Algorithmic Trading & Quantitative Strategies

Lecture 7 (4/23/2024)

Giuseppe Paleologo (gardener)

Today's Session

- No quiz today
- Today is entirely devoted to backtesting

Recap on Last Quiz

1. Why do we perform factor/idio short-term updating? Wrong answer only
   a. **There are transient, mean-reverting shocks**
   b. Volatility is higher in January
   c. The world is not stationary
   d. It lowers QLIKE loss
2. What can go wrong with thresholding cov matrices?
   a. **Non pos-definite**
   b. Performance can worsen
   c. Matrix storage
   d. Hard to interpret
3. What operation you **don't** do on idio cov matrices
   a. Shrinkage
   b. **Rotation**
   c. Short-term updating
   d. Exponential Weighting
4. Which constraint here is not like the others?
   a. Long-only
   b. Max Style Vol
   c. Max GMV
   d. **Factor Vol > Idio Vol**
5. If my covariance matrix is ill-conditioned:
   a. **Alpha error is more damaging**
   b. Covariance error is more damaging
   c. Max GMV constraint is more damaging
   d. Max variance constraint is more damaging

Backtesting

- Financial Researchers cannot design experiments. History is all we have
- We still live in a small data kind of world
- Backtesting is an (the) essential tool for confirmatory analysis
- Many challenges
- Two parts to this lecture.
  - Folk wisdom stuff
  - Some hardcore analysis

Best Practices: Data (1)

- What does the data mean? The most basic question
  - Definition
  - Physical dimensions
- Provenance
  - Where are the data coming from?
  - Does the vendor collect the data themselves?
  - Collection criterion?
  - Does the vendor sample data or collects the data exhaustively?
  - Is the population sampling methodology sound?
  - If originate elsewhere, who is originating the data?
-

Best Practices: Data (2)

- Completeness
  - Missingness analysis
  - X-sec, serial completeness
- Quality assurance
  - Vendor procedures
  - Change point detection
- Point-in-time vs. restated
- Transformation
- Substitutes and complements

Best Practices: Process (1)

- **Data Leakage**: presence in the training data, the data available up to time t, of information contained in the target, i.e., returns in periods $t+1$ and later
  - Survivorship bias
  - Financial statements as-of-date or lagged (check for robustness)
  - Price adjustments. Adjusted prices are subtly informative!
  - Missingness. Can be informative!
  - Stupid mistakes. E.g., off-by-one. They happen.

Best Practices: Process (2)

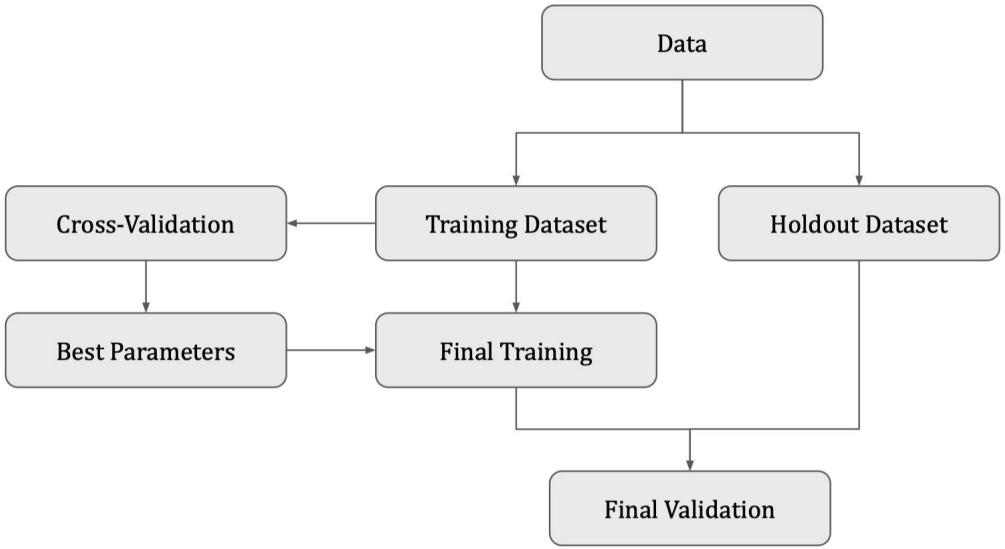
- **Strategy development**
  - Have a theory (if you can)
  - Enforce reproducibility
  - Use as much as possible the same setting in research and production
  - Calibrate the market impact
  - Include borrow costs (and maybe/maybe not dividend taxation)
  - Define beforehand the dataset
  - Define beforehand the backtesting protocol

Backtesting Protocol

Two main approaches: cross-validation and walk-forward

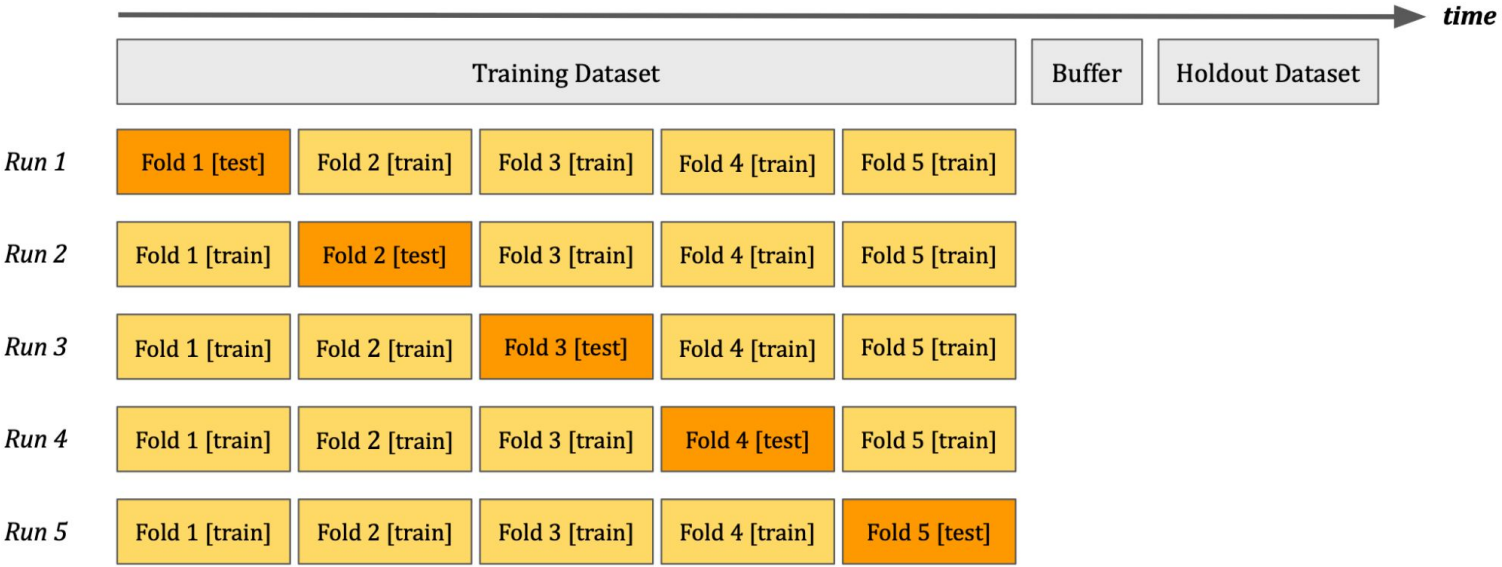
Some

Cross-Validation



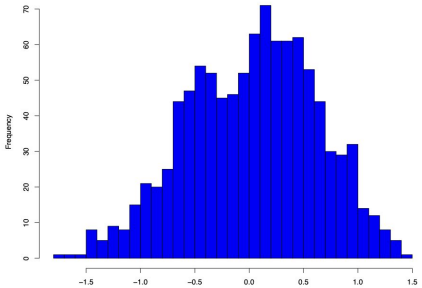
Cross-Validation

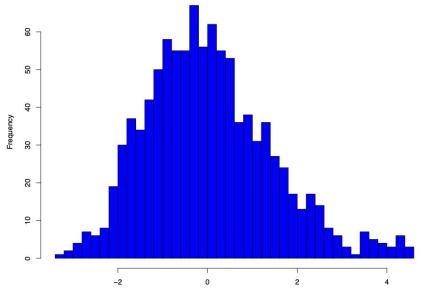
- The intended goal is to estimate the expected loss (performance) of a model
- Usage to select large numbers of models would be very different from practice
- Do not screen factors before or by cross-validation
- Challenges (partially addressable):
  - Serial dependencies
  - Data leakage



An Example

- N=1000, T variable (20 years, 5 years), number of random factors variable (2, 500). True SR zero
- 5-fold cross-validation
- Run 1000 simulations

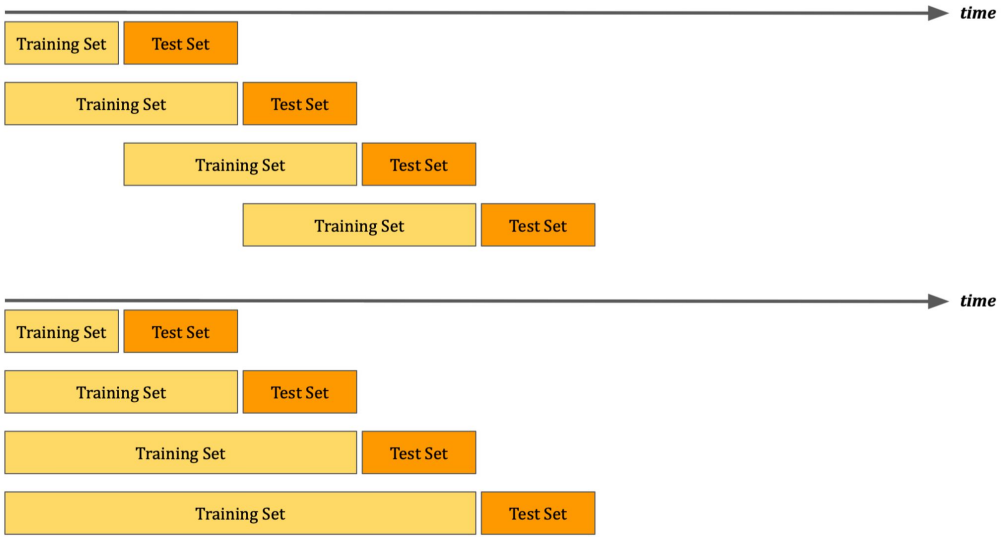




T	p	Mean(SR)	Stdev (SR)	% passing
5000	2	0.07	0.6	1.2
1250	500	0.04	1.4	19

Walk Forward

- Basic idea: align research and production
  - No data leakage
  - Serial dependencies are accounted for
  - Comes in variants



Walk-Forward

It is a necessary step in the research process. Often the chain is:

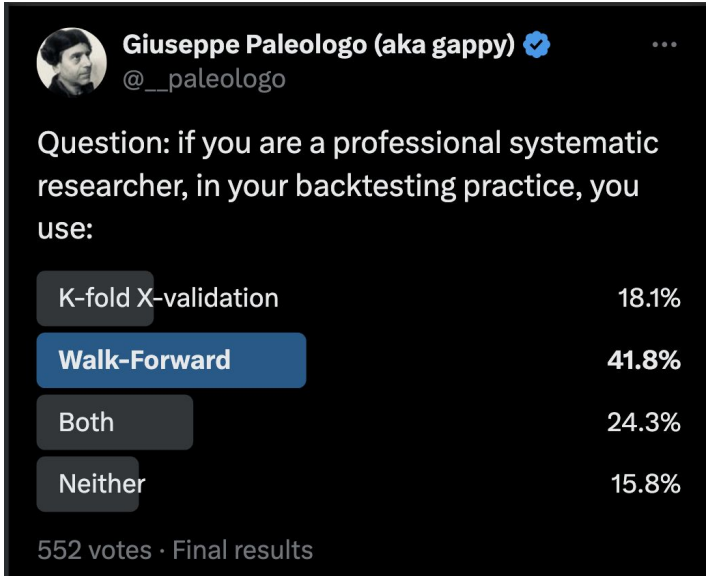
K-fold cross-validation (screening)

=> walk-forward

=> test production

=> deploy at at scale

Drawback: data usage. You can't select among large sets of signal using a single walk-forward run. In practice each signal is tested separately



A Wish List

1. non-anticipative/immune from data leakage
2. taking into account serial dependency
3. using all data
4. allowing for multiple testing of a very large number of signals
5. Providing a rigorous decision rule

A Modest Proposal: The Rademacher AntiSerum (**RAS**) Against Backtesting Bites

Driving ideas:

1. Use a data-based measure of complexity
2. Reinterpret and extend known results in Machine Learning
3. Answer the question quants are asking (what is the haircut?), not the question you like to answer (hypothesis tests)
4. Computations are cheap

Setup

- Primitive data: T x N matrix X.
- X[t, i] is the performance at time t

  a are signals, r are return, w are portfolios, and S are covariance matrices
  - Signals: Information Coefficient x[t,i]:

    x[t,i] :=⟨a[i,t], r[t]⟩/ (||a[i,t]|| ||r[t]||)
  - Strategy: z-scored strategy return:

    x[t,i] =⟨w[i,t]',r[t]⟩/ sqrt (w[i,t]' * S[t] * w[i,t])
- Use x_t:=x[t,:] and x^n:=x[:,n]
- Define the empirical performance:

$$\hat{\boldsymbol{\theta}}(\mathbf{X}) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t$$

Moar Machinery

Rademacher random vector $\boldsymbol{\epsilon}$ in R^T: iid binary -1/1 with prob $\frac{1}{2}$

The star of the show: **Rademacher Complexity**

$$\hat{R} = E_{\boldsymbol{\epsilon}} \left(\sup_n \frac{|\boldsymbol{\epsilon}' \mathbf{x}^n|}{T} \right)$$

Interpretations of Rademacher Complexity

As the covariance to random noise: ε as a random covariate. We can interpret R as the expected value of the highest covariance of the performance measure of a strategy to random noise. If, on average, for every set of +/-1 indicators, there is at least a strategy that covaries with it, then "we can do no wrong": for every realization of a random series, there's a strategy that would do well matching it. If we interpret the x[t,n] as predictions for epoch t, then this means that for every sequence of events ε[t] we have a strategy that predicts them well.

Interpretations of Rademacher Complexity

- **As the covariance to random noise**: Consider ϵ as a random covariate. We can interpret $\hat{R}$ as the expected value of the highest covariance of the performance measure of a strategy to random noise. If, on average, for every set of +/-1 indicators, there is at least a strategy that covaries with it, then "we can do no wrong": for every realization of a random series, there's a strategy that would do well matching it. If we interpret the $x_{t,n}$ as predictions for epoch t, then this means that for every sequence of events ϵ_t we have a strategy that predicts them well.

- **As generalized 2-way cross-validation**: For sufficiently large T, the sets of positive elements in $\boldsymbol{\epsilon}_t$ concentrates around size $T/2$. We denote S^+ the set of $T/2$ periods where $\epsilon_t = 1$, and S^- the other periods. Rewrite the term inside the sup as

$$\frac{|\boldsymbol{\epsilon}'\mathbf{x}^n|}{T} = \frac{1}{2}\left|\frac{2}{T}\sum_{s\in S^+1} x_{s,n} + \frac{2}{T}\sum_{s\in S^-} x_{s,n}\right| = \frac{1}{2}|\hat{\theta}_n^+ - \hat{\theta}_n^-|$$

  For strategy n, this is the discrepancy in average performance measured on two equal-sized random subsets of the observations. By taking the sup across strategies, we are estimating the worst case: we estimate performance on a subset, and get a very different result on the remaining subset! And if the discrepancy is high for each random subset, this will indicate that performance is not consistent: there's always at least a strategy that performs comparatively well *somewhere* and poorly in the remaining periods. The associated $\hat{R}$ is high, and means that the set of strategies has unreliable performance.

- **As measure of span over possible performances**: We interpret $\boldsymbol{\epsilon}$ as a "random direction" chosen at random in $\mathbb{R}^T$. The vector has Euclidean norm equal to $\sqrt{T}$. In the case where the performance measure is the standardized return, $E(\|\mathbf{x}^n\|)$ is also equal to $\sqrt{T}$, and is strongly concentrated around this value. The empirical Rademacher $\hat{R}$ is then approximately equal to

$$E_{\boldsymbol{\epsilon}}\left(\sup_n \left|\frac{\boldsymbol{\epsilon}'\mathbf{x}^n}{\|\boldsymbol{\epsilon}\|\,\|\mathbf{x}^n\|}\right|\right)$$

  This can be interpreted in the following way. We have a set of N vectors $x_{\cdot n} \in \mathbb{R}^T$. We pick a random direction in the ambient space, and observe the maximum collinearity (expressed as the cosine distance) of this random direction to our vectors. The expected value is this collinearity measures how much our set of strategy vectors span $\mathbb{R}^T$. If we have n vectors that are copies of the same vector, the answer is: not very well. If conversely these vectors are all orthogonal, we have maximum collinearity. The Rademacher complexity if a geometric measure of how much the vectors $\mathbf{x}^n$ "span" $\mathbb{R}^T$.

Main Result: Signals

For all signals, with probability at least equal to 1-δ:

$$\theta_n > \hat{\theta}_n - \underbrace{2\hat{R}}_{\textit{(data snooping)}} - \underbrace{2\sqrt{\frac{\log(2/\delta)}{T}}}_{\textit{(estimation error)}}$$

Main Result: Strategies

For all signals, with probability at least equal to 1-δ:

$$\theta_n - \hat{\theta}_n \geq -\underbrace{2\hat{R}}_{\textit{(data snooping)}} \underbrace{-3\sqrt{\frac{2\log(2/\delta)}{T}} - \sqrt{\frac{2\log(2N/\delta)}{T}}}_{\textit{(estimation error)}}$$

Discussion: Strategy/Signal Complexity Space

- These are uniform bounds, and you can check whether a strategy's Sharpe > 0 with high probability
- The first is the term 2R. This is the **data snooping term**
- The larger the number of strategies => the higher the R
- The higher the dependency among strategies => the lower R.
- In the limit case where we test multiple replicas of the same strategy R is zero.

Discussion: Estimation Error

The second is the **estimation term**. It is the unavoidable

Intuition: consider T iid rv standard normal. Their average is approximately distributed as a normal distribution with standard deviation 1/sqrt(T).

What is the δ-quantile of the distribution? Approximation:

$$P(\hat{\theta} < -b) = \bar{F}(b) \leq \exp(-b^2 T/2) \quad \Rightarrow \quad b > \sqrt{2\log(1/\delta)/T}$$

Practical Considerations

1. IC is bounded by 1, but in practice it's $\leq$ 0.1. Apply this (non-rigorous) correction, and the estimation error is reduced by a factor 10
2. And the estimation error constant is probably non-optimal. If you consider the stand-along Sharpe estimation Error, it's 3x too big. Effectively, we can do better.
3. The real issue is the data mining impact, and that is unchanged and equal to 2R