FE670 Algorithmic Trading Strategies

Lecture 2. Common Pitfalls in Financial Engineering

Sheung Yin Kevin Mo

Stevens Institute of Technology

September 6, 2018

Outline

- Science is able to provide the framework to perform analysis but is often ill-equipped to perform synthesis. Our ability to synthesize purposeful artifacts, crucial for practical application, is much more limited.
- Engineering is to use existing knowledge to sythesize an artifact such as an airplane or a portfolio.

- Financial engineering relies on finance theory, in particular on the ability to forecast financial quantities.
- Sometimes solution fails, and one possible cause is the failure to recognize that the solution to the problem calls for a true theoretical advance.

Engineering and Theoretical Science

- Modern science is based on the concept of laws of nature formulated in mathematical language and (for the most part) expressed through differential equations.
- A differential equation is an expression that links quantities and their rates of change. For example:

$$\frac{dC}{dt} = C_r \tag{1}$$

  Given initial or boundary conditions, a differential equation allows infering the behavior of a system in the future or in other regions of space.
- When probablistic laws are involved, differential equations, or their discrete counterpart, describe the evolution of probablity distributions. For example, the price of an asset can be expressed as:

$$dS_t = \mu S_t dt + \sigma S_t dW_t \tag{2}$$

Engineering and Theoretical Science

- Engineering is a process of synthesis in the sense that the objective of the engineering process is to construct purposeful artifacts, such as airplanes, trains or, in finance, portfolios, investment strategies or derivative products.
- We are given a problem formulated in terms of design specifications and we attempt to synthesize a design or an artifact that meets these specifications.
- The process of engineering is based on iterating cycles of synthesis and analysis - we start by sythesizing and overall design and then we analyze the design with analytical tools based on our scientific knowledge.
- Problem-solving abilities can be formalized and mimicked by computer programs - Herbert Simon, 1978 Nobel Prize in Economic Science.

Engineering and Theoretical Science

- Automatic problem-solving works well when solutions can be expressed as the maximization of some goal function; that is, when the problem can be cast in an optimization framework.
- Constructive methodoloties are available only when we arrive at the point where we can optimize, that is, codify our design in terms of variables and express the quality of our design in terms of a goal function defined on the design variables.
  - *Science is analytic:* We have the models to analyze a given system.
  - *Design is a constructive process:* We need to synthesze a design starting from general high-level specification.
  - *Constructive design is performed iteratively:* We make an approximate design and analyze it.
  - *Design automation:* The process of design can be automated only when we arrive at the stage of expressing the design quality in terms of a goal function.

Finanical Engineering can be defined loosely as an engineering process whose objective is to create financial products with specific characteristics.

- Designing a derivative-based product to meet specific client needs is engaged in financial engineering;
- Managing a portfolio with desired properties such as a given risk return profile is also engaged in financial engineering;
- Design a tradging strategy with specified risk return profile is also engaged in financial engineering.

- Indeed most of the financial engineering processes, including portfolio construction and derivative-based strategies, can be cast, at least theoretically, in an optimization framework.
- Optimization depends critically on the ability to make forecasts and to evaluate the risk of those forecasts. Much of its success can be atrributed to the following two reasons:
  - We have learned how to make forecast more effective;
  - We have the technology necessary to make the optimization process more robust to measurement errors and uncertainty in the inputs.

Learning, Theoretical, and Hybrid Approaches to Portfolio Management

There are three basic approaches to finanical modeling: **the learning approach, the theoretical approach, and the learning-theoretical approach**.

- **The learning approach** to financial modeling is in principle a consequence of the diffusion of low-cost high-performance computers. It is based on using a family of models that
  1. include an unlimited number of parameters and
  2. can approximiate sample data with high precision.

- Neural networks are a classical example. With an unrestricted number of layers and nodes, a neural network can approximate any function with arbitrary precision.
- However, practice has shown that if we represent sample data with very high precision, we typically obtain poor forecasting performance.

Overfitting: In general, the main features of the data can be described by a simple structural model plus unpredictable noise. As the noise is unpredictable, the goal of a model is to capture the structural components. A very precise model of a sample data (in-sample) will also try to match the unpredictable noise. This leads to poor (out-of-sample) forecasting abilities.

- To avoid over-fitting, the learning approach constrains the complexity of models. This is typically done by introducing what is called a penalty function.
- The central idea in learning theory is to add a penalty term to the objective function that grows with the number of parameters but gets smaller if the number of sample points increases.

- **The Theoretical Approach** to financial modeling is based on human creativity. In this approach, models are the result of new scientific insights that have been embodied in theories.
  - Laws such as the Maxwell equations of electromagnetism were discovered not through a process of learning but by a stroke of genius.
  - The Capital Asset Pricing Model (CAPM) is the most well-known example of a theoretical model in financial economics.
- **The Hybrid Approach** to financial modeling retains characteristics of both the theoretical and learning approaches. It uses a theoretical foundation to identify families of models but uses a learning approaches to choose the correct model within the family.
  - For example the ARCH/GARCH family of models is suggested by theoretical considerations while the right model is selected through a learning approach that identifies the model parameters.

Biases

- **Survivorship Bias** is exhibited by samples selected on the basis of criteria valid at the last point in the sample population.

- In the presence of survivorship biases in our data, return processes relative to firms that ceased to exist prior to that date are ignored.
- For example, while poorly performing mutual funds often close down (and therefore drop out of the sample), better performing mutual funds continue to exist (and therefore remain in the sample).
- In this situation, estimating past returns from the full sample would result in overestimation due to survivorship bias.

Biases

- **Selection Bias** is an error in choosing the individuals or groups to take part in a scientific study.

- Intrinsic in common indexes such the Russell 1000 universe (large-cap stocks). In order to understand the selection bias, we can apply a selection rule similar to that of the Russell 1000 to artificially generated random walks.
- Assume we have 10,0000 independent random walk price processes, each representing the price of a company's stock, over 1,000 periods using the recursive formula:

$$\begin{aligned} P_i(2) &= (1 + R_i(2)) \times P_i(1) = 1 + 0.007 \times \epsilon_i(2) \\ P_i(3) &= (1 + R_i(3)) \times P_i(2) = (1 + 0.007 \times \epsilon_i(3)) \times (1 + 0.007 \times \epsilon_i(2)) \\ &\dots \\ P_i(n) &= (1 + 0.007 \times \epsilon_i(n)) \times \dots \times (1 + 0.007 \times \epsilon_i(3)) \times (1 + 0.007 \times \epsilon_i(2)) \end{aligned} \tag{3}$$

Pitfalls in Choosing from Large Datasets

STATEMENT: Any statistical test, regardless of its complexity and power, will fail in a certain number of cases simply by chance.

- *For example, pairs trading is based on selecting pairs of stocks that stay close together. Suppose we know that the price paths of two stocks will stay close together. When they are at their maximum distance, we can go long in the stock with the highest value and short in the other stock. When their distance is reduced or changes sign a profit is realized.*
  - Given a large universe of stocks, a pairs trading strategy will look for cointegrate pairs. A typical approach will consist in running a cointegration test on each pair. Actually test can consist of multiple tests that each pair has to pass in order to be accepted as cointegrated.
  - However, a pair can appear cointegrated in a sample period purely by chance. Or a truly cointegrated pair may fail the test.

Pitfalls in Choosing from Large Datasets

To illustrate this phenomenon, let's consider a set of 1,000 artificial arithmetic random walk paths that are 1,000 steps long. Consider that is the sample set there are $(1,000 \times 1,000 - 1,000)/2 = 1,000 \times 999 \times 0.5 = 499,500$ different pairs of processes.

- The random walk is defined by the following recursive equation:

$$P_i(t) = P_i(t-1) + 0.007 \times \epsilon_i(t)$$

where the $\epsilon_i(t)$ are independent draws from a standard normal distribution $N(0,1)$.

  - Using the ADF test at 1% significance level, in run 1, 1.1% pass the cointegration test, in run 2, 0.8%.
  - Using the Johansen test at 99% significance level, in run 1, 2.7% pass the cointegration test, in run 2, 1.9%.
  - Using the Johansen maximum eigenvalue test, in run 1, 1.7% pass the cointegration test, in run 2, 1.1%.

Pitfalls in Selection of Data Frequency

- In financial theory, we have both discrete-time and continuous-time models. For example, the Black-Scholes option pricing equation, under certain assumptions, can be solved in a closed-form format. In other cases, we have to look for numeric solutions.
- Let's look at a discrete-time models. For example a vector autoregressive model of order 1:

$$X_t = AX_{t-1} + E_t$$

  Such a model is characterized by a time step. If the X are returns, the time steps could be days, weeks, or months.
- Given a process that we believe is described by a given model, can we select the time step arbitrarily? Or are different time steps characterized by different models?
- There is no general answer to these questions. Most models currently used are not invariant after time aggregation.

Model Risk and Its Mitigation

- We have to conclude that errors in choosing and estimating models cannot be avoided. This is because models are inevitably misspecified as they are only an approximation, more or less faithful, of the true data generating process (DGP).
- Model risk means that we cannot be certain that the model that we have selected to represent the data is correctly specified. If models are misspecified, forecasting errors might be significant.
- The notion of model risk entered science with the engineering of complex artifacts, the study of complex systems, and the widespread adoption of statistical learning methods.

Source of Model Risk

When modeling complex systems such as financial markets, we might encounter one of the following:

- The phenomena under study might be very complex and thus only a simplified description is possible; this leaves open the possibility that some critical aspect is overlooked.
- The phenomena under study can be very noisy; as a consequence, the scientific endeavor consists in extracting small amounts of information from highly noisy environments.
- Being not a law of nature but the behavior of an artifact, the object under study is subject to unpredictable changes.

Source of Model Risk

Existing techniques to reduce sources of error in model selection and estimation:

- Information theory, to assess the complexity and the limits of the predictability of time series.
- Bayesian modeling, which assumes that models are variations of some a priori model.
- Shrinkage, a form of averaging between different models.
- Random coefficient models, a technique that averages models estimated on clusters of data.

Information Theory Approach to Model Risk

The critical questions to be asked:

- Is it possible to estimate the maximum information extractable from a financial time series?
- Can we prescribe an information boundary such that sound robust models are not able to yield information beyond that boundary?
- Is it possible to assess the intrinsic complexity of empirical time series?

In a finite probability scheme, with N outcomes each with probability $p_i, i = 1, 2, ..., N$ information is defined as

$$I = \sum_{i=1}^{T} p_i log(p_i)$$

Information Theory Approach - Entropy

The quality I, which always negative, assumes a minimum

$$I = log\left(\frac{1}{N}\right)$$

if all outcomes have the same probability; it assumes a maximum $I = 0$; if one outcome has probability 1 and all other outcomes probability 0, that is, in the case of certainty of one outcome.

Entropy is a measure of disorder in physics. And we define:

$$I = -H$$

If we can associate the quantity of information to physical processes, we can establish laws that make sense empirically.

Information Theory Approach - Coarse Graining and Symbolic Dynamics

Coarse graning means dividing the possible outcome x_t of the series into discrete segments (or partitions) and associating a symbol to each segment.

For example the symbol a_i is associated to values x_t in the range $v_{i-1} < x_t < v_i$. In doing so, the original DGP of the time series entails a discrete stochastic dynamics of the corresponding sequence of symbols.

Given the probabilistic dynamics of the symbol sequence, we can associate a probability to any sequence of n symbols $p(i_1, ..., i_n)$. The entropy H can be defined as

$$H = -\sum_{i=1}^{T} p_i log(p_i)$$

Information Theory Approach - Kolmogorow-Sinai Entropy

We can therefore define the entropy per block of length n (or block entropy) as follows:

$$H_n = -I_n = -\sum p(i_1, ..., i_n) log(p_1, ..., p_n)$$

From the block entropy, we can now define the conditional entropy h_n as the difference of the entropies per blocks of length $n+1$ and n:

$$h_n = H_{n+1} - H_n = -\sum p(i_{n+1}|i_1, ..., i_n) log(p_{n+1}|p_1, ..., p_n)$$

Finally, we can define the Kolmogorow-Sinai entropy, or *entropy of the source*, as the limit for large n of the conditional entropy. The conditional entropy is the information on the following step conditional on the knowledge of the previous n steps. The quantity $r_n = 1 - h_n$ is called the predictability of the series.

Information Theory Approach - Entropy Applications

The concepts of conditional entropy and entropy of the source are fundamental to an understanding of the complexity of a series.

They supply a model-free methodology for estimating the basic predictability of a time series.

It establishes a reasonable boundary to the performance of models. Models that seem to exceed by a large measure the predictability level of entropy-based estimation are also likely to exhibit a high level of model risk.

In general, the conditional entropy and the entropy of source of coarse graining models give an assessment of the complexity of a series and its predictability.

Information Theory Approach - Other Methods

- *Transfer entropy* gauges the information flow from one series to another. It is defined as the information about future observation $I(t+1)$ gained from past observations of I and J minus the information about future observation $I(t+1)$ gained from past observations of I only:

$$T_{I\to J} = \sum p(i_1, ..., i_{m+1}, j_1, ..., j_l) log \left[\frac{p_(i_{m+1}|(i_1, ..., i_{m+1}, j_1, ..., j_l))}{p(i_{m+1}|i_1, ..., i_{m+1})} \right]$$

  This quantity evaluates the amount of information that flows from one series to another.

- Vapnik and Chervonenkis (VC) theory establishes limits to the ability of given models to learn in a sense made precise by concepts such as Vapnik entropy, empirical risk, structural risk, and the VC dimension.

Bayesian Approach to Model Risk

- *Bayesian Statistics:*
  - Statistical models are uncertain and subject to modification when new information is acquired.
  - There is distinction between prior probability (or prior distribution), which conveys the best estimate of probabilities given initial available information, and the posterior probability, which is the modification of the prior probability consequent to the acquisition of new information.
  - The mathematical link between prior and posterior probabilities is given by Bayes' Theorem.

Given two events A and B, we have Bayes' Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

We replace the event A with a statistical hypothesis H and event B with the data:

$$P(H|data) = \frac{P(data|H)P(H)}{P(data)}, P(H|data) \propto P(data|H)P(H)$$

Bayesian Approach to Model Risk

- The Bayesian handling of model risk is based on *Bayesian dynamic modeling*. It assumes that though there is uncertainty as regards the model, we have a good idea of a basic form of the model.

$$p(\theta|y) \propto L(y|\theta)\pi(\theta)$$

where, y represents the data, θ is the parameter set, $p(\theta|y)$ is the posterior distribution, $L(y|\theta)$ is the likelihood function, and $\pi(\theta)$ is the prior distribution.

- Uncertainty is expressed as a prior distribution of the model parameters where the means of the distribution determine the basic model.
- The estimation process does not determine the model from the data but uses the data to determine deviations of the actual model from a standard idealized model.

Bayesian Approach to Model Risk

- The key issue in Bayesian statistics is how to determine the prior. Though considered subjective, the prior is not arbitrary. The prior represents the basic knowledge before specific measurements are taken into account. Two types of priors are often used: diffuse priors and conjugate priors.

1. The *diffuse prior* assumes that a uniform distribution over an unspecified range.
2. The *conjugate prior* is a prior such that, for a given likelihood, the prior and the posterior distribution coincide.

Bayesian Analysis of an Univariate AR(1) Model

- Consider the following simple autoregressive model:

$$y_t = \rho y_{t-1} + \epsilon_t$$

  Assume that the preceding model is Gaussian so that the likelihood is also Gaussian. The model being linear, Gaussian innovations entail Gaussian variables. The likelihood is a given, not a prior.
- We can write the likelihood, which is a function of data parameterized by the initial conditions y_0, the autoregressive parameters ρ, and the variance σ of the innovation process as follows:

$$L(y|\rho, \sigma, y_0) = \frac{1}{\sqrt{(2\pi)^T}} \sigma^{-T} \exp - \frac{\sum_{t=1}^{T} \epsilon_t^2}{2\sigma^2}, -1 < \rho < 1, \sigma > 0$$

  Assume a flat prior for (ρ, σ), that is $\pi(\rho, \sigma) \propto \frac{1}{\sigma}$.

Bayesian Analysis of an Univariate AR(1) Model

- The the joint posterior distribution is the following:

$$p(\rho, \sigma | y, y_0) = \sigma^{-T-1} \exp - \frac{\sum_{t=1}^{T} \epsilon_t^2}{2\sigma^2}$$

$$= \frac{1}{\sqrt{(2\pi)^T}} \sigma^{-T} \exp - \frac{\sum_{t=1}^{T} (y_t - \rho y_{t-1})^2}{2\sigma^2}$$

Let

$$\hat{\rho} = \frac{\sum_{t=1}^{T} y_t y_{t-1}}{\sum_{t=1}^{T} y_{t-1}^2}$$

be the OLS estimator of the regressive parameter and call $Q = \sum y_{t-1}^2$ and $R = \sum (y_t - \hat{\rho} y_{t-1})^2$
Then the marginal distribution of (ρ, σ) are

$$p(\rho | y, y_0) \propto \quad (R + (\rho - \hat{\rho})^2 Q)^{-0.5T}$$
$$p(\sigma | y, y_0) \propto \quad \sigma^{-T} \exp\left(-\frac{R}{2\sigma^2}\right)$$

Model Averaging and the Shrinkage Approach to Model Risk

Model Averaging: Reliable estimations and forecasts from different models should be highly correlated. When they are not, this means that the estimation and forecasting processes have become dubious and averaging can substantially reduce the forecasting error. *If model averaging has a strong impact on forecasting performance, it is a sign that forecasts are uncorrelated and thus unreliable.*

Shrinkage: The method of shrinkage can be generalized to averaging between any number of models. The weighting factors can be determined by Bayesian principles if one has an idea of the relative strength of the models. Shrinkage is averaging between possibly different models. In Bayesian terms this would call for multiple priors.

Random Coefficient Models

- *Random coefficient models* are based on the idea of segmenting data in a number of clusters and estimating models on multiple clusters.

- Consider an ordinary linear regression, the regression parameters can be estimated with OLS methods using *fully pooled data*. This means that all the available data are pooled together and fed to the OLS estimator.
- However, this strategy might not be optimal if the regression data come from entities that have slightly different characteristics.
- To reduce model risk, we might decide to segment data into clusters that reflect different types of firms, and estimate regression for each cluster, and combine the estimates.

Random Coefficient Models - Example

- Random coefficient modeling techniques perform estimates assuming that clusters are randomly selected from a population of clusters with normal distributions.

- We write a regression equation for the j-th cluster:

$$\mathbf{y}_i = \mathbf{X}_j \beta_j + \epsilon_j$$

where n_j is the number of elements in the j-th cluster and ϵ_j are mutually independent, normally distributed vectors. We can rewrite the regression as follows:

$$\mathbf{y}_i = \mathbf{X}_j \beta_j + \mathbf{X}_j \gamma_j + \epsilon_j$$

where γ_j are the deviations of the regression coefficients from their expectations: $\gamma_j = \beta_j - \beta \sim N(0, \Sigma)$.