor*
- **DSR**: Deflated Sharpe Ratio — *raportul Sharpe deflatat*
- **ELW**: Exact Local Whittle (estimator) — *estimatorul local Whittle exact*
- **EWMA**: Exponentially Weighted Moving Average — *medie mobilă ponderată exponențial*
- **FF3**: Fama–French three-factor model — *modelul Fama–French cu trei factori*
- **FF5**: Fama–French five-factor model — *modelul Fama–French cu cinci factori*
- **FFD**: Fixed-width window Fractional Differentiation — *diferențiere fracționară cu fereastră fixă*

Acronime folosite în acest material (2/3)

- **FFN**: FeedForward Network — *rețea neuronală feedforward (fără bucle)*
- **FN**: False Negative — *fals negativ*
- **GAN**: Generative Adversarial Network — *rețea generativă adversarială*
- **GB**: Gradient Boosting — *boosting pe gradient*
- **GW**: Giacomini–White (test) — *testul Giacomini–White*
- **HAC**: Heteroskedasticity and Autocorrelation Consistent (standard errors) — *erori standard robuste la heteroscedasticitate și autocorelație*
- **IS**: In-Sample — *în eșantion*
- **LASSO**: Least Absolute Shrinkage and Selection Operator — *regresie penalizată L1 cu selecția variabilelor*
- **LS**: Linear SDF estimated by unregularised Least Squares — *SDF liniar estimat prin cele mai mici pătrate, fără regularizare*
- **LW**: Local Whittle (estimator) — *estimatorul local Whittle*
- **MDA**: Mean Decrease Accuracy — *scăderea medie a acurateței (importanța prin permutare)*
- **MDI**: Mean Decrease Impurity — *scăderea medie a impurității*
- **OLS**: Ordinary Least Squares — *metoda celor mai mici pătrate*
- **OOS**: Out-Of-Sample — *în afara eșantionului*

Acronime folosite în acest material (3/3)

- **PBO**: Probability of Backtest Overfitting — *probabilitatea de overfitting a backtest-ului*
- **PSR**: Probabilistic Sharpe Ratio — *raportul Sharpe probabilistic*
- **PT**: Pesaran–Timmermann (test of directional accuracy) — *testul Pesaran–Timmermann al acurateței direcționale*
- **RC**: Reality Check (White's test for data snooping) — *testul Reality Check al lui White pentru data snooping*
- **RF**: Random Forest — *pădure aleatoare*
- **ROC**: Receiver Operating Characteristic (curve) — *curba caracteristică de operare a receptorului*
- **SDF**: Stochastic Discount Factor — *factorul stocastic de actualizare*
- **SE**: Standard Error — *eroare standard*
- **SHAP**: SHapley Additive exPlanations — *explicații aditive Shapley*
- **SPA**: Superior Predictive Ability (test) — *testul abilității predictive superioare*
- **SR**: Sharpe Ratio — *raportul Sharpe*
- **SUA**: Statele Unite ale Americii
- **TP**: True Positive — *adevărat pozitiv*
- **VIX**: CBOE Volatility IndeX — *indicele de volatilitate CBOE*

Bibliografie (1/5)

- 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.
- 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.
- 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.
- Broadie, M., Glasserman, P., Kou, S. (1997). A continuity correction for discrete barrier options. *Mathematical Finance*, 7(4), 325–349.

Bibliografie (2/5)

- Chen, L., Pelger, M., Zhu, J. (2024). Deep learning in asset pricing. *Management Science*, 70(2), 714–750.
- Clark, T.E., West, K.D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1), 291–311.
- 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.
- Dickey, D.A., Fuller, W.A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366), 427–431.
- Diebold, F.X., Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.
- Fama, E.F., French, K.R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1), 1–22.
- 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.

Bibliografie (3/5)

- Hanley, J.A., McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1), 29–36.
- Hansen, P.R. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4), 365–380.
- Hosking, J.R.M. (1981). Fractional differencing. *Biometrika*, 68(1), 165–176.
- Karatzas, I., Shreve, S.E. (1998). *Brownian Motion and Stochastic Calculus*, 2nd ed. Springer.
- Künsch, H.R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17(3), 1217–1241.
- Lo, A.W. (2002). The statistics of Sharpe ratios. *Financial Analysts Journal*, 58(4), 36–52.

Bibliografie (4/5)

- López de Prado, M. (2018). *Advances in Financial Machine Learning*. Wiley.
- Lundberg, S.M., Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
- Newey, W.K., West, K.D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703–708.
- 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.
- 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.

Bibliografie (5/5)

- 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.
- White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5), 1097–1126.