Overfitting: Causes and Solutions

Prof. Marcos López de Prado
Cornell University – School of Engineering
ORIE 5256

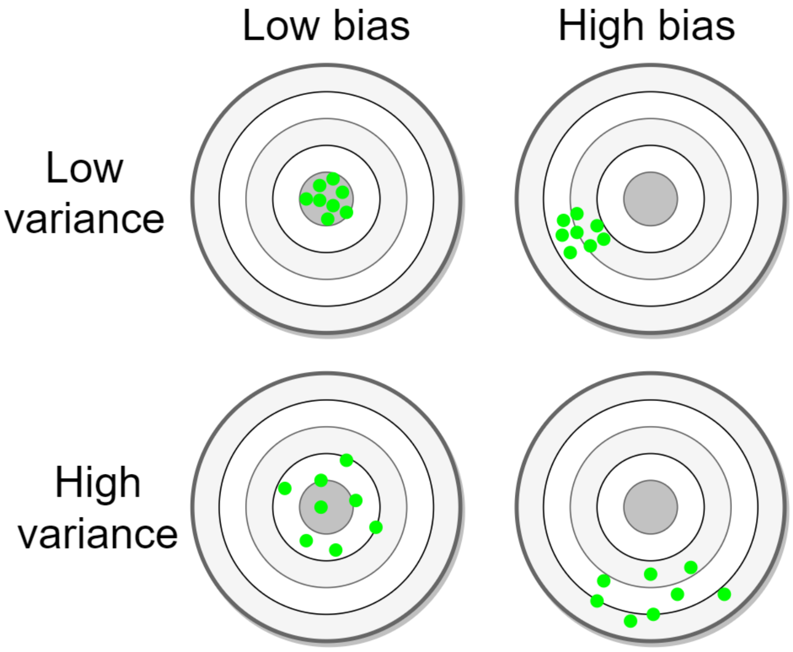
Key Points

- Scientific disciplines have successfully applied Machine Learning (ML) methods for decades
- In recent years, investment managers have begun to replace or complement classical statistical methods (e.g., Econometrics) with computer-based statistical methods (e.g., ML)
  - Well-known ML firms include RenTec, Two Sigma, DE Shaw, TGS, Capital Fund Management, etc.
- Classical methods are prone to overfitting due to their
  - reliance on train-set error estimates
  - assumption that only one trial has taken place
- **When used incorrectly, the risk of ML overfitting is higher than with classical methods**
- **This presentation reviews the causes and solutions wrt overfitting**

What is Overfitting?

Error Decomposition

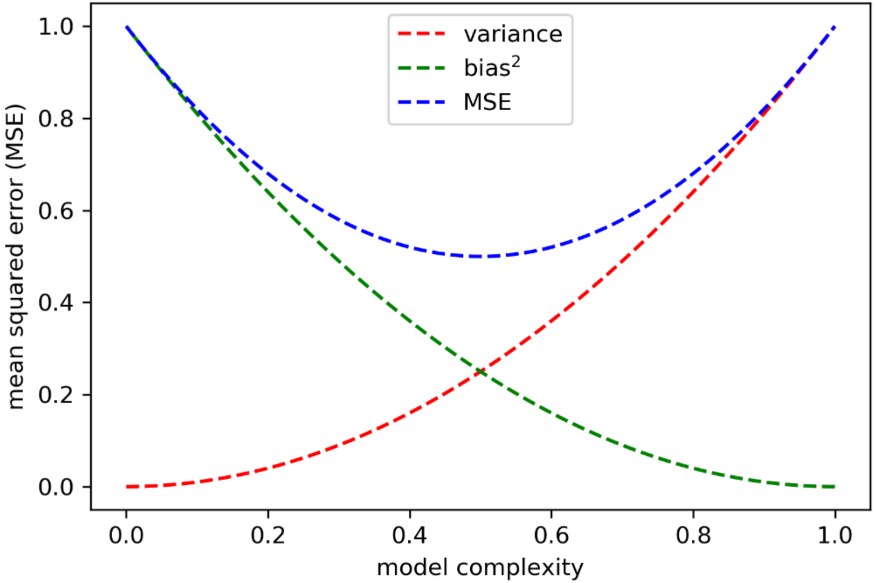
- Consider a function $f[x]$ that predicts outcomes y, such that errors $\varepsilon = y - f[x]$ are unpredictable, with $\mathrm{E}[\varepsilon] = 0$ and minimal $\mathrm{V}[\varepsilon]$
- A statistical model proposes a function $\hat{f}[x]$ that approximates $f[x]$
- The mean squared error (MSE), $\mathrm{E}\left[\left(y - \hat{f}[x]\right)^2\right]$, is the sum of:
  - Bias squared: $\left(\mathrm{E}\left[\hat{f}[x] - f[x]\right]\right)^2$
  - Variance: $\mathrm{V}\left[\hat{f}[x]\right]$
  - Noise: $\mathrm{V}[\varepsilon]$



Combination of bias and variance in an estimator.

Bias-Variance Trade-Off

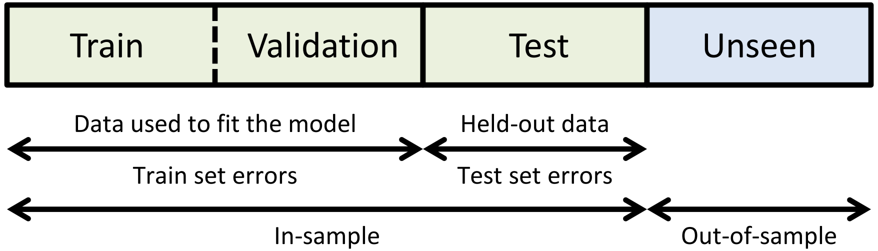
- Bias occurs when $\hat{f}[x]$ underfits the data
  - The model confounds signal for noise
- Variance occurs when $\hat{f}[x]$ overfits the data
  - The model confounds noise for signal
- In general, bias can only be reduced at the expense of increasing variance
- **Overfitting** causes model variance, because a model overfit on one set does not *generalize* well outside that set
  - The model tries to forecast noise



A good statistical model minimizes the mean squared error (MSE) by finding the optimal balance between bias and variance.

Two Kinds of Errors

- We can split a dataset into two subsets
  - **Train set**: Used to select features and fit model parameters
    - This may include a **validation set**, used to find the optimal hyper-parameters
  - **Test set**: Hold-out data, not used for fitting the model
- We can estimate two in-sample errors:
  - **Train set errors**: Errors estimated on the train set (the same data used to fit the model)
  - **Test set errors**: Errors estimated on the test set
- Overfitting can occur when we try to minimize one or both of these errors



Train

Validation

Test

Unseen

Data used to fit the model

Held-out data

Train set errors

Test set errors

In-sample

Out-of-sample

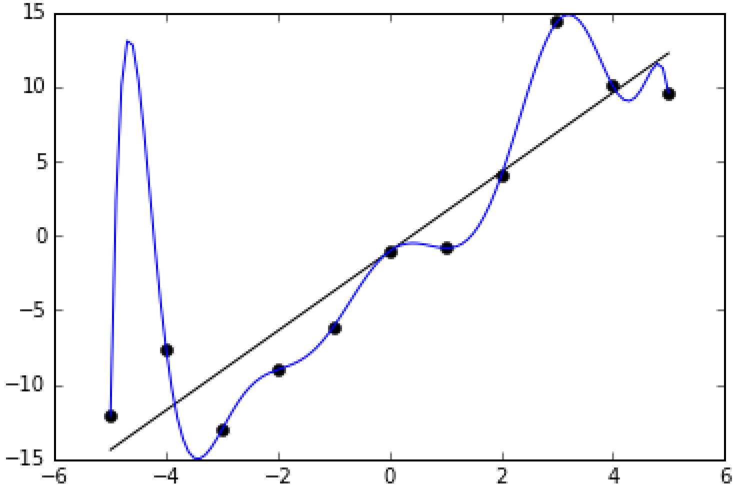
Overfitting can occur on the train set and/or the test set.

The amount of overfitting can be estimated through the **generalization error**: the model's error on data not used to choose the model.

The Two Kinds of Overfitting

Train Set Overfitting

- Train set overfitting occurs when
  - a model is chosen to minimize train set errors
  - at the expense of higher variance on test set errors
- Train set overfitting is related to model complexity
  - This overcomplexity attempts to fit signal, but it ends up fitting noise
- Train set overfitting can be easily diagnosed by estimating the generalization error on the test set, via
  - Resampling methods (e.g., cross-validation)
  - Monte Carlo
- Solutions: Simplify the model; get more data

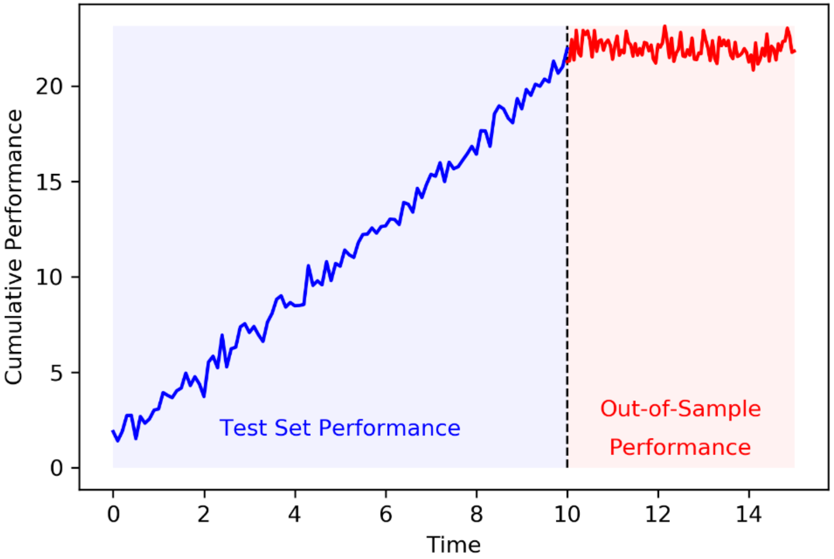


Source: Wikimedia Commons

A typical example of overfitting: The (complex) polynomial function provides a perfect fit because it explains all the noise, however it will generalize worse than a (simpler) line.

Test Set (or Backtest) Overfitting

- Test set overfitting occurs when
  - a model is chosen to minimize test set errors
  - at the expense of higher out-of-sample variance
- Test set overfitting is related to selection bias under multiple testing (SBuMT)
- Test set overfitting can be diagnosed by
  - estimating the generalization error on unseen data (out-of-sample)
  - controlling for the number and variance of the independent trials involved in the model selection
- Solutions:
  - start all over with a new (unseen) dataset
  - adjust the probability of a false positive

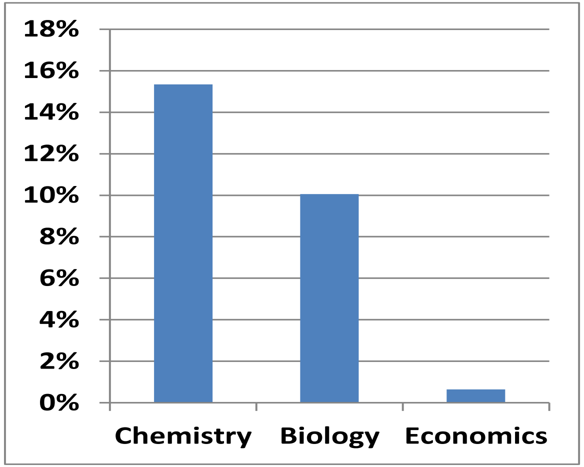


A strategy overfit on the test set will fail to perform on unseen data (out-of-sample). Note that this kind of overfitting is **entirely unrelated to model complexity**.

Classical Statistical Methods

What are Classical Statistical Methods?

- Classical statistical methods follow the research program initiated by Ronald Fisher's
  - Statistical Methods for Research Workers (1925)
  - The Design of Experiments (1935)
- This program is founded on
  - Correlation, method of moments
  - Goodness of fit, maximum likelihood estimation
  - Statistical significance, tests of hypothesis, *p*-values, ANOVA
  - Strong assumptions, needed for asymptotic properties
- This program
  - was developed before the computer age
  - was adopted by the Econometric Society (est. 1930)
  - is the backbone of the most popular Econometrics textbooks
  - has become the canon accepted/required by Financial journals

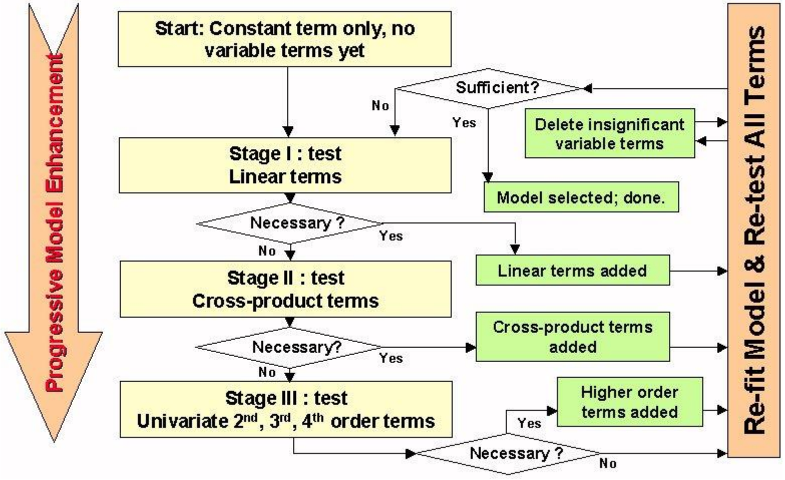


Source: The Web of Science

Fewer than 1% of journal articles in economics mention ML-related terms, such us: classifier, clustering, neural network, machine learning.

Train Set Overfitting

- Classical statistical models try to deal with train set overfitting via regularization
  - penalizing complexity (e.g., degrees of freedom)
  - reducing complexity (e.g., stepwise regression)
- However, classical models
  - do not split the data between train, validation and test sets
  - do not estimate generalization errors
- The train set is also the validation set, and the test set
  - As a result, **classical regularization fails to prevent train & test set overfitting**

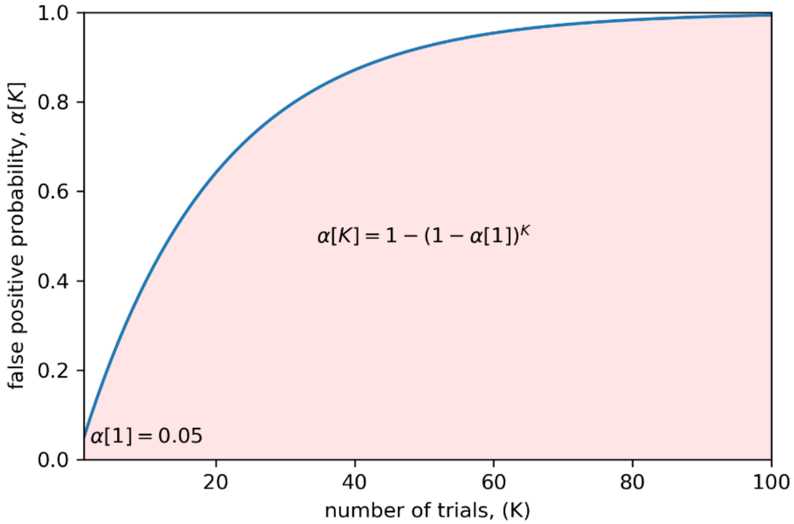

Progressive Model Enhancement
Start: Constant term only, no variable terms yet
Sufficient?
No
Yes
Delete insignificant variable terms
Stage I : test Linear terms
Model selected; done.
Necessary ?
Yes
No
Linear terms added
Stage II : test Cross-product terms
Cross-product terms added
Necessary?
Yes
No
Stage III : test Univariate 2nd, 3rd, 4th order terms
Higher order terms added
Yes
Necessary ?
No
Re-fit Model & Re-test All Terms

Source: Wikimedia Commons

Stepwise regression is often used in econometrics software and papers to reduce the complexity of a model and thus limit train set overfitting. Unfortunately, this makes it all but certain that the econometric model will suffer from test set overfitting.

Test Set Overfitting

- Classical statistical models were devised
  - before the invention of computers (e.g., Pearson-Neyman Lemma [1933])
  - to be run only once
- Classical statistics rarely controls for SBuMT
- One ubiquitous instance of test set overfitting is *p*-hacking
  - A researcher applies multiple statistical tests on the same data
  - Each test has a false positive rate of 5%
  - The combined false positive rate quickly converges to 100%

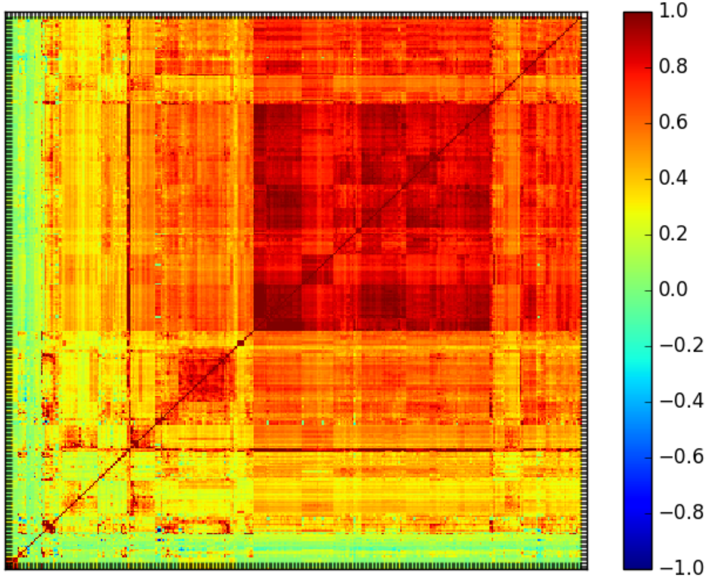


The false positive probability quickly rises after the first trial. Financial journal articles almost always present findings as if they had been the result of a single trial. Because that is rarely the case, most discoveries in finance are false.

Computer-based Statistical Methods (ML)

What is ML?

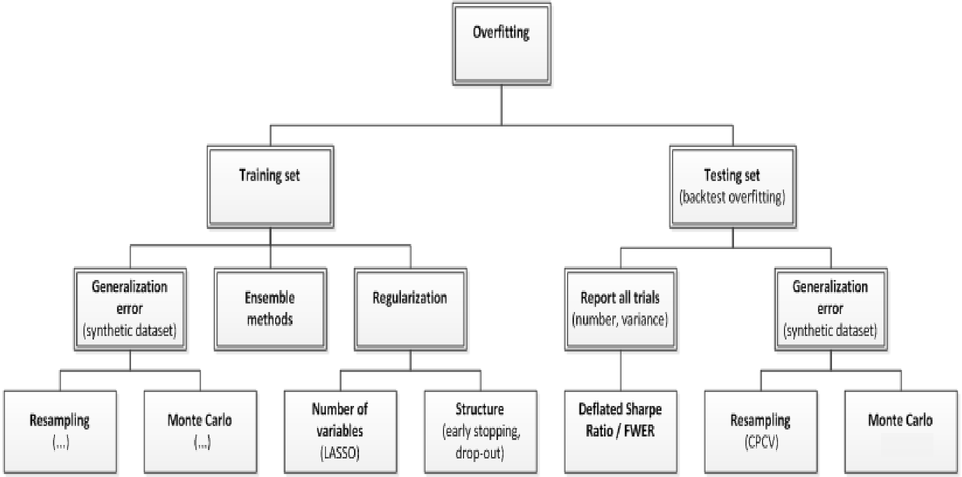
- An ML algorithm learns complex patterns in a high-dimensional space without being specifically directed
  - The ML algorithm may find a pattern that cannot be easily represented with a finite set of equations
  - Solutions often involve a large number of variables and the interactions between them
  - Unlike with other empirical tools, researchers do not impose a particular structure on the data
- ML algorithms rely on computationally-intensive methods, such as
  - estimation of the generalization error
  - ensembles, heuristics
  - experimental hypothesis testing, with minimal assumptions



Suppose that you have a 1000x1000 correlation matrix... A clustering algorithm finds that there are 3 blocks: Highly correlated, low correlated, uncorrelated.

ML Solutions to Overfitting

- There are several ML solutions for each type of overfitting
- Train set overfitting is addressed with
  - Ensembles methods
  - Regularization methods
  - Generalization errors (test set)
- Test set overfitting is addressed by
  - Reporting/controlling for all trials
  - Generalization errors (out-of-sample)
- All of these approaches require more computing power than what was available when classical methods were [developed]



A summarized description of various ML methods specifically designed to prevent both types of overfitting. There is no need to choose one method, and all of them can be applied simultaneously.

Train Set: Ensemble Methods

- Ensemble methods combine a set of low-correlated weak learners in order to create a learner that performs better than the individual ones
- The three main types of ensemble methods are
  - Bagging (Bootstrap aggregation)
  - Boosting
  - Stacking
- In addition, there are hybrid methods
  - E.g., random forests combine bagging with random subspaces (random sampling of features at each split, without replacement)

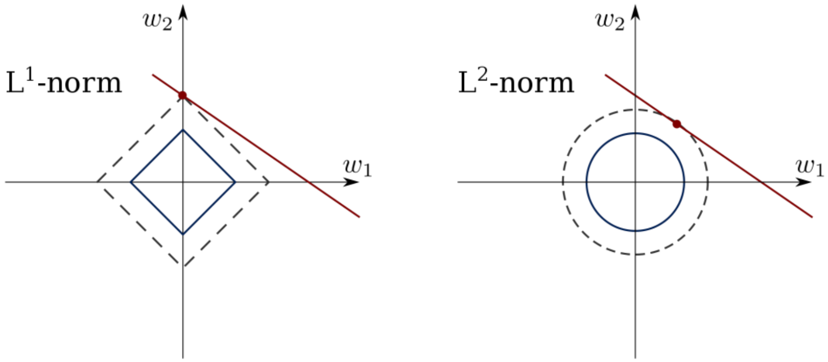
Method	Same algorithm	Parallel training	Aggregation	Primary error reduction
Bagging	Often	Yes	Deterministic	Variance
Boosting	Often	No	Deterministic	Bias
Stacking	Seldom	Yes	Meta-model (K-Fold cross-training)	Variance

Most ML algorithms can be used in ensembles. For instance, with proper parallelization, a SVC algorithm can be “bagged” to reduce train set overfitting, with minimal extra computing time.

If the weak classifiers have a minimum accuracy, bagging can also reduce bias.

Train Set: Regularization Methods

- Regularization prevents overfitting by introducing additional information to the model
- This additional information takes the form of a penalty for complexity
  - The optimization algorithm that fits the data only adds complexity if it is warranted by a certain amount of gain in explanatory power
- Three main types of regularization:
  - Tikhonov: ℓ^2 norm on the coefficients
  - LASSO: ℓ^1 norm on the coefficients
  - Elastic Net: It combines Tikhonov and LASSO

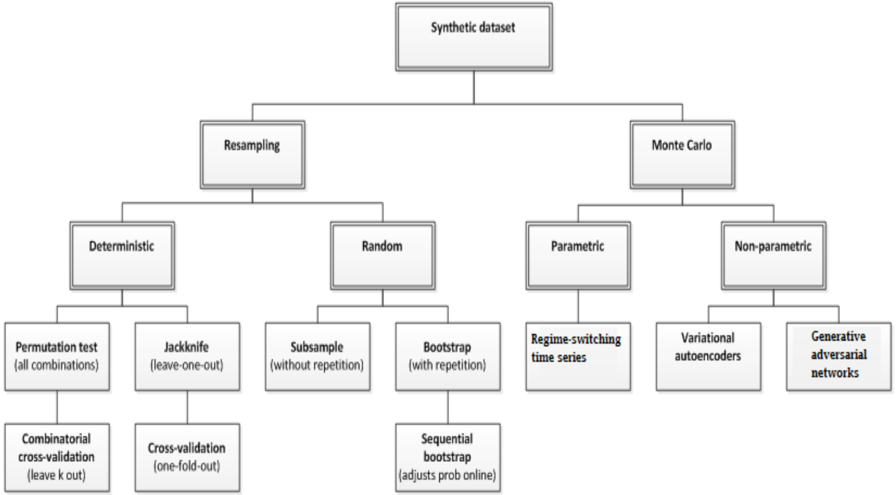


Source: Wikimedia Commons

The constraint region defined by a ℓ^1 norm is more likely to set some weights to exactly zero. In contrast, the constraint region defined by a ℓ^2 norm rarely sets any weight to zero. Elastic Nets overcome two limitations of LASSO: (a) They do not saturate when there are more variables than observations; and (b) they do not select one out of multiple multicollinear variables, discarding the rest.

Train Set: Generalization Error

- There are two main ways of estimating the generalization error on the test set: Resampling and Monte Carlo
- **Resampling** generates synthetic datasets by sampling from the observed dataset
  - Deterministic sampling (E.g., K-fold CV)
  - Random sampling (E.g., bootstrap)
- **Monte Carlo** generates synthetic datasets by running a Monte Carlo on a data-generating process
  - Parametric (E.g., Regime-switch Markov chain)
  - Non-parametric (E.g., GAN)

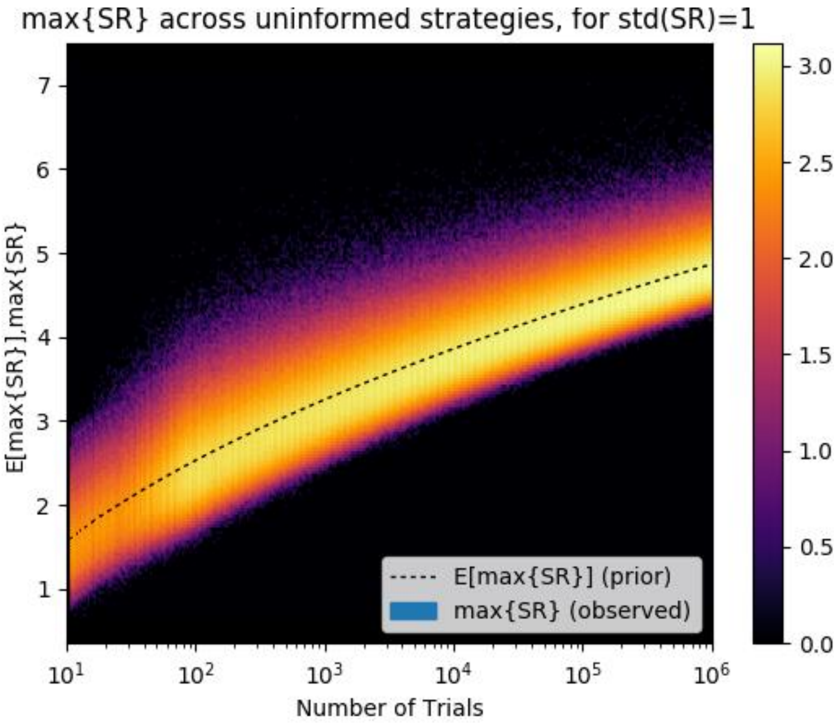


Synthetic dataset
Resampling
Monte Carlo
Deterministic
Random
Parametric
Non-parametric
Permutation test (all combinations)
Jackknife (leave-one-out)
Subsample (without repetition)
Bootstrap (with repetition)
Regime-switching time series
Variational autoencoders
Generative adversarial networks
Combinatorial cross-validation (leave k out)
Cross-validation (one-fold-out)
Sequential bootstrap (adjusts prob online)

A summary of ML methods that control for train set overfitting, by estimating the generalization error.

Test Set: Controlling for All Trials

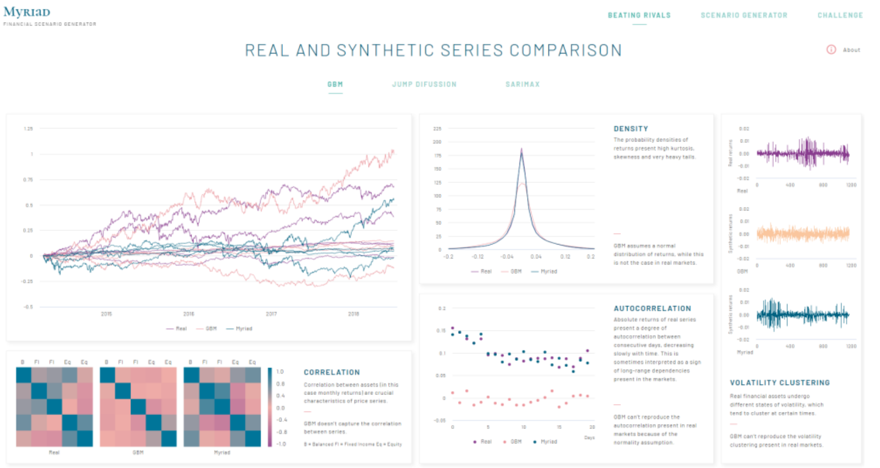
- SBuMT inflates a model's performance statistics
  - The model will perform out-of-sample worse than it did on in-sample data
- Two main approaches to control for performance inflation
  - **Parametric**: Derive the adjusted *p*-value
    - Family-wise error rate (FWER)
    - False discovery rate (FDR)
  - **Non-parametric**: Deflate the model's performance while controlling for the number and variance of the trials
    - E.g., deflated Sharpe ratio (DSR)



Non-parametric methods for SBuMT rely on fewer assumptions and tend to be more reliable.

Test Set: Generalization Error

- Once the researcher has chosen the final model, we can further estimate its generalization error on unseen data
- In order to do that, we can produce new synthetic datasets, using the same techniques described for train set generalization error
- For instance:
  - **Combinatorial cross-validation** can be used to
    - generate new test sets, different from those used by the researcher, and
    - bootstrap the entire distribution of the test set error (not only its mean), which is harder to overfit than its mean
  - **Monte Carlo methods** enable the production of arbitrarily-large new (unseen) datasets



REAL AND SYNTHETIC SERIES COMPARISON

DENSITY

The probability densities of returns present high kurtosis, skewness and very heavy tails.

GBM assumes a normal distribution of returns, while this is not the case in real markets.

CORRELATION

Correlation between assets (in this case monthly returns) are crucial characteristics of price series.

GBM doesn't capture the correlation between series.

B = Balanced FI = Fixed Income Eq = Equity

AUTOCORRELATION

Absolute returns of real series present a degree of autocorrelation between consecutive days, decreasing slowly with time. This is sometimes interpreted as a sign of long-range dependencies present in the markets.

GBM can't reproduce the autocorrelation present in real markets because of the normality assumption.

VOLATILITY CLUSTERING

Real financial assets undergo different states of volatility, which tend to cluster at certain times.

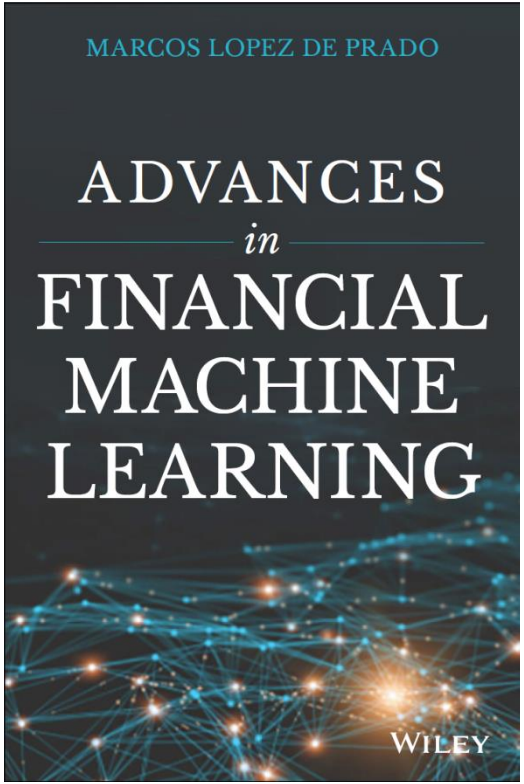
GBM can't reproduce the volatility clustering present in real markets.

Myriad is one example of a non-parametric Monte Carlo tool that generates synthetic datasets that match the statistical properties of the observed data.

Conclusions

- When used *incorrectly*, the risk of ML overfitting is extremely high
  - Given ML's power, that risk is higher than with classical statistical methods
- However, ML counts with sophisticated methods to prevent
  - Train set overfitting
  - Test set overfitting
- Thus, the popular belief that ML overfits is false
- A more accurate statement would be that
  - **in the wrong hands, ML overfits**
  - **in the right hands, ML is more robust to overfitting than classical methods**
- **When it comes to modelling unstructured data, ML is the only choice**
  - Classical statistics should be taught as a preparation for ML courses, with a focus on overfitting prevention

For Additional Details



The first wave of quantitative innovation in finance was led by Markowitz optimization. Machine Learning is the second wave and it will touch every aspect of finance. López de Prado's Advances in Financial Machine Learning is essential for readers who want to be ahead of the technology rather than being replaced by it.
— Prof. **Campbell Harvey**, Duke University. Former President of the American Finance Association.

Financial problems require very distinct machine learning solutions. Dr. López de Prado's book is the first one to characterize what makes standard machine learning tools fail when applied to the field of finance, and the first one to provide practical solutions to unique challenges faced by asset managers. Everyone who wants to understand the future of finance should read this book.
— Prof. **Frank Fabozzi**, EDHEC Business School. Editor of The Journal of Portfolio Management.

Disclaimer

- The views expressed in this document are the authors' and do not necessarily reflect those of the organizations he is affiliated with.
- No investment decision or particular course of action is recommended by this presentation.
-

www.QuantResearch.org