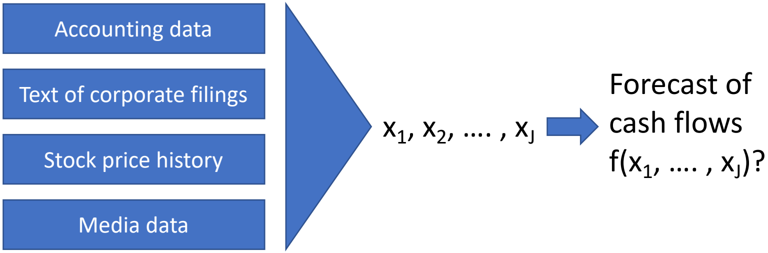
Evaluating market efficiency in a high-dimensional world

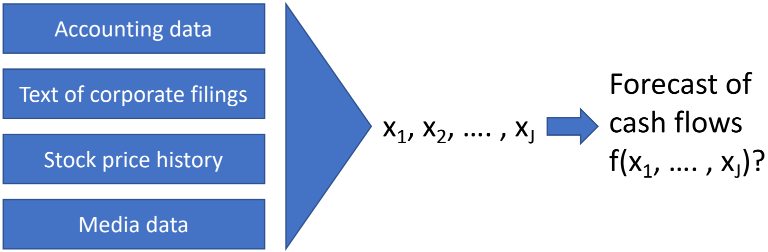
Stefan Nagel

August 2022

Investors' high-dimensional prediction problem



Investors' high-dimensional prediction problem



- Example: SEC Edgar database of corporate filings alone receives 3,000 filings per day, $\approx 3,000$ terabytes of data annually

Market efficiency in a high-dimensional world

Accounting data

Text of corporate filings

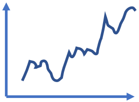
Stock price history

Media data

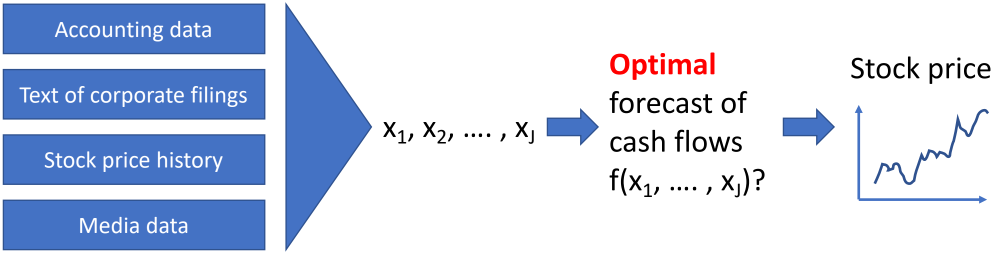
$x_1, x_2, \ldots., x_J$

Optimal forecast of cash flows $f(x_1, \ldots., x_J)$?

Stock price



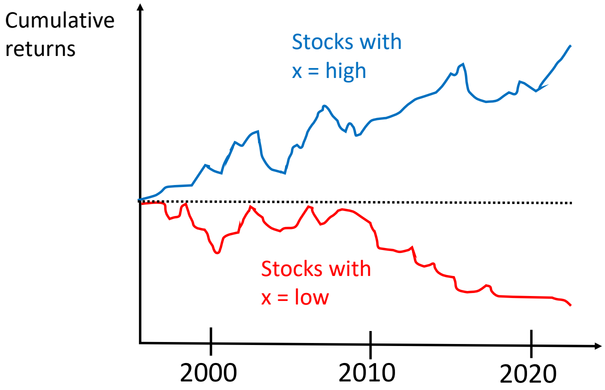
Market efficiency in a high-dimensional world



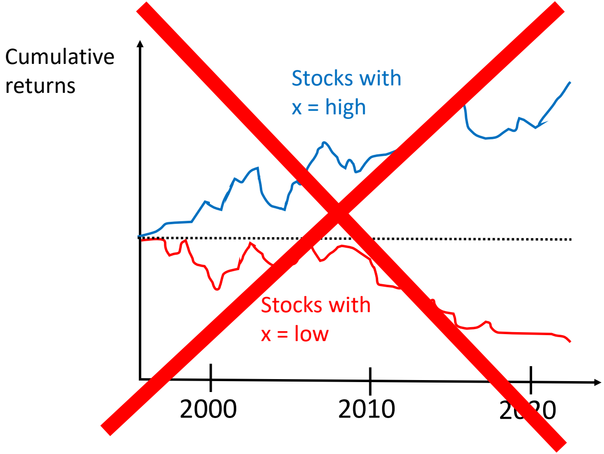
- Questions: In a high-dimensional world
  - what is the benchmark for forecast optimality, and hence market efficiency?
  - how can we detect deviations from this benchmark?

Evaluating market efficiency: Typical approach

Track relative performance of stocks with different firm characteristic x_j



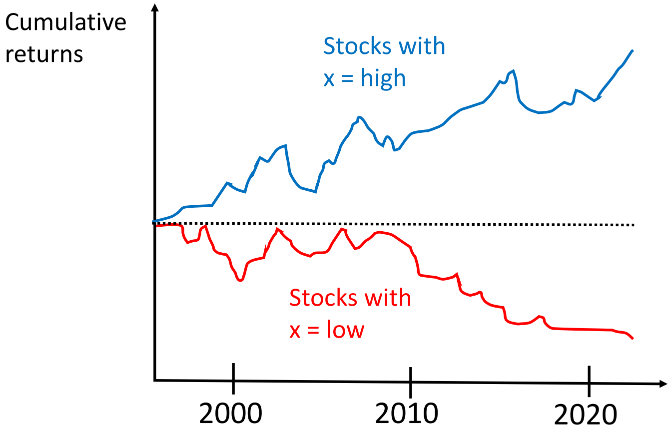
Hypothesis in standard market efficiency tests



NB: Abstract from classic joint hypothesis problem

Market efficiency rejections: Factor zoo

Finding that x_j predicts stock returns $\Rightarrow$ declare a new "factor"



Evaluating market efficiency: Alternative methods

- **Portfolio sorts**: Group stocks with similar x_j and track their performance
  - Example: small- / mid- / large-capitalization stocks

Evaluating market efficiency: Alternative methods

- **Portfolio sorts**: Group stocks with similar x_j and track their performance
  - Example: small- / mid- / large-capitalization stocks
- **Regressions**: Estimate statistical return prediction model

$$R_{t+1} = a + b_1 x_{1,t} + b_2 x_{2,t} + \dots + e_t$$

Evaluating market efficiency: Alternative methods

- **Portfolio sorts**: Group stocks with similar x_j and track their performance
  - Example: small- / mid- / large-capitalization stocks
- **Regressions**: Estimate statistical return prediction model

$$R_{t+1} = a + b_1 x_{1,t} + b_2 x_{2,t} + ... + e_t$$

- **Machine learning (ML)**: Accommodate very large number of predictors and nonlinearity
  - Ridge, lasso
  - Random forests
  - Neural networks

Evaluating market efficiency: The problem of investor learning

- Example: Suppose x_{jt} is found to predict r_{t+1} in historical data from 1990 to 2020.

Evaluating market efficiency: The problem of investor learning

- Example: Suppose x_{jt} is found to predict r_{t+1} in historical data from 1990 to 2020.
- Problems in interpreting this fact: For an investor in years before 2020, predictive power of x_{jt} was not necessarily knowable yet.

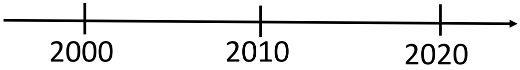
Evaluating market efficiency: The problem of investor learning

- Example: Suppose x_{jt} is found to predict r_{t+1} in historical data from 1990 to 2020.
- Problems in interpreting this fact: For an investor in years before 2020, predictive power of x_{jt} was not necessarily knowable yet.
- Problem is magnified in a high-dimensional world with many potential predictors.

Learning can generate seemingly predictable returns

x = high $\Rightarrow$ Value = $150

x = low $\Rightarrow$ Value = $50



Learning can generate seemingly predictable returns

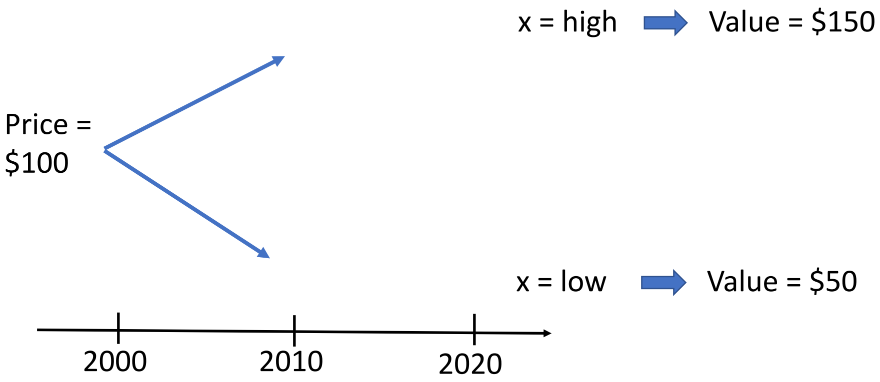
x = high ➡ Value = \$150

Price = \$100

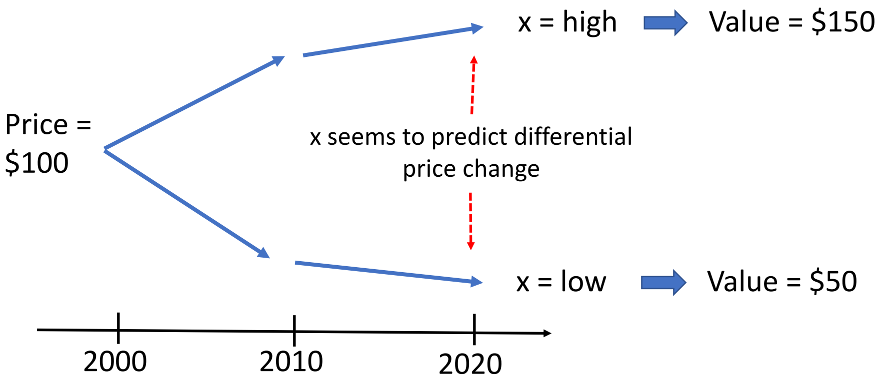
x = low ➡ Value = \$50

2000 2010 2020

Learning can generate seemingly predictable returns

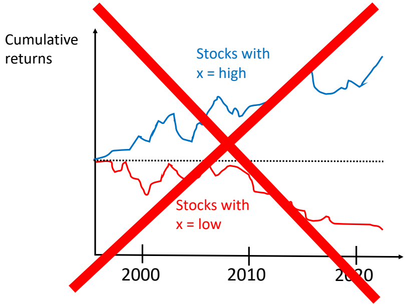


Learning can generate seemingly predictable returns



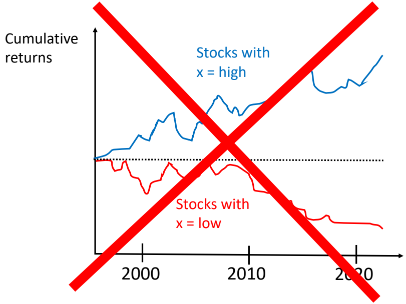
Implicit assumption in typical market efficiency tests

- Standard market efficiency tests assume absence of learning effects: investors assumed to know perfectly how predictor variables map into future cash flows



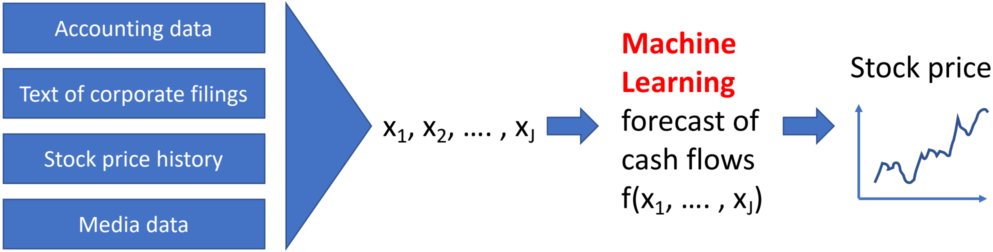
Implicit assumption in typical market efficiency tests

- Standard market efficiency tests assume absence of learning effects: investors assumed to know perfectly how predictor variables map into future cash flows



- Not a useful benchmark in high-dimensional settings where investors are faced with thousands or millions of potential predictors

Quantifying learning effects in high-dimensional environments: Modeling investors as "machine learners"



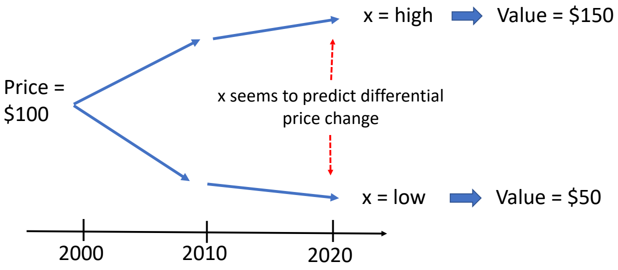
- Investors in this model face large number of potentially relevant predictor variables and learn over time how to use them for forecasting cash flows

Learning effects in asset returns

- Investors' learning problem in this model is hard $\Rightarrow$ substantial unavoidable errors

Learning effects in asset returns

- Investors' learning problem in this model is hard $\Rightarrow$ substantial unavoidable errors
- As a consequence, lots of contamination of returns with errors that look predictable with hindsight



In-sample return prediction backtest in model-generated data

- Now consider a researcher running an in-sample backtest with a regression

$$R_{t+1} = a + b_1 x_{1,t} + b_2 x_{2,t} + ... + b_3 x_{J,t} + e_t$$

in a panel of stocks.

In-sample return prediction backtest in model-generated data

- Now consider a researcher running an in-sample backtest with a regression

$$R_{t+1} = a + b_1 x_{1,t} + b_2 x_{2,t} + \dots + b_3 x_{J,t} + e_t$$

in a panel of stocks.

- How likely is that the researcher will find that returns are predictable according to conventional statistical criteria?

In-sample return prediction backtest in model-generated data

- Now consider a researcher running an in-sample backtest with a regression

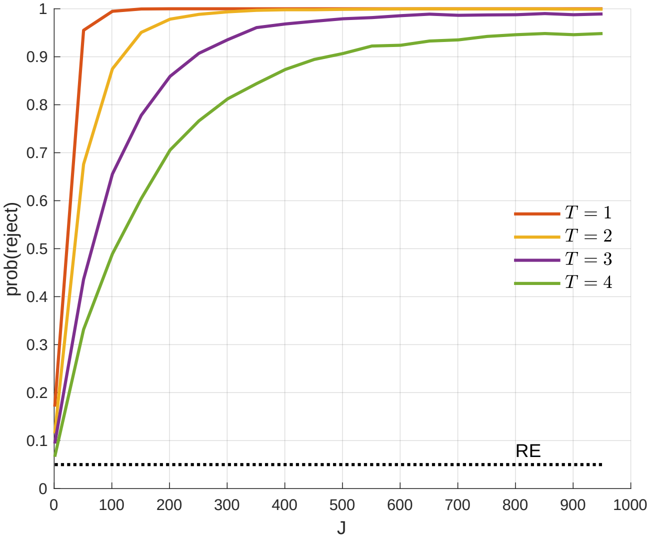
$$R_{t+1} = a + b_1 x_{1,t} + b_2 x_{2,t} + ... + b_3 x_{J,t} + e_t$$

in a panel of stocks.

- How likely is that the researcher will find that returns are predictable according to conventional statistical criteria?

- Compare with case (RE) where stocks are priced by investors with perfect knowledge of the cash-flow process parameters

Overrejection of no-return-predictability null hypothesis



Implication for market efficiency tests

- Rejection of no-predictability null hypothesis in in-sample tests can be artifact of look-ahead advantage of researcher rather than market efficiency violation

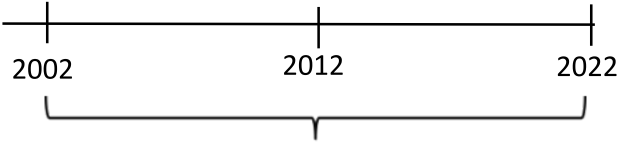
Implication for market efficiency tests

- Rejection of no-predictability null hypothesis in in-sample tests can be artifact of look-ahead advantage of researcher rather than market efficiency violation
- Researchers' look-ahead advantage vis-a-vis investors is magnified in high-dimensional setting

Implication for market efficiency tests

- Rejection of no-predictability null hypothesis in in-sample tests can be artifact of look-ahead advantage of researcher rather than market efficiency violation
- Researchers' look-ahead advantage vis-a-vis investors is magnified in high-dimensional setting
- How can one test market efficiency in a high-dimensional setting with investor learning?

Ideal, but infeasible: True out-of-sample test



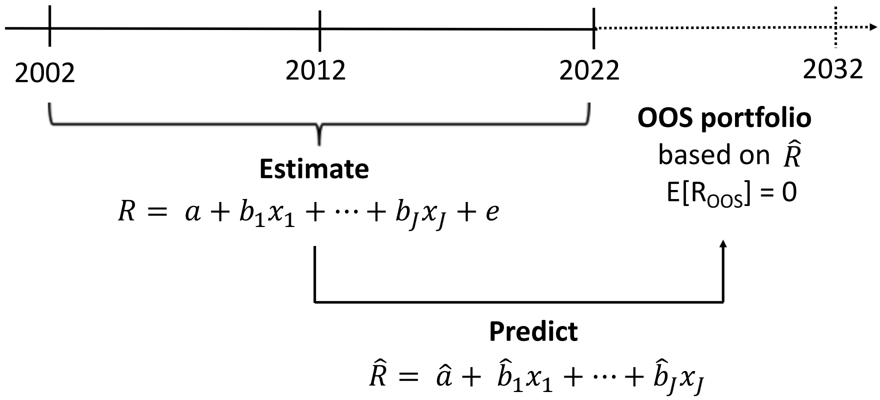
Estimate

$$R = a + b_1 x_1 + \cdots + b_J x_J + e$$

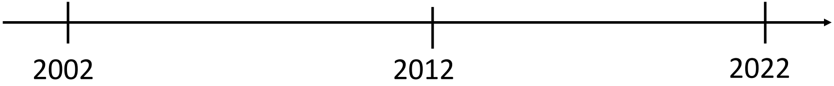
Predict

$$\hat{R} = \hat{a} + \hat{b}_1 x_1 + \cdots + \hat{b}_J x_J$$

Ideal, but infeasible: True out-of-sample test



Feasible: Pseudo-OOS backtest



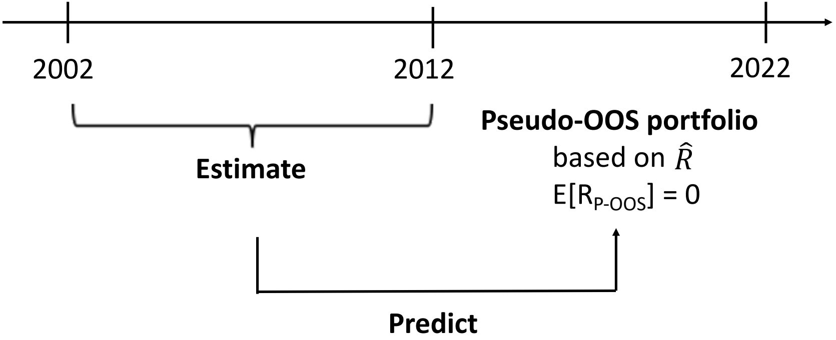
Estimate

$$R = a + b_1 x_1 + \cdots + b_J x_J + e$$

Predict

$$\hat{R} = \hat{a} + \hat{b}_1 x_1 + \cdots + \hat{b}_J x_J$$

Feasible: Pseudo-OOS backtest



$$\hat{R} = \hat{a} + \hat{b}_1 x_1 + \cdots + \hat{b}_J x_J$$

What does not work: Backtest for ex-post selected predictors

- Use full data set of returns until now to find “significant” predictors

What does not work: Backtest for ex-post selected predictors

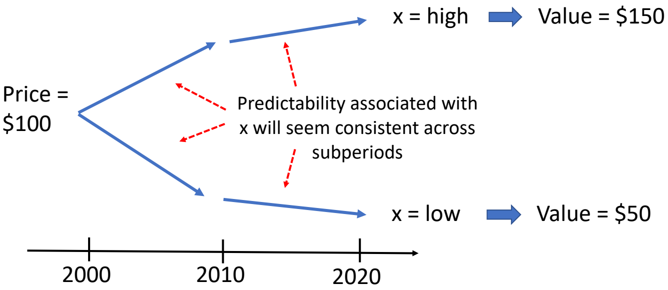
- Use full data set of returns until now to find "significant" predictors
- Then backtest these selected predictors with pseudo-OOS test, e.g.,
  1. Split data into subperiods
  2. Evaluate whether return predictability consistent across subperiods

What does not work: Backtest for ex-post selected predictors

- Use full data set of returns until now to find "significant" predictors

- Then backtest these selected predictors with pseudo-OOS test, e.g.,
  1. Split data into subperiods
  2. Evaluate whether return predictability consistent across subperiods

- Does not remove look-ahead bias:



Uncovering market inefficiencies: Shrinkage regression

- So far: testing market efficiency

Uncovering market inefficiencies: Shrinkage regression

- So far: testing market efficiency
- Now: if there are inefficiencies, how can we estimate
  - their relation to predictor variables?
  - the magnitude of inefficiencies?

Uncovering market inefficiencies: Shrinkage regression

- So far: testing market efficiency
- Now: if there are inefficiencies, how can we estimate
  - their relation to predictor variables?
  - the magnitude of inefficiencies?
- Estimation with shrinkage that maximizes pseudo-OOS predictive performance

Uncovering market inefficiencies: Shrinkage regression

- So far: testing market efficiency
- Now: if there are inefficiencies, how can we estimate
  - their relation to predictor variables?
  - the magnitude of inefficiencies?
- Estimation with shrinkage that maximizes pseudo-OOS predictive performance
- Machine learning tools allow consideration of large numbers of predictors jointly, without focusing on on arbitrary subsets or pre-selecting based on hindsight information
  - Here: Ridge regression or lasso for linear models

Uncovering market inefficiencies: Shrinkage regression

2002 2012 2022

Estimate

$$R = a + b_1 x_1 + \cdots + b_J x_J + e$$

Pseudo-OOS portfolio

based on $\hat{R}$

E[$R_{\text{P-OOS}}$]

Choose degree of shrinkage to maximize

Shrink

$$\hat{b}_1 \to 0, \ldots, \hat{b}_J \to 0$$

Predict

$$\hat{R} = \hat{a} + \hat{b}_1 x_1 + \cdots + \hat{b}_J x_J$$

Intuition: Shrinking away researchers hindsight advantage vis-a-vis investors

Application: Past returns as predictors

- Stock characteristics that have already appeared in published asset pricing studies are a selected sample, subject to look-ahead bias

Application: Past returns as predictors

- Stock characteristics that have already appeared in published asset pricing studies are a selected sample, subject to look-ahead bias
- Therefore: Use an entire class of predictor variables without pre-selecting particular variables in this class

Application: Past returns as predictors

- Stock characteristics that have already appeared in published asset pricing studies are a selected sample, subject to look-ahead bias
- Therefore: Use an entire class of predictor variables without pre-selecting particular variables in this class
- Here: Linear prediction based on lagged monthly past stock returns $r_{it-1}, r_{it-2}, \ldots, r_{it-120}$

Application: Past returns as predictors

- Stock characteristics that have already appeared in published asset pricing studies are a selected sample, subject to look-ahead bias
- Therefore: Use an entire class of predictor variables without pre-selecting particular variables in this class
- Here: Linear prediction based on lagged monthly past stock returns $r_{it-1}, r_{it-2}, \ldots, r_{it-120}$
- Construct 120 portfolios, each weighted by market-adjusted returns lagged k months

Past-return-based anomalies

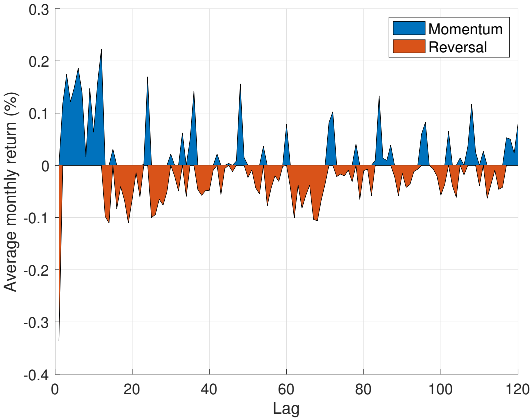
- Prior research has selectively focused on subsets and did not adjust for learning effects
  - DeBondt and Thaler (1985): 3- to 5-year reversals (1926-1982)
  - Jegadeesh (1990): one-month reversals (1926-1982)
  - Jegadeesh and Titman (1993): 3- to 12-month momentum (1965-1987)
  - Heston and Sadka (2008): Autocorrelation at 12-month lags (1945-2002)
  - Novy-Marx (2012): Momentum at 7- to 12-month lags (1926-2010)

Past-return-based anomalies

- Prior research has selectively focused on subsets and did not adjust for learning effects
  - DeBondt and Thaler (1985): 3- to 5-year reversals (1926-1982)
  - Jegadeesh (1990): one-month reversals (1926-1982)
  - Jegadeesh and Titman (1993): 3- to 12-month momentum (1965-1987)
  - Heston and Sadka (2008): Autocorrelation at 12-month lags (1945-2002)
  - Novy-Marx (2012): Momentum at 7- to 12-month lags (1926-2010)
- Here:
  - Tests robust to investor learning
  - Characterize predicted Sharpe ratio based on using many lags of returns jointly as predictors

Full-sample historical average returns

Average returns of 120 portfolios that weight stocks by their market-adjusted returns in month $t-1$, $t-2$, ... , $t-120$



Sample period: 1926 to 2021; first portfolio returns in January 1936.

Modified Gaussian process regression

- Every month t, estimate expected returns of each of the 120 portfolios using data up to t

Modified Gaussian process regression

- Every month t, estimate expected returns of each of the 120 portfolios using data up to t
- Apply
  - **Shrinkage**: shrink away learning effects

Modified Gaussian process regression

- Every month t, estimate expected returns of each of the 120 portfolios using data up to t
- Apply
  - **Shrinkage**: shrink away learning effects
  - **Smoothing**: increase statistical power

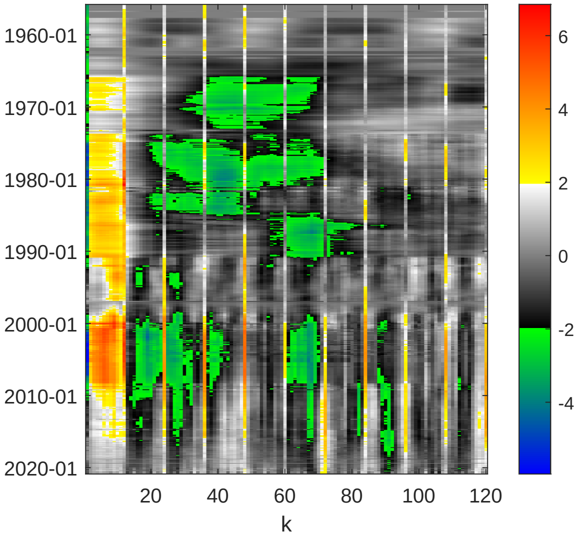
Modified Gaussian process regression

- Every month t, estimate expected returns of each of the 120 portfolios using data up to t
- Apply
  - **Shrinkage**: shrink away learning effects
  - **Smoothing**: increase statistical power
  - **Exponential weighing**: allow downweighting of data in distant past

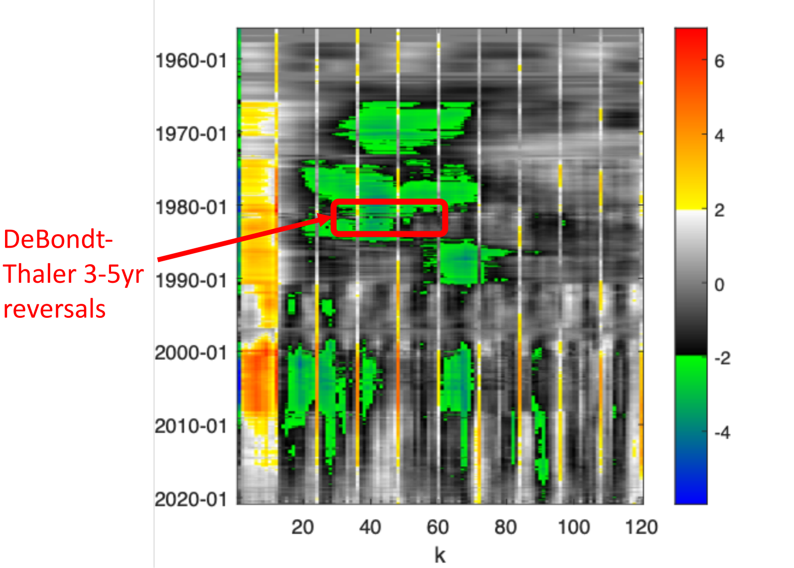
Modified Gaussian process regression

- Every month t, estimate expected returns of each of the 120 portfolios using data up to t
- Apply
  - **Shrinkage**: shrink away learning effects
  - **Smoothing**: increase statistical power
  - **Exponential weighing**: allow downweighting of data in distant past
- All optimized to achieve maximum pseudo-OOS predictive performance in data until month t

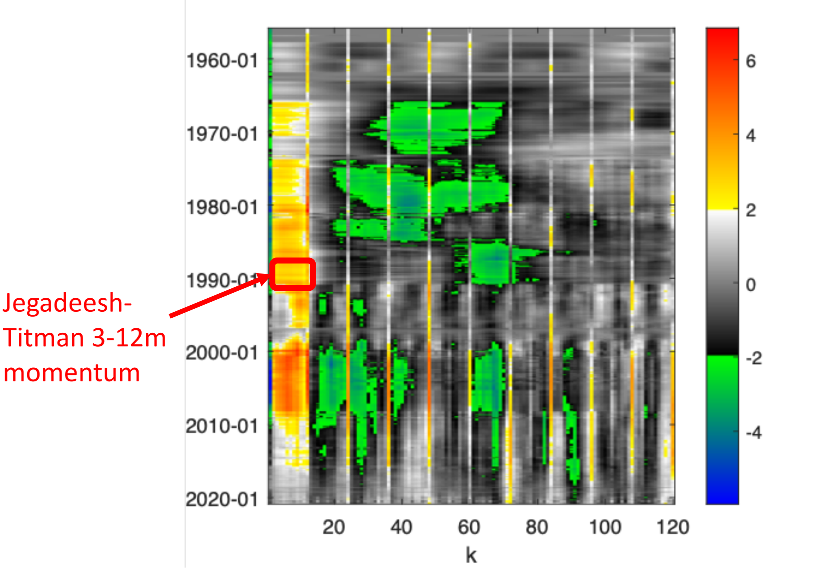
Posterior t-statistics



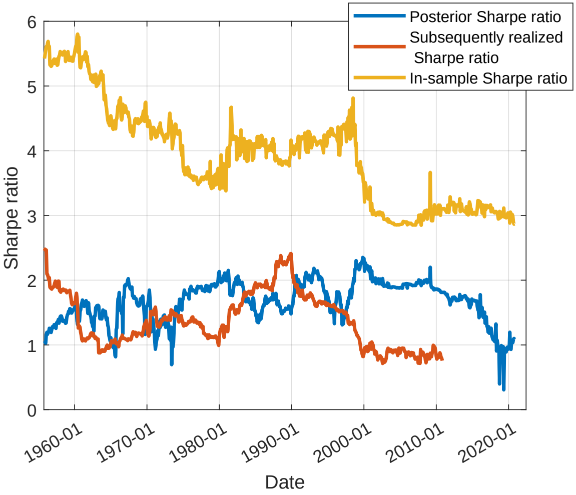
Posterior t-statistics



Posterior t-statistics



Sharpe ratios



Conclusion

- Learning contaminates stock returns with components that appear predictable with researchers' look-ahead advantage relative to investors

Conclusion

- Learning contaminates stock returns with components that appear predictable with researchers' look-ahead advantage relative to investors
- Investors' learning problem is particularly difficult in a high-dimensional setting

Conclusion

- Learning contaminates stock returns with components that appear predictable with researchers' look-ahead advantage relative to investors
- Investors' learning problem is particularly difficult in a high-dimensional setting
- Standard market efficiency tests misleading in a high-dimensional setting

Conclusion

- Learning contaminates stock returns with components that appear predictable with researchers' look-ahead advantage relative to investors
- Investors' learning problem is particularly difficult in a high-dimensional setting
- Standard market efficiency tests misleading in a high-dimensional setting
- Shrinkage methods can remove this hindsight bias from backtests, if
  - applied to universe of all predictors within a certain class
  - without selection based on full-sample returns

Conclusion

- Learning contaminates stock returns with components that appear predictable with researchers' look-ahead advantage relative to investors
- Investors' learning problem is particularly difficult in a high-dimensional setting
- Standard market efficiency tests misleading in a high-dimensional setting
- Shrinkage methods can remove this hindsight bias from backtests, if
  - applied to universe of all predictors within a certain class
  - without selection based on full-sample returns
- ML tools allow embracing of high-dimensionality in empirical asset pricing rather than forcing artificially low-dimensional models