INFO 7374 **Yizhen Zhao**

Lecture 2 Machine Learning in Market Making

Algorithm Trading Basics

Market Making

1. **Market makers** are agents who stand ready to buy and sell securities in the financial markets. The rest of the market participants are therefore always guaranteed counterparty for their transactions.
2. Traditional market makers are usually under contractual arrangements with the stock exchange and are incentivized to achieve benchmark quoting requirements. Nowadays, **High Frequency Trading (HFT)** firms play the role of market makers by creating bid-ask spreads, churning mostly low priced, high volume stocks (typical favorites for HFT) many times in a single day.
3. Prior to the Volcker Rule (July 21, 2015), many investment banks had segments dedicated to HFT. Post-Volcker, no commercial banks can have proprietary trading desks or any such hedge fund investments. All major banks have shut down their HFT shops.
4. Nowadays, the HFT world still has players ranging from small firms to medium sized companies and big players. A few names from the industry (in no particular order) are Chopper Trading, Virtu Financial, Jump Trading, Jane Street, etc.
5. HFT firms make money from **two sources**: (1) proprietary trading and (2) getting paid for providing liquidity by Electronic Communications Networks (ECNs) and some exchanges.

Algorithm Trading

1. **Algorithmic trading** is a method of executing a large order (too large to fill all at once) using automated pre-programmed trading instructions accounting for variables such as time, price, and volume to send small slices of the order (child orders) out to the market over time.
2. Any strategy for algorithmic trading requires an identified opportunity that is profitable in terms of improved earnings or cost reduction. The following are **common trading strategies** used in algorithm-trading:

- **Trend-following Strategies:** The most common algorithmic trading strategies follow trends in moving averages, oscillators, price momentum and related technical indicators. These strategies do not involve making any predictions or price forecasts. Trades are initiated based on the occurrence of desirable trends.
- **Relative Value Arbitrage:** Buying a dual-listed stock or bond at a lower price in one market and simultaneously selling it at a higher price in another market offers the price differential

as risk-free profit or arbitrage. The same operation can also be done for stocks vs. futures instruments, as price differentials do exist from time to time.

- **Mean Reversion Strategy:** Mean reversion strategy is based on the idea that the high and low prices of an asset are a temporary phenomenon that revert to their mean value (average value) periodically. Identifying and defining a price range and implementing an algorithm based on that allows trades to be placed automatically when the price of asset breaks in and out of its defined range.

$$Range = Midprice \pm Spread$$

Here, $midprice$ doesn't have to be the statistical mean, but an approximation of the price trend. It can be a simple moving average (SMA), an exponential moving average (EMA) or an OLS estimate. It can also be a latent time-series estimated using $Filtering$ method. $Spread$ is an estimation about the range in which the price will move in the next time period.It is usually a function of market volatility and can incorporate any information that may impact the volatility, such as volume and news.

Note: Leverage Effect refers to that market volatility increases in response to bad news, but falls in response to good news.

- **Volume Weighted Average Price (VWAP):** Volume weighted average price strategy breaks up a large order and releases dynamically determined smaller chunks of the order to the market using stock-specific historical volume profiles. The aim is to execute the order close to the Volume Weighted Average Price (VWAP):

$$VWAP = \frac{\sum Number\ of\ Shares\ Bought\ \times Share\ Price}{Total\ Shares\ Bought}$$

If the price of a buy order is lower than the VWAP, or if the price of a sell order is higher than VWAP, it is a good trade.

- **Time Weighted Average Price (TWAP):** Time weighted average price strategy breaks up a large order and releases dynamically determined smaller chunks of the order to the market using evenly divided time slots between a start and end time. The aim is to execute the order close to the average price between the start and end times, thereby minimizing market impact.

- **Percentage of Volume (POV):** Until the trade order is fully filled, this algorithm continues sending partial orders, according to the defined participation ratio and according to the volume traded in the markets. The related “steps strategy” sends orders at a user-defined percentage of market volumes and increases or decreases this participation rate when the stock price reaches user-defined levels.

- **Statistical Arbitrage:** is a computationally intensive approach to algorithmically trading financial market assets such as equities and commodities. It involves the simultaneous

buying and selling of security portfolios according to predefined or adaptive statistical models. Statistical arbitrage techniques are modern variations of the classic Cointegration-based ***Pairs Trading Strategy:***

Pairs Trading is a market-neutral trading strategy that matches a long position with a short position in a pair of highly correlated instruments such as two stocks, exchange-traded funds (ETFs), currencies, commodities or options.

Parameter Estimation Methods

- Method of Moments (GMM)
- Maximum Likelihood Estimation (MLE)
- Bayesian Estimation

Choice of Quantitative Models

- Reduced Form Models
- Structure Models

Construction of Test Statistics

The Holy Trinity of Tests: Wald, LR and LM.

Suppose that θ is a parameter, $l(\theta)$ is the log-likelihood function and

$$\tilde{\theta} = argmax_{\theta} l(\theta)$$

is the MLE. We want to test $\theta = \theta_0$ against $\theta \neq \theta_0$.

- Wald: compare $\tilde{\theta}$ and θ_0.
- LR: compare $l(\tilde{\theta})$ and $l(\theta_0)$.
- LM: See how close $l'(\theta_0)$ is to zero.

Predictive Accuracy

Diebold Mariano Test

Out-of-sample root mean square prediction error (RMSPE) is a natural metric for the quality of point forecasts. Given two competing forecasts, we can work out their out-of-sample RMSPEs in recursive or rolling-window schemes. Let $\hat{u}_{1t}$ and $\hat{u}_{2t}$ denote the two prediction errors. The idea of the Diebold-Mariano test is to apply a $t-test$ to the series $z_t = u_{1t}^2 - u_{2t}^2$ and see if the mean is zero or not. Concretely, take the test statistic

$$u_t = \widehat{Y}_t - Y_t = f(X_t, \beta) - Y_t$$

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T} \hat{u}_t^2}$$

$$DM\ test\ stat = \frac{\sum_{t=1}^{T}(\hat{u}_{1t}^2 - \hat{u}_{2t}^2)}{\hat{\sigma}/\sqrt{T}}$$

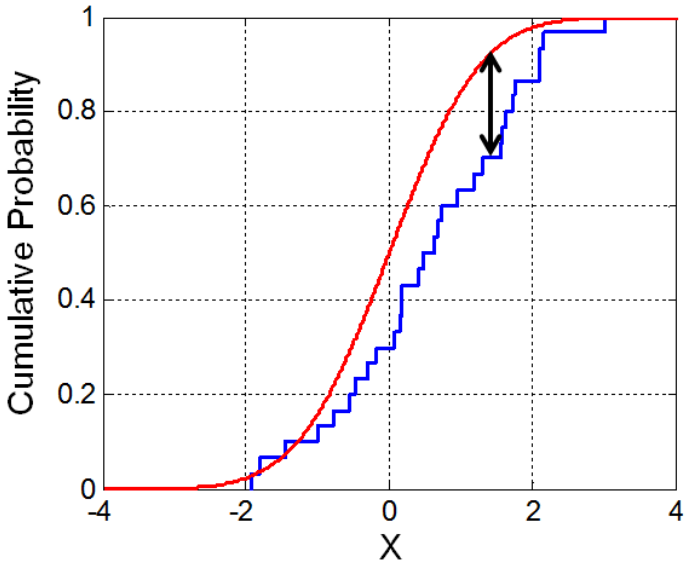
where T is the number of time periods for the out-of-sample forecast comparison and $\hat{\sigma}^2$ is the sample variance of $\hat{u}_{1t}^2 - \hat{u}_{2t}^2$. This is simply a $t-statistic$ testing the hypothesis that

$$E(u_{1t}^2) = E(u_{2t}^2)$$

and it has a standard normal null limiting distribution. This all works well for "non-nested" forecast comparisons, that is where the neither model is nested in the other.

Kolmogorov–Smirnov Test

The Kolmogorov–Smirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution. The null distribution of this statistic is calculated under the null hypothesis that the sample is drawn from the reference distribution (in the one-sample case). The distributions considered under the null hypothesis are continuous distributions but are otherwise unrestricted.



Note: Illustration of the Kolmogorov–Smirnov statistic. Red line is CDF, blue line is an ECDF (empirical CDF), and the black arrow is the K–S statistic.

Amisano Giacomini Test

The measure of accuracy for a particular density forecast uses ***Predictive Likelihood***, which is the sum of logarithmic predictive densities over the out-of-sample forecasting periods:

$$g(Y_t^*|x,\theta,T) = \sum_{t=1}^{T} \log f(Y_t^*|x,\theta,T)$$

$f(Y_t^*|x,\theta,T)$ is the predictive density of Y_t^* implied by model at date t.

To decide whether one density forecast outperforms the other, one can use the difference-in-predictive-likelihood test proposed by Amisano and Giacomini(2007). The test pair wisely compares the predictive accuracy of one model with the other. The advantage ofone density forecast over another is measured by the difference of the logarithmic predictive densities in every period:

$$\Delta L(Y_t^*) = \log[g(Y_t^*|x,\theta,T)] - \log[p(Y_t^*|x,\theta,T)]$$

The test statistic takes the form of a t-statistic:

$$AG - test\ stat = \frac{\Delta\bar{L}(Y_t^*)}{\hat{\sigma}/\sqrt{T-k}}$$

where $\Delta\bar{L}(Y_t^*)$ is the sample average of $\Delta L(Y_t^*)$, and $\hat{\sigma}^2$ is the sample variance of $\Delta L(Y_t^*)$.

Review on Time-Series Econometrics

The simplest time series model is an AR(1) - the *Ornstein–Uhlenbeck process*

$$y_t = \alpha y_{t-1} + u_t$$

In the case $|\alpha| < 1$, we have

$$T^{\frac{1}{2}}(\hat{\alpha} - \alpha) \underset{d}{\rightarrow} N(0, 1 - \alpha^2)$$

But this breaks down in the case $\alpha = 1$, which is a random walk. The knife-edge case where α is exactly equal to one arguably isn't that interesting per se. More importantly, this result doesn't work well unless the sample size is enormous if α is close to, but less than 1.

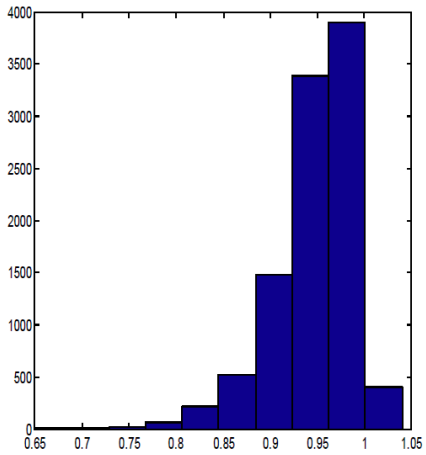
Stationary /Non-stationary Time Series

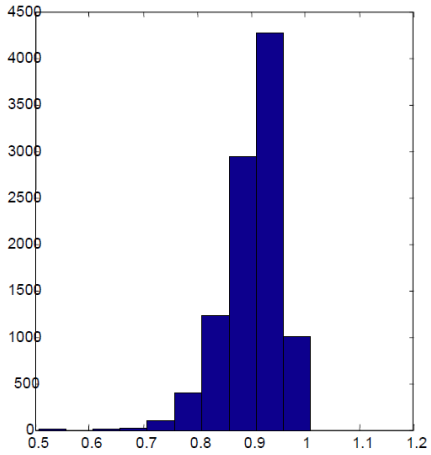
- A time series is a **random walk** if $y_t = y_{t-1} + u_t$ where u_t is iid.
- A time series is a **martingale** if $E(y_t) = y_{t-1}$.
- A time series is (weakly) **stationary** if its first two moments exist and do not change over time.
- A time series is **invertible** if it can be written as an autoregression.
- A time series is I(0) if it is both stationary and invertible. A time series is **I(d)** if its' d-th difference are I(0).
- If a time series is I(1), it is said to have a **unit root**.
- An **ARIMA(p,d,q)** model is a time series the d-th differences of which form a stationary and invertible ARMA(p,q) model.
- Suppose that x_t and y_t are two unrelated random walks. In a regression of one on the other, the coefficient is likely to be significant and the R-squared is likely to be high. But there is in fact no relation between the series. It is called a **spurious regression**.
- The fact that two time series have unit roots, does not mean that a relationship between them is a spurious regression. It is also possible that they are **cointegrated**.

Unit Root

When $\alpha \rightarrow 1$, the simulated distribution of OLS estimator of α when $T = 100$.

$\alpha = 1$ $\alpha = 0.95$





Both are skewed to the left.
Here, normal distribution doesn't work and for this and many non-standard problems in econometrics, we need to introduce new tools --- Brownian motion.

Brownian Motion

The stochastic process $B(t)$ is a **Brownian motion** if

1. $B(0) = 0$
2. $B(t) - B(s) \sim N(0, \sigma^2(t-s))$ for any $t > s$
3. If $t_1 < t_2 < t_3 < t_4$ then
   $B(t_2) - B(t_1)$ is independent of $B(t_4) - B(t_3)$

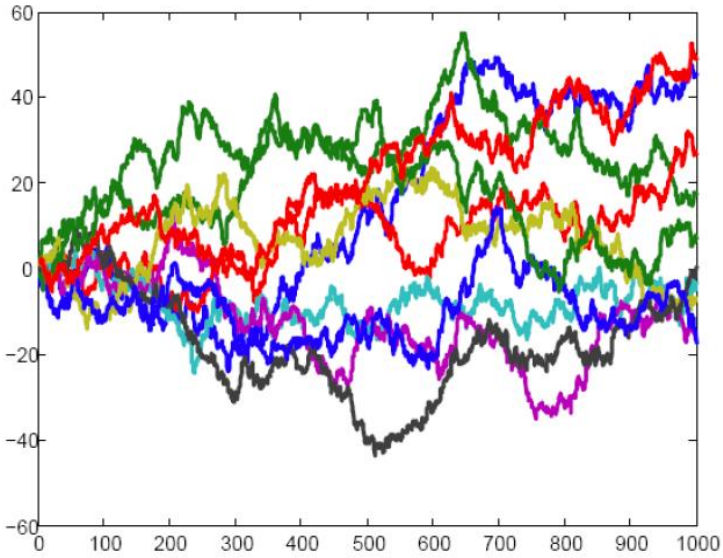
Functional Central Limit Theorem with Non-i.i.d. Errors

Suppose that $\varepsilon_1, \varepsilon_2, \cdots \varepsilon_T$ are stationary with mean 0 and (average) zero-frequency spectral density ω^2 satisfying suitable conditions. Let $S_t = \sum_{s=1}^{t} \varepsilon_s$. Define the function

$$S_T(r) = \frac{1}{T^{\frac{1}{2}}\omega} S_{Tr}$$

Then $S_T(r) => B(r)$.

Example: Ten Brownian Motions



Each time-series you see is actually one of the many possible realizations.

Continuous Mapping Theorem

Suppose that $X_T \underset{d}{\to} X$ (uniformly in r, if applicable). Then $f(X_T) \underset{d}{\to} f(X)$ where $f(\cdot)$ is any continuous function.

Combining these pieces,

$$T(\hat{\alpha} - 1) \underset{d}{\to} \frac{\frac{1}{2}\{B(1)^2 - 1\}}{\int_0^1 B(r)^2 dr}$$

Unit Root Tests

1. Augmented Dicky-Fuller Test (1979)

Null: Unit Root, $\sqrt{\mathrm{T}}(\hat{\rho}_{\mathrm{T}} - 1) \xrightarrow{\mathrm{p}} 0$

$$y_t = y_{t-1} + \varepsilon_t$$

Alternatives:

(i) No Drift

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

(ii) With Drift

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

(iii) With Drift and Time Trend

$$y_t = c + \delta t + \rho y_{t-1} + \varepsilon_t$$

Notes: Failure to reject the null (insignificant test stat or to say, large p -value) indicates that the time series y_t is a unit root process.

2. Philips-Perron Test (1988)

Null: Unit Root, $\sqrt{T}(\hat{\rho}_T - 1) \xrightarrow{p} 0$

$$y_t = y_{t-1} + \varepsilon_t$$

Alternatives:

(i) No Drift

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

(ii) With Drift

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

(iii) With Drift and Time Trend

$$y_t = c + \delta t + \rho y_{t-1} + \varepsilon_t$$

Notes: Failure to reject the null (insignificant test stat or to say, large p -value) indicates that the time series y_t is a unit root process.

3. Kwiatkowski, Philips, Schmidt, and Shin (KPSS) Test (1992)

Null: Trend Stationary

$$y_t = c + \delta t + u_{1t}$$

Alternatives:

(i) no deterministic trend:$y_t = c_t + u_{1t}$ and $c_t = c_{t-1} + u_{2t}$

(ii) with trend: $y_t = c_t + \delta t + u_{1t}$ and$c_t = c_{t-1} + u_{2t}$

where $u_{1t}$is a stationary process.$u_{2t}$is an independent and identically distributed process with mean 0 and variance σ^2.

The null says $\sigma^2 = 0$, which implies the random walk term c_t is constant. The alternative says $\sigma^2 > 0$, which indicates that the above unit root is a random walk. Rejection of the null indicates y_t is a non-stationary process.

Cointegration

Definition: Two non-stationary time series x_t and y_t are said to be cointegrated, if they are both I(1) but if there exists some linear combination $u_t = y_t - \beta x_t$, for $0 < k < \infty$, that is I(0).

We can rewrite the definition of cointegration as $y_t = \beta x_t + u_t$ where the regressor is I(1) and the error term is I(0). This model has intriguing statistical properties:

- OLS is super-consistent (meaning $T(\hat{\beta} - \beta)$ converges to a distribution, that is a function of Brownian motions).
- If x_t is strictly exogeneous (independent of the error at all leads and lags), then t- and F-statistics associated with OLS have their usual normal and χ^2 limiting distributions.
- If x_t is not strictly exogeneous, there are estimators other than OLS such that t- and F-statistics have normal and χ^2 limiting distributions. A popular choice is dynamic OLS which estimates the relationship

$$y_t = \beta x_t + d(L)\Delta x_t + u_t$$

where $d(L)$ is a two-sided polynimial. Another choice is the maximum likelihood estimator proposed by Soren Johansen.

1. Engle-Granger Two-Step Method

The Engle-Granger Two-Step method starts by creating residuals based on the static regression and then testing the residuals for the presence of unit-roots. It uses the Augmented Dickey-Fuller Test (ADF) or other tests to test for stationarity units in time series. If the time series is cointegrated, the Engle-Granger method will show the stationarity of the residuals.

The limitation of the Engle-Granger method is that if there are more than two variables, the method may show more than two cointegrating relationships. Another limitation is that it is a single equation model. However, some of the drawbacks have been addressed in recent cointegration tests like Johansen's and Phillips-Ouliaris tests.

2. Johansen Test

The Johansen test is used to test cointegrating relationships between several non-stationary time series data. Compared to the Engle-Granger test, the Johansen test allows for more than one cointegrating relationship. However, it is subject to asymptotic properties (large sample size) since a small sample size would produce unreliable results. Using the test to find cointegration of several time series avoids the issues created when errors are carried forward to the next step.

Johansen's test comes in two main forms, i.e., Trace tests and Maximum Eigenvalue test.

- **Trace tests**

Trace tests evaluate the number of linear combinations in a time series data, i.e., K to be equal to the value K_0, and the hypothesis for the value K to be greater than K_0. It is illustrated as follows:

$$H_0: K = K_0$$

$$H_1: K > K_0$$

When using the trace test to test for cointegration in a sample, we set K_0 to zero to test whether the null hypothesis will be rejected. If it is rejected, we can deduce that there exists a cointegration relationship in the sample. Therefore, the null hypothesis should be rejected to confirm the existence of a cointegration relationship in the sample.

- **Maximum Eigenvalue test**

An Eigenvalue is defined as a non-zero vector which, when a linear transformation is applied to it, changes by a scalar factor. The Maximum Eigenvalue test is similar to the Johansen's trace test. The key difference between the two is the null hypothesis.

$$H_0: K = K_0$$

$$H_1: K = K_0 + 1$$

In a scenario where $K=K_0$ and the null hypothesis is rejected, it means that there is only one possible outcome of the variable to produce a stationary process. However, in a scenario where K_0 = m-1 and the null hypothesis is rejected, it means that there are M possible linear combinations. Such a scenario is impossible unless the variables in the time series are stationary.

Paris Trading

- Pair Selection：Use β from CAPM
- Introduce CAPM

$$R_{i,t}^{e} = \beta \cdot R_{M,t}^{e} + u_{i,t}$$

- Benchmark Pair：Exxon Mobil Corporation (XOM) and Chevron Corporation (CVX)
- Test two things：(1) the difference between two β from CAPMis at most 0.15.
  (2) whether the two stocks are cointegrated.
- Reason to choose β：If the ratio β_i/β_j for two separate securities, i and j is close to one, then we expect them to be affected by market movements in the same fashion, a condition that favors pairs-trading compatibility.
- Test Cointegration:
  1. Engle-Granger two-step cointegration test.
  2. Cointegration vector:
     (a) Trace Test (see Johansen (1988))
     (b) Maximum Eigen Value Test (see Johansen and Juselius' (1990)).

Procedure

- Regression: $\log(P_1) = \beta_{coint} \log(P_2)$, β_{coint} is the cointegration ratio. We constrain the intercept to 0 since if pair is cointegrated, then we expect 0 returns on one asset to predict 0 returns on the other.
- Spread $S_t = \log(P_1) - \beta_{coint} \log(P_2)$.
- Test spread of pair for stationarity using an Augmented-Dickey Fuller (ADF) Test, which tests the null hypothesis that a process has a unit root (is not stationary). If the pair is cointegrated, then the spread should be stationary.

  Spread = P1- P2
  Mean(Spread)~0
  Vol(Spread)~ constant

Strategy Design

Allocation Ratio	Calculation of Spread	Strategy Design
1:1 Dollar	$\log(P_1) - \log(P_2)$.	long one share of P_2, short one share of P_1
CAPM β	$\log(P_1) - \beta_1/\beta_2 \log(P_2)$.	long one share of P_2, short β_1/β_2 share of P_1
Cointegration β	$\log(P_1) - \beta_{coint} \log(P_2)$	long one share of P_2, short β_{coint} shares of P_1

1. For each time point in the time series, calculate the risk-adjusted spread between the two assets of the pair.
2. If the spread is above its historical mean, then we expect that stock 1 is overpriced and stock 2 is underpriced. Thus, we short-sell stock 1 and buy stock 2. On the other hand, if the spread is under its historical mean, we short-sell stock 2 and buy stock 1.
3. If the signal is less than the closing threshold, close any existing position in the pair.
4. If the signal is greater than the stop-loss threshold, we close the position.

Note: the open threshold is set to 1σ, the close threshold is set to 0.5σ, and the stop-loss threshold to 4σ.

Pairs Examples

Below are some examples of asset pairs based on market observations and intuitions. The descriptions of the ticker symbols and associated exchange trading hours are shown in Table 1.

- USDJPY / NK : The choice is motivated by the positive correlation between these two assets. Given that Japan is export driven, when JPY strengths against USD (i.e. USDJPY weakens), Nikkei is expected to decline because of lower consumption. Conversely,

when Nikkei, which is considered a risk sentiment indicator, weakens, JPY strengthens and so USDJPY weakens.

- USDCAD / CL : Canada is the 7th largest oil producer in the world.2 This suggests a positive correlation between the price of oil and CAD. When crude futures price increases, Canadian dollar appreciates and so USDCAD decreases.
- CL / USO : USO is the largest crude oil ETF that consists of a mixture of West Texas Intermediate (WTI) futures of different maturities. As such, USO and crude oil futures prices are expected to have consistent positive correlation.
- GC / SI : Pairs trading between gold spot and silver spot, while staying neutral to metals exposure.
- AUDJPY / SPX : This is pairs trading between risk-on currency AUDJPY and equity S&P500 futures. The latter is commonly viewed as a barometer of general risk-on appetite of investors.
- USDCHF / GC : Swiss National Bank backs up a portion of their Swiss franc holdings with gold thus suggesting a correlation between the two assets. When gold increases in its value, CHF is expected to strengthen (i.e. USDCHF weakens), and vice versa.
- C / GS : Citigroup and Goldman are two large-cap stocks are in the banking industry.

- AAPL / FB : Apple and Facebook are two large-cap stocks in the technology sector.

List of assets with their ticker symbols, descriptions, and trading hours (in Eastern Standard Times).

Symbol	Description	Trading Hours
USDJPY	US Dollar valued against Japanese Yen	Sun 17:00 to Fri 17:00
AUDJPY	AUD Dollar valued against Japanese Yen	Sun 17:00 to Fri 17:00
USDCAD	US Dollar valued against Canadian Dollar	Sun 17:00 to Fri 17:00
USDCHF	US Dollar valued against Swiss Franc	Sun 17:00 to Fri 17:00
NK	Nikkei 225 on TSE	Sun to Fri, 20:00 to 16:30
CL	Crude futures on CME	Sun to Fri, 18:00 to 17:00
USO	United States Oil Fund ETF	Mon to Fri, 9:30 to 16:00
GC	Gold spot valued against US Dollar	Sun to Fri, 18:00 to 17:00
SI	Silver spot valued against US Dollar	Sun to Fri, 18:00 to 17:00
SPX	E-mini S&P 500 Futures traded on CME	Sun to Fri, 17:00 to 16:00
C	Citigroup Inc	Mon to Fri, 9:30 to 16:00
GS	Goldman Sachs Group Inc	Mon to Fri, 9:30 to 16:00
AAPL	Apple Inc	Mon to Fri, 9:30 to 16:00
FB	Facebook Inc	Mon to Fri, 9:30 to 16:00

List of assets with their ticker symbols, descriptions, and trading hours (in Eastern Standard Times).

While pairs trading is an intuitive strategy, any serious pairs trading system must include a procedure for optimizing the positions along with timing for entry and exit.

While observing the prevailing market prices, a trader can choose to establish a pairs trading position immediately or wait. After starting a pairs trade, the trader will need to decide when to close the positions.

How does using an optimal exit rule improve the profitability of pairs trading? We consider the analytically optimal strategy, with formulae derived by Leung & Li (2015), where the position is liquidated at an optimal exit price. We then compare its performance against the baseline strategy with entry and exit at a standard deviation of the spread. No attempts have been made to optimize for performance other than the pair ratio and parameter selection as part of the OU model fitting.

Conditional Heteroskedasticity

It is very common for financial time series to exhibit bursts of volatility. Modeling this is important for forecasting and many other purposes. The original model was autoregressive conditional heteroskedasticity (ARCH) which specifies that

$$r_t = (\mu + \sigma_t \varepsilon_t)$$
$$\sigma_t^2 = \omega + \alpha(r_{t-1} - \mu)^2 + \beta\sigma_{t-1}^2$$
$$\sigma_1^2 = \omega/(1-\alpha)$$

where ε_t is iid standard normal.

- The kurtosis of above process is

$$\frac{3(1-\alpha^2)}{(1-3\alpha^2)} > 3$$

The above model not only allows for burst of volatility, but it also accounts for fat tails.

- The estimation is fairly easy by maximum likelihood as the log-likelihood function is

$$-\frac{1}{2}\sum_{t=1}^{T}\log(\sigma_t^2) - \frac{1}{2}\sum_{t=1}^{T}\frac{r_t - \mu}{\sigma_t^2}$$

and can be numerically maximized with respect to the parameters α, μ and ω.

- The model has been extended in a great many ways. Three in particular are:
  1. GARCH: Generalized ARCH. A GARCH(p, q) model is

$$r_t = \mu + \sigma_t \varepsilon_t$$
$$\sigma_t^2 = \omega + \sum_{i=1}^{q}\alpha_i(r_{t-i} - \mu)^2 + \sum_{i=1}^{p}\beta_i\sigma_{t-i}^2$$

  2. GARCH in mean

$$r_t = \mu + \lambda\sigma_t + \sigma_t \varepsilon_t$$
$$\sigma_t^2 = \omega + \sum_{i=1}^{q}\alpha_i(r_{t-i} - \mu)^2 + \sum_{i=1}^{p}\beta_i\sigma_{t-i}^2$$

  3. Exponential GARCH

$$r_t = \mu + \sigma_t \varepsilon_t$$

$$log\sigma^2 = \omega + \alpha \log \sigma_{t-1}^2 + \beta[\theta \varepsilon_t + (|\varepsilon_t| - E(|\varepsilon_t|))]$$

Since ε_t is standard normal, $E(|\varepsilon_t|) = \sqrt{2/\pi}$. This model is particularly useful for representing stock returns, because they not only show burst of volatility, but volatility tends to rise when returns are low. This model can capture a skewness effect of this sort.

- All above models can be estimated by maximum likelihood (MLE). Any number of extensions to this framework have been proposed. One can add in explanatory variables or have nonlinear or multivariate specifications. A general challenge is ensuring that the variances remain positive. Of course, any parameterization which gives negative variances will in turn have a likelihood of minus infinity.

Filtering Methods

The filtering problem is that there is unobserved variable (a "state" variable) that evolves by some law of motion and there are observed variables that are related to the unobserved state. It might sound an arcane problem, but as we have seen, it has an enormous number of important applications.

The basic filtering problem is a linear model in what is known as state space form. This is the model where we observe y_t while α_t is an unobserved state and

$$y_t = Z_t \alpha_t + \varepsilon_t$$

$$\alpha_t = T\alpha_{t-1} + \eta_t$$

where $\varepsilon_t \sim i.i.d.N(0, H)$ and $\eta_t \sim i.i.d.N(0, Q)$.

$$QQQ_t = SPY_t a_t + \varepsilon_t$$

The Kalman Filter

The Kalman Filter allows inference to be done in the basic state space model. In this model, $a_t|Y_s$ is normal; let $u_{t|s}$ and $P_{t|s}$ denote its mean and variance. We then have

Updating Equations

$$u_{t|t} = u_{t|t-1} + P_{t|t-1} Z_t^{'} F_t^{-1} (y_t - Z_t u_{t|t-1})$$

$$P_{t|t} = P_{t|t-1} - P_{t|t-1}Z_t^{'}F_t^{-1}Z_tP_{t|t-1}$$

where

$$F_t = Z_tP_{t|t-1}Z_t' + H$$

Prediction Equations

$$u_{t+1|t} = Tu_{t|t}$$

$$P_{t+1|t} = TP_{t|t}T' + Q$$

if α_t is stationary, can initialize from the unconditional mean and variance-covariance matrix of α_t. The mean is just zero. The variance is $P_{t|t}$ which is the solution to the equation

$$P_{0|0} = TP_{0|0}T' + Q$$

the solution to which is

$$vec\left(P_{0|0}\right) = (I - T \otimes T)^{-1}vec(Q)$$

So, in Python, $P_{0|0}$is simply

$$reshape(inv\left(eye(n^2) - kron(T,T)\right) * reshape(Q, n^2, 1), n, n)$$

where n is the number of elements in the state vector α_t.

Then we iterate through the updating the predictive equations to get $u_{t|t-1}$ and u_t. The Kalman filter has two potential purposes:
(i) estimation and
(ii) inference about the state vector.
For the first of these, we have log-likelihood:

$$l = \sum_{t=1}^{T} \log\left(f(y_t|Y_{t-1})\right)$$

$$y_t|Y_{t-1} \sim N\left(Z_tu_{t|t-1}, Z_tP_{t|t-1}Z_t' + H\right) = N(Z_tu_{t|t-1}, F_t)$$

$$l = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\log|F_t| - \frac{1}{2}\sum_{t=1}^{T}v_t'F_t^{-1}v_t$$

where $v_t = y_t - Z_tu_{t|t-1}$.
For inference about the sate vector, we already have $u_{t|t}$, the filtered estimates. But we might want $u_{t|T}$, the "smoothed" estimates. These are obtained with one more set of recursions known as the Kalman smoother:

$$u_{t|T} = u_{t|t} + P_{t|t}T'P_{t+1|t}^{-1}(u_{t+1|T} - Tu_{t|t})$$

$$P_{t|T} = P_{t|t} + P_{t|t}T^{'}P_{t+1|t}^{-1}\left(P_{t+1|T} - P_{t+1|t}\right)P_{t+1|t}^{-1}TP_{t|t}$$

The Kalman filter/smoother need values of the parameters. We can use numerical methods to find the parameter values that maximize the likelihood (MLE), and then plug these in to get filtered and smoothed estimates of the states, which are also of interest.

Example: Use Kalman Filter for Pairs Trading

The pairs-trading strategy could be applied to a two ETFs that both track the performance of varying duration US Treasury bonds. They are:

- **TLT** - iShares 20+ Year Treasury Bond ETF
- **IEI** - iShares 3-7 Year Treasury Bond ETF

The goal is to build a mean-reverting strategy from this pair of ETFs.

The synthetic "spread" between TLT and IEI is the time series that we are actually interested in longing or shorting. The Kalman Filter is used to dynamically track the hedging ratio between the two in order to keep the spread stationary (and hence mean reverting).

Trading Rule: we go "long the spread" if the forecast error drops below the negative one standard deviation of the spread. Respectively we can go "short the spread" if the forecast error exceeds one positive standard deviation of the spread. The exit rules are simply the opposite of the entry rules.

The dynamic hedge ratio is represented by one component of the hidden state vector at time t, which we will denote as θ_t. This is the "beta" slope value that is from linear regression:

$$R_{i,t}^e = \alpha + \beta \cdot R_{M,t}^e + u_{i,t}$$

Longing the spread here means purchasing (longing) N units of TLT and selling (shorting) $\theta_t^0 N$, which must take the "floor" value integer. The latter is necessary as we must transact a whole number of units of the ETFs. "Shorting the spread" is the opposite of this. N controls the overall size of the position.

e_t represents the *forecast error* or *residual error* of the prediction at time t, while Q_t represents the variance of this prediction at time t.

For completeness, the rules are specified here:

1. $e_t < -\sqrt{Q_t}$- Long the spread: Go long N shares of TLT and go short $\theta_t^0 N$ units of IEI
2. $e_t \geq -\sqrt{Q_t}$- Exit long: Close all long positions of TLT and IEI
3. $e_t > \sqrt{Q_t}$ - Short the spread: Go short N shares of TLT and go long $\theta_t^0 N$ units of IEI
4. $e_t \leq \sqrt{Q_t}$ - Exit short: Close all short positions of TLT and IEI

The role of the Kalman filter is to help us calculate θ_t, as well e_t and Q_t. θ_t represents the vector of the intercept and slope values in the linear regression between TLT and IEI at time t. It is estimated by the Kalman filter. The forecast error/residual $e_t = y_t - \hat{y}_t$ is the difference between the predicted value of TLT *today* and the Kalman filter's estimate of TLT *today*. Q_t is the variance of the predictions and hence $\sqrt{Q_t}$ is the standard deviation of the prediction.

Hamilton Switching Model

This is an important nonlinear filtering model

$$y_t = \alpha + \beta S_t + \varepsilon_t$$
$$e.g. y_t = \alpha I(Vol_t > 30MM) + \beta S_t + \varepsilon_t$$

where S_t is a Markov switching process.

$$P(S_t = 1|S_{t-1} = 1) = p$$
$$P(S_t = 0|S_{t-1} = 0) = q$$

$$f(y_t|Y_{t-1}) = N(\alpha, \sigma^2)P(S_t = 0|Y_{t-1}) + N(\alpha + \beta, \sigma^2)P(S_t = 1|Y_{t-1})$$
$$= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{(y_t - \alpha)^2}{2\sigma^2}\right) P(S_t = 0|Y_{t-1})$$
$$+ \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{(y_t - \alpha - \beta)^2}{2\sigma^2}\right) P(S_t = 1|Y_{t-1})$$

a mixture of normals.

Updating equations (from Bayes Theorem)

$$P(S_t = 0|Y_t) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{(y_t - \alpha)^2}{2\sigma^2}\right) P(S_t = 0|Y_{t-1})/f(y_t|Y_{t-1})$$

$$P(S_t = 1|Y_t) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{(y_t - \alpha - \beta)^2}{2\sigma^2}\right) P(S_t = 1|Y_{t-1})/f(y_t|Y_{t-1})$$

Prediction Equations

$$P(S_t = 1|Y_{t-1}) = pP(S_{t-1} = 1|Y_{t-1}) + (1 - q)P(S_t = 0|Y_{t-1})$$
$$P(S_t = 0|Y_{t-1}) = (1 - p)P(S_{t-1} = 1|Y_{t-1}) + qP(S_{t-1} = 0|Y_{t-1})$$

Starting iterations

Only need $P(S_1 = 1|Y_0) = P(S_1 = 1)$, the unconditional probability. We know that

$$P(S_1 = 1) = \frac{1 - p}{2 - p - q}$$

and this allows us to start the recursions.

The model is useful for fitting business cycles (Hamilton (1989)). A few last comments:
(i) In practice, we would replace ε_t by an AR process.
(ii) Also, it is possible to let p and q depend on variables at time $t - 1$.

(iii) There is also a smoother that can be run backwards to get state probabilities conditional on the whole sample. This uses the recursions

$$P(S_t = 1|Y_T) = \frac{P(S_{t+1} = 0|Y_T)P(S_t = 1|Y_t)P(S_{t+1} = 0|S_t = 1)}{P(S_{t+1} = 0|Y_t)} + \frac{P(S_{t+1} = 1|Y_T)P(S_t = 1|Y_t)P(S_{t+1} = 1|S_t = 1)}{P(S_{t+1} = 1|Y_t)}$$

for $t = T-1, T-2, \ldots$ starting from $P(S_T = 1|Y_T)$.

Particle Filtering

Particle Filtering is a numeric device that is useful for filtering in a very general context (nonlinear and non-Gaussian). But it is based on simulation. There are many versions of the particle filter. Here are the steps for a simple illustrative implementation.

1. Generate n draws from the steady state distribution of α_0.
2. For each of the draws in (1), use the transition equation to get draws from the distribution of α_1. Call these draws $\{\tilde{\alpha}_{1,i}\}^n{}_{i=1}$. This gives the density of α_1 conditional on Y_0.
3. For each of the draws in (2), compute $q_1^i = p(y_1|\tilde{\alpha}_{1,i})$. Normalize these probabilities so that $\sum_{i=1}^n q_1^i = 1$.
4. Resample with replacement from $\{\tilde{\alpha}_{1,i}\}^n{}_{i=1}$ with probability q_1^i of selecting $\tilde{\alpha}_{1,i}$. Call these new draws $\{\tilde{\alpha}_{1,i}\}^n{}_{i=1}$. This gives the density of α_1 conditional on Y_1.
5. Repeat step 2-4 cycling through the whole sample.
   The density of y_t conditional on Y_{t-1} can be approximated by
   $$p(y_t|Y_{t-1}) = \frac{1}{n}\sum_{i=1}^{n} p(y_t|\tilde{\alpha}_{1,i})$$
   and the likelihood is $L = \prod_{t=1}^T p(y_t|Y_{t-1})$.

Particle Filter could be estimated using simulated Maximum Likelihood method or Markov Chain Monte Carlo (MCMC).

Both Hamilton Switching Model and the MCMC for Particle Filtering can be estimated by MLE.

Maximum Likelihood Estimation

Say $X_1, X_2, \cdots, X_n$ is $i.i.d.$ from a density $f(x, \theta)$ where θ is a $k \times 1$ vector of parameters. The joint probability density of the data is $\prod_{i=1}^n f(X_i, \theta)$.

Idea of maximum likelihood estimation. View this as a function of θ called the likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i, \theta)$$

$$\log\left(L(\theta)\right) = \sum_{i=1}^{n} \log\left[f(X_i, \theta)\right]$$

The MLE is the value of θ that maximizes the likelihood function:

$$\hat{\theta}_{MLE} = argmax_{\theta} L(\theta)$$

Because it is easier to work with sums than products, we generally write the MLE as

$$\hat{\theta}_{MLE} = argmax_{\theta} L(\theta)$$

where $l(\theta) = \log L(\theta) = \prod_{i=1}^{n} \log f(X_i, \theta)$

Example 3: **Generate Trading Signals Using Price/Volume Momentum**

TA Indicators for Simple Price Momentum Signals

Moving Averages

Double Crossover

- A buy signal is produced when the shorter average crosses above the longer. +1
- A sell signal is produced when the shorter averages moves below the longer average. -1
- Pairs Choice: 5day - 20day, 10day - 50day

Triple Crossover

- 4-9-18 method is used mainly in futures trading. 5-10-20 day moving averages are widely used in commodity circles.
- A buying alert takes place in a downtrend when the 4 day crosses above both the 9 and the 18.
- A confirmed buying signal occurs when the 9 day then crosses above the 18.
- When the uptrend reverses to downside, the first thing that should take place is that the shortest (and the most sensitive) average – the 4 day - dips below the 9 day and the 18 day. This is only a selling alert. Some traders, however, might use that initial crossing as reason

enough to begin liquidating long positions. Then, if the next longer average - the 9 day – drops below the 18 day, a confirmed sell short signal is given.

Moving Average Envelope

- Percentage envelopes can be used to help determine when a market has gotten overextended in either direction. They tell us when prices have strayed too far from their moving averages line.
- Short term traders often use 3% envelopes around a simple 21 day moving average. When prices reach one of the envelopes (3% from the average), the short-term trend is considered to be overextended. For long range analysis, some possible combinations include 5% envelopes around a 10-week average or a 10% envelope around a 40-week average.

Bollinger Bands

- Two trading bands are placed around a moving average similar to the envelope technique.
- Bolling Bands are placed two standard deviations above and below the moving average, which is usually 20 days (*What's the underlying assumption here? Normality*).
- Using two standard deviations ensures that 95% of the price data will fall between the two trading bands. Each touch of the lower band signaled an important market bottom and a buying opportunity.

Oscillators and Contrary Opinion

Momentum

- Market momentum is measured by continually taking price differences for a fixed time interval. The formula for momentum is

$$M = V - V^X$$

where V is the latest closing price and V^X is the closing price X days ago.

- While the 10 day momentum is a commonly used time period for reasons discussed later, any time period can be employed. A shorter time period produces a more sensitive line with more pronounced oscillations. A longer number of days (such as 40 days) results in a much another line in which the oscillator swings are less volatile.
- If prices are rising and the momentum line is above zero line and rising, this means the upward trend is accelerating. If the up-slanting momentum line begins to flatten out, this means the new gains being achieved by the latest closes are the same as the gains 10 days earlier. While prices may still be advancing, the rate of ascent (or the velocity) has leveled off. When the momentum line begins to drop toward the zero line, the uptrend in prices is still in force, but at a decelerating rate. The uptrend is losing momentum.

- The momentum chart has zero line. One could use the crossing of the zero line to generate buy and sell signals. A crossing above the zero line would be a buy signal, and a crossing below the zero line, is a sell signal.

The Relative Strength Index (RSI)

$$RSI = 100 - \frac{100}{1 + RS}$$

$$RS = \frac{Average\ of\ x\ days'up\ closes}{Average\ of\ x\ days'\ down\ closes}$$

- The 80 level usually becomes the overbought level in bull markets and 20 level the oversold level in bear markets.
- 14 days are usually used in the calculation.
- The 50 level is the RSI midpoint value, and will often act as support during pullbacks and resistance during bounces. Some traders treat RSI crossings above and below the 50 level as buying and selling signals respectively.

Moving Average Convergence/Divergence (MACD)

- The faster line (called the MACD line) is the difference between two exponentially smoothed moving averages of closing prices (usually the last 12 and 26 days or weeks). The slower line (the signal line) is usually a 9 period exponentially smoothed average of the MACD line.
- Most traders utilize the default values of 12, 26, and 9 in all instances. That would include daily and weekly values.
- The actual buy and sell signals are given when the two line cross. A crossing by the faster MACD line below the slower is a sell signal. In that sense, MACD resembles a dual moving average crossover method.
- MACD values also fluctuate above and below a zero line. That's where it begins to resemble an oscillator. An overbought condition is present when the lines are too far above the zero line. An oversold condition is present when the lines are too far below the zero line. The best buy signals are given when prices are well below the zero line (oversold).
- The two MACD lines can be turned into an MACD histogram. The histogram has a zero line of its own. When the MACD lines are in positive alignment (faster line over the slower), the histogram is above its zero line. Crossings by the histogram above and below its zero line coincide with actual MACD crossover buy and sell signals.
- The real value of the histogram is spotting when the spread between the two lines is widening or narrowing. When the histogram is over its zero line (positive) but starts to fall toward the zero line, the uptrend is weakening. When the histogram is below its zero line (negative) and starts to move upward toward the zero line the downtrend is losing its momentum.

- Histograms are best used for spotting early exit signals from existing positions. It is much more dangerous to use histogram turns as an excuse to initiate new positions against the prevailing trend.

On Balance Volume (OBV)

On Balance Volume (OBV) is a momentum indicator that uses volume flow to predict changes in stock price. The underlying assumption is when volume increases sharply without a significant change in the stock price, the price will eventually jump upward, and vice versa.

$$OBV_t = OBV_{t-1} + Volume_t \cdot I(Close_t > Close_{t-1}) - Volume_t \cdot (Close_t \leq Close_{t-1})$$

Note: All above signals are just reference values. You may alter them to fit your own trading needs and assets.

Signal Modeling

- Given a threshold θ, the signal is

$$signal_t = \begin{cases} 1 & buy & for\ P_t < -\theta \\ 0 & hold & for\ -\theta < P_t < \theta \\ -1 & sell & for\ P_t > \theta \end{cases}$$

A Toy Model of Neural Networks

Assume $s_t^{n_1,n_2}$ is the trading signals generated from the short n_1 and the long n_2 moving averages. Under general regularity conditions, a sufficiently complex single hidden layer feed-forward network can approximate any number of a class of functions to any desired degree of accuracy.

The Linear Test Regression

$$r_t = \alpha + \sum_{i=1}^{p} \beta_i r_{t-i} + \sum_{i=1}^{p} \eta_i s_{t-i}^{n_1,n_2} + \epsilon_t \; \epsilon_t \sim ID(0, \sigma_t^2)$$

Single Layer Feed-forward Network Model

$$r_t = \alpha + \sum_{i=1}^{p} \beta_i r_{t-i} + \sum_{j=1}^{d} \eta_j G(\alpha_j + \sum_{i=1}^{p} \gamma_i s_{t-i}^{n_1,n_2}) + \epsilon_t \; \epsilon_t \sim ID(0, \sigma_t^2)$$

where G is the **activation function** which is chosen to be a sigmoidal function:

$$G(x) = \frac{1}{1 + e^{-\alpha x}}$$

Single Layer Feed-forward Network Model with lagged returns alone:

$$r_t = \alpha + \sum_{i=1}^{p} \beta_i r_{t-i} + \sum_{j=1}^{d} \beta_j G(\alpha_j + \sum_{i=1}^{p} \gamma_i r_{t-i}) + \epsilon_t \; \epsilon_t \sim ID(0, \sigma_t^2)$$

d is the total types of signals you want to enclose in the prediction.
p is the total numbers of lags you choose to enclose in the information set of prediction.

Random Forest

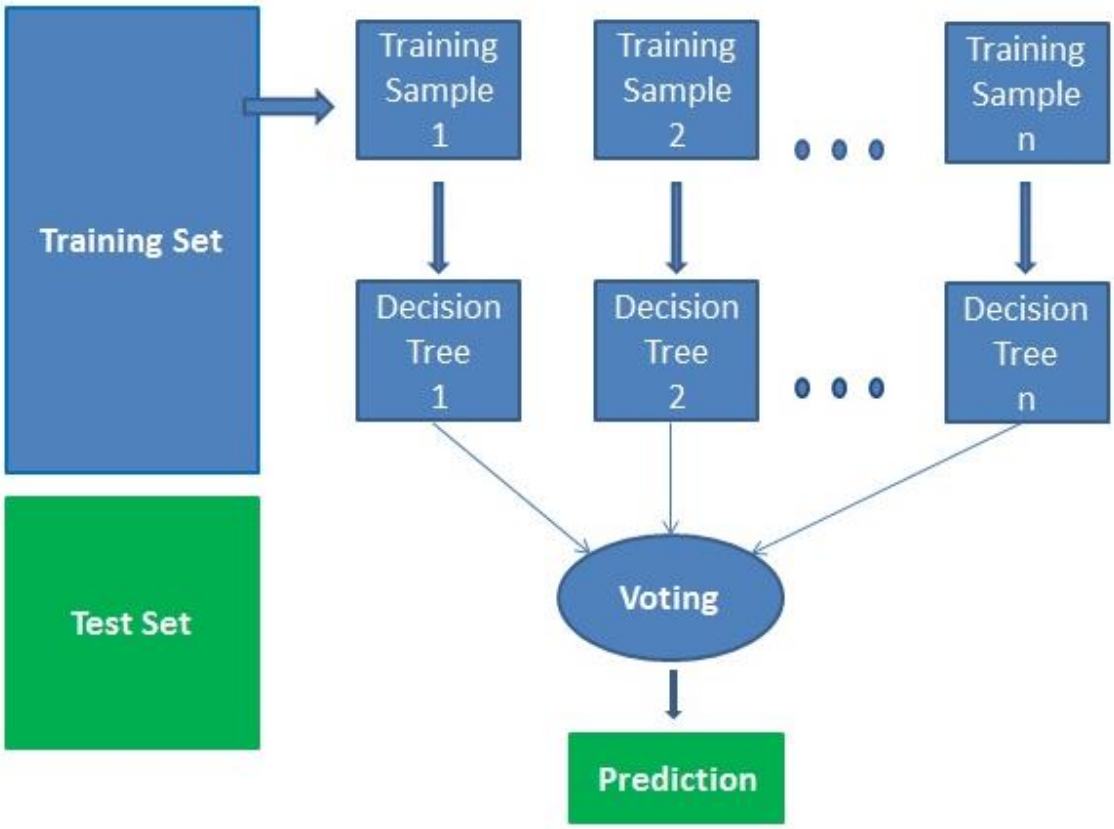
You have learnt

- Bootstrap
- Time-Series Regression
- Measurement of Predictive Accuracy

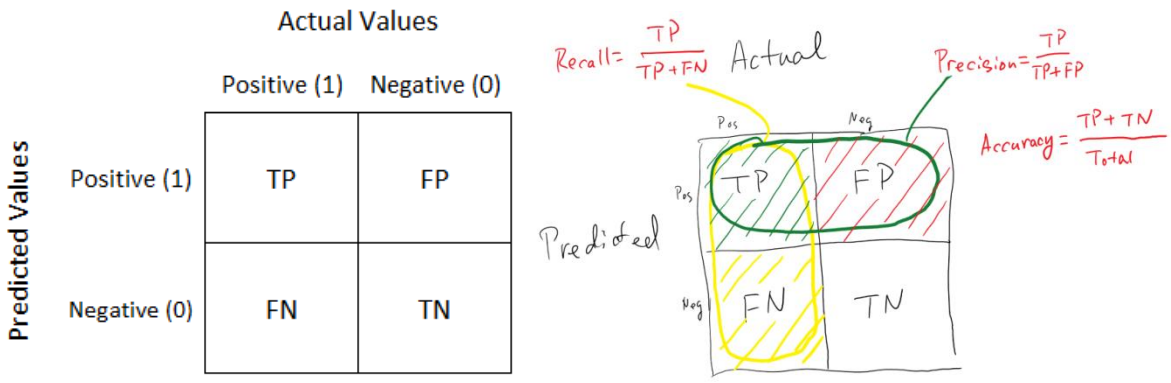
Now you are ready to enter **Random Forest,**

Random Forest works in four steps:

1. Select random samples from a given dataset.
2. Construct a decision tree for each sample and get a prediction result from each decision tree.
3. Perform a vote for each predicted result.
4. Select the prediction result with the most votes as the final prediction.


Training Set
Test Set
Training Sample 1
Training Sample 2
Training Sample n
Decision Tree 1
Decision Tree 2
Decision Tree n
Voting
Prediction

Confusion Matrix



True Positive:

Interpretation: You predicted positive and it's true.

You predicted that a woman is pregnant and she actually is.

True Negative:

Interpretation: You predicted negative and it's true.

You predicted that a man is not pregnant and he actually is not.

False Positive: (Type 1 Error)

Interpretation: You predicted positive and it's false.

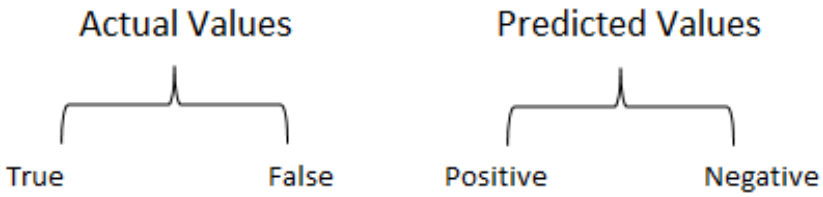
You predicted that a man is pregnant but he actually is not.

False Negative: (Type 2 Error)

Interpretation: You predicted negative and it's false.

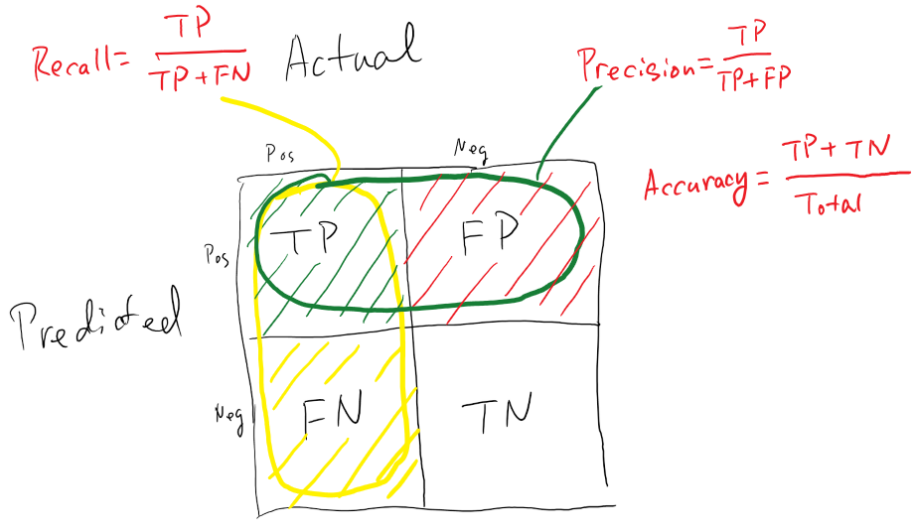
You predicted that a woman is not pregnant but she actually is.

Just Remember, We describe predicted values as Positive and Negative and actual values as True and False.



Example:

y	y pred	output for threshold 0.6	Recall	Precision	Accuracy
0	0.5	0			
1	0.9	1			
0	0.7	1			
1	0.7	1	**1/2**	**2/3**	**4/7**
1	0.3	0			
0	0.4	0			
1	0.5	0			



Recall

Out of all the positive classes, how much we predicted correctly. It should be high as possible.

Precision

Out of all the positive classes we have predicted correctly, how many are actually positive.

Accuracy

Out of all the classes, how much we predicted correctly, which will be, in this case 4/7. It should be high as possible.

F-measure

$$F\text{ - measure} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall} + \text{Precision}}$$

It is difficult to compare two models with low precision and high recall or vice versa. So to make them comparable, we use F-Score. F-score helps to measure Recall and Precision at the same time. It uses Harmonic Mean in place of Arithmetic Mean by punishing the extreme values more.