Global Rates and FX Primer

# Learning about machine learning: applications for rates and FX

Primer

Bank of America Merrill Lynch

09 October 2018

## A primer on machine learning applications in rates and FX

We discuss machine learning (ML) applications in rates and FX. The report includes examples of supervised learning, where input and output data are labelled to form a learning framework and unsupervised learning, algorithms that act on information without guidance. We also detail the data involved and the math behind the analysis. It is meant to be an introduction to a new and still evolving field.

## K-nearest neighbours

A supervised learning method for pattern recognition that can be used for classification and regression. Example: we use economic and market indicators to determine the part of history that most resembles the US today, to assess the current phase of the hiking cycle.

## Naïve Bayes

A supervised learning technique to classify inputs based on Bayes' theorem, which describes the relationship between probabilities of events given current evidence. Example: sentiment analysis to gauge the severity of trade war rhetoric.

## K-means clustering

Unsupervised learning, partitioning a dataset into k clusters according to each observation's distance from the cluster mean. Example: to understand the behaviour of EURUSD in the context of the Italian bond selloff, we apply the K-means classification algorithm to European and US macroeconomic data that we deem relevant for FX.

## Support Vector Machine (SVM)

Supervised learning for classification and regression analysis, using a set of training data to construct boarders that optimally separate the input data. Example: classifying currency spot and volatility into states based on economic fundamentals.

## Kalman filter

Supervised learning, an algorithm that provides an efficient recursive means to estimate the state of a process by producing a joint probability distribution over the variables in question. Example: we improve carry trades with dynamically estimated hedging ratios that adapt to the price action over time.

## And more

We give a high level discussion of the potential role for machine learning in finance, provide a dictionary on machine learning jargon, as well as a list for further reading. We also brush on additional methods such as principal component analysis (an unsupervised learning technique) which we apply to construct leading indicators of economic activity.


G10 FX Strategy
Global

**Alice Leng**
FX Strategist
MLPF&S
alice.leng@baml.com

**Carol Zhang**
Rates Strategist
MLPF&S
carol.zhang@baml.com

**Samuel Mann**
FX Strategist
MLI (UK)
samuel.mann@baml.com

**Athanasios Vamvakidis**
FX Strategist
MLI (UK)
athanasios.vamvakidis@baml.com






**Refer to important disclosures on page 35 to 36.**


**11920075**

Timestamp: 09 October 2018 06:51AM EDT

# The basics: What is machine learning?

The term "machine learning" is generally understood to refer either to statistical techniques that give computers the ability to "learn", or to the field of study concerned with the design and development of these techniques. This definition begs an obvious question: how should we imagine "learning" of computers to happen? In the context of machine learning, it is understood as the ability to progressively improve performance on a specific task, or to modify processing on the basis of newly acquired information. While a more "traditional" computer programme might solve a problem and stop there, a machine learning algorithm might use the result to go back and evaluate the starting point, perhaps adjusting it and solving the problem again.

This is nicely illustrated by algorithms that recognise handwriting. Assume we define what the number "4" looks like and then ask a computer whether a set of handwritten numbers are 4s or not. A "traditional" programme might simply compare the pre-defined and hand-written shapes and, given the amount of overlap, come to a conclusion. The programme would then stop. A machine learning algorithm, on the other hand, might, once it has recognised a handwritten number, update its original knowledge of what 4s can look like, implicitly acknowledging that they come in different shapes. Having done so, the algorithm can then re-evaluate the original dataset, perhaps finding another handwritten 4 that was too dissimilar at first to be recognised. Through these iterations, the algorithm builds up a more comprehensive knowledge of the shapes, gradually increasing its ability to recognise a handwritten number. In this sense, machine learning tries to emulate and formalise the heuristic circle employed intuitively by humans to understand the world.

## Is this new to macroeconomics and finance? Yes and no.

Yes, because some algorithms have only been constructed, and made easily accessible, in the very recent past and have not been applied to macroeconomics and finance until recently. No, because they are often based on processes that – both formally and informally – have been employed by researchers for decades. Financial analysts, for example, like to draw parallels between current and historical episodes to infer future developments. Is today's rise in tech stocks reminiscent of the dot-com bubble? Are Europe's inflation dynamics similar to those of Japan at the start of the lost decade? Can we compare today's rise in nationalism to developments before the outbreak of the First World War? Given a dataset with historical information, some of the most popular machine learning techniques – we will talk about "classification" and "clustering" below – try to answer exactly these questions, in an efficient and mathematical way.

### Demystifying ML

At the danger of potentially irritating a few flag bearers of machine learning, it is worth pointing out that machine learning jargon is somewhat misleading, seemingly presenting the reader with novel ideas when these are sometimes just old tricks in new clothes. Terms such as "input feature", "output feature", "training set" and "test set" might cause confusion among the uninitiated and make machine learning texts inaccessible. Were they replaced with "independent variable", "dependent variable", "in sample" and "out of sample", people with a basic knowledge of statistics would understand perfectly well. We encourage the reader not to be put off by this linguistic hurdle and we point out that many of the underlying methods are very intuitive and promise fruitful results.

In what follows, we will wherever possible provide a translation from machine learning jargon to statistics terms to help the uninitiated navigate the new land.

Leaving aside jargon, there are in fact a number of machine learning algorithms which are already well known in econometrics, such as probit regressions, Kalman filters or principal component analysis.

## Machine learning can be extremely useful for rates and FX

Intuitively, a lot of machine learning algorithms that are popular in an FX or rates context do nothing other than provide a marginal improvement on traditional and well-known processes. In that sense, they simply offer a way of doing something that has been done previously in a slightly better way – typically faster, more efficiently or more comprehensive. For example, a machine learning algorithm might offer a comprehensive way of establishing whether the market is in a "risk-on" or "risk-off" environment. Of course, there are established ways for market participants to evaluate this – looking at volatility, safe haven currency returns, emerging market flows, etc. – but a machine learning algorithm can do so while incorporating vastly more information and classify a data point as either risk on or off with more precision and rigour.

### But it does not replace humans

Note that solving such a problem with a machine learning algorithm is not unambiguously superior to human judgement. Sometimes experience is hard to beat. But even in such a scenario, machine learning can help to inform the decision of an experienced analyst.

## Machine learning is not automation, but the line is blurry

Simply automating a task is not machine learning, because the computer "only" follows strict instructions without improving its understanding of the problem in any way. That said, the line is blurry, as can be illustrated with the following example.

In econometrics, one of the main challenges faced by the econometrician is model specification, such as selection of the number of variables to improve model fit without overfitting. Typically, econometricians would start with an informed choice of at least two specifications and then compare statistics such as the adjusted R-squared (i.e. the percentage of explained variation in the dependent variable, adjusted for the number of explanatory variables), or a slightly more sophisticated criterion such as the Akaike Information Criterion (AIC) or Bayes Schwartz Information Criterion (BIC). Whichever specification performs better along these criteria would then be selected. Poignantly, this has often been described as an art as much as a science.

A machine learning algorithm might be programmed in a way to internalise the model evaluation as part of the learning process. The algorithm could, for example, iterate through model specification, evaluating the model (according to a pre-specified criterion), and use the results to inform the next iteration of model specification. Critical voices might say that there is no real "learning", because the computer at each step follows a clearly defined instruction. In practice, however, the bar for an algorithm to pass as machine learning is much lower and many of the most powerful machine learning tools are surprisingly simple. By internalising layers of decision-making that were traditionally done by humans, machine learning can also help to move statistical analysis closer towards the realm of science, rather than art (although there is still a good dose of the latter in machine learning).

# Cutting to the chase: methods & applications

Machine learning tasks have been traditionally separated into two distinct classes: *supervised* and *unsupervised* learning. There are also newer approaches that fall in neither category, such as deep and reinforcement learning. For the purpose of this report, however, we will focus on the more wide-spread traditional approaches as an introduction.

In supervised learning, the researcher provides the computer with both inputs and sample output. The algorithm is then tasked to find a rule that maps the two. Econometric regression would, for example, fall into this category. By means of illustration, we might want to find a connection between economic indicators and recession probabilities. Under supervised learning, the researcher provides historical

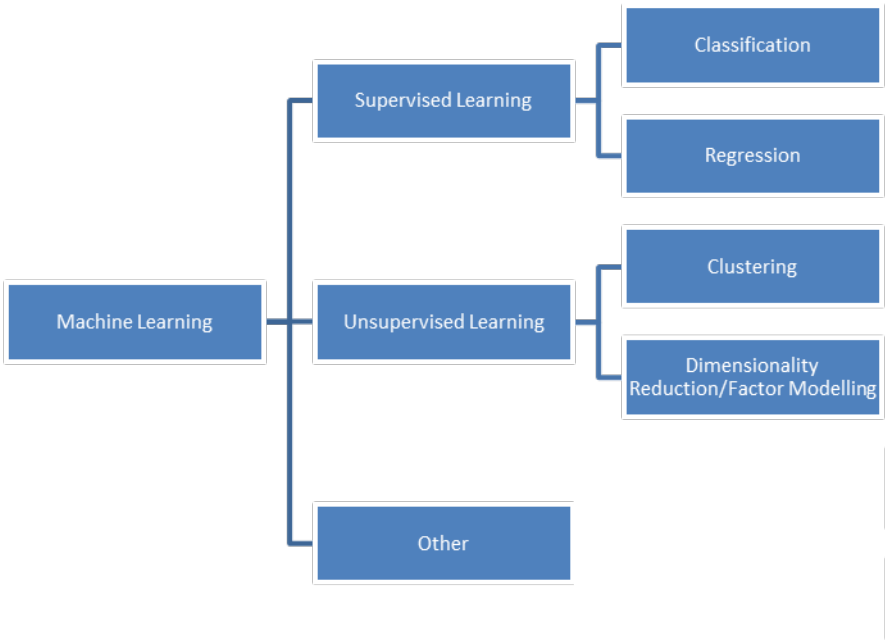
data, already labelled as either “recession” or “no recession” and the algorithm then tries to uncover which economic variables are informative to predict recessions and how.

In unsupervised learning, inputs are provided to the computer without corresponding outputs (i.e. without any labels or pre-categorisation). The algorithm then uncovers structures in the data on its own. Linking to the example about recession probabilities above, if the researcher feeds the economic variables into an unsupervised classification algorithm, the computer might on its own group data into “recession” or “no recession”, but might as well come up with completely different groupings, whether they are corresponding to “high/low volatility”, “risk on/off”, “accelerating/decelerating cycle”, etc.

Intuitively, in supervised learning, the researcher “teaches” the computer what a certain group/class/label is, while in unsupervised learning, no such information is provided to the computer. The most important applications of supervised learning are classification and regression, while the prime applications of unsupervised learning are clustering and dimensionality reduction/factor modelling.

In what follows we will first go over what each of these applications are, followed by examples of algorithms for each.

**Exhibit 1: Overview of main machine learning classes and applications**




Source: BofA Merrill Lynch Global Research

## Supervised learning: Classification

As the name suggests, classification has the aim of classifying inputs into a set of categories. Applications are far-ranging with anything from an email system deciding if an incoming message is a spam, or a bank determining if a credit card transaction is fraudulent. All these examples are using some kind of classification algorithm. Importantly, this task aims at identifying to which category a *new* observation should belong, based on other observations that have already been classified. Because the researcher provides these already classified observations (the so-called “training set”), classification falls under supervised learning. A very straightforward question that an FX strategist might ask of a classification algorithm would be whether we are in a week (year, month, etc.) that sees a currency pair, say EURUSD, rise or fall. The researcher can provide the algorithm with a training set that contains information on the currency pair over past episodes (“output”) as well as explanatory data that might be informative

("input"). The algorithm then finds a mapping from input to output and uses the mapping to classify the current period, in effect producing a forecast about whether EURUSD will go up or down. Examples of classification algorithms – some of which we discuss in detail below – include Logistic, Support Vector Machine, Random Forest or Hidden Markov.

## Supervised learning: Regression

The concept of regression will be familiar to anyone with a foundation in classic statistics. While the term is used in a slightly wider sense in machine learning, the fundamental idea carries over. In particular, machine learning sees regression as similar to classification, but with a continuous, rather than discreet, output. Taking the example from above on EURUSD, while the classification algorithm would predict EURUSD up or down, a regression algorithm would predict *by how much* EURUSD is expected to move. In fact, there are some approaches that have characteristics of both a regression and a classification. A logistics algorithm, for example, is an approach where, in a first step, a regression produces a continuous output variable between 0 and 1, which is then typically transformed into a binary classification, depending on whether the output is above or below 0.5. Because of the last step, "logistic regressions" are usually seen as a classification algorithm, rather than a regression within machine learning. Examples for regression algorithms include Lasso, Ridge, Loess, K-nearest neighbours, Spline and XGBoost.

## Unsupervised learning: Clustering

Clustering can be understood as a variation of classification where no prior labels are provided by the researcher. Hence, the algorithm is one of unsupervised learning. Rather than trying to map inputs to pre-defined outputs, the algorithm itself tries to establish notions of similarity or sameness within the input data. In doing so, however, it is up to the algorithm to establish the measure of similarity, based on the inputs provided. Naturally, the researcher has an influence on the kind of similarity that is being established through the data that is provided. An intuitive example of this might be data on a country's population. If we feed an algorithm with only the body height of individuals, the algorithm might group the population into children and grown-ups, or men and women. If, however, we add data on the number of siblings, hours worked, income and education, the algorithm might for example establish urban/rural groups, or cluster along socio-economic lines. The important characteristic of clustering algorithms is that it is the computer, not the researcher, who establishes the clusters. Examples for clustering algorithms include K-means, Birch, Ward and Spectral Cluster.

## Unsupervised learning: Dimensionality Reduction/Factor Analysis

As the name suggests, dimensionality reduction is the process of reducing the dimensions of a set of variables, by identifying a set of "principal" variables. This can be done in one of two ways. Either, a subset of the original set of variables is selected, which, according to some criterion, is particularly informative. Alternatively, new variables can be identified that explain as much as possible of the variation within the original set of variables.

Again, this is a process that is very similar to what is often done informally by analysts in financial market. Taken to the extreme, an FX strategist might proclaim that a particular currency moves only with the price of oil. Intuitively, the analyst reduced the dimensionality of the set of variables that might matter for the currency to one. A machine learning algorithm can do this in a particularly efficient way. One of the most wide-spread approaches is so-called Principal Component Analysis (PCA), a statistical method that uncovers the series which, in descending order, explain the most variation within a dataset. By, for example, conducting a PCA on a collection of hundreds of US economic indicators, one is able to identify the main underlying factors for economic activity within the US. Interestingly, in such a context, a small number of series, typically 3-7, tend to explain a large share of the variation within the larger dataset. Naturally,

working with a handful of variables in any further analysis is a lot easier than working with hundreds. Examples for dimensionality reduction and factor analysis include Principal Component Analysis, Individual Component Analysis and Non-negative matrix factorization.

While not advertising it as machine learning, factor analysis has been used in macroeconomics and macro strategy for a long time. Bernanke, Boivin and Eliasz[1] (2005) include factors – extracted from a large macro dataset – into a vector autoregressive model (VAR) to analyse the effects of monetary policy. Their so-called Factor Augment VAR (FAVAR) manages to avoid the curse of dimensionality, deal with problems of mismeasurement and avoid biased results due to a lack of information in the model. In a more recent contribution, Corsetti, Duarte and Mann[2] (2018) use a parsimonious dynamic factor model to analyse the transmission of ECB monetary policy to over 300 economic series across the euro area.

## Other approaches

Besides the techniques listed above, there are a number of approaches that do not fall under the above categories. They cover methods such as deep learning, reinforcement learning and active learning. While these methods are extremely powerful in certain areas – deep learning, for example, is particularly suitable to evaluating pictures – they tend to only find secondary application in rates and FX for now. As they go beyond the introductory level we aim to provide in this primer, we leave them for the interested reader to explore on their own.

[1] Bernanke, B., Boivin, J. and Eliasz P. (2005) Measuring the Effects of Monetary Policy: a Factor-Augmented Vector Autoregressive (FAVAR) Approach. *The Quarterly Journal of Economics*, February 2005.
[2] Corsetti, G., Duarte, J. and Mann, S. (2018) One Money Many Markets: a Factor Model Approach to Monetary Policy in the Euro Area with High-Frequency Identification. *CFM Discussion Paper DP2018-05*. Centre For Macroeconomics, February 2018.

# Jargon buster

We pointed out above that a number of expressions used in machine learning have close relatives in classical statistical analysis. The following list translates some of the common machine learning jargon into statistics terms. It is followed by a list of common machine learning-related terms and buzzwords.

**Table 1: Translating machine learning terminology to statistics**

| Machine learning | Statistics |
|---|---|
| training set | in sample |
| test set | out of sample |
| classifier | hypothesis |
| learning | fitting |
| response | Label |
| weights | parameters |
| Instance / example | data point |

Source: BofA Merrill Lynch Global Research

**Table 2: Selection of common jargon and buzzwords used in machine learning**

| | |
|---|---|
| Alternative data | Non-traditional data used in the investment process. Contrary to traditional datasets (e.g. PMI, CPI), alternative data tends to be unstructured, hard to find and/or large in volume (e.g. twitter feed, satellite images, news articles) that requires significant computational power to analyse. |
| Artificial intelligence(AI) | A computer science field that tries to teach machines to solve problems the way humans do (e.g. speech recognizing, image recognition etc. |
| Big data | Usually used to describe datasets that are big and complex that traditional techniques and software are inadequate to deal with them. Increasingly, it also refers to techniques that extract value from data that aren't necessarily huge in size. |
| ChatBot | Software that converses with a human. The chatbot tries to simulate human conversation with the help of artificial intelligence. SIRI and Alexa are examples of them. Bank of America also launched Erica early in 2018 that answers client's daily banking questions. |
| Classification | It is supervised learning method where the output variable is a category, such as “positive” or “negative” or “Yes” and “No”. |
| Cloud computing | Cloud computing refers to services provided remote servers over the internet, as opposed to on a local server, and its infrastructure is usually managed by a third party vendor. |
| Clustering | Clustering is an unsupervised learning method used to discover the inherent groupings in the data. For example: Grouping customers on the basis of their purchasing behaviour. |
| Computer vision | A computer science field that tries to teach computers to visualize, process and identify images/videos in the same way human vision does. Applications include facial recognition, self-driving cars and more. |
| Cross validation | Cross Validation is a technique which involves reserving a particular sample of a dataset which is not used to train the model. Later, the model is tested on this sample to evaluate the performance. |
| Data science | The combination of computer science, statistics, modelling, and artificial intelligence. A cross-discipline focused on getting the most from modern data-rich technical environments. |
| Deep learning | A branch of machine learning. It often uses Artificial Neural Network (ANN) which adopts the concept of human brain to facilitate modelling. ANN requires a vast amount of data and this algorithm is highly flexible when it comes to model multiple outputs simultaneously. |
| Feature engineering | Converting available real life information into datasets that can be fed into standard algorithms. |
| Feature selection | The process that selects the key subset of original data features in an attempt to reduce the dimensionality of the training problem. |
| Hadoop | A collection of open-source software utilities that facilitate using a network of many computers to solve problems involving massive amounts of data and computation. |
| Neural Networks (or ANN) | Inspired by the biology of the brain, these ANN’s are a huge network of numerous interconnected conceptualized artificial neurons which pass data between themselves. The neurons are “fired” if an activation threshold is met. The combination of these fired neurons makes up the learning process. |
| NLP | Natural Language Processing. A field which aims to make computer systems understand human speech, including techniques to process, structure and categorize raw text and extract information. |
| Recommender system | Algorithms that attempts to recommend the most relevant items to a particular user, often based on their needs and interests. Real world examples include Netflix's movie recommendations, Bank of America Mercury's "Recommended for you" section. |
| Reinforcement Learning | A branch of Machine Learning. Unlike traditional algorithms that minimize loss, these techniques focus on actions that maximize reward. |
| scikit-learn | One of the most popular free machine learning libraries for the Python programming language |
| TensorFlow | An open source software library developed by Google. It is mainly used for machine learning applications such as building neural networks. |
| Transfer learning | Transfer learning refers to applying a pre-trained model on a new dataset to solve a similar problem. |

Source: BofA Merrill Lynch Global Research

# Machine learning applied to Rates and FX

## K-nearest neighbour

- Supervised learning → Classification or Regression

**What is it? And what does it do?**

A k-nearest neighbour (k-NN) algorithm is a method for pattern recognition that can be used for classification and regression. The non-parametric method forecasts data by looking at a pre-classified dataset (hence supervised learning) and establishing which class is most similar to the data point in question. Those similar situations within the historical sample are then used as the best forecast.

When k-NN is used as a classification tool, the output of the algorithm is class membership, while in a regression setup the output is the so-called "property value" of the object in question. In practice, that is the average of the values of its k nearest neighbours.

The "k" in k-NN specifies how many observations near the data point in question are taken into consideration when classifying it. This is best illustrated with an example. Let's suppose we are in front of a restaurant and want to know whether it is "good" or "bad". To find out more, we might look at a map of restaurants in the vicinity of which we already know whether they are "good" or "bad" (the pre-classified historical sample), assuming that closeness to other good restaurants increases chances that the restaurant in question is good as well. By specifying k, we decide how many of the nearest restaurants we look at for a guide on how to classify the restaurant in question. For k=1, we look at only the restaurant nearest to the one in question. If it is "good", then we classify the restaurant in question as "good" too – or "bad" if the nearest neighbour is "bad". If we set k=3 instead, we look at the three restaurants that are closest to the one in question. If, for example, two of them are "bad" and only one is "good", we might classify it as "bad". (Note, in a regression setup, we would instead get a result along the lines of "0.33% good" – the average of the k nearest neighbours.) Naturally, all of this rests on the highly contentious assumption that geographical closeness between restaurants implies closeness in quality. But we could use other measures of "distance" such as the difference in the number of guests per month, the number of years' experience of the kitchen staff, etc. In fact, the algorithm could be set up in a way that takes into account all of these dimensions and more, to establish an informative measure of distance or similarity.

**Example: what part of cycle are we in?**

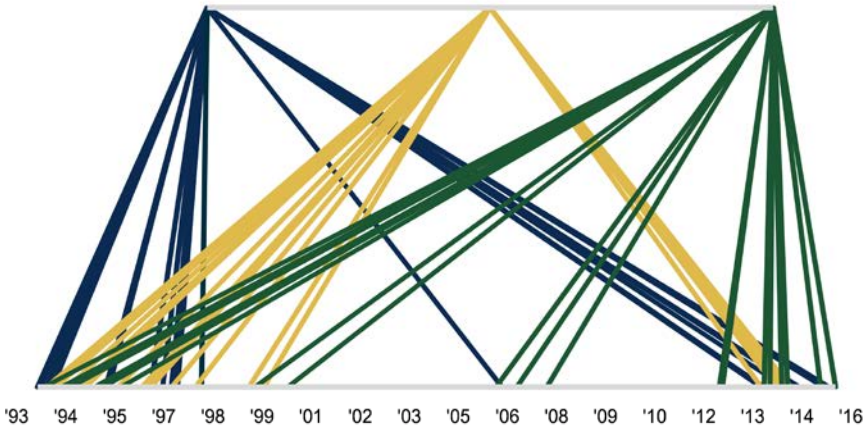
In the following example, we used over 140 monthly economic and market indicators in the US to algorithmically determine the months in history that most closely resemble today. This example was originally published as Liquid Insight: We like big data 10 May 2018.

Specifically, we deconstruct economic history (as summarized by the time series of economic and market indicators) into individual economic configurations (i.e. monthly snapshots). Instead of arbitrarily comparing the present to past hiking cycles (a popular exercise among strategists), we algorithmically quantify the closest historical parallels.

First, we collect the data, encompassing various aspects of the economy going back to the beginning of 1993. We then apply Principal Component Analysis (see section on Principal Component Analysis below) to reduce this dataset to eight factors (which roughly explain 50% of the variance in the original dataset). Intuitively, we now have a world where each month is characterized by eight factors. Visual inspection indicates that the first principal component could be interpreted as representing the overall direction of the economy (Chart 2). That said, there are a number of theoretical issues connected to interpreting factors and we recommend not focusing too much on trying to find meaning in the factors themselves.

We then employ a k-nearest neighbour algorithm, which mathematically computes the distance for a given month to every historical snapshot. "Distance" in this context is measured along the 8 dimensions given by the factors that characterise the various episodes. The best forecast is then given by the months with the shortest distance to the present.

**Chart 1: top 20 closest historical parallels for each of the past three months**




Note: For each of the last three months, we show the mapping of today's market environment to 20 of its close historical analogues in the past 25 years. We experimented with k-nearest neighbours algorithm in machine learning to determine the "closeness"
Source: BofA Merrill Lynch Global Research

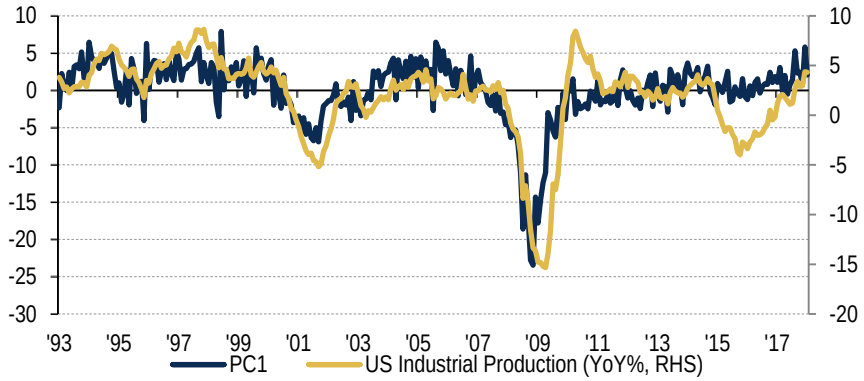
### In this exercise, big data says the US economy is doing just fine

To us, the most interesting observations from this exercise are two-fold.

- First, just because we are (at the time of analysis in May 2018) in a hiking cycle, it does not automatically make the last two hiking cycles the best reference points. Looking at the top 20 historical parallels for each of the last three months, the periods that came up more often were months in '93-'94, '96-97 and '13-'14 (Chart 1).

- Second, with the benefit of hindsight, these episodes did see relatively strong economic growth, despite the fact that most of these periods were not part of a consistent hiking cycle. For example, 2014 saw significant improvements in broad economic data, as indicated by data surprise indices. On average, these historical parallels saw Industrial Production grow at 3.8% yoy and a relatively strong manufacturing PMI of 53.5.

**Chart 2: The 1st principal component of our dataset and US industrial production**




Source: BofA Merrill Lynch Global Research

### Big data in this illustration says the curve may continue to flatten

Given the diverging views on the US yield curve among investors, we also use the classification tool to try and assess the most likely move in rates over the following month. To this end, we look at the directional moves of the 5y-30y curve in the months following the historical "neighbours". We find that, typically, the curve is flattening in the month following the selected historical episodes, with an average move of about -3.5bp.

### Data

For our analysis, we use the St. Louis Fed FRED-MD database, a monthly frequency dataset summarizing key economic and market indicators in the US. The data cover various aspects of the economy including input and output, the labor market, inflation, consumption, money and credit. In addition, we added market variables such as changes in Fed pricing, VIX and MOVE indices, non-commercial net positioning in ED and TY contracts (as a percentage of total open interest) and an economic policy uncertainty index. Because of the limited history of the added data, our working dataset starts in 1993. We summarize the data categories used in Table 3.

**Table 3: Summary of data by category**

| Output and income | Labor market | Housing | Consumption, orders and inventories | Money and credit | Prices | Rates, FX & Stocks | Additional Market variables |
|---|---|---|---|---|---|---|---|
| • Personal income<br>• Aggregate and sector Industrial Production<br>• Manufacturing Production Index<br>• Capacity utilization | • Aggregate and sector NFP<br>• Avg Weekly Hours<br>• Avg Hourly Earnings<br>• ISM Manufacturing: Employment Index<br>• Initial Claims<br>• Unemployment rate<br>• Duration of Unemployment<br>• Help-Wanted Index | • Aggregate and by region Housing Starts<br>• Aggregate and by region New Private Housing Permits | • Real Personal Consumption Expenditures<br>• Industries Sales<br>• Retail Sales<br>• ISM<br>• New Orders<br>• Unfilled Orders<br>• Inventories<br>• Consumer Sentiment | • Money Stock<br>• Monetary Base<br>• Reserves of Depository Institutions<br>• C&I loans<br>• Real Estate Loans<br>• Consumer loans | • PPI & CPI<br>• Crude oil prices<br>• Manufacturing Prices Index<br>• Personal Consumption Expenditure | • Fed Funds Rate<br>• Short interest rates<br>• Key UST yields and curves<br>• Corporate bond yields<br>• Trade weighted Dollar<br>• Major USD-pairs<br>• SPX indices, dividend yield and PE ratio | • 1y- ahead Fed hike expectation<br>• MOVE Index<br>• VIX Index<br>• Non-commercial net positioning in futures<br>• Monetary policy uncertainty index |

There were 139 variables used in this analysis, we summarize the key categories here. Source: St. Louis Fed, BofA Merrill Lynch Global Research

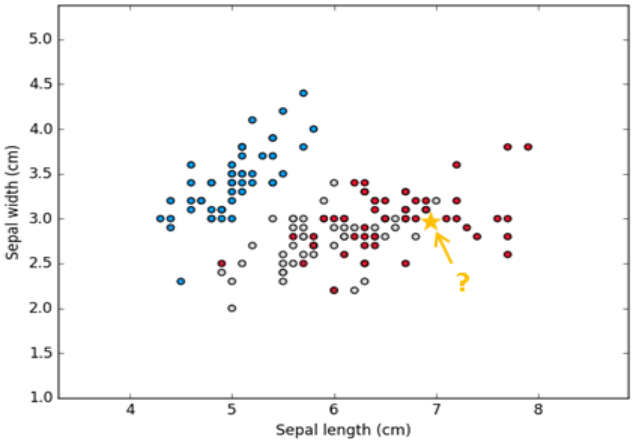
### The Math

The goal for the K-nearest neighbour (KNN) algorithm is to assign a new observation to a group, given a set of pre-defined groups and the data points each group contains. It essentially boils down to forming a majority vote between the K most similar data points to the new observation. In most cases, similarity is defined according to a distance metric between two data points. A common choice for distance calculation is the Euclidean distance, though other distance measures such as Manhattan, Chebyshev and Hamming distance are also used depending on the problems at hand and the nature of the data.

The concepts and math of similarity and distance, are in our opinion best explained with a visual aid. A classical textbook example is to apply KNN on the Iris flower dataset. This famous dataset, collected by Edgar Anderson in the 1930s, has become one of the most popular datasets on which to test machine learning techniques. It contains 50 samples for each of the three Iris species, and measurements for four features of each sample: the length and width of sepals and petals. Consequently, the dataset is a 150 by 5 matrix, where the 5$^{th}$ column contains the species label. For illustration purposes, we use two variables (sepal length and width) so that we can plot the data in two-dimensional space.

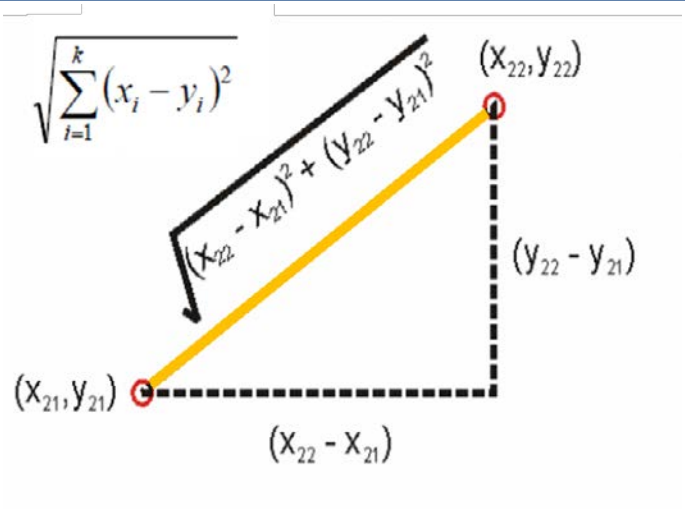
The question we are trying to answer with KNN is to find out what the most likely species of a new Iris sample (sepal length and width) is, given what we know about the flowers from existing samples (Exhibit 2)? Similarly, recall in our finance application

**Exhibit 2: Iris measurement: sepal width vs length, color of the circles represents different species. The goal of the KNN analysis is to assign the new observation (gold star in the chart) to the most likely species**




Source: BofA Merrill Lynch Global Research

**Exhibit 3: Calculating the Euclidean distance**



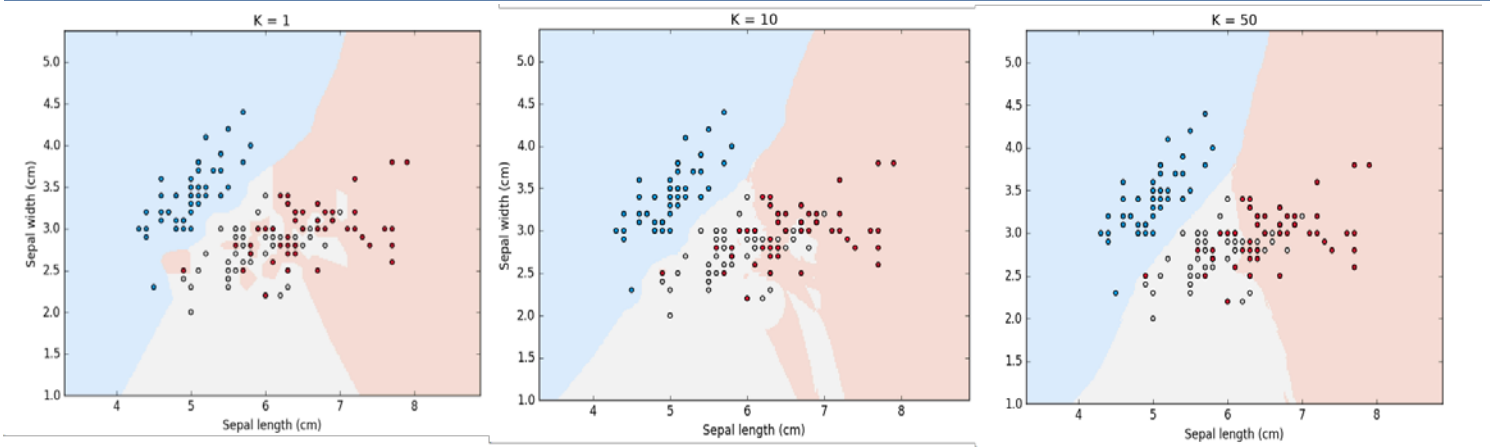

Source: BofA Merrill Lynch Global Research

above, we were trying to find the historical episodes that are were most similar to the economic conditions prevalent in the latest observation.

The key to the question is to find out which existing samples have the most similar (closest) combination of sepal length and width. To do this, we need to compute the distance between our unknown Iris sample and every existing data points. We use the Euclidean distance to measure closeness for this purpose - a common default choice. The calculation for Euclidean distance is straightforward (in two-dimensional space it is given by the Pythagorean Theorem), and illustrated in Exhibit 3. In our finance application above, we use the same methodology, except applied to a higher dimension, as our dataset contains more than two characteristics.

The next step is to determine K, the number of neighbouring data points to search for, and take the majority vote. In our sample, the species represented by the most neighbours among the K points make up the prediction. The choice of K is half science and half art (unsurprisingly), a balancing act between overfitting and under fitting (Exhibit 4). A standard method to determine the optimal K is to search for a range of possibilities and pick the number with the lowest misclassification rate (Exhibit 5).

**Exhibit 4: Different choice of K would result in different classifications results.**




Note: The background color of each chart (light blue, light red, and light gray) represents classification results produced by choosing different level of K (number of neighbours)
Source: BofA Merrill Lynch Global Research

The steps of the algorithm are summarised in Exhibit 5.

**Exhibit 5: Stylised steps of k-NN algorithm**

Step 1. Load the data
Step 2. Initialize the value of K
Step 3. Iterate from 1 to total number of training data points
    a. Calculate the distance between the query-instance and all the training examples
    b. Sort the distance in non-decreasing order
    c. Determine the K nearest neighbors based on sorted list
    d. Use simple majority of the category of nearest neighbors as the prediction value of the query instance
    e. Return the predicted class

## Naïve Bayes

- Supervised learning → Regression

### What is it? And what does it do?

Naïve Bayes is a supervised learning technique using regression to classify inputs. As the name suggests, Naïve Bayes is based on Bayes' theorem, which describes the relationship between probabilities of events given current evidence. Naïve Bayes is a family of probabilistic classification algorithms that take advantage of probability theory and Bayes' Theorem to predict the tag (i.e. character/class) of an item such as, for example, a piece of text (like a news headline or a customer review). The algorithms are probabilistic, which means that they calculate the probability of a tag for a given item, and then output the tag with the highest probability.

Intuitively, the way algorithms get these probabilities is by using Bayes' Theorem, which describes the probability of a feature based on prior knowledge of conditions that might be related to that feature. The "naïve" part in Naïve Bayes comes from the fact that the researcher makes strong independence assumptions between features. An algorithm might for example predict whether an animal is a sheep if it has four legs, weighs 80kg and is hairy. In doing so, however, the algorithm will not consider any correlation between the attributes ("features") and uses the additional information independently.[3]

A Naive Bayesian model is relatively easy to build, with no complicated iterative parameter estimation. This makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier does surprisingly well in many applications and often outperforms more sophisticated classification methods.

### Example: sentiment analysis

Market participants are constantly navigating between pricing and rhetoric on a variety of topics, which has become increasingly challenging in today's news cycle. Having an algorithm that helps filter through what's known and what's new regarding the sentiment of each topic (similar to an economic data surprise index) can be extremely useful in understanding how an issue evolves. Doing this manually would be a huge undertaking given the mass of the information available, making it a natural target for automation and machine learning. As a toy example, we apply Naïve Bayes to gauge the severity of trade war rhetoric in President Trump's Twitter comments.

The goal of this algorithm is to assign a probability to each new sentence, indicating how positive or negative the rhetoric is. In order to do this, we need some training data to tell us what kind of information could be classified as a positive or negative comment. For illustration purposes, we take the six sentences listed in Table 4 as training data, classifying them as either positive ("pos") or negative ("neg"). After training the algorithm with these sentences, we then want to see which group other sentences such as "The days of the U.S. being ripped-off by other nations is OVER!" belong to.

**Table 4: Toy example training data: Trump's comments regarding Trade issues from Twitter**

| Tag | Text |
|---|---|
| Neg | The U.S. has been ripped-off by other countries for years on Trade, time to get smart! |
| Neg | China, which is a great economic power, is considered a Developing Nation within the World Trade Organization. They therefore get tremendous perks and advantages, especially over the U.S. |
| Neg | Russia and China are playing the Currency Devaluation game as the U.S. keeps raising interest rates. Not acceptable! |
| Pos | China has agreed to buy massive amounts of ADDITIONAL Farm/Agricultural Products - would be one of the best things to happen to our farmers in many years! |
| Pos | European Union representatives told me that they would start buying soybeans from our great farmers immediately. |
| Pos | Deal with Mexico is coming along nicely. Autoworkers and farmers must be taken care of or there will be no deal. |

Source: BofA Merrill Lynch Global Research,

[3] Ironically, this implies that a Naïve Bayes algorithm typically disregards Bayesian probability theory, which would make the information from one feature contingent on the information from other features.

Since Naive Bayes is a probabilistic classifier, we want to calculate both the probability that the sentence is negative in sentiment and the probability that the sentence is positive. We can then compare the two probabilities and take the larger one as the final prediction. Mathematically, we want P(statement=neg | words="The days of the U.S. being ripped-off by other nations is OVER!") — the probability that the sentiment of the statement is negative given the words in the sentence are "The days of the U.S. being ripped-off by other nations is OVER!" – and the corresponding probability that the statement is positive. This involves three steps.

### Step 1: Feature engineering

Unlike with numerical datasets, the first step needed when dealing with text data is to convert words into numbers. To do this, one simple approach is to just count the frequency of words. We ignore features such as capitalisation, word order and sentence construction, treating every document as a set of the words it contains. Our features are purely the count of each of the words. Based on the training data, we collect this information in a frequency table, of which an excerpt is shown below.

**Exhibit 6: Excerpt of the matrix showing word frequencies in the training data, grouped by pre-labelled sentiment**

| | Sentiment | |
|---|---|---|
| Word | Pos | Neg |
| acceptable | 0 | 1 |
| additional | 1 | 0 |
| advantages | 0 | 1 |
| agreed | 1 | 0 |
| agricultural | 1 | 0 |
| along | 1 | 0 |
| am | 0 | 2 |
| amounts | 1 | 0 |
| and | 1 | 2 |
| are | 0 | 1 |
| as | 0 | 1 |
| autoworkers | 1 | 0 |
| be | 3 | 0 |
| … | … | … |
| Word count | 65 | 63 |

Source: BofA Merrill Lynch Global Research

**Exhibit 7: Calculation of prior and conditional probability**

| | Prior probability |
|---|---|
| P(statement=positive) | 3/6 |
| P(statement=negative) | 3/6 |

| | conditional probability | |
|---|---|---|
| Word | P(word \| positive) | P(word \| negative) |
| the | (1 + 1) / (65 + 128) | (5 + 1) / (63 + 128) |
| days | (1 + 1) / (65 + 128) | (0 + 1) / (63 + 128) |
| of | (3 + 1) / (65 + 128) | (0 + 1) / (63 + 128) |
| US | (0 + 1) / (65 + 128) | (3 + 1) / (63 + 128) |
| be | (3 + 1) / (65 + 128) | (0 + 1) / (63 + 128) |
| ripped | (0 + 1) / (65 + 128) | (1 + 1) / (63 + 128) |
| off | (0 + 1) / (65 + 128) | (1 + 1) / (63 + 128) |
| by | (0 + 1) / (65 + 128) | (1 + 1) / (63 + 128) |
| other | (0 + 1) / (65 + 128) | (1 + 1) / (63 + 128) |
| nation | (0 + 1) / (65 + 128) | (1 + 1) / (63 + 128) |
| is | (1 + 1) / (65 + 128) | (2 + 1) / (63 + 128) |
| over | (0 + 1) / (65 + 128) | (1 + 1) / (63 + 128) |

Source: BofA Merrill Lynch Global Research

### Step 2: Calculating likelihoods

This step again mostly involves counting words in our training data.

First, we calculate the prior probability of a sentence belonging to either class, based on our training data, i.e. the probability that a statement is positive, P(statement=positive), or negative, P(statement=negative). Since we have 3 positive statements and 3 negative statements in the training data, both probabilities are ½.

Next, we compute the conditional probability of a specific word appearing, given sentiment: P(word | statement=positive) and P(word | statement=negative). This involves counting how many times a word appears among positive and negative statements, compared to the total number of words in statements of a given sentiment. For example, P("US" | statement=positive ) is simply the number of times "US" appears in positive statements divided by the total number of words in these positive sentences.

However, the one problem we run into here is that not all words appear in every class – a consequence of using a fairly small training sample. Because a zero term in Bayes' formula would make the overall result zero, we need an additional step to mitigate this issue: Laplace smoothing. This is done by adding 1 to every numerator, giving them a

value that is always larger than zero. To balance out the additional count, we add the total number of possible words to the denominator, making sure the probability will never be greater than 1. For example, applying this method to the word "US" (a word that does not appear in any of our positive training sentences) would result in P("US" | statement=pos) = (0+1) / (65 + 128) = 1/193.

We calculate this probability for all words in the sentence "The days of the U.S. being ripped-off by other nations is OVER!", and summarise it in a likelihood table (Exhibit 7), These likelihoods are then used to compute the conditional probability of the whole sentence being either positive or negative, P (statement | sentiment=positive) and P (statement | sentiment=negative).

### Step 3: The naïve part

Then comes the "naïve" part: we assume every word in a sentence is independent of the other words, allowing us to focus on individual words rather than whole sentences. This assumption is very strong but highly useful, ultimately helping the model to work well with little data or data that may be mislabelled.

As probability theory tells us, the joint probability of independent random events is the product of their probabilities. Consequently, we compute the probability of a particular statement (given sentiment) as the product of the probabilities of particular words (given sentiment): P("The days of the U.S. being ripped-off by other nations is OVER!" | sentiment=positive) = P("the" | sentiment=positive) x P("days" | sentiment=positive)) x … x P("over" | sentiment=positive).

Finally, we compute the posterior probabilities, taking into account the probabilities calculated previously. According to Bayes' formula, the overall probability of sentiment being positive (or negative), given the statement is P(sentiment=pos | sentence) = P(sentence | sentiment = positive) x P(sentiment) / P(sentence). In both formulas for positive and negative sentiment, the denominator – P(sentence) – is the same. Consequently, we can ignore the denominator altogether and simply compare the numerators. Exhibit 8 summarises the calculation. The result shows that the probability of sentiment being positive (given the sentence in question) is lower than the probability of sentiment being negative. In other words, the algorithm predicts that sentiment is negative - in our view a fair judgement.

**Exhibit 8: Estimating the posterior likelihood of a sentence being "positive" or "negative"**

$$P(positive \mid sentence) \propto P(positive) \times P("the"|positive) \times P("days"|positive) \cdots \times P("is"|positive) \times P(\text{over}|positive) = 2.4 \times 10^{-26}$$

$$P(negative \mid sentence) \propto P(negative) \times P("the"|negative) \times P("days"|negative) \cdots \times P("is"|negative) \times P(\text{over}|negative) = 9.8 \times 10^{-25}$$

Source: BofA Merrill Lynch Global Research

### The Math

Bayes theorem provides a way of calculating the posterior probability of "c given x", P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Where

- $P(c|x)$ is the posterior probability of *class* (*target*) given *predictor* (or *attribute*).
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

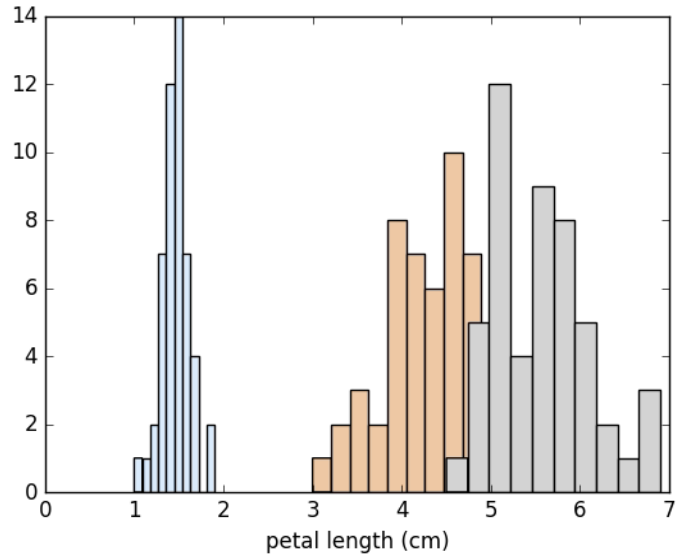
Since we already discussed the calculation steps in our sentiment analysis toy example above using unconventional data, we here demonstrate the Naïve Bayes logic using conventional data by applying the concept to the Iris dataset that we introduced in the previous section. Recall that the data set contains samples for 3 iris species with 50 samples each. For each sample, we have information on sepal and petal length and width, as well as species names: setosa, versicolor, and verginica. Let's say we are now given a new iris sample with the four measurements: sepal length = 6cm, sepal width = 2.5, petal length = 4cm and petal width = 1cm. What's the most likely species this new iris flower belongs to?

Let c denote the species and x denote the sepal and petal measurements. Then the question translates to finding P (c = species of the iris flower | x = the sepal and petal measurements ). From Bayes' formula above, we need to find out the prior probability of coming across a species, P(c), and the likelihood of observing certain measurements among a particular species P (x | c). Put differently, this means calculating P(flower being a particular species) and P(measurements | iris species).

The prior probabilities are straightforward since we have an equal number of samples for each species: P(setosa) = P(versicolor) = P(verginica) = 1/3. However, the conditional likelihood function is slightly different from our Twitter example above since the data we have are continuous. The standard approach is to assume the data are normally distributed, and compute the mean and standard deviation for each feature and each iris species. Means and standard deviations fully characterise normal distributions, making the calculation of conditional probabilities uncomplicated. Looking at the data where we plot the distribution of petal lengths for the three species, the assumption of normality does not seem unreasonable (Exhibit 9). Once we have defined the shape of the normal distribution, we can easily compute the probability mapped to each attribute measure.

Finally, we can estimate the posterior probability by plugging into Bayes' formula. As mentioned before, since all three calculation share the same denominator, we only need to compute the numerators, e.g. P (measurements | flower is species A) x P (flower is species A). The full calculations are shown in Exhibit 10. Based on the results, the species of our new sample is predicted to be a versicolor.

**Exhibit 9: Distribution of petal length by species**




Source: BofA Merrill Lynch Global Research

**Exhibit 10: Posterior estimation of iris species**

$$p(C = setosa \mid iris\ measurements\ ) \propto$$
$$p(C = setosa) \times p(sepal\ length = 6 | C = setosa)$$
$$\times p(sepal\ width = 4 | C = setosa)$$
$$\times p(petal\ length = 2.5 | C = setosa)$$
$$\times p(petal\ width = 1 | C = setosa) = 2.78 \times 10^{-21}$$

$$p(C = versicolor \mid iris\ measurements\ ) \propto$$
$$p(C = versicolor\ ) \times p(sepal\ length = 6 | C = versicolor\ )$$
$$\times p(sepal\ width = 4 | C = versicolor\ )$$
$$\times p(petal\ length = 2.5 | C = versicolor\ )$$
$$\times p(petal\ width = 1 | C = versicolor\ ) = 5.93 \times 10^{-8}$$

$$p(C = verginica \mid iris\ measurements\ ) \propto$$
$$p(C = verginica) \times p(sepal\ length = 6 | C = verginica)$$
$$\times p(sepal\ width = 4 | C = verginica)$$
$$\times p(petal\ length = 2.5 | C = verginica)$$
$$\times p(petal\ width = 1 | C = verginica) = 2.4 \times 10^{-13}$$

Source: BofA Merrill Lynch Global Research

## K-means

- Unsupervised learning → Clustering

**What is it? And what does it do?**

K-means clustering is a method that aims at partitioning a dataset into k clusters. The clustering is done in such a way, that each observation belongs to the cluster whose mean is closest to that observation. The methodology is highly versatile and can be used for cluster analysis on a wide variety of input data.

K-means falls under unsupervised learning, as the researcher does not classify data in advance (i.e. she does not "train" the algorithm). Instead, the algorithm itself decides on the clusters. Naturally, the researcher exerts influence on the process by deciding what data to provide the machine with, and how many clusters, K, the machine should form. It is then up to the machine to find the best way of clustering the data.

The main "learning" element of K-means is given by the iterative process used to find clusters. The machine starts off with an initial set of K means within the space spanned by the data. Next, each data point is assigned to the mean that is closest to it. "Closeness" is typically defined by Euclidean distance in this context, although other measures of distance can also be used. After the assignment step, new means are defined by calculating the mean (centroid) of each group. These new means form the start for the next iterative round of assignment and updating, until the process has converged.

To illustrate this iterative process, we can think of a highly stylised example where the aim is to cluster a dataset containing three observations on body height: 135cm, 160cm and 183cm. Specifying K=2 and starting off with two randomly chosen means K_1=190 and K_2=200, we first assign each observation in the dataset to the nearest mean. In our example all three observations are assigned to K_1. Next, the new mean for each cluster is calculated. The mean of the first cluster, containing all three observations, is (135+160+183)/3=159.3, giving K_1 an updated value of 159.3. K_2 stays unchanged as no new information has been provided for the second cluster. At this point, the iterative process begins . Each observation is assigned to the nearest mean, but given the new values for K_1 and K_2, assignments are different compared the first round. In fact, the third data point (183) is now closest to K_2. The other two observations are still closest to K_1. In the updating step, we find the new K_1=147.5 ((135+160)/2) and K_2=183. Repeating the process one more time, we would find that the means do not change – the algorithm has converged.

Naturally, the above example is trivial. But the allure of the algorithm becomes apparent when we apply it to multidimensional dataset, where human inspection alone would have difficulties in detecting similarities. Note that the clusters may or may not be open to interpretation in a sensible way, but they provide a basis upon which a human researcher can build with experience and creativity.

In the following example, we illustrate the technique in the context of currency research.

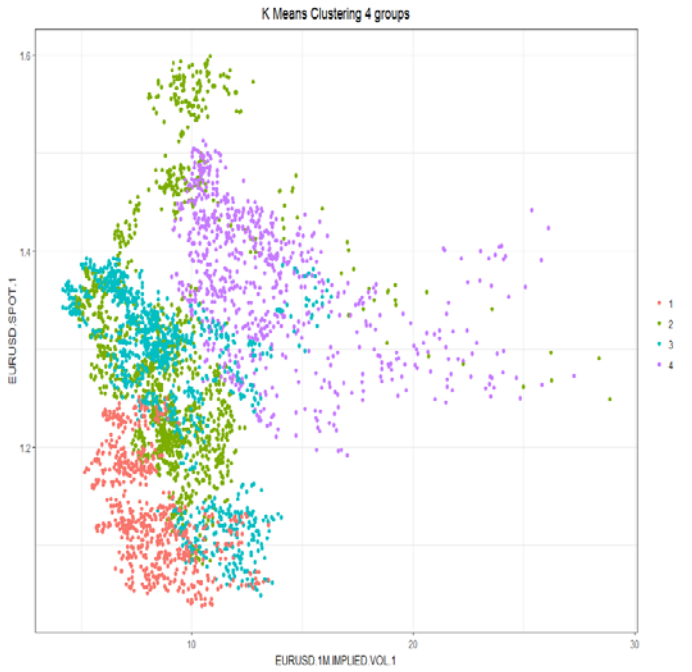
**Example: determine economic regime via K-means clustering**

To find out more about the behaviour of the EUR/USD exchange rate in the context of the Italian bond selloff in May 2018, we study economic regimes based on a variety of macroeconomic data (this example has originally been published in Liquid Insight: EUR/USD Le Rêve 31 May 2018). We apply the K-means classification algorithm to European and US macroeconomic data that we deem relevant for FX, aiming to offer a new perspective on the outlook for the exchange rate. The selection of data series is listed at the end of this section.

By running a K-means algorithm on the whole dataset, machine learning provides us with a number of clusters that share certain characteristics. More specifically, we run 2 variations of the algorithm, one with K=4 and another with K=10.

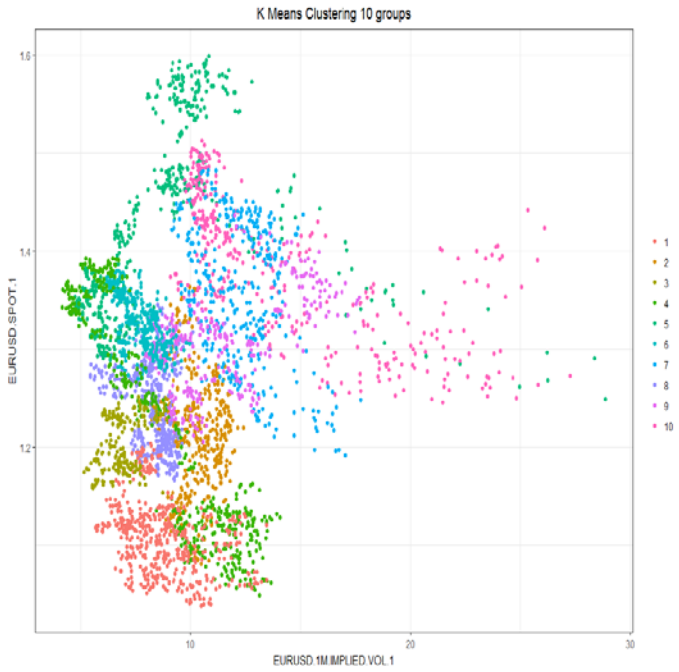
Exhibit 11and Exhibit 12 show the resulting clusters, mapped against EUR/USD spot and volatility. Each colour represents one particular cluster as defined by the K-means algorithm. As can be seen in both specifications, episodes that are similar in terms of spot and vol also tend to be grouped in the same cluster, implying that even in a more general sense, they share the same macroeconomic environment (note that if the macroeconomic environment was not similar, the algorithm would not have clustered them). That said, there is no perfect separation between clusters in the spot/vol space, with the clusters having significant overlap in certain regions.

**Exhibit 11: K-Means clustering with 4 groups**




Source: BofA Merrill Lynch Global Research, Bloomberg

**Exhibit 12: K-Means clustering with 10 groups**




Source: BofA Merrill Lynch Global Research, Bloomberg

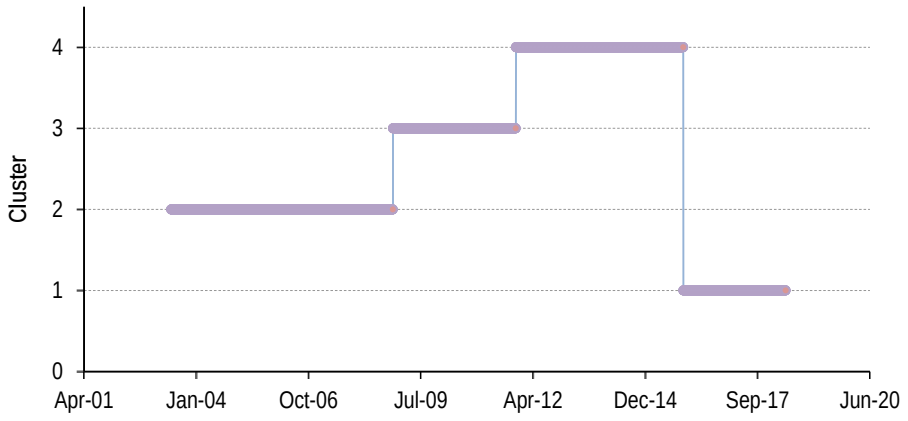
## Implications

Interestingly, we find that regimes based on unsupervised K-means clustering tend to be highly persistent over time. This is particularly evident in the 4-means case (Chart 3). A first look at the clusters over time reveals the regime shift that took place at the start of the Global Financial Crisis in 2007, as well as the shift around the height of the European crisis in 2012. The latest points in our sample have fallen into a different cluster yet again.

Looking at the spot-vol charts, we see that the latest cluster (1 - orange) has been characterised by a low EUR/USD exchange rate and low volatility. Finding that the current state is fundamentally different compared to previous episodes of sell-offs, such as the Eurozone crisis of 2012, is in itself an interesting observation. It could be interpreted as an indication that the elevated volatility (at the time of analysis in May 2018) is not justified by fundamentals (which are used to create the clusters), but rather market sentiment and behavioural factors. This in turn could hint at a possible correction.

Note, however, that K-means clustering itself does not provide predictions. The algorithm itself simply tells us whether certain time-periods are similar to each other or not given the information it is fed. In the face of a highly complex environment, this can in itself be a powerful and enlightening message.

**Chart 3: K-Means clusters over time: 4 group case**




Source: BofA Merrill Lynch Global Research, Bloomberg

### The Math

K-means clustering is a method to categorize n observations, x, into k ≤ n sets S = {S1, S2, ..., Sk} to minimize distance based on pre-set metrics. Mathematically, given a set of observations (x1, x2, ..., xn), the algorithm solve the below optimization problem:

$$\arg_S \min \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 = \arg_S \min \sum_{i=1}^{k} |S_i| \mathrm{Var}\, S_i$$

Where $\mu_i$ is the mean of observations in set $S_i$.

Given a predetermined number of clusters k, an iterative two-step algorithm is used to find the clusters. It consists of an assignment step and an update step, which are repeated until the algorithm converges.

Starting with an initial set of k centroids (a form of midpoint), each observation is assigned to the cluster with the closest centroid. Proximity to the centroid is typically measured by the Euclidean distance, but other measures can be used as well depending on the complexity of convergence.

The next step is to calculate the new mean of every cluster and use this new mean as the updated centroid.

These steps are repeated until the algorithm converges.

## Data

We employ fundamental and survey data from both the US and Eurozone. The list of series is shown in the table below. 4

**Table 5: Data Summary**

| National Account | Consumer Prices | Producer Prices | Labor Market | Economic Activity | Business Cycle Indicator | Business Conditions | Leading Indicators | Housing Market | Retail Sector | Consumer Confidence | Personal Sector | External Sector | Gov't Sector | Monetary Sector | Financial Sector |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| · Personal Consumption<br>· Real GDP<br>· Private Investment<br>· Government spending<br>· GDP Price Deflator | · CPI<br>· CPI ex-Food& Energy<br>· PCE | · PPI Final Demand<br>· PPI Final Demand ex Foods&Energy | · Unemployment Rate<br>· Initial Jobless Claims<br>· Continuing Claims<br>· NFP<br>· Job openings | · Industrial Production<br>· Capacity Utilization<br>· Factory Orders<br>· Durable Goods Orders<br>· Business Inventories<br>· Manufacturing Inventories | · Real Personal Income<br>· Industrial Production Index | · ISM Manufacturing<br>· ISM Non-Manufacturing<br>· Chicago Fed National Activity Index<br>· Markit PMI<br>· Euro-Area Business Climate Index<br>· EC Economic Sentiment Index<br>· EC Industrial Confidence Index<br>· EC Service Confidence Index<br>· Eurozone Economic Expectations | · OECD Leading Indicator<br>· Conf. Board Leading Indicator | · Housing Starts<br>· Building Permits<br>· Existing Home Sales<br>· Mortgage Applications | · Retail Sales<br>· Retail Sales ex-Autos<br>· Total Vehicle Sales | · Confidence Board Index<br>· U. of Mich Consumer Confidence Index | · Personal Income<br>· Personal Expenditure<br>· Personal savings rate<br>· Consumer Credit | · Current Account<br>· Trade Balance<br>· Exports<br>· Imports | · Gov't Budget Balance<br>· Total Public Debt Outstanding | · Monetary Base<br>· M1<br>· M2<br>· M3 | · Fed Funds Rate<br>· 3-Month T-Bill Rate<br>· 10-Year Gov't Bond Yield<br>· Equity market<br>· MOVE Index<br>· VIX<br>· GFSI<br>· German CDS SR 5Y<br>· Italy CDS SR 5Y<br>· Spain CDS SR 5Y<br>· ECB Main Refinancing Rate<br>· 3-month Euribor Rate |

Source: BofA Merrill Lynch Global Research

## Support Vector Machine

- Supervised learning → Classification

**What is it? And what does it do?**

Support vector machines (SVM) are supervised learning models used for classification and regression analysis. Given a set of training data (observations for which the categorisation is known in advance), an SVM builds a model that assigns a new observations to one of the pre-established categories.

Given a particular set of training data, the algorithm tries to place a line or plane through the data in such a way that data points belonging to one category are on one side of the plane, while data points belonging to the other category are on the other side of the plane. More specifically, this classifier is a so-called hyperplane, i.e. a subspace whose dimension is one less than that of the space it is in. In 2-dimensional space (think of a standard graph with horizontal and vertical axes), the hyperplane is simply a line. In 3-dimensional space, the hyperplane would be a 2-dimensional plane, and so on. Once this hyperplane is established, classification of new data points is given simply by the side of the hyperplane that the data point falls on. Classification is typically done in a non-probabilistic way, meaning that observations are assigned in an absolute sense and not with probabilities attached.

A stylised example: Assume we have a dataset comprising information about lorries and cars. We know the top speed of each vehicle, as well as fuel consumption. Plotting the vehicles in a chart where top speed is on one axis and fuel consumption is on the other, an SVM would attempt to draw a line (the “hyperplane”) through the data points such that all lorries are on one side and all cars are on the other. Given a new observation of speed and fuel consumption, the SVM predicts whether it is a car or a lorry depending on which side of the hyperplane the observation falls in the graph.

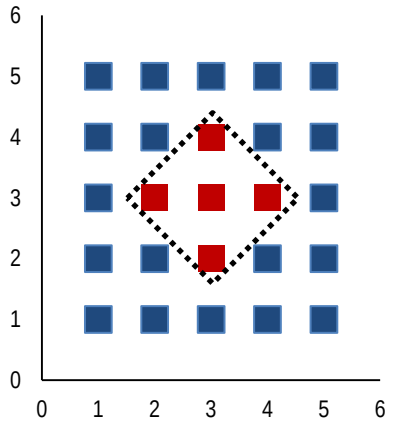
Of course there can be many different hyperplanes that separate particular groups of data, but typically an algorithm is designed to choose the one that maximises the margin between the data points closest to the hyperplane. The data samples closest to the hyperplane are called “support vectors”. These data points typically have the largest influence on what the hyperplane looks like.

What happens if no straight line – or, more generally speaking, no linear classifier – can be found that to separate the data points? In this case, the researcher has two options. Option 1 is the introduction of a so-called "soft margin" which allows for data points to lie on the "wrong side" of the line. This solution is particularly useful when a dataset contains outliers or measurement errors, as the soft margin is less affected by them than a hard margin. In practice, the researcher specifies the "softness" of the margin by choosing the size of the penalty incurred by having data points on the wrong side of the hyperplane.

Option 2 takes a different, but extremely useful approach. Using the so-called “kernel trick”, data is projected into a higher-dimensional space (i.e. dimensions are added to the data) until a linear separation is possible. Chart 4 shows a typical example for a dataset where the two classes cannot be separated by a linear classifier (i.e. by a straight line). Nonetheless it seems that the two classes fall into distinct areas which we would like to separate. Using the kernel trick, we can map the data onto a higher-dimensional space. Chart 5 shows the same data, but with an added third dimension. In three dimensions it is easy to see that a horizontal plane around height 5 would perfectly separate the red from the blue data points. This would be our "hyperplane". Note that, as mentioned above, the space spanned by the hyperplane is one dimension below the space spanned by the data. In our simplified example, the hyperplane is two-dimensional, to separate the three-dimensional data. In fact, we can map the data back into two-dimensional space, plotting the added third dimension on the vertical axis (Chart 6). In this space, the hyperplane is once again a straight line. Along the original

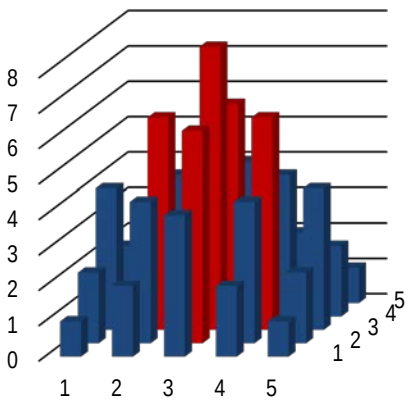
dimensions in Chart 4, on the other hand, the hyperplane would take the shape of a square around the red data points – even though it is a linear classifier. Note that by mapping into infinite dimensions, any dataset can be perfectly separated with a linear classifier, no matter how convoluted the clusters may seem. However, by choosing to perfectly separate clusters in higher dimensions, a researcher runs the risk of overfitting. Similar to the trade-off when fitting curves in classical statistics, a perfect in-sample fit can imply lack of generality and poor out-of-sample properties.

**Chart 4: Data not linearly separable...**



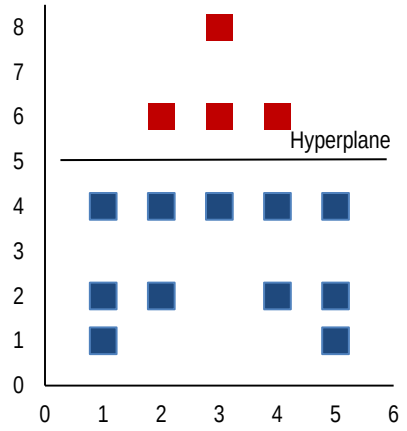

Source: BofA Merrill Lynch Global Research

**Chart 5: ...can be mapped to higher dimension...**




Source: BofA Merrill Lynch Global Research

**Chart 6: ...to find linear hyperplane**



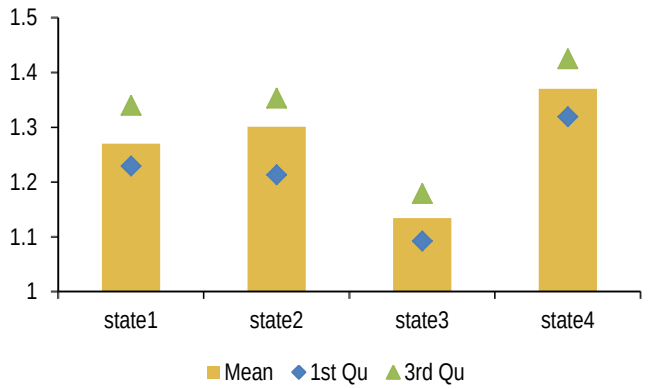

Source: BofA Merrill Lynch Global Research

In practice, this leaves the researcher with two dimensions along which to make an informed judgement: The "softness" of the margin and the dimension for the kernel trick. These choices exemplify that machine learning still requires experience and human judgement. As we pointed out in the introduction, an element of art remains within the science.

### Example: The state of the economy vs FX spot and volatility

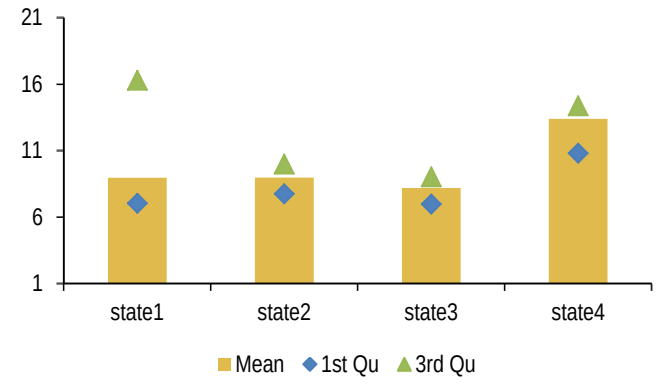
This example has originally been published in Liquid Insight: EUR/USD Le Rêve 31 May 2018. Following the previous example on K-means clustering, we now want to use a support vector machine to tell us which state of the economy we are in, given observations about the exchange rate. More specifically, we use the clusters from the K-means exercise to train the SVM (as a supervised method, SVM cannot cluster data itself and requires a pre-classified dataset). The SVM is then employed to find hyperplanes (i.e. draw borders between the clusters) in such a way, that new data points can be classified simply by checking which side of the hyperplane they lie on.

**Chart 7: EUR/USD spot distribution by state**




Source: BofA Merrill Lynch Global Research, Bloomberg

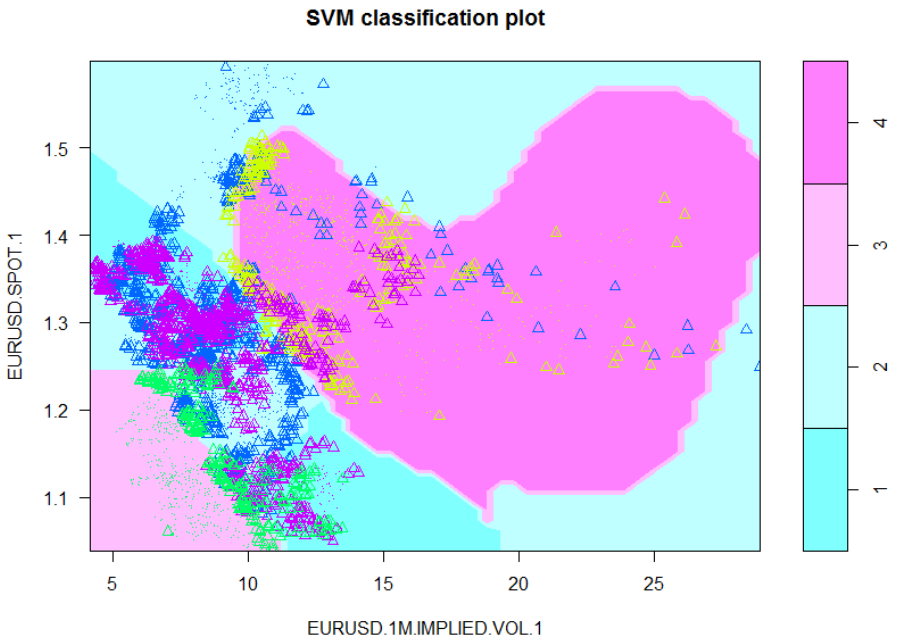
**Chart 8: EUR/USD vol distribution by state (1M Implied volatility)**




Source: BofA Merrill Lynch Global Research, Bloomberg

The exchange rate tends to behave differently across different states of the economy as can be seen in Chart 7 and Chart 8. Notably, state 3 is characterised by a low average EUR/USD spot price, while the opposite is the case for state 4, which also exhibits high average volatility. Support vector machines can help us draw this link by giving us a way of classifying the state of the economy based on observations of the exchange rate and volatility.

**Exhibit 13: Classifying the state of the economy with a support vector machine**




Source: BofA Merrill Lynch Global Research

In Exhibit 13, we plot the classification result of the SVM in spot/vol space. Triangles and dots mark historical observations, with marker colours indicating classification according to the k-means exercise above. This corresponds to the data in Exhibit 11 of the k-means example. Within a particular cluster, a triangle represents a “support vector”, i.e. an observation that is close to (and thereby defines the position of) the hyperplane. Other data points which do not directly influence the hyperplane are marked with small dots. The background colours are the result of the SVM algorithm itself. They outline the areas separated by hyperplanes and thereby provide the basis for classification of new observations.

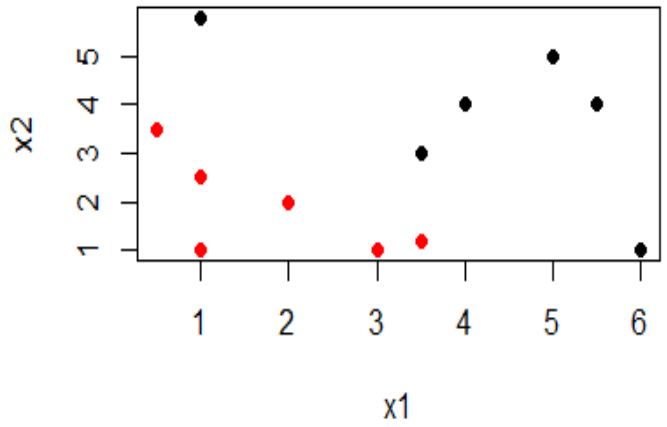
With EUR/USD at 1.20 and 1M implied volatility at 6.4 at the time of analysis in April 2018, the SVM classified the economy as being in state 3. This observation is close to the hyperplane separating state 3 from state 2.

### The Math

What we are looking for is a hyperplane that separates the two classes of data such that the margin of the hyperplane is as large as possible. To visualize the process, we generate 12 observations and assign them to two classes as shown in Exhibit 14. The aim is to find the linear boundary that maximizes the total distance between the line and the closest points in each class (the "support vectors"). Exhibit 15 shows the linear classifier found by the support vector machine for our sample observations. In an environment where data cannot be separated by a linear classifier, a "soft margin" can be introduced, which allows for misclassification. Typically, functions in the optimisation

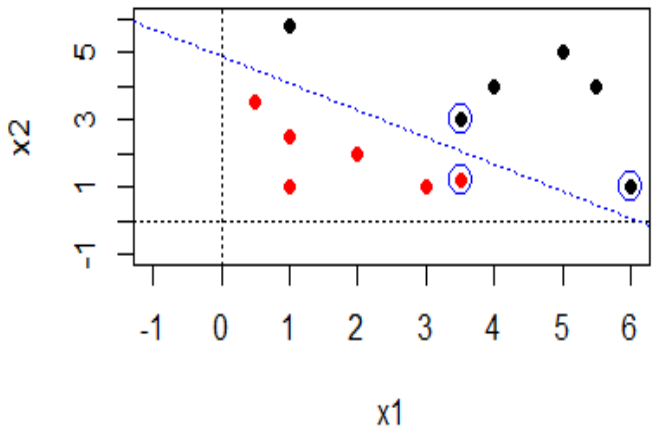
process then feature a penalisation term with a coefficient that trades off the width of the margin against the amount of misclassification done by the hyperplane. Going beyond linear classification, the kernel trick of mapping data into higher dimensions allows for a non-linear classification of the data. The following brief example illustrates the difference.

**Exhibit 14: SVM example**




Source: BofA Merrill Lynch Global Research

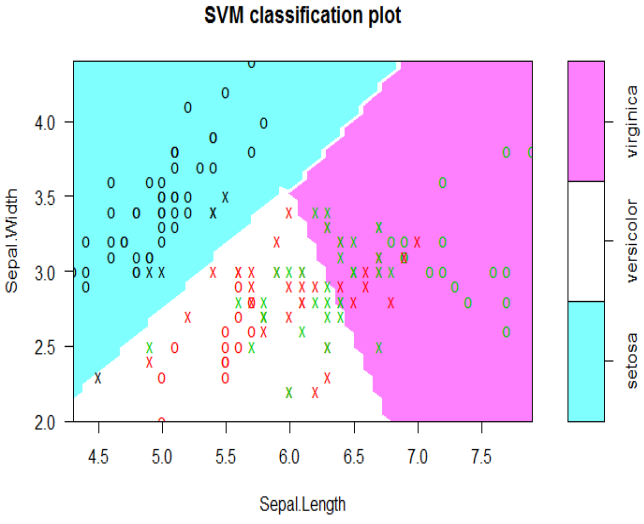
**Exhibit 15: SVM example with hyperplane**




Source: BofA Merrill Lynch Global Research

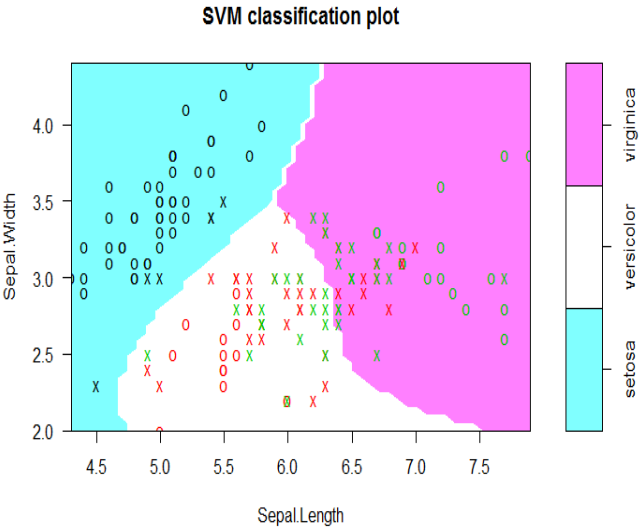
Using the Iris data set introduced in earlier sections, Exhibits 16 and 17 show data separated using an SVM algorithm. For Exhibit 16, we employed a linear classifier. The data points marked by crosses are "support vectors" - they directly influence the hyperplane. Data points marked with circles are other observations that do not directly affect the hyperplane. Because species cannot be separated perfectly, a soft margin is required. For Exhibit 17, on the other hand, we employed a non-linear classifier through the kernel trick. Examples of kernels include polynomial, Gaussian radial basis function or hyperbolic tangent.

**Exhibit 16: SVM classification plot with linear kernel**




Source: BofA Merrill Lynch Global Research

**Exhibit 17: SVM classification plot with polynomial kernel**




Source: BofA Merrill Lynch Global Research

The actual optimisation process for finding the SVM classifier is typically done through a so-called sub-gradient descent algorithm or a coordinate descent algorithm.

## Kalman filter

- Supervised learning → Regression

**What is it? And what does it do?**

The Kalman filter is an algorithm for optimally combining different imperfect measurements of unknown variables, such that the resulting estimate has the highest possible precision given the available information. It does so by producing a joint probability distribution over the variables in question. The Kalman filter was introduced by and named after Rudolf E. Kalman in the 1960s in the context of signal processing problems. It has since been heavily researched and its applications range from trajectory tracking for the Apollo missions to dealing with noisy data in econometrics.

The Kalman filter operates recursively on streams of noisy input data to produce estimates of a system's underlying state. Among the benefits of the algorithm are the properties that it does not require a long history of observations, and reacts to new data faster than a traditional rolling regression.

The algorithm is best explained with a short, intuitive example. Let's imagine we are operating a remote-controlled airplane with a set of simple controls. At any point in time, what we are interested in is the plane's position and speed – the so-called "state variables" in our system. The plane has a GPS system, which is accurate to 50 metres, and we can control the rotations of the plane's propeller. But, because there might be wind, clouds and other influences on the plane, the propeller's rotations do not perfectly relate to the speed of the plane.

Now, if we want to know where exactly our plane is at a specific moment in time, we might do one of two things. First, we could simply use the GPS, accepting that the reading is inaccurate. Second, we could take whatever position we had for the plane in the past and predict where it is now given our choice of propeller rotations – accepting that this will be inaccurate given that the speed is only roughly related to the rotations. The Kalman filter offers a third option: to combine both estimates in the best possible way. In essence, the filter combines the probability distributions for the planes position which we obtained from the GPS and from the speed-based prediction into a new, more precise, joint probability distribution. Intuitively, the Kalman filter allows us to predict a distribution (the "prediction step") and then use additional information to tell us whether some points on the distribution might be more likely than others (the "updating step"). In fact, the Kalman filter doesn't stop there. Having updated the original estimates for the states, it goes back to the start and creates a new prediction on the basis of these new estimates. Again, the predictions are updated, leading to yet more precise estimates, and so on.

This type of iterative process has turned out to be extremely versatile and adaptive, lending itself to a large variety of applications. Laubach and Williams (2003) famously use a Kalman filter to jointly estimate the neutral real interest rate ("r-star"), potential output and the trend growth of the economy. In another application, Bahmani-Oskooee and Brown (2004) use a Kalman filter to estimate central bank demand for international reserves. Modelling reserve demand as a function of the exchange rate and the oil price, and noting that the relationship is likely to be time-varying, the authors use the Kalman filter to estimate the (unobservable) coefficients in a time-dependent way. In a closely related set-up, we demonstrate the Kalman filter below to estimate dynamic hedging ratios.

**Example: improving carry trades with dynamic hedging ratio**

A common strategy to find hedge ratios is via linear regression. However, the key drawback of linear regression is that two assets may move together in general, but that relationship could deviate, and sometimes for a long time. Kalman filter takes linear regression one step further. It tries to find the hedge ratio dynamically by adapting to the price action of the two assets over time. In our airplane location example above, the

Kalman filter can be used to estimate the location of the plane. In this case, we use Kalman filter to estimate the hedge ratio. This following example has originally been published as Liquid Insight: The quest for carry 12 July 2018.

Carry trades typically work well in a low volatility environment. Not long ago, as volatility hit record lows in 2017, carry trades became go-to positions in global markets. However, after a few equity market corrections, a European peripheral scare, and a looming trade war, even bonds that in January convinced with outstanding risk-adjusted yields did not fare well.

Stellar economic data combined with heightened risks in the second half of the year left the market with little outright conviction. In our view, in the current environment, it is possible to still capture carry while limiting market exposure. Buying a high-yielding bond, and dynamically hedging with a low yielding bond can help offset mark-to-market moves. We illustrate the idea with an example where we apply the Kalman filter to dynamically estimate hedge ratios.

**Table 6: The most attractive carry trade at the beginning of the year fared the worst**

| | US | Germany | France | Italy | Spain | UK | Canada | Japan | Australia |
|---|---|---|---|---|---|---|---|---|---|
| **FX hedged yield as of 7/11 (%)** | | | | | | | | | |
| 2y | 2.58 | 2.51 | 2.68 | 3.93 | 2.88 | 2.52 | 2.54 | 2.99 | 2.44 |
| 5y | 2.75 | 2.56 | 2.86 | 4.66 | 3.20 | 2.59 | 2.53 | 3.15 | 2.47 |
| 10y | 2.85 | 2.65 | 2.93 | 4.97 | 3.58 | 2.59 | 2.47 | 3.22 | 2.47 |
| 30y | 2.95 | 2.49 | 3.00 | 4.95 | 3.94 | 2.94 | 2.51 | 3.35 | 2.82 |
| **Vol adj net yield** | | | | | | | | | |
| 2y | 83.7 | 52.1 | 64.7 | 11.7 | 80.0 | 68.8 | 69.5 | 103.7 | 74.5 |
| 5y | 71.5 | 50.5 | 60.9 | 18.2 | 64.6 | 65.9 | 67.7 | 89.5 | 76.8 |
| 10y | 76.3 | 53.4 | 70.3 | 41.4 | 74.8 | 60.2 | 64.8 | 91.1 | 67.9 |
| 30y | 87.1 | 50.4 | 74.1 | 72.1 | 76.3 | 81.8 | 68.1 | 97.6 | 78.2 |
| | **US** | **Germany** | **France** | **Italy** | **Spain** | **UK** | **Canada** | **Japan** | **Australia** |
| **FX hedged yield as of 1/4 (%)** | | | | | | | | | |
| 2y | 1.95 | 2.05 | 2.20 | 2.32 | 2.25 | 1.99 | 2.01 | 2.46 | 2.04 |
| 5y | 2.27 | 2.23 | 2.44 | 3.11 | 2.78 | 2.16 | 2.06 | 2.70 | 2.12 |
| 10y | 2.45 | 2.44 | 2.80 | 4.02 | 3.54 | 2.49 | 2.13 | 2.86 | 2.17 |
| 30y | 2.79 | 2.61 | 3.10 | 4.59 | 4.18 | 3.02 | 2.48 | 3.15 | 2.61 |
| **Vol adj net yield** | | | | | | | | | |
| 2y | 90.09 | 78.67 | 88.22 | 101.10 | 98.98 | 66.11 | 84.79 | 104.04 | 70.20 |
| 5y | 74.68 | 67.92 | 73.29 | 87.20 | 69.80 | 64.17 | 66.92 | 86.15 | 65.54 |
| 10y | 73.19 | 64.54 | 67.20 | 94.33 | 85.32 | 63.48 | 58.39 | 88.03 | 63.62 |
| 30y | 74.15 | 63.00 | 75.52 | 104.60 | 98.75 | 81.29 | 60.28 | 91.08 | 60.67 |

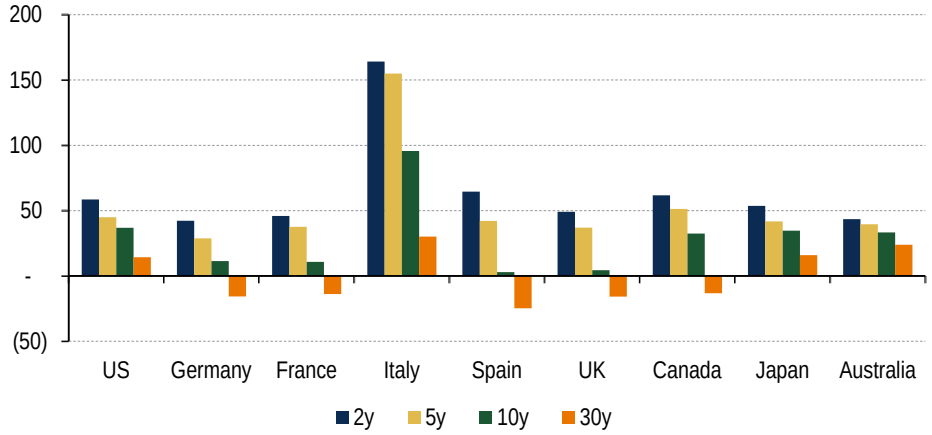
Note: We calculate synthetic yields in USD terms via asset swap across key G10 markets and tenors, to highlight the relative attractiveness across G10 rates. Vol adjusted yields are calculated as the fraction of the FX hedged yields divided by the 3m rolling standard deviation of daily changes. Green cells represent top three attractive assets, and yellow cells represent the least attractive three. For more details of the calculation please contact the author. Source: BofA Merrill Lynch Global Research, Bloomberg

## Keep calm and carry on

Carry trades falling out of favor is one dimension in which 2018 is different to 2017. With the G10 duration sell-off in 1Q and the European peripheral spreads blowout in 2Q, most carry positions were in limbo as yields moved markedly higher YTD (Chart 9). Both from an absolute yield or risk-adjusted perspective, Italian government bonds were the go-to carry trades at the beginning of the year, but now rank the lowest in vol adjusted terms given the sharp move in peripheral spreads (Table 6)

The key risks with fixed income carry trades are directionality and volatility. In a rising rate environment, the income accumulated from being long bonds can be cannibalized quickly by a sell-off in rates or a move in FX if currency risk is unhedged. Market events this year suggest that even after taking into account realized volatility in the calculation, market dynamics can quickly shift.

**Chart 9: YTD, FX hedged yields in key G10 markets moved up markedly, putting carry trades at risk**




Source: BofA Merrill Lynch Global Research, Bloomberg

### Market neutral carry trades

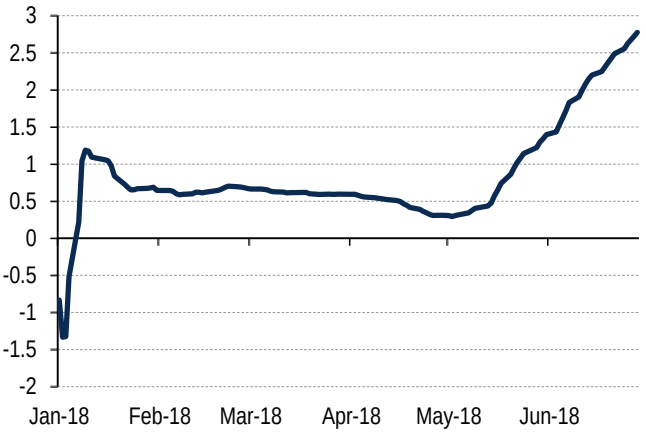
Given the low level of conviction in duration and potential risks ahead, our role is to recommend ways for carry-seeking investors to hedge their positions to limit market exposure. Common practices of such trades tend to involve having diversified or long/short positions such that market fluctuations are cancelled out. In rates, one way to do this is to go long high-yielding bonds, and hedge with matched maturity low yielding bonds in an FX hedged way.

The key issue is the choice of hedge ratio. As two bonds do not always move in the same direction and the magnitude of the moves change over time in both legs, we like hedging the long position with dynamic hedge ratios that evolve with market conditions. In the present example, we apply the Kalman filter algorithm to find optimal hedge ratios. In this context, the Kalman filter functions similarly to a linear regression, but responds more quickly to new data. Simplistically speaking, the Kalman filter recursively estimates and updates the variable (in this case, the hedge ratio) every time new information comes in. Much like a regression, the Kalman filter is applied here to find the beta between two assets so that the cumulative yield changes on both sides of the trade can be largely offset.

To illustrate the idea, we pick two bonds from the perspective of a USD-based investor. We go long the bond that has been consistently more attractive in yield terms, and hedge with a matched duration bond that has been consistently the least attractive. We note that, as shown in the example below, not all hedging positions are to short the bond, but we pick the lowest yielding position so that the negative carry would be the lowest when we do go short. From Table 6, after adjusting for FX hedge, 30y BTPs have consistently been the most attractive bond in terms of yield pickup, and in the same maturity bucket, 30y Canadian bonds have been consistently the least attractive.

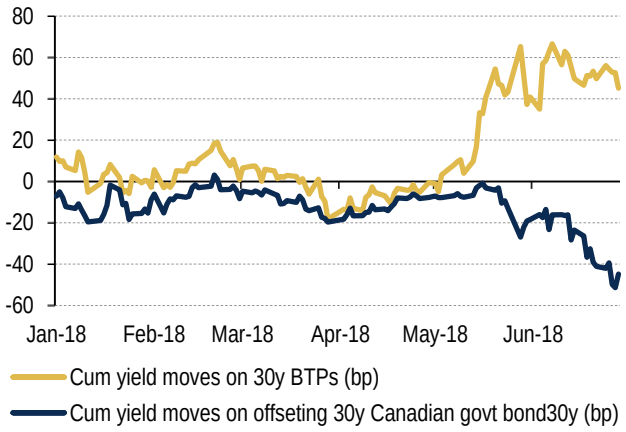
Assuming we go long 30y BTPs at the beginning of this year, and hedge using 30y Canadian bonds, applying the Kalman filter gives us a time series of what would have been the optimal hedge ratio on the 30y Canadian bond as shown in Chart 10Chart 10. The vertical scale shows the amount of Canadian bonds that need to be held for each BTP. An increasing number indicates the need to increase exposure to CAD bonds. As the peripheral risks started to escalate in late May, the amount of the 30y Canadian bond quickly increases from below 50% to almost three times the notional of 30y BTPs.

**Chart 10: Hedge ratio on 30y Canada government bond over time**




Source: BofA Merrill Lynch Global Research

**Chart 11: By dynamically adjusting hedge ratio, the hedge to a long 30y BTP position mitigates exposure to market direction.**




Source: BofA Merrill Lynch Global Research

From a capital gain/loss perspective, while not a perfect match, dynamically changing the size of the hedging leg more or less removes exposure to the market environment, as illustrated in Chart 11. From a carry perspective, this trade makes over 3% in two quarters, over 2% from the long position in 30y BTPs, and about another 1% in the 30y Canada government bond, since the hedge was in long positions most of the time. In comparison, an outright long position in 30y BTPs would suffer a loss given the almost 9% price correction.

Interestingly, our example shows that the hedge to a long position in a bond even within G10 markets is not always to short another bond. Intuitively, bonds with higher yields could have an inherent risk premium built in, and vice versa for a low yielding bond. When market dynamics change and volatility picks up, the two positions are likely to act like a risk asset and haven asset, respectively.

### The Math

The Kalman filter involves two stages, an estimation stage (first equation below) and an update stage (second equation below). The estimation stage describes how the state variables (hedge ratios in our example) evolve from one period to the next, while the update stage takes account of new information (performance of the two assets) to update the estimation. Algebraically, this can be expressed as:

$$x_{t+1} = \beta_t x_t + w_t$$

$$z_t = H_t x_t + e_t$$

Where:

- $x_t$ is the current hidden state (in our case, the hedge ratio),
- $\beta_t$ is the transition matrix characterising the evolution of the state over time (in our case, this is the identity matrix assuming the hedge ratio follows a random walk)
- $z_t$ and $H_t$ are observed variables (eg, change in yields of bond 1and 2)
- $w_t$ and $e_t$ are normally distributed white noise error terms.

## Principal Component Analysis

- Unsupervised learning → Factor modelling/dimensionality reduction

**What it is? And what does it do?**

Principal component analysis is a statistical procedure that transforms a number of observations into a set of values, called principal components, that are linearly uncorrelated. This is achieved by applying an orthogonal transformation, defined in such a way that the principal components explain the largest degree of variation possible from the original observations, in descending order.

Intuitively, what PCA does is the following: Given a set of observations, for example time series of asset prices, PCA finds a set of series that explain as much of the original dataset as possible, while at the same time not duplicating any of the information. If our original dataset only consists of one series, than the principal component is exactly that series, as it is the series that best explains our original observation. If our original dataset includes multiple series, then the first principal component is the series that explains as much as possible of the variation within the whole original dataset, while the second component explains as much of the *remaining* variation as possible, taking into account what is already explained by the first factor. The third factor explains as much as possible of the variation that is left after this, and so on. Importantly, the principal components are by construction uncorrelated, meaning they do not duplicate any information.

Even though the interpretation of the principal components (sometimes also called "factors") is not always straightforward, the procedure can be extremely helpful in distilling underlying information from within a dataset. Given high degrees of correlation within macroeconomic and financial datasets, a small number of extracted series can often explain a surprisingly large amount of the information within a large dataset.[4] Analysing and working with the small number of factors often offers more direct insight and technically easier, than working with a larger dataset. In fact, for some applications such as vector autoregressive models (VARs), working with anything but a small number of series is technically no feasible. Being able to nonetheless capture the information from a larger dataset can remove omitted variable bias.

The approach falls under "unsupervised learning" because the researcher does not classify the dataset in any way before running the algorithm.

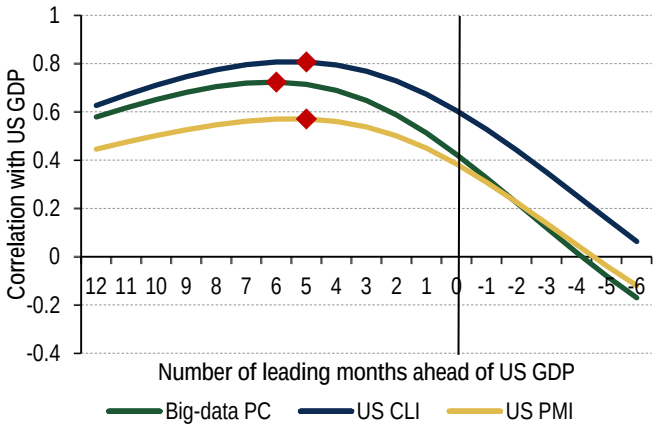
**Example: leading indicator and asset allocation**

This example is an excerpt of a report originally published in Global Rates & Currencies 2016 Year Ahead: The "Great Divorce" 23 November 2015, where we use PCA as leading indicators and examine investment implications in the rates and FX market. To get an early read on the economy, investors often turn to leading indicators such as the Purchasing Managers Index (PMI) or the OECD Composite Leading Indicator (CLI). While both are valid leading indicators, they only take into account a small number of macro variables. To make better use of the large amounts of US data available to us, we constructed a new leading indicator using PCA, basing it on the first principal component (PC) of a large panel of 135 economic variables (similar to Stock and Watson (2002) and McCracken and Ng (2015)). Comparing our indicator to the two leading indicators mentioned above, we find that the former outperforms the others on a number of dimensions. In particular, we evaluate the leading indicators according to the following criteria:

[4] See, for example, Bernanke, Boivin and Eliasz (2005), who use a Factor Augmented VAR to measure the effect of monetary policy. More recently, Corsetti, Duarte and Mann (2018) have shown that as few as 5 principal components can explain more than 80% of the variation within a dataset containing roughly 200 economic and financial indicators from the eurozone.

- **Leading period** ahead of GDP: Cross-correlations between leading indicators and GDP suggest that all three are valid leading indicators, as correlations peak when indicators lead. Among the three, our big-data PC leads one month ahead of US CLI and US PMI (Chart 12).
- **Forecasting power** for future GDP: When we include these indicators in a regression to predict GDP growth in the next quarter, all of them are significant and boost the R-squared. That said, the model with our big-data PC has the highest R-squared of 50% (Table 7).
- **Hit ratio** for predicting turning points and asset allocation: Investors often look at peaks and troughs in leading indicators to predict turning points in the economy. While all three have more than a 60% success ratio, US CLI ranks first, with our big-data PC runner-up (Table 8). For cross-asset allocation, our big-data PC has the highest average hit ratio (Table 9).

**Chart 12: Big-data PC shows longer lead on US GDP than CLI and PMI**




Source: BofA Merrill Lynch Global Research, FRED

**Table 7: Big-data PC has the highest forecasting power for US GDP**

| | R-square |
|---|---|
| Big-data PC | 50% |
| US CLI | 39% |
| US PMI | 23% |
| GDP lag only | 13% |

We regress US GDP on a leading indicator from previous quarter and four GDP lags, 1960-2015.
Source: BofA Merrill Lynch Global Research

The big advantage of our leading indicator lies in the fact that it summarises information from a data set many times the size of what the other indicators take into account. Moreover, the procedure avoids any judgement call from the side of the researcher regarding what is important and what is not. Principal component analysis simply looks at the totality of information (variation) in the provided data and tries to explain as much of it as possible. This is particularly relevant if we believe that many of the less prominent, but very reactive series in our dataset hold information about the development of GDP. By summarising this information in a parsimonious way, PCA helps us to gain an insight on GDP well ahead of its publication.[5]

[5] While our example is kept as simple as possible for illustrative purposes, highly sophisticated models for "now-casting" the economy are typically also a variation of a factor model.

**Table 8: Peak/trough in GDP follows indicator peak/trough within 1 year**

| | | Success split as | |
|---|---|---|---|
| | Correct% in 1yr | 0-6m | 6-12m |
| PC for US | 65% | 68% | 32% |
| CLI for US | 67% | 83% | 17% |
| PMI for US | 61% | 73% | 27% |
| CLI for EU | 71% | 50% | 50% |
| CLI for Japan | 62% | 56% | 44% |
| CLI for G10 | 66% | 60% | 40% |

Note: We define peak/trough signals to be correct when it precedes GDP peak/trough within the next 12 months, and compute the hit ratio of the CLIs for different countries/areas.
Back-testing is hypothetical in nature and reflects application of the model prior to its Inception Date as if the model had been in existence at that time. It is not intended to be indicative of actual or future performance. Source: BofA Merrill Lynch Global Research

**Table 9: Big-data PC works consistently well for asset allocation**

| | 10y UST | Oil | SPX | AUDUSD | NZDUSD |
|---|---|---|---|---|---|
| US PC | 66% | 52% | 58% | 56% | 63% |
| US CLI | 53% | 48% | 53% | 53% | 58% |
| US PMI | 55% | 52% | 49% | 63% | 60% |
| OECD total CLI | 60% | 64% | 54% | 54% | 48% |

Note: We looked at the probability that the asset (US10y yield, AUD/USD, NZD/USD, Brent oil, S&P 500) goes up over 3m after positive signal and goes down over 3m after negative signal. We report the average hit ratio and highlight the indicator with the highest value. Data covers 1962-2015. Similar results since 1990 and for 6m holding period.
Back-testing is hypothetical in nature and reflects application of the model prior to its Inception Date as if the model had been in existence at that time. It is not intended to be indicative of actual or future performance. Source: BofA Merrill Lynch Global Research

### Alternative example

For another example, we refer the reader to a report analysing FX vol using principal component analysis: Liquid Insight: Understanding FX vol using PCA 18 April 2016.

### The Math

PCA analysis is done by performing an eigendecomposition of the matrix containing the data. The key steps are shown in Exhibit 20.

**Exhibit 18: Key steps in performing PCA analysis**

Suppose we have an N x P matrix X, where N is the number of observations and P is the number of variables (or features)

Step 1 Center the data by subtracting the mean

Step 2 Calculate the P x P covariance matrix $C = \frac{1}{N-1} X^T X$

Step 3 Calculate the eigenvectors of the covariance matrix

Step 4 Select m eigenvectors that correspond to the largest m eigenvalues to be the new, transformed matrix

Source: BofA Merrill Lynch Global Research

# Conclusion

Having gone through some of the most widely used machine learning concepts and approaches in this primer, we hope the reader has gotten not only a taste of what machine learning is, but also how it can affect research on rates and FX.

By nature, the primer is meant to be a first introduction to the subject and only covers a small fraction of the topic. While we leave it to the reader to explore further, we would like to conclude our introduction with a few general considerations and a word of caution based on our own experience.

Machine learning in finance is most commonly applied in a micro setting (e.g. predicting prices based on high frequency data such as trading volumes, order books etc.). The lack of (or difficulty to observe) higher frequency data in a macro setting might suggest that these techniques do not have much to offer. On the contrary, we find that machine learning tools can help to make better sense of the limited data that is available and increase the timeliness of analysis. Our example on tracking economic activity using PCA is a case in point. As ever, though, it is important to ask the right question ex ante, as opposed to mindlessly collecting data. We would always prioritise this over the adoption of fancy techniques.

Modelling with machine learning techniques is subject to many of the issues that are encountered in traditional statistical and econometric models. The temptation to include as much data as possible, for example, makes machine learning models highly susceptible to overfitting. To contain the scope of the primer, we spent little time on the process of training and testing the models we apply. In full-scale application, however, testing and validating models would be crucial to avoid issues like overfitting. A good separation of data into a training sample and test sample is equally important to ensure the validity of the model.

As we like to remind ourselves: garbage in, garbage out. The quality of data is as important in a machine learning context as it is in any other statistical applications. As readers familiar with modelling might have experienced themselves, we find that the actual "modelling" part of the work is relatively straightforward in most cases. The data processing stage, however, usually requires hard work. Dealing with missing data, checking the accuracy of the inputs, or pre-processing the data to satisfy the assumptions of the model (e.g. stationarity in time series analysis) are just a few that a quantitative researcher is frequently faced with. This is also where domain expertise and familiarity with the subject matter become invaluable.

In this environment, what role is left to play for macro strategists? Given the complexities (and irrationalities) of macroeconomic developments, we believe that, even with the advances of machine learning, we are still far away from having a model that can replace human judgement. At the same time, the complexities imply that making use of the best available techniques to analyse information is as important as ever. We believe machine learning has the potential to significantly enhance our understanding of the world and become a standard tool used by macro researchers on a daily basis. As such, it should be embraced for what it is. It is then up to the macro strategist to integrate the newly gained insights into her overall analysis in a constructive way. To do so, knowledge of machine learning techniques is essential. We hope this primer has provided a helpful step in this direction.

# Further Reading


Bishop, C. M. (2011) Pattern Recognition and Machine Learning. Springer.

Goodfellow, I., Bengio, Y., Courville, A. and Bach, F. (2017) Deep Learning. MIT Press.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017) An Introduction to Statistical Learning: with Applications in R. Springer, 7th edition.

Kelleher, J. D., Mac Namee, B. and D'Arcy, A. (2015) Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. The MIT Press, 1st edition.

Lantz, B. (2015) Machine Learning with R: Expert techniques for predictive modeling to solve all your datda analysis problems. Packt Publishing, 2nd edition

Lopez de Prado, M. (2018) Advances in Financial Machine Learning. Wiley, 1st edition.

Muller, A. C. and Guido, S. (2016) Introduction to Machine Learning with Python: A Guide for Data Scientists. O'Reilly Media, 1st edition.

Rajaraman, A. and Ullman, J. D. (2011) Mining of Massive Datasets. Cambridge University Press.

Shalev-Shwartz, S. and Ben-David, S. (2014) Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press.

Theobald, O. (2018) Machine Learning For Absolute Beginners: A Plain English Introduction. Independently published.

# Disclosures

## Important Disclosures

BofA Merrill Lynch Research Personnel (including the analyst(s) responsible for this report) receive compensation based upon, among other factors, the overall profitability of Bank of America Corporation, including profits derived from investment banking. The analyst(s) responsible for this report may also receive compensation based upon, among other factors, the overall profitability of the Bank's sales and trading businesses relating to the class of securities or financial instruments for which such analyst is responsible.
BofA Merrill Lynch fixed income analysts regularly interact with sales and trading desk personnel in connection with their research, including to ascertain pricing and liquidity in the fixed income markets.

## Other Important Disclosures

Prices are indicative and for information purposes only. Except as otherwise stated in the report, for the purpose of any recommendation in relation to: (i) an equity security, the price referenced is the publicly traded price of the security as of close of business on the day prior to the date of the report or, if the report is published during intraday trading, the price referenced is indicative of the traded price as of the date and time of the report; or (ii) a debt security (including equity preferred and CDS), prices are indicative as of the date and time of the report and are from various sources including Bank of America Merrill Lynch trading desks.
The date and time of completion of the production of any recommendation in this report shall be the date and time of dissemination of this report as recorded in the report timestamp.

This report may refer to fixed income securities that may not be offered or sold in one or more states or jurisdictions. Readers of this report are advised that any discussion, recommendation or other mention of such securities is not a solicitation or offer to transact in such securities. Investors should contact their BofA Merrill Lynch representative or Merrill Lynch Global Wealth Management financial advisor for information relating to fixed income securities.
Rule 144A securities may be offered or sold only to persons in the U.S. who are Qualified Institutional Buyers within the meaning of Rule 144A under the Securities Act of 1933, as amended.
SECURITIES DISCUSSED HEREIN MAY BE RATED BELOW INVESTMENT GRADE AND SHOULD THEREFORE ONLY BE CONSIDERED FOR INCLUSION IN ACCOUNTS QUALIFIED FOR SPECULATIVE INVESTMENT.
Recipients who are not institutional investors or market professionals should seek the advice of their independent financial advisor before considering information in this report in connection with any investment decision, or for a necessary explanation of its contents.
The securities discussed in this report may be traded over-the-counter. Retail sales and/or distribution of this report may be made only in states where these securities are exempt from registration or have been qualified for sale.
Officers of MLPF&S or one or more of its affiliates (other than research analysts) may have a financial interest in securities of the issuer(s) or in related investments.
This report, and the securities discussed herein, may not be eligible for distribution or sale in all countries or to certain categories of investors.
**BofA Merrill Lynch Global Research policies relating to conflicts of interest are described at https://go.bofa.com/coi.**
**"BofA Merrill Lynch" includes Merrill Lynch, Pierce, Fenner & Smith Incorporated ("MLPF&S") and its affiliates. Investors should contact their BofA Merrill Lynch representative or Merrill Lynch Global Wealth Management financial advisor if they have questions concerning this report. "BofA Merrill Lynch" and "Merrill Lynch" are each global brands for BofA Merrill Lynch Global Research.**
**Information relating to Non-US affiliates of BofA Merrill Lynch and Distribution of Affiliate Research Reports:**
MLPF&S distributes, or may in the future distribute, information of the following non-US affiliates in the US (short name: legal name, regulator): Merrill Lynch (South Africa): Merrill Lynch South Africa (Pty) Ltd., regulated by The Financial Service Board; MLI (UK): Merrill Lynch International, regulated by the Financial Conduct Authority (FCA) and the Prudential Regulation Authority (PRA); Merrill Lynch (Australia): Merrill Lynch Equities (Australia) Limited, regulated by the Australian Securities and Investments Commission; Merrill Lynch (Hong Kong): Merrill Lynch (Asia Pacific) Limited, regulated by the Hong Kong Securities and Futures Commission (HKSFC); Merrill Lynch (Singapore): Merrill Lynch (Singapore) Pte Ltd, regulated by the Monetary Authority of Singapore (MAS); Merrill Lynch (Canada): Merrill Lynch Canada Inc, regulated by the Investment Industry Regulatory Organization of Canada; Merrill Lynch (Mexico): Merrill Lynch Mexico, SA de CV, Casa de Bolsa, regulated by the Comisión Nacional Bancaria y de Valores; Merrill Lynch (Argentina): Merrill Lynch Argentina SA, regulated by Comisión Nacional de Valores; Merrill Lynch (Japan): Merrill Lynch Japan Securities Co., Ltd., regulated by the Financial Services Agency; Merrill Lynch (Seoul): Merrill Lynch International, LLC Seoul Branch, regulated by the Financial Supervisory Service; Merrill Lynch (Taiwan): Merrill Lynch Securities (Taiwan) Ltd., regulated by the Securities and Futures Bureau; DSP Merrill Lynch (India): DSP Merrill Lynch Limited, regulated by the Securities and Exchange Board of India; Merrill Lynch (Indonesia): PT Merrill Lynch Sekuritas Indonesia, regulated by Otoritas Jasa Keuangan (OJK); Merrill Lynch (Israel): Merrill Lynch Israel Limited, regulated by Israel Securities Authority; Merrill Lynch (Russia): OOO Merrill Lynch Securities, Moscow, regulated by the Central Bank of the Russian Federation; Merrill Lynch (DIFC): Merrill Lynch International (DIFC Branch), regulated by the Dubai Financial Services Authority (DFSA); Merrill Lynch (Spain): Merrill Lynch Capital Markets Espana, S.A.S.V., regulated by Comisión Nacional del Mercado De Valores; Merrill Lynch (Brazil): Bank of America Merrill Lynch Banco Multiplo S.A., regulated by Comissão de Valores Mobiliários; Merrill Lynch KSA Company, Merrill Lynch Kingdom of Saudi Arabia Company, regulated by the Capital Market Authority.
This information: has been approved for publication and is distributed in the United Kingdom (UK) to professional clients and eligible counterparties (as each is defined in the rules of the FCA and the PRA) by MLI (UK) and Bank of America Merrill Lynch International Limited, which are authorized by the PRA and regulated by the FCA and the PRA, and is distributed in the UK to retail clients (as defined in the rules of the FCA and the PRA) by Merrill Lynch International Bank Limited, London Branch, which is authorized by the Central Bank of Ireland and subject to limited regulation by the FCA and PRA - details about the extent of our regulation by the FCA and PRA are available from us on request; has been considered and distributed in Japan by Merrill Lynch (Japan), a registered securities dealer under the Financial Instruments and Exchange Act in Japan, or its permitted affiliates; is issued and distributed in Hong Kong by Merrill Lynch (Hong Kong) which is regulated by HKSFC is issued and distributed in Taiwan by Merrill Lynch (Taiwan); is issued and distributed in India by DSP Merrill Lynch (India); and is issued and distributed in Singapore to institutional investors and/or accredited investors (each as defined under the Financial Advisers Regulations) by Merrill Lynch International Bank Limited (Merchant Bank) (MLIBLMB) and Merrill Lynch (Singapore) (Company Registration Nos F 06872E and 198602883D respectively). MLIBLMB and Merrill Lynch (Singapore) are regulated by MAS. Bank of America N.A., Australian Branch (ARBN 064 874 531), AFS License 412901 (BANA Australia) and Merrill Lynch Equities (Australia) Limited (ABN 65 006 276 795), AFS License 235132 (MLEA) distribute this information in Australia only to 'Wholesale' clients as defined by s.761G of the Corporations Act 2001. With the exception of BANA Australia, neither MLEA nor any of its affiliates involved in preparing this information is an Authorised Deposit-Taking Institution under the Banking Act 1959 nor regulated by the Australian Prudential Regulation Authority. No approval is required for publication or distribution of this information in Brazil and its local distribution is by Merrill Lynch (Brazil) in accordance with applicable regulations. Merrill Lynch (DIFC) is authorized and regulated by the DFSA. Information prepared and issued by Merrill Lynch (DIFC) is done so in accordance with the requirements of the DFSA conduct of business rules. Bank of America Merrill Lynch International Limited, Frankfurt Branch (BAMLI Frankfurt) distributes this information in Germany and is regulated by BaFin.
This information has been prepared and issued by MLPF&S and/or one or more of its non-US affiliates. The author(s) of this information may not be licensed to carry on regulated activities in your jurisdiction and, if not licensed, do not hold themselves out as being able to do so. MLPF&S is the distributor of this information in the US and accepts full responsibility for information distributed to MLPF&S clients in the US by its non-US affiliates. Any US person receiving this information and wishing to effect any transaction in any security discussed herein should do so through MLPF&S and not such foreign affiliates. Hong Kong recipients of this information should contact Merrill Lynch (Asia Pacific) Limited in respect of any matters relating to dealing in securities or provision of specific advice on securities or any other matters arising from, or in connection with, this information. Singapore recipients of this information should contact Merrill Lynch International Bank Limited (Merchant Bank) and/or Merrill Lynch (Singapore) Pte Ltd in respect of any matters arising from, or in connection with, this information.
**General Investment Related Disclosures:**
Taiwan Readers: Neither the information nor any opinion expressed herein constitutes an offer or a solicitation of an offer to transact in any securities or other financial instrument. No part of this report may be used or reproduced or quoted in any manner whatsoever in Taiwan by the press or any other person without the express written consent of BofA Merrill Lynch.
This document provides general information only, and has been prepared for, and is intended for general distribution to, BofA Merrill Lynch clients. Neither the information nor any opinion

expressed constitutes an offer or an invitation to make an offer, to buy or sell any securities or other financial instrument or any derivative related to such securities or instruments (e.g., options, futures, warrants, and contracts for differences). This document is not intended to provide personal investment advice and it does not take into account the specific investment objectives, financial situation and the particular needs of, and is not directed to, any specific person(s). This document and its content do not constitute, and should not be considered to constitute, investment advice for purposes of ERISA, the US tax code, the Investment Advisers Act or otherwise. Investors should seek financial advice regarding the appropriateness of investing in financial instruments and implementing investment strategies discussed or recommended in this document and should understand that statements regarding future prospects may not be realized. Any decision to purchase or subscribe for securities in any offering must be based solely on existing public information on such security or the information in the prospectus or other offering document issued in connection with such offering, and not on this document.

Securities and other financial instruments referred to herein, or recommended, offered or sold by BofA Merrill Lynch, are not insured by the Federal Deposit Insurance Corporation and are not deposits or other obligations of any insured depository institution (including, Bank of America, N.A.). Investments in general and, derivatives, in particular, involve numerous risks, including, among others, market risk, counterparty default risk and liquidity risk. No security, financial instrument or derivative is suitable for all investors. In some cases, securities and other financial instruments may be difficult to value or sell and reliable information about the value or risks related to the security or financial instrument may be difficult to obtain. Investors should note that income from such securities and other financial instruments, if any, may fluctuate and that price or value of such securities and instruments may rise or fall and, in some cases, investors may lose their entire principal investment. Past performance is not necessarily a guide to future performance. Levels and basis for taxation may change.

Futures and options are not appropriate for all investors. Such financial instruments may expire worthless. Before investing in futures or options, clients must receive the appropriate risk disclosure documents. Investment strategies explained in this report may not be appropriate at all times. Costs of such strategies do not include commission or margin expenses.

BofA Merrill Lynch is aware that the implementation of the ideas expressed in this report may depend upon an investor's ability to "short" securities or other financial instruments and that such action may be limited by regulations prohibiting or restricting "shortselling" in many jurisdictions. Investors are urged to seek advice regarding the applicability of such regulations prior to executing any short idea contained in this report.

This report may contain a trading idea or recommendation which highlights a specific identified near-term catalyst or event impacting a security, issuer, industry sector or the market generally that presents a transaction opportunity, but does not have any impact on the analyst's particular "Overweight" or "Underweight" rating (which is based on a three month trade horizon). Trading ideas and recommendations may differ directionally from the analyst's rating on a security or issuer because they reflect the impact of a near-term catalyst or event.

Foreign currency rates of exchange may adversely affect the value, price or income of any security or financial instrument mentioned in this report. Investors in such securities and instruments effectively assume currency risk.

UK Readers: The protections provided by the U.K. regulatory regime, including the Financial Services Scheme, do not apply in general to business coordinated by BofA Merrill Lynch entities located outside of the United Kingdom. BofA Merrill Lynch Global Research policies relating to conflicts of interest are described at https://go.bofa.com/coi.

MLPF&S or one of its affiliates is a regular issuer of traded financial instruments linked to securities that may have been recommended in this report. MLPF&S or one of its affiliates may, at any time, hold a trading position (long or short) in the securities and financial instruments discussed in this report.

BofA Merrill Lynch, through business units other than BofA Merrill Lynch Global Research, may have issued and may in the future issue trading ideas or recommendations that are inconsistent with, and reach different conclusions from, the information presented herein. Such ideas or recommendations reflect the different time frames, assumptions, views and analytical methods of the persons who prepared them, and BofA Merrill Lynch is under no obligation to ensure that such other trading ideas or recommendations are brought to the attention of any recipient of this information.

In the event that the recipient received this information pursuant to a contract between the recipient and MLPF&S for the provision of research services for a separate fee, and in connection therewith MLPF&S may be deemed to be acting as an investment adviser, such status relates, if at all, solely to the person with whom MLPF&S has contracted directly and does not extend beyond the delivery of this report (unless otherwise agreed specifically in writing by MLPF&S). If such recipient uses the services of MLPF&S in connection with the sale or purchase of a security referred to herein, MLPF&S may act as principal for its own account or as agent for another person. MLPF&S is and continues to act solely as a broker-dealer in connection with the execution of any transactions, including transactions in any securities referred to herein.



Materials prepared by BofA Merrill Lynch Global Research personnel are based on public information. Facts and views presented in this material have not been reviewed by, and may not reflect information known to, professionals in other business areas of BofA Merrill Lynch, including investment banking personnel. BofA Merrill Lynch has established information barriers between BofA Merrill Lynch Global Research and certain business groups. As a result, BofA Merrill Lynch does not disclose certain client relationships with, or compensation received from, such issuers. To the extent this material discusses any legal proceeding or issues, it has not been prepared as nor is it intended to express any legal conclusion, opinion or advice. Investors should consult their own legal advisers as to issues of law relating to the subject matter of this material. BofA Merrill Lynch Global Research personnel's knowledge of legal proceedings in which any BofA Merrill Lynch entity and/or its directors, officers and employees may be plaintiffs, defendants, co-defendants or co-plaintiffs with or involving issuers mentioned in this material is based on public information. Facts and views presented in this material that relate to any such proceedings have not been reviewed by, discussed with, and may not reflect information known to, professionals in other business areas of BofA Merrill Lynch in connection with the legal proceedings or matters relevant to such proceedings.

This information has been prepared independently of any issuer of securities mentioned herein and not in connection with any proposed offering of securities or as agent of any issuer of any securities. None of MLPF&S, any of its affiliates or their research analysts has any authority whatsoever to make any representation or warranty on behalf of the issuer(s). BofA Merrill Lynch Global Research policy prohibits research personnel from disclosing a recommendation, investment rating, or investment thesis for review by an issuer prior to the publication of a research report containing such rating, recommendation or investment thesis.

Any information relating to the tax status of financial instruments discussed herein is not intended to provide tax advice or to be used by anyone to provide tax advice. Investors are urged to seek tax advice based on their particular circumstances from an independent tax professional.

The information herein (other than disclosure information relating to BofA Merrill Lynch and its affiliates) was obtained from various sources and we do not guarantee its accuracy. This information may contain links to third-party websites. BofA Merrill Lynch is not responsible for the content of any third-party website or any linked content contained in a third-party website. Content contained on such third-party websites is not part of this information and is not incorporated by reference. The inclusion of a link does not imply any endorsement by or any affiliation with BofA Merrill Lynch. Access to any third-party website is at your own risk, and you should always review the terms and privacy policies at third-party websites before submitting any personal information to them. BofA Merrill Lynch is not responsible for such terms and privacy policies and expressly disclaims any liability for them.

All opinions, projections and estimates constitute the judgment of the author as of the date of publication and are subject to change without notice. Prices also are subject to change without notice. BofA Merrill Lynch is under no obligation to update this information and BofA Merrill Lynch's ability to publish information on the subject issuer(s) in the future is subject to applicable quiet periods. You should therefore assume that BofA Merrill Lynch will not update any fact, circumstance or opinion contained herein.

Certain outstanding reports may contain discussions and/or investment opinions relating to securities, financial instruments and/or issuers that are no longer current. Always refer to the most recent research report relating to an issuer prior to making an investment decision.

In some cases, an issuer may be classified as Restricted or may be Under Review or Extended Review. In each case, investors should consider any investment opinion relating to such issuer (or its security and/or financial instruments) to be suspended or withdrawn and should not rely on the analyses and investment opinion(s) pertaining to such issuer (or its securities and/or financial instruments) nor should the analyses or opinion(s) be considered a solicitation of any kind. Sales persons and financial advisors affiliated with MLPF&S or any of its affiliates may not solicit purchases of securities or financial instruments that are Restricted or Under Review and may only solicit securities under Extended Review in accordance with firm policies.

Neither BofA Merrill Lynch nor any officer or employee of BofA Merrill Lynch accepts any liability whatsoever for any direct, indirect or consequential damages or losses arising from any use of this information.