Portfolio Optimization

Portfolio Backtesting

Daniel P. Palomar (2025). *Portfolio Optimization: Theory and Application.*
Cambridge University Press.

portfoliooptimizationbook.com

Latest update: March 10, 2025

Outline

Abstract

A backtest is a historical simulation of how a strategy would have performed should it have been run over a past period of time. It is an essential step prior to the actual live trading with real money. Nevertheless, backtesting is one of the least understood techniques in the quant toolbox. The reality is that backtesting is full of dangers and virtually impossible to execute properly. These slides will explore portfolio backtesting, so that we become aware of all the potential pitfalls (Palomar 2025, chap. 8).

Outline

A typical backtest

- **Backtest definition:**
  - A historical simulation of a strategy's performance over a past period.
  - Commonly seen in academic publications, fund brochures, and practitioner blogs.

- **Data splitting:**
  - Data is split into in-sample and out-of-sample datasets.
  - In-sample data is used to estimate parameters (e.g., expected returns, covariance matrix).
  - Out-of-sample data is used to assess the strategy's performance.

A typical backtest

Example of a backtest result in the form of cumulative P&L and drawdown plots:

A typical backtest

Example of a backtest result in the form of performance measures:

Portfolio	Sharpe ratio	annual return	annual volatility	Sortino ratio	max drawdown	CVaR (0.95)
1/N	3.23	117%	36%	5.40	11%	5%
GMRP	2.19	138%	63%	4.09	19%	7%
IVolP	3.35	113%	34%	5.61	11%	4%

A typical backtest

- **Many other performance measures in the form of plots and tables:**
  - Rolling Sharpe ratio plot over time.
  - Monthly performance measures instead of overall annualized values.

- **Global Investment Performance Standards (GIPS):**
  - GIPS is a set of standardized, industry-wide ethical principles for presenting investment performance to clients, regulators, and other stakeholders.
  - Provides guidance on how to calculate and report investment results to prospective clients.
  - Compliance with GIPS standards demonstrates a firm's commitment to ethical best practices and strong internal control processes.
  - Promotes transparency, fair representation, and full disclosure of investment performance.
  - Facilitates comparison of investment performance across firms and strategies.

A typical backtest

- **Limitations and cautions:**
  - Backtest results provide limited information on real performance.
  - Results are often faulty and misleading due to various biases and issues.
  - We will explore: the pitfalls of backtesting and best practices for backtesting (Palomar 2025, chap. 8).

- Quote (Harvey, Liu, and Zhu 2016):

  "Most claimed research findings in financial economics are likely false."

- Quote (López de Prado 2018):

  "Most backtests published in journals are flawed, as the result of selection bias on multiple tests."

Outline

The seven sins of fund management

In 2005, a practicioner report compiled the "Seven Sins of Fund Management" (Montier 2005) (of which sins #1 and #5 are the most directly related to backtesting):

- Sin #1: Forecasting (Pride)
- Sin #2: The illusion of knowledge (Gluttony)
- Sin #3: Meeting companies (Lust)
- Sin #4: Thinking you can out-smart everyone else (Envy)
- Sin #5: Short time horizons and overtrading (Avarice)
- Sin #6: Believing everything you read (Sloth)
- Sin #7: Group based decisions (Wrath)

The seven sins of quantitative investing

In 2014, a team of quants at Deutsche Bank published a study under the suggestive title "Seven Sins of Quantitative Investing" (Luo et al. 2014). These seven sins are a few basic backtesting errors that most journal publications make routinely:

- Sin #1: Survivorship bias
- Sin #2: Look-ahead bias
- Sin #3: Storytelling bias
- Sin #4: Overfitting and data snooping bias
- Sin #5: Turnover and transaction cost
- Sin #6: Outliers
- Sin #7: Asymmetric pattern and shorting cost

The seven sins of quantitative investing: Survivorship bias

- **Definition:**
  - Survivorship bias occurs when only the surviving or successful entities are considered, ignoring those that have failed or underperformed.
  - In investing, it involves backtesting strategies using only companies that are currently in business and performing well, often listed in major indices.

- **Impact:**
  - Ignores stocks that have left the investment universe due to bankruptcy, delisting, acquisition, or underperformance.
  - Leads to an overestimation of returns and an underestimation of risk.
  - Provides a misleading and overly optimistic view of the strategy's performance.

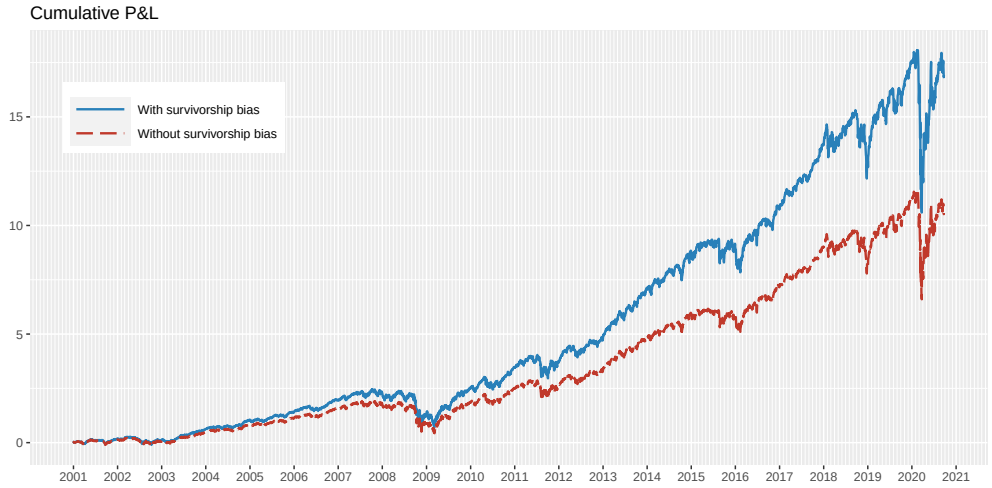
- **Causes:**
  - Practitioners backtest using only current index constituents (e.g., S&P 500).
  - Removing stocks with missing data from the universe.
  - Using databases that exclude delisted or underperforming companies.

The seven sins of quantitative investing: Survivorship bias

- **Center for Research in Security Prices (CRSP):**
  - Maintains a comprehensive database of historical stock market data.
  - Considered highly reliable and accurate, widely used by researchers and professionals.
  - Helps mitigate survivorship bias by including delisted and underperforming stocks.
- **Survivorship bias in other areas:**
  - Not limited to financial investments, but a persistent phenomenon across various domains.
  - Example: Success stories of college dropouts who became billionaires (e.g., Bill Gates, Mark Zuckerberg) distort perceptions by ignoring the majority of unsuccessful dropouts.
- **Significance:**
  - Survivorship bias can lead to significant overestimation of returns and underestimation of risk.
  - Accounting for survivorship bias is crucial for accurate and reliable backtesting and investment strategy evaluation.
  - Failure to address this bias can result in flawed decision-making and suboptimal investment outcomes.

The seven sins of quantitative investing: Survivorship bias

Effect of survivorship bias on the S&P 500 stocks:



The seven sins of quantitative investing: Look-ahead bias

- **Definition:**
  - Look-ahead bias occurs when information or data that were unknown or unavailable at the time of backtesting are used in the analysis.
  - It is a very common bias in backtesting and can lead to overly optimistic results.

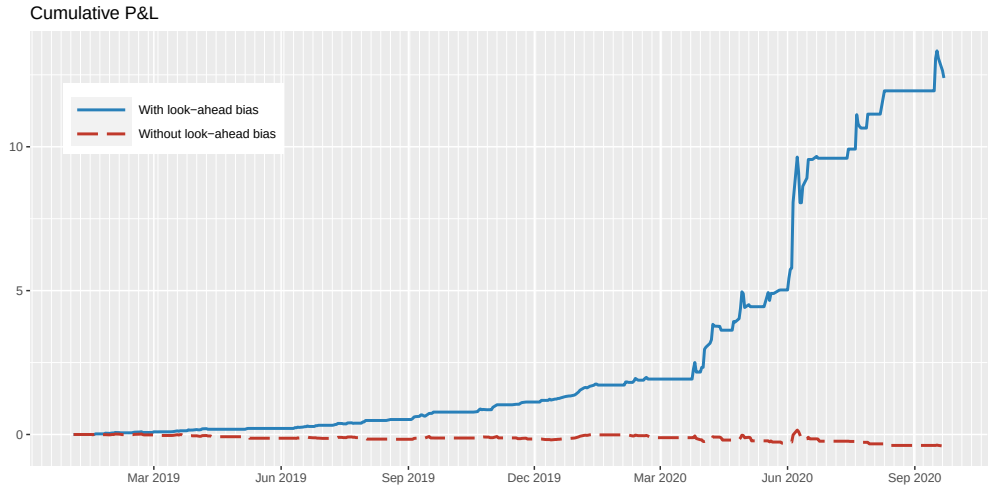
- **Examples:**
  - Using financial statement data with incorrect timestamps or without considering release dates and distribution delays.
  - Coding errors, such as training parameters using future information or pre-processing data with statistics from the entire dataset.
  - Time alignment errors in backtesting code, e.g., computing portfolio returns using future asset prices.

The seven sins of quantitative investing: Look-ahead bias

- **Time alignment error:**
  - Incorrectly computing portfolio returns as $R_t^{\text{portf}} = \boldsymbol{w}_t^{\mathsf{T}} \boldsymbol{r}_t$, where $\boldsymbol{w}_t$ uses information up to time t, and $\boldsymbol{r}_t$ assumes positions were executed at $t-1$.
  - This error can lead to seemingly amazing but unrealistic performance.
- **Mitigation:**
  - Carefully handle timestamps and data release dates, ensuring information is available at the time of backtesting.
  - Rigorously check backtesting code for time alignment errors and proper use of information.
  - Consider using backtesting frameworks or libraries that handle time alignment and data availability correctly.
  - Validate backtesting results against known benchmarks or theoretical expectations.
- Addressing look-ahead bias is crucial for obtaining reliable and realistic backtesting results. Failure to account for this bias can lead to overly optimistic performance estimates and flawed investment decisions.

The seven sins of quantitative investing: Look-ahead bias

Effect of look-ahead bias (from a time alignment mistake) trading a single stock:



The seven sins of quantitative investing: Storytelling bias

- **Definition:**
  - Storytelling bias occurs when we create a narrative or story to explain a random pattern or event after it has occurred (ex-post).
  - It is related to confirmation bias, where we favor information that supports our pre-existing beliefs and ignore contradictory evidence.
- **Prevalence:**
  - Storytelling is pervasive in financial news, where "experts" often justify random patterns after the fact.
  - Popular science best-selling books often rely on anecdotal stories rather than solid statistical foundations, as stories appeal more to the general public.
- **Antidote:**
  - Collecting more historical data to see if the story passes statistical tests or the test of time.
  - However, in economics and finance, the limited number of observations hinders the resolution of storytelling bias.
  - Compelling stories alone should not drive investment decisions without proper validation.

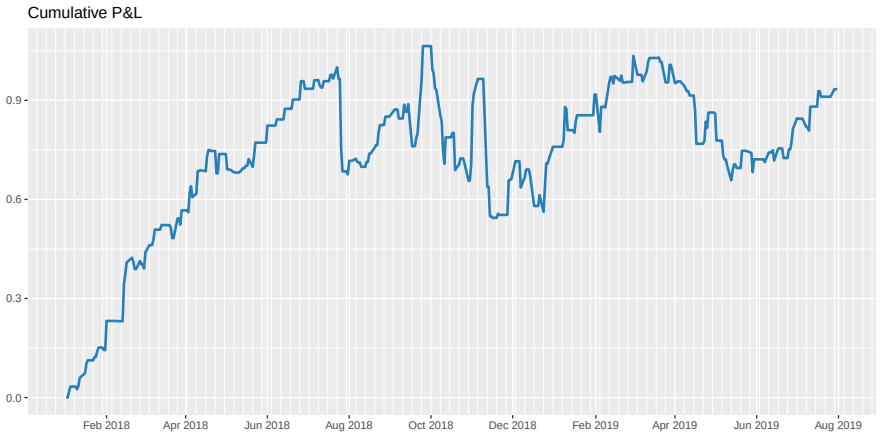
The seven sins of quantitative investing: Storytelling bias

- **Implications:**
  - Storytelling bias can lead to false conclusions and flawed decision-making in finance and investing.
  - It can cause investors to attribute meaning to random patterns or events, leading to suboptimal investment strategies.
  - Recognizing and mitigating storytelling bias is crucial for making rational and data-driven investment decisions.

- **Illustration:**
  - The next figure demonstrates the effect of storytelling bias by trading a single stock using a random binary sequence.
  - Before August 2018, one could be inclined to believe the story that the random sequence was a good predictor of the stock trend.
  - However, with more data collected over time, it becomes evident that the apparent predictive power was merely a fluke.

The seven sins of quantitative investing: Storytelling bias

Effect of story-telling bias in the form of a random strategy that performs amazingly well until August 2018, but not afterwards:



The seven sins of quantitative investing: Overfitting and data snooping

- **Definition:**
  - Data snooping bias, also known as data mining bias or overfitting, refers to the practice of extensively searching for patterns or rules in data until a model fits perfectly.
  - It involves manipulating data or models to find the desired pattern an analyst wants to show.

- **Causes:**
  - Fine-tuning model parameters to achieve optimal performance on the available data.
  - Iteratively adjusting the strategy based on its evaluation on the test data.
  - Using the test data too many times, effectively making it part of the training data.

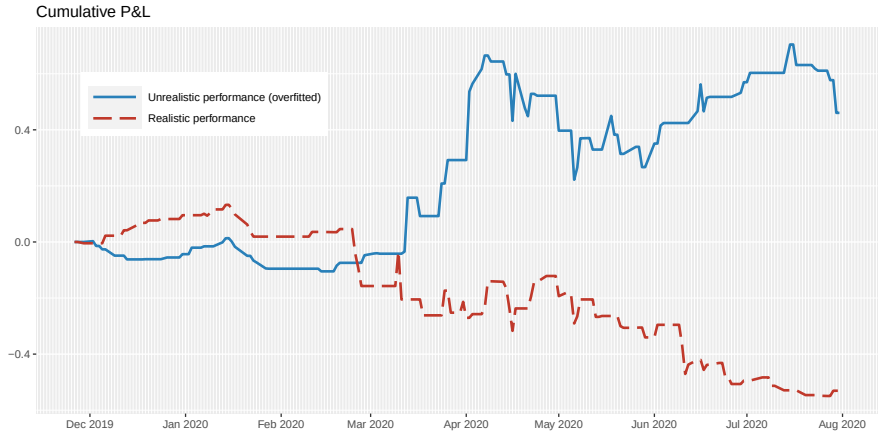
- **Consequences:**
  - Models or strategies that appear promising on the available data but are actually spurious.
  - Inability to trust published backtest results due to selection bias from multiple tests.
  - Overfitted models fail to generalize and perform poorly on new, unseen data.

The seven sins of quantitative investing: Overfitting and data snooping

- **Mitigation strategies:**
  - Splitting data into separate training and test sets.
  - Avoiding excessive fine-tuning of parameters and conducting sensitivity analysis.
  - Evaluating the strategy's performance on unseen data through live or paper trading.
- **Importance:**
  - Data snooping bias is considered one of the most difficult biases to address in finance.
  - Not accounting for this bias can result in flawed investment strategies and major losses.
  - Rigorous data handling, model validation, and continuous evaluation on new data are crucial to mitigate this bias.
- **Illustrative example:**
  - The next figure demonstrates the effect of data snooping bias by trading a single stock using a machine learning strategy.
  - The strategy is trained on either the training data alone (correct approach) or the combined training and test data (overfitting).
  - The overfitted backtest appears to have predictive power on the out-of-sample data, but this is misleading and does not reflect the true performance.

The seven sins of quantitative investing: Overfitting and data snooping

Effect of data snooping or overfitting on a backtest after tweaking the strategy too many times:



The seven sins of quantitative investing: Turnover and transaction cost

- **Turnover:**
  - Turnover refers to the overall amount of orders to be executed when rebalancing a portfolio from $\boldsymbol{w}_t$ to $\boldsymbol{w}_t^{\text{reb}}$.
  - It is calculated as $\|\boldsymbol{w}_t^{\text{reb}} - \boldsymbol{w}_t\|_1$.

- **Transaction costs:**
  - Transaction costs are often modeled as proportional to the turnover.
  - If liquidity is insufficient compared to the turnover size, slippage may have a significant effect.
  - Simulating transaction costs at the limit order book level can be extremely challenging.

- **Factors affecting overall transaction cost:**
  - Turnover per rebalancing.
  - Rebalancing frequency.

The seven sins of quantitative investing: Turnover and transaction cost

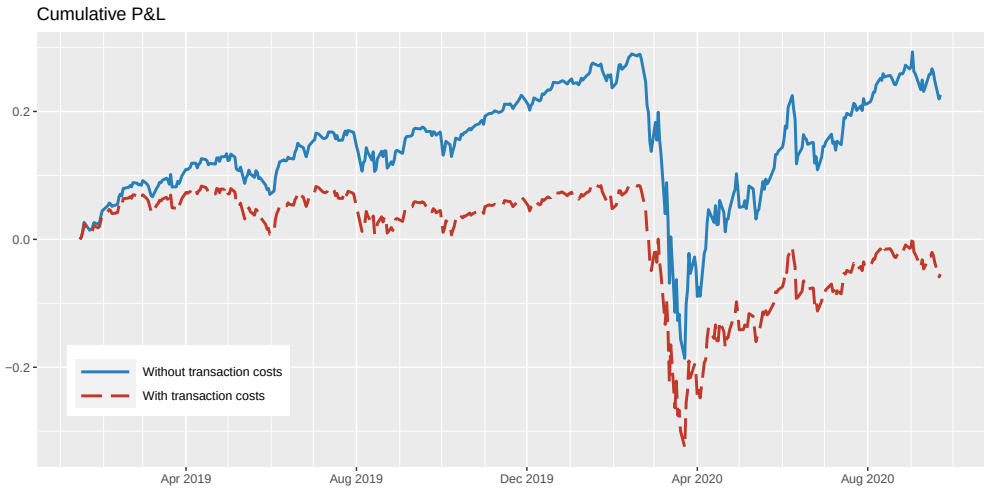
- **Considerations for rebalancing frequency:**
  - High rebalancing frequency leads to higher overall turnover and transaction costs.
  - Low rebalancing frequency may cause the portfolio to fail to adapt to changing signals.
  - Deciding the appropriate rebalancing frequency is a critical step in practice.

- **Mitigation strategies:**
  - Keep rebalancing frequency to a minimum.
  - Control the turnover per rebalancing.
  - Consider the trade-off between transaction costs and the ability to adapt to changing signals.

- **Impact of transaction costs:**
  - The next figure illustrates the detrimental effect of transaction costs on the daily-rebalanced inverse volatility portfolio (IVolP) on the S&P 500 stocks with fees of 60 bps.
  - The effect of transaction costs accumulates slowly over time.

The seven sins of quantitative investing: Turnover and transaction cost

Effect of transaction costs on a portfolio (with daily rebalancing and fees of 60 bps):



The seven sins of quantitative investing: Outliers

- **Definition:**
  - Outliers are events that do not fit the normal and expected behavior in financial data.
  - They can occur due to various reasons, such as historical events, lack of liquidity, large execution orders, or data errors.
- **Importance:**
  - Outliers cannot be predicted, and one can only try to be robust to them.
  - Robust estimation methods and robust portfolio techniques are important in practice to handle outliers.
  - Accidentally benefiting from a few outliers in backtesting can distort the realistic assessment of a portfolio's performance.
- **Approaches to handling outliers:**
  1. **Outlier control:**
     - Traditional techniques include winsorization (capping data at certain percentiles) and truncation/trimming/censoring (removing outliers from the data sample).
     - Data normalization processes are closely related to outlier control.
  2. **Keeping outliers, but ensuring robustness:**
     - Keep the outliers in the data, but ensure that strategies do not rely solely on them.

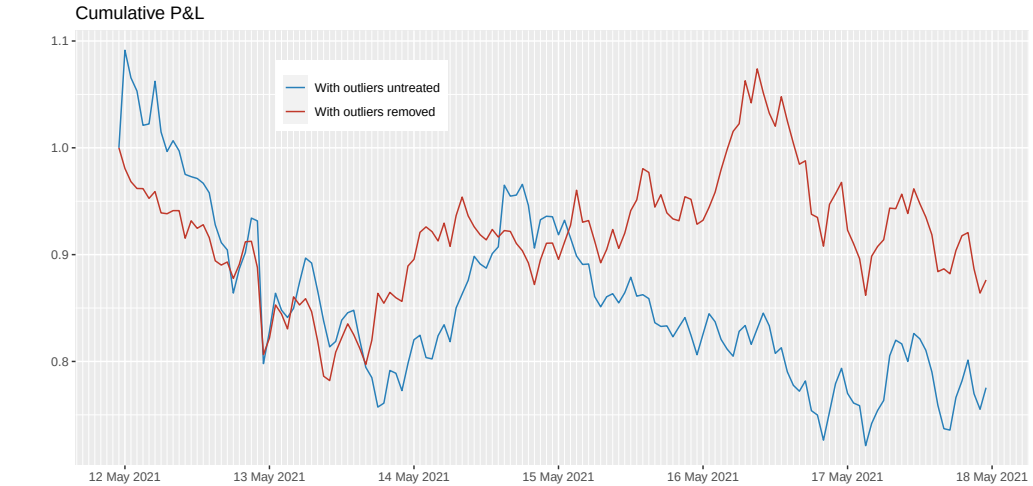
The seven sins of quantitative investing: Outliers

- **Considerations:**
  - Outliers caused by data errors or specific events are unlikely to be repeated in the future.
  - Basing a portfolio's success on a few historical outliers can lead to unrealistic future performance expectations.
  - Robust portfolio construction and evaluation techniques should be employed to mitigate the impact of outliers.

- **Illustration:**
  - The next figure demonstrates the effect of outlier control in the design phase of a quintile portfolio with hourly cryptocurrency data.
  - Outliers larger than 5% (hourly returns) are removed.

The seven sins of quantitative investing: Outliers

Effect of outliers on a backtest with hourly cryptocurrency data:



The seven sins of quantitative investing: Asymmetric pattern and shorting cost

- **Unrealistic assumptions in backtesting:**
  - Analysts often assume the ability to short any stock at no cost or the same level of cost as going long.
  - However, borrowing costs can be prohibitively high for some stocks, or it may be impossible to locate borrowable shares.
- **Regulatory and market constraints:**
  - Some countries or exchanges prohibit short selling entirely or limit its extent.
  - During periods of market stress, such as the 2008 financial crisis, borrowing costs can skyrocket, and certain stocks may be banned from being shorted.
- **Impact on long positions:**
  - Short availability constraints not only affect portfolios that short sell but can also impact long positions due to the "limited arbitrage" argument.
  - Arbitrageurs may be prevented from immediately forcing prices to fair values.

The seven sins of quantitative investing: Asymmetric pattern and shorting cost

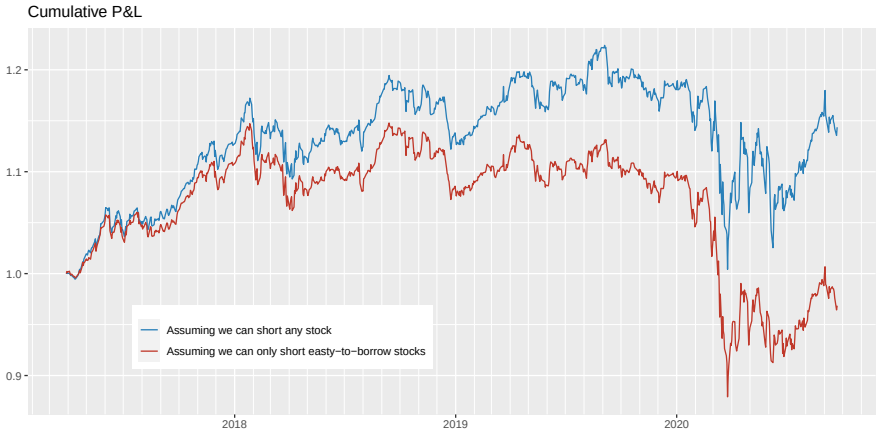
- **Implications:**
  - Ignoring short selling constraints and borrowing costs skews performance estimates.
  - Strategies relying on short selling are greatly affected.
  - Considering short availability and costs is vital for accurate evaluation.

- **Mitigation strategies:**
  - Include realistic borrowing costs and constraints in backtesting.
  - Create strategies less dependent on short selling or use alternatives.
  - Monitor and adapt to changes in regulations and market conditions.

- **Illustration:** The next figure shows two long-short quintile portfolios:
  1. An unrealistic portfolio perfectly longs the top 20% and shorts the bottom 20%.
  2. A realistic portfolio shorts only easy-to-borrow stocks (for illustration).

The seven sins of quantitative investing: Asymmetric pattern and shorting cost

Effect of shorting availability in a long-short quintile portfolio:



Outline

The dangers of backtesting

- We have learned from the "seven sins of quantitative investing" that backtesting is a dangerous process.
- There are more than these seven types of errors one can make (López de Prado 2018):
  "A full book could be written listing all the different errors people make while backtesting."
- The most common mistake in backtesting involves overfitting or data snooping.
- John von Neumann made a humorous quote about the general concept of overfitting:
  "With four parameters I can fit an elephant, and with five I can make him wiggle his trunk."

Backtest overfitting

- **Common mistake:**
  - Overfitting is arguably the most common error in backtesting.
  - It involves developing a model or strategy that captures random noise instead of a general pattern.
- **Definition:**
  - Overfitting occurs when a model is tailored to fit particular noisy observations rather than the underlying data structure.
  - In investing, it refers to creating strategies that perform well on historical data by exploiting random patterns.
- **Consequences:** (D. H. Bailey et al. 2014)
  - Strategies developed through overfitting are likely to fail in the future as random historical patterns may not recur.
  - Performance on in-sample data can be misleading, leading to unfounded confidence in a strategy's future success.
  - Even out-of-sample data performance can be misleading if the data is overused in the model development process.

Backtest overfitting

- **Adjusting parameters:**
  - Researchers often adjust strategy parameters to improve backtest performance.
  - Repeated adjustments can indirectly transform the test data into training data.
  - This process can lead to the belief that a portfolio will perform well, only to disappoint during live trading.
- **False positives:**
  - A small number of trials can identify strategies with spuriously high backtested performance, especially for complex strategies.
  - Not reporting the number of trials involved in identifying a successful backtest can be misleading.
- **Difficulty in assessment:**
  - The probability of false positives increases with each new test on the same dataset.
  - This information is often unknown to the researcher or not disclosed to investors or referees.
  - Typically, only successful backtests are shared, skewing perceptions of investment strategy success.

Backtest overfitting

- **Mitigation strategies:**
  - Use a separate and untouched dataset for final validation of the strategy.
  - Limit the number of trials and parameter adjustments during the development phase.
  - Disclose the number of trials and methodology used to arrive at the final strategy.
  - Employ statistical techniques to correct for multiple testing and selection bias.

- Understanding and avoiding overfitting is crucial for developing robust investment strategies. It requires discipline, transparency, and rigorous validation processes to ensure that backtested performance is indicative of a strategy's potential in live trading.

p-hacking

- **p-value in hypothesis testing:**
  - In statistics, the p-value assesses the probability of observing the given results under the null hypothesis.
  - A small p-value (< 0.01 - 0.05) suggests strong evidence to reject the null hypothesis.
  - Used across scientific disciplines to support or reject hypotheses.
- **p-hacking defined:**
  - The practice of performing multiple hypothesis tests and only reporting those with significant p-values.
  - Can lead to misleading conclusions by cherry-picking data that appears statistically significant.
  - Increases the risk of false positives, presenting unreliable findings as true.
- **Examples of p-hacking:**
  - Reporting a portfolio's excellent results over a specific period while omitting weaker results from a broader timeframe.
  - Highlighting profitability using a specific stock universe without disclosing that variations yield poorer performance.

p-hacking

- **Consequences of *p*-hacking:**
  - Often, the number of experiments conducted is not disclosed, misleading readers to believe results came from a single trial.
  - Most claimed research findings in financial economics may be false due to *p*-hacking.

- **Quotes:** (Harvey 2017)

  "Empirical research in financial economics relies too much on p-values, which are poorly understood in the first place."

  "Journals want to publish papers with positive results and this incentivizes researchers to engage in data mining and p-hacking."

p-hacking

- **Criticism of empirical research:**
  - Overreliance on *p*-values, which are often misunderstood.
  - Journals' preference for publishing papers with positive results incentivizes data mining and *p*-hacking.

- **Mitigation strategies:**
  - Transparency in reporting the number of tests conducted and the selection criteria for reported results.
  - Adoption of stricter statistical standards, including correction methods for multiple testing.
  - Encouraging the publication of negative results to counteract publication bias.

Backtests are not experiments

- **Backtests are not experiments:**
  - Experiments can be controlled and repeated in a lab setting to isolate variables.
  - Backtests simulate historical performance of a strategy and cannot be repeated in the same way.

- **Limitations of backtesting:**
  - A backtest does not prove the effectiveness of a strategy.
  - It cannot guarantee future performance or even replicate past performance if the past were to occur again.
  - Randomness and unique historical events mean the past will not exactly repeat itself.

- **Implications:**
  - Backtesting should be used cautiously and not be seen as conclusive evidence of a strategy's success.
  - Investors and researchers should be aware of the limitations and not rely solely on backtest results for decision-making.

The paradox of flawless backtests

- **Inherent irony:** (López de Prado 2018)
  - Even a backtest that appears flawless can still be fundamentally wrong.
  - A flawless backtest implies reproducibility, consideration of slippage, transaction costs, etc., yet its positive outcome may not be reliable.

- **Expertise and false discovery:**
  - Crafting a flawless backtest requires significant expertise and experience.
  - An expert likely has conducted numerous backtests, increasing the risk of encountering a statistical fluke or false discovery due to overfitting.

- **Overfitting and false discoveries:**
  - The more backtests run on the same dataset, the higher the likelihood of stumbling upon patterns that are merely coincidental.
  - This phenomenon is exacerbated by the expertise and proficiency gained in backtesting over time.

The paradox of flawless backtests

- **Implications:**
  - A good performance in a backtest, even if flawlessly executed, does not guarantee future success.
  - The process of backtesting, especially when done extensively, inherently increases the risk of identifying misleading patterns.

- **Mitigation strategies:**
  - Awareness of the limitations and potential pitfalls of backtesting is crucial.
  - Employing rigorous statistical methods to correct for multiple testing and selection bias.
  - Validation of backtest results with out-of-sample data or through live trading to confirm robustness.

- **Paradox of flawless backtests:** Even expertly conducted backtests can mislead by mistaking statistical anomalies for genuine patterns. Recognizing backtesting limitations and employing strategies to mitigate false discoveries is crucial for developing robust investment strategies.

Limitations of backtesting insights

- **Limited insight:**
  - Backtesting often fails to explain why a strategy would have been profitable, similar to attributing skill to lottery winnings.
- **Ex-post rationalization:**
  - Like lottery winners crafting narratives of deserving luck, backtesters may construct stories to justify past success.
- **The "alpha" illusion:**
  - Claims of discovering numerous "alphas" or "factors" are often akin to highlighting winning lottery tickets without acknowledging the vast number of attempts.
- **Omission of failures:**
  - Authors rarely disclose the multitude of simulations and failed attempts behind the few successful "alphas."
- **Implications:**
  - The true value of backtesting is limited by its inability to provide genuine insights into strategy effectiveness.
  - Recognizing the inherent limitations and biases in backtesting is crucial for a realistic assessment of investment strategies.

Purpose and limitations of backtesting

- **Identifying underperforming strategies:**
  - Backtesting's primary utility is in eliminating strategies that show poor historical performance, rather than guaranteeing future success.
- **Sanity checks:**
  - Provides a reality check on various aspects such as bet sizing, turnover, cost resilience, and scenario behavior.
- **Discarding vs. improving models:**
  - The goal is to discard bad models, not to improve them based on results (overfitting).
- **Model adjustment warning:**
  - Adjusting a model based on backtesting is discouraged as it risks overfitting and is considered ineffective.
- **Strategy development:**
  - Emphasis should be on developing a sound strategy upfront; modifications post-backtesting are too late and not advised.
- **Backtesting challenges:**
  - While valuable, conducting effective and reliable backtesting is challenging.

Recommendations to avoid overfitting in backtesting

From (López de Prado 2018):

- **Broad model development:** Focus on entire asset classes or investment universes instead of specific securities to lower false discovery chances.
- **Model averaging:** Use model averaging to prevent overfitting and reduce forecasting error variance.
- **Complete research before backtesting:** Avoid the cycle of tweaking parameters and repeatedly backtesting.
- **Track backtest count:** Record the number of backtests on a dataset to estimate overfitting probability and adjust the Sharpe ratio accordingly.
- **Scenario simulation:** Perform stress tests and scenario simulations, ensuring profitability across various conditions, not just historical paths.

Recommendations to avoid overfitting in backtesting

From (Arnott, Harvey, and Markowitz 2019):

- **Foundation and hypothesis:** Develop and adhere to hypotheses before testing to prevent fitting models to data without a theoretical basis.
- **Documentation and transparency:** Document all testing trials and predefine sample data and transformations to maintain integrity and evaluate statistical significance.
- **Out-of-sample awareness:** Acknowledge the limitations of out-of-sample data due to prior market knowledge; recognize live trading as the true test.
- **Cost consideration:** Incorporate all trading costs and fees to ensure realistic strategy evaluation.
- **Model discipline:** Avoid adjusting the model based on backtest results to prevent overfitting.
- **Simplicity and regularization:** Prioritize model simplicity and employ regularization techniques to enhance robustness.

Mathematical tools to combat overfitting

- **Probability of overfitting framework:** (D. H. Bailey et al. 2017) introduced a framework to assess the likelihood of backtest overfitting.
- **Minimum backtest length:** (D. H. Bailey et al. 2014) suggested a metric for determining the necessary backtest length to avoid overfitting.
- **Probabilistic Sharpe ratio (PSR):** (D. H. Bailey and López de Prado 2012) created the PSR to estimate the probability of a Sharpe ratio outperforming a benchmark.
- **Deflated Sharpe ratio (DSR):** (D. Bailey and López de Prado 2014) developed the DSR to adjust the Sharpe ratio for multiple testing, especially with non-normal returns.
- **Online overfitting tools:** (D. H. Bailey et al. 2016) provided tools to illustrate the ease of overfitting and its financial impact.
- **Multiple randomized backtests:** Later we discuss executing multiple randomized backtests to prevent overfitting (Palomar 2025, chap. 8).

Outline

Backtesting with historical market data

- A backtest evaluates the out-of-sample performance of an investment strategy using past observations.
- These observations can be used in a multitude of ways from the simplest method to more sophisticated versions.
- Two approaches to use the past historical observations:
  1. Directly to assess the historical performance as if the strategy had been run in the past.
  2. Indirectly to simulate scenarios that did not happen in the past, such as stress tests.

- We now focus on using the historical data directly to assess the performance. Four types of backtest methods:
  - vanilla (one-shot) backtest
  - walk-forward backtest
  - k-fold cross-validation backtest
  - multiple randomized backtest

Vanilla backtest

Overview:

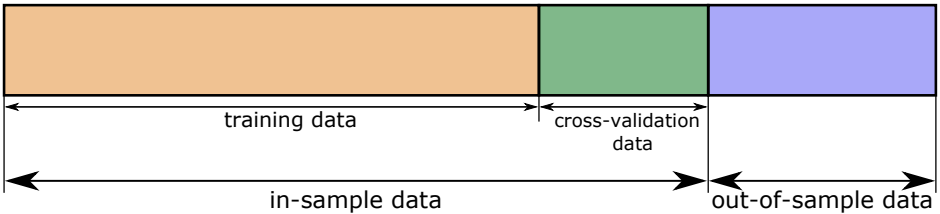
- A vanilla backtest is the simplest form of backtesting.
- It involves splitting data into in-sample and out-of-sample sets.
- The in-sample data is used for strategy design, while the out-of-sample data is for evaluation.

Data splitting:

- In-sample data is typically divided into training data and cross-validation (CV) data.
- Out-of-sample data is also referred to as test data.
- Training data is for fitting the model; CV data is for selecting hyper-parameters.
- A gap may be left between in-sample and out-of-sample data to simulate execution delay.

Vanilla backtest

Data splitting in a vanilla backtest:



Vanilla backtest

Example process:

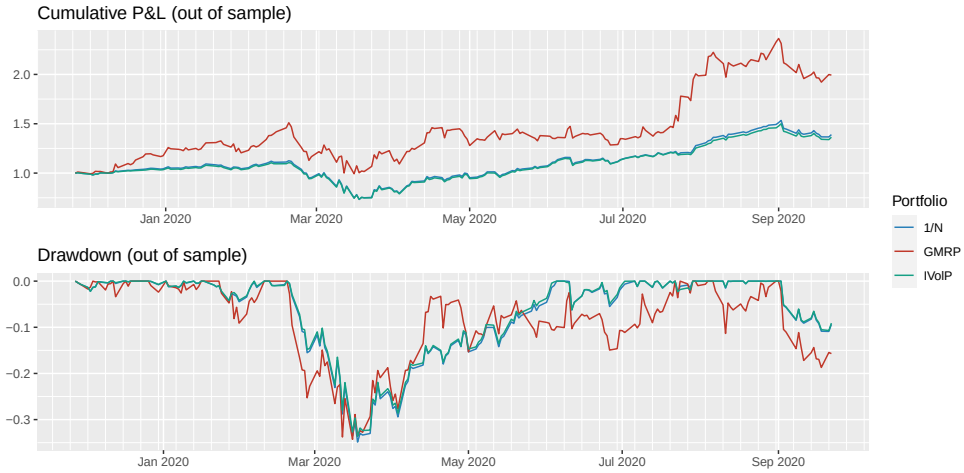
- Data might be split 70% in-sample (further split into 70% training, 30% CV) and 30% out-of-sample.
- Training data could estimate mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
- CV data might choose hyper-parameter λ in a mean-variance portfolio.
- After selecting λ, the model is refitted with all in-sample data.
- Test data is then used to assess the portfolio's performance.

Limitations:

1. A single historical path is evaluated, which may not represent different market conditions.
2. The backtest execution does not reflect real-life portfolio management, where strategies are updated with new data.

Vanilla backtest

Vanilla backtest: cumulative P&L and drawdown:



Vanilla backtest

Vanilla backtest: performance measures:

Portfolio	Sharpe ratio	annual return	annual volatility	Sortino ratio	max drawdown	CVaR (0.95)
1/N	1.18	49%	42%	1.63	35%	7%
GMRP	1.62	105%	65%	2.58	34%	9%
IVolP	1.14	46%	41%	1.58	34%	6%

Walk-forward backtest

Overview:

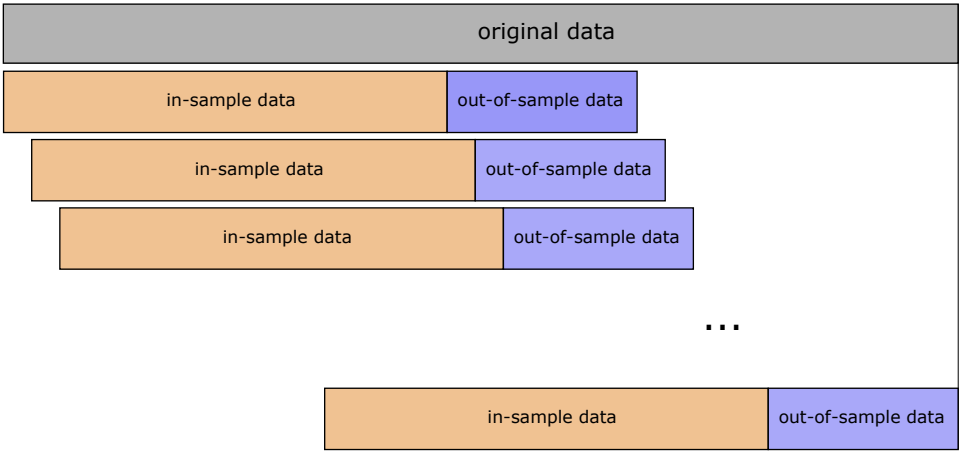
- The walk-forward backtest enhances the vanilla backtest by updating the portfolio with new data, simulating live trading practices.
- It's a historical simulation reconcilable with paper trading and is widely used in financial literature.

Methodology:

- **Rolling-window basis:** At time t, uses a lookback window of the past k samples $(t-k, \ldots, t-1)$ as in-sample data.
- **Expanding-window variation:** Uses all previous data up to time $t-1$, allowing the window to expand over time.
- A gap may be left between the last observation and portfolio execution to avoid look-ahead bias.

Walk-forward backtest

Data splitting in a rolling-window or walk-forward backtest:



Walk-forward backtes

Limitations:

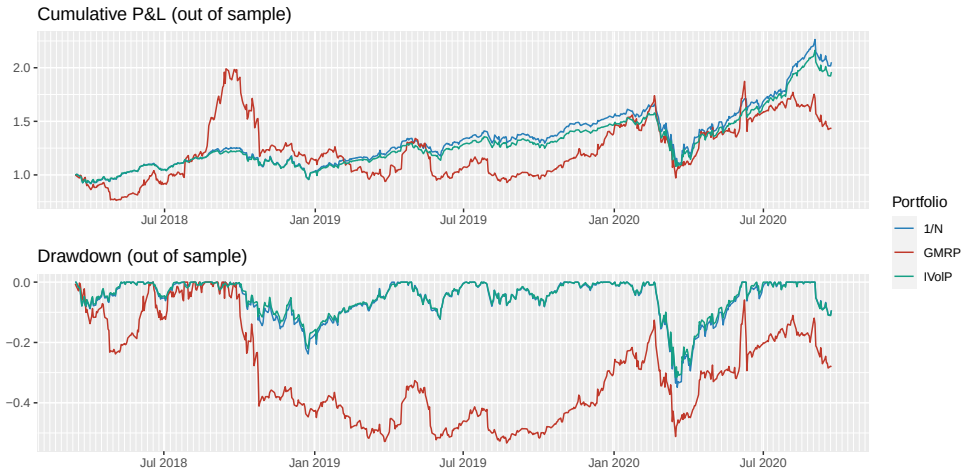
1. Evaluates only one historical scenario, raising overfitting concerns.
2. Risk of look-ahead bias due to incorrect use of future information.

Summary:

- The walk-forward backtest closely mirrors real trading by periodically updating the strategy with new data.
- Despite its realism and common use in finance, it's crucial to be aware of its limitations, including the potential for overfitting and the risk of time alignment errors.

Walk-forward backtest

Walk-forward backtest: cumulative P&L and drawdown:



Walk-forward backtest

Walk-forward backtest: performance measures:

Portfolio	Sharpe ratio	annual return	annual volatility	Sortino ratio	max drawdown	CVaR (0.95)
1/N	1.10	33%	30%	1.54	35%	5%
GMRP	0.54	27%	50%	0.75	53%	8%
IVolP	1.09	31%	28%	1.52	32%	4%

k-fold cross-validation backtest

Overview:

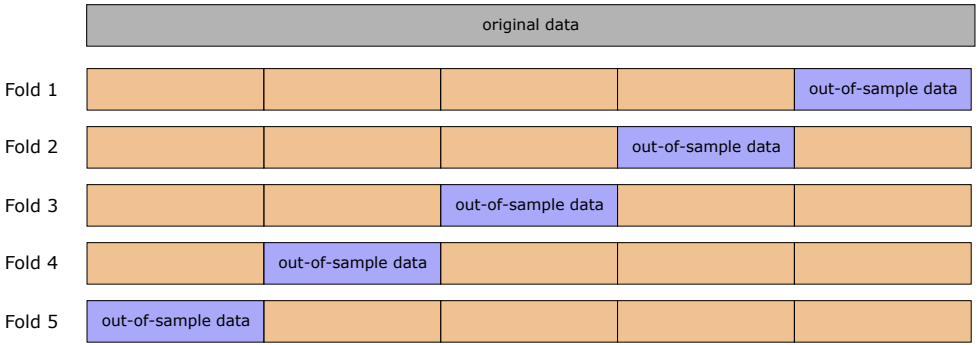
- The k-fold cross-validation backtest addresses the limitation of evaluating a single historical path, inherent in vanilla and walk-forward backtests, by testing k alternative scenarios.
- k-fold cross-validation is a widely used technique in machine learning to evaluate the generalizability of a model across different subsets of data.
- It divides data into k subsets for training and testing. However, its application in finance is problematic due to the unique characteristics of financial data.

Methodology:

- **Partitioning:** The dataset is divided into k subsets.
- **Cross-validation steps:**
  1. **Training:** For each subset $i = 1, \ldots, k$, the model is trained on all subsets except for subset i.
  2. **Testing:** The model, trained excluding subset i, is then tested on subset i.

k-fold cross-validation backtest

Data splitting in a k-fold cross-validation backtest (with $k = 5$):



k-fold cross-validation backtest

Benefits:

- **Multiple scenarios:** By testing the model on k different scenarios, this method provides a more robust assessment of its performance.
- **Reduced overfitting:** The approach helps mitigate the risk of overfitting by ensuring the model is not overly tailored to a specific subset of data.

Issues:

- **Implicit assumption:** The order of data blocks in k-fold cross-validation is assumed to be irrelevant, suitable for i.i.d. (independent and identically distributed) data.
- **Financial data reality:** Financial returns, while possibly uncorrelated, are not independent. Phenomena like volatility clustering indicate a temporal structure, making the i.i.d. assumption invalid.

k-fold cross-validation backtest

Issues with k-fold cross-validation in finance:

1. **Single data path:** Still relies on a single historical data path, limiting scenario diversity.
2. **Historical interpretation:** Lacks a clear historical interpretation due to the non-sequential testing of data subsets.
3. **Leakage risk:** High likelihood of leakage since training data may not strictly precede test data, compromising the test's integrity.

Summary:

- While k-fold cross-validation is valuable in many machine learning contexts, its application in financial backtesting is fraught with challenges.
- The temporal structure of financial data and the high risk of leakage make it a potentially dangerous method in practice.
- Alternative backtesting approaches that respect the temporal order of financial data are recommended to avoid these pitfalls.

Multiple randomized backtests

Overview:

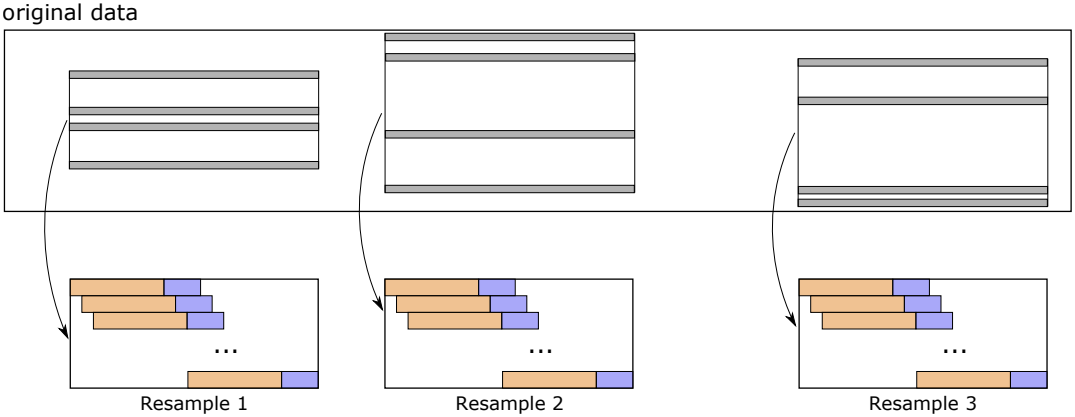
- Multiple randomized backtests address the limitations of evaluating a single historical path, as seen in vanilla and walk-forward backtests, and the leakage issue in k-fold cross-validation.
- This method generates various backtests, each representing a different historical scenario, while maintaining the chronological order of data to prevent leakage.

Process:

1. **Data preparation:** Start with extensive historical data across a broad time frame and asset range.
2. **Execution:** Repeat k times:
   a. Resample dataset by randomly selecting a subset of assets and a contiguous time period.
   b. Conduct a walk-forward or rolling-window backtest on this resampled dataset.
3. **Analysis:** Compile and analyze statistics from the k backtests to assess overall strategy performance.

Multiple randomized backtests

Data splitting in multiple randomized backtests:



Multiple randomized backtests

Advantages:

- **Diverse historical scenarios:** By simulating multiple scenarios, this method offers a broader evaluation of strategy performance across different market regimes.
- **Reduced leakage risk:** Maintaining the chronological order of data minimizes the risk of leakage, enhancing the reliability of backtest results.

Summary:

- While not guaranteeing future performance, multiple randomized backtests provide a more comprehensive and accurate assessment of a strategy's potential, making it a preferred method for backtesting in finance.
- This approach effectively mitigates the main drawbacks of other backtesting methods by incorporating randomness and respecting data temporality.

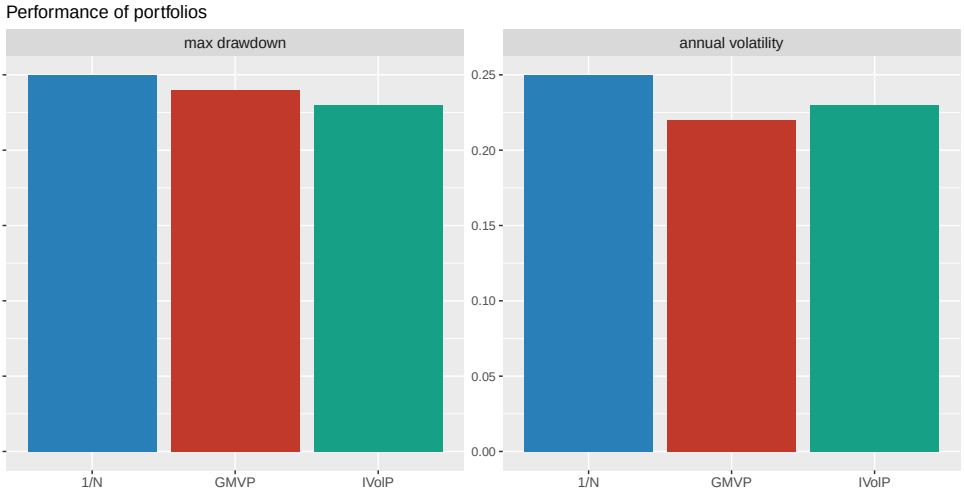
Multiple randomized backtests

Multiple randomized backtest: performance measures:

Portfolio	Sharpe ratio	annual return	annual volatility	Sortino ratio	max drawdown	CVaR (0.95)
1/N	1.01	28%	25%	1.42	25%	4%
GMVP	0.72	16%	22%	1.01	24%	3%
IVolP	0.94	24%	23%	1.33	23%	3%

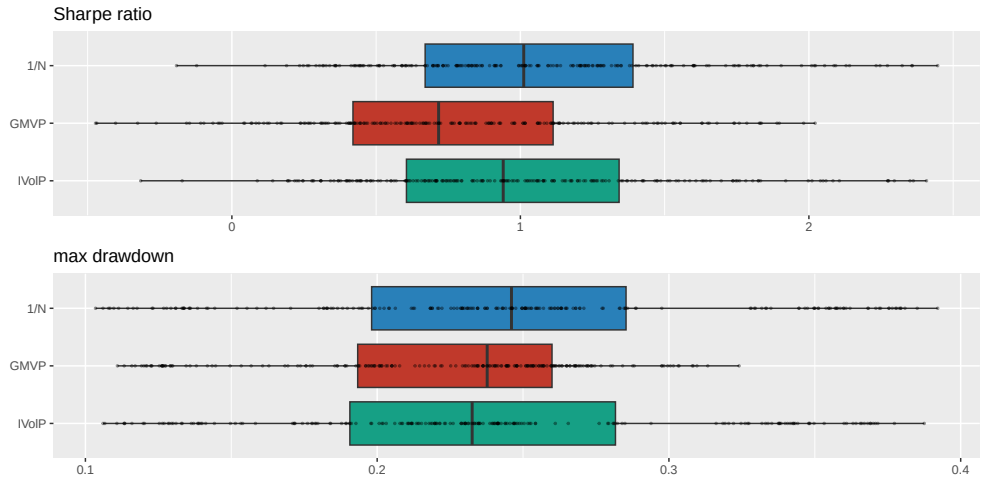
Multiple randomized backtests

Multiple randomized backtests: barplots of maximum drawdown and annualized volatility:



Multiple randomized backtests

Multiple randomized backtests: boxplots of Sharpe ratio and maximum drawdown:



Outline

Backtesting with synthetic data

- **Monte Carlo simulations**
  - Generate synthetic yet realistic data to simulate scenarios not present in historical data
  - Allows backtesting strategies on large number of unseen testing sets
  - Reduces likelihood of overfitting to a particular dataset

- **Advantages of synthetic data**
  - Stress test strategies under different market scenarios
  - Evaluate performance in extreme or rare events
  - Increase the number of testing sets for more robust backtesting

Backtesting with synthetic data

- **Approaches to Monte Carlo simulations**
  - **Parametric methods**
    - Postulate and fit a model to the data
    - Generate synthetic data from the fitted model
    - Quality depends on model assumptions
  - **Nonparametric methods**
    - Directly resample historical data without modeling
    - More robust but may destroy temporal structure
  - **Hybrid methods**
    - Combine modeling and resampling for model residuals
    - Model as much structure as possible
    - Generate residuals using parametric or nonparametric approach

Synthetic data: i.i.d. assumption

- **Parametric method**
  - Model returns as i.i.d. (independent and identically distributed)
  - Fit a distribution function (e.g., Gaussian or heavy-tailed)
  - Generate synthetic data from the fitted distribution
- **Nonparametric method**
  - Resample original returns with replacement (assuming i.i.d.)
- **Limitations of i.i.d. assumption**
  - Financial data violates the i.i.d. assumption
  - Absolute returns exhibit volatility clustering (not captured by i.i.d.)
- **Issues with i.i.d. methods**
  - Destroy volatility clustering structure present in original data
  - Parametric method assumes Gaussian distribution (fails to capture heavy tails)
  - Nonparametric method disperses deep spikes instead of clustering them
- **Need for incorporating temporal structure**
  - More realistic synthetic data generation requires modeling temporal dependencies
  - Capture volatility clustering and other stylized facts of financial data

Synthetic data: i.i.d. assumption

Example of an original sequence and two synthetic sequences generated with i.i.d. parametric and nonparametric methods:

Synthetic data: Temporal structure

- **Sophisticated modeling approach**
  - Expected returns modeled based on past values:

$$\boldsymbol{\mu}_t = \boldsymbol{f}(\boldsymbol{r}_{t-k}, \ldots, \boldsymbol{r}_{t-1})$$

  - Returns expressed as forecast plus residual error:

$$\boldsymbol{r}_t = \boldsymbol{\mu}_t + \boldsymbol{u}_t$$

    where $\boldsymbol{u}_t$ is a zero-mean residual error with covariance matrix $\boldsymbol{\Sigma}$.
  - Advanced model includes covariance model for $\boldsymbol{\Sigma}_t$:

$$\boldsymbol{r}_t = \boldsymbol{\mu}_t + \boldsymbol{\Sigma}_t^{1/2}\boldsymbol{\epsilon}_t$$

    where $\boldsymbol{\epsilon}_t$ is a standardized zero-mean identity-covariance residual error.
- **Generating synthetic residuals and returns**
  - **Parametric approach**
    - Model residuals with an i.i.d. model
    - Generate new synthetic residuals
  - **Nonparametric approach**
    - Resample residuals from historical data

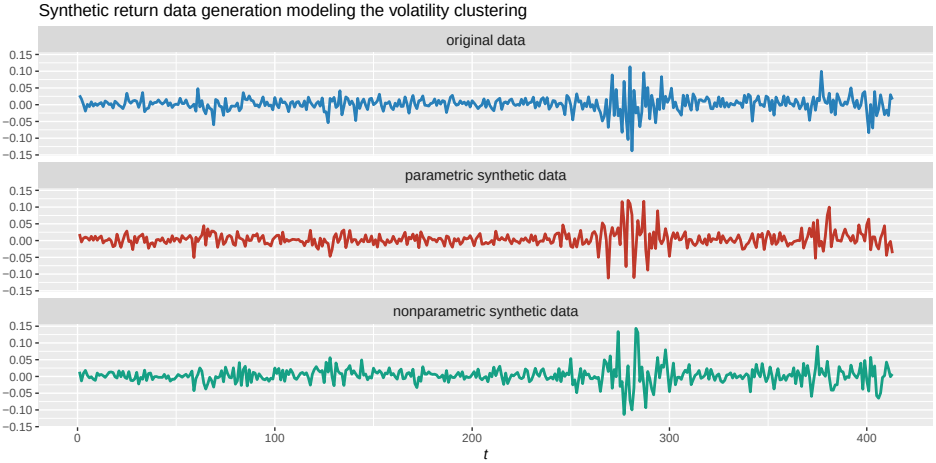
Synthetic data: Temporal structure

- **Preservation of volatility clustering**
  - Both methods aim to preserve volatility clustering in synthetic data
  - **Parametric method issues**
    - Assumes Gaussian distribution for residuals
    - Could be improved with a heavy-tailed distribution
  - **Nonparametric method advantages**
    - More robust to modeling errors
    - Directly resamples original residuals

- **Conclusion**
  - Hybrid method combines sophisticated modeling of returns and residuals
  - Allows for more accurate synthetic data generation
  - Preserves important financial data characteristics like volatility clustering

Synthetic data: Temporal structure

Example of an original sequence and two synthetic sequences generated by modeling the volatily clustering and the residuals with parametric and nonparametric methods:



Synthetic data: Stress tests

- **Purpose of stress tests**
  - Generate realistic synthetic data for different market scenarios
  - Test investment portfolio resilience against potential future financial situations

- **Customized market scenarios**
  - Strong bull market
  - Weak bull market
  - Side market (range-bound)
  - Weak bear market
  - Strong bear market

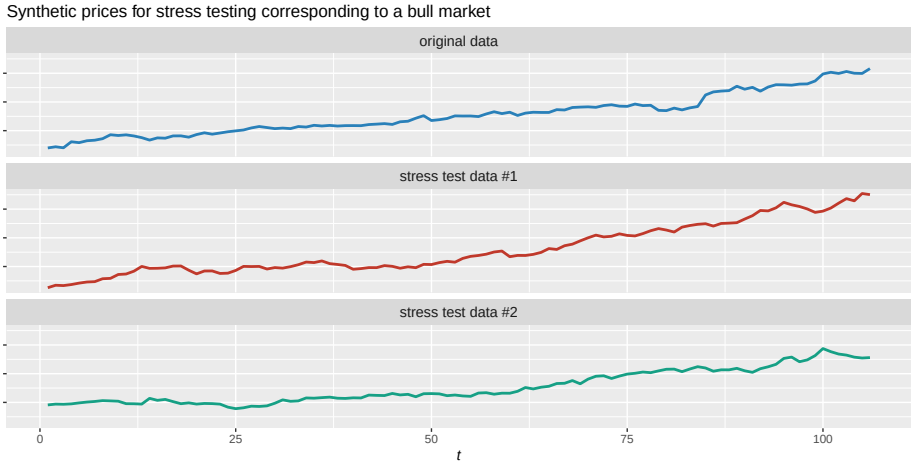
- **Historical crisis periods for stress testing**
  - Stock market crash of October 1987
  - Asian crisis of 1997
  - Tech bubble burst in 1999-2000

Synthetic data: Stress tests

- **Stress testing process**
  - Recreate specific market conditions or periods
  - Assess strategy performance under these extreme conditions
- **Significance of stress testing**
  - Provides insights into how strategies might perform during market extremes
  - Helps in understanding potential risks and tailoring risk management strategies
- **Illustrations of stress test scenarios**
  - **Bull market stress test**
    - Synthetic data generation based on a reference period (e.g., April-August 2020)
  - **Bear market stress test**
    - Synthetic data generation based on a reference period (e.g., September-December 2018)

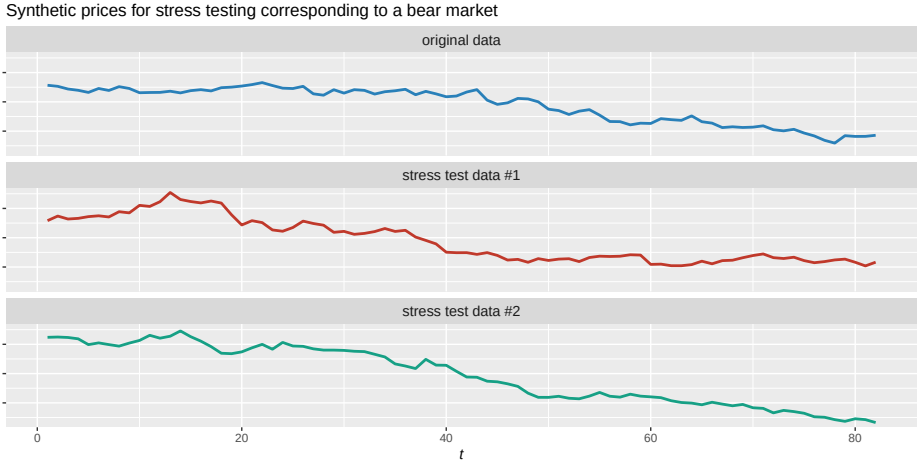
Synthetic data: Stress tests

Example of original data corresponding to a bull market and two synthetic generations of bull markets for stress testing:



Synthetic data: Stress tests

Example of original data corresponding to a bear market and two synthetic generations of bear markets for stress testing:



Outline

Summary

Backtesting is crucial for strategy development and evaluation but is often misunderstood and its risks underestimated. Key considerations include:

- Backtest results are questionable due to biases like survivorship, look-ahead, storytelling, overfitting, transaction costs, outliers, and shorting costs.
- Overfitting is the primary reason backtests can be deceptive.
- Given these pitfalls, backtest results from publications and marketing materials should be viewed skeptically.
- It's advisable to perform varied backtests and stress tests to gauge a strategy's robustness.

This cautionary note aims to equip the reader with awareness and guidance for future backtesting endeavors.

References I

Arnott, R., C. R. Harvey, and H. Markowitz. 2019. "A Backtesting Protocol in the Era of Machine Learning." *The Journal of Financial Data Science*, 64–74.

Bailey, D. H., J. M. Borwein, M. López de Prado, A. Salehipour, and Q. J. Zhu. 2016. "Backtest Overfitting in Financial Markets." *Automated Trader* 39 (2): 52–57.

Bailey, D. H., J. M. Borwein, M. López de Prado, and Q. J. Zhu. 2014. "Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance." *Notices of the American Mathematical Society* 61 (5): 458–71.

———. 2017. "The Probability of Backtest Overfitting." *Journal of Computational Finance (Risk Journals)* 20 (4): 458–71.

Bailey, D. H., and M. López de Prado. 2012. "The Sharpe Ratio Efficient Frontier." *The Journal of Risk* 15 (2): 3–44.

Bailey, D., and M. López de Prado. 2014. "The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting and Non-Normality." *Journal of Portfolio Management* 40 (5): 94–107.

Harvey, C. R. 2017. "Presidential Address: The Scientific Outlook in Financial Economics." *The Journal of Finance* 72 (4): 1399–1440.

References II

Harvey, C. R., Y. Liu, and H. Zhu. 2016"... And the Cross-Section of Expected Returns." *Review of Financial Studies* 29 (1): 5–68.

López de Prado, M. 2018. *Advances in Financial Machine Learning*. Wiley.

Luo, Y., M. Alvarez, S. Wang, J. Jussa, A. Wang, and G. Rohal. 2014. "Seven Sins of Quantitative Investing." *White Paper, Deutsche Bank Markets Research*, September.

Montier, J. 2005. "Seven Sins of Fund Management." *Dresdner Kleinwort Wasserstein - Global Equity Strategy*, September.

Palomar, D. P. 2025. *Portfolio Optimization: Theory and Application*. Cambridge University Press.