Marko Kolanovic, PhD
(1-212) 272-1438
marko.kolanovic@jpmorgan.com

Non-Parametric Regression: K-Nearest Neighbor and LOESS

Supervised learning algorithms are sometimes classified as being either parametric or non-parametric techniques. In parametric techniques, the model is described by a set of parameters such as linear regression beta (that is estimated from historical data). In non-parametric techniques, we do calibrate parameters of a model, but directly identify similar historical instances, and assume the behavior will be the same. Given a new datapoint, we search through historical data and identify a number 'K' of similar instances that we call "nearest neighbors" 'NN' (hence name K-NN). Then we then make predictions by averaging historical outcomes for these "nearest neighbors". Two examples of non-parametric regressions are.

- K-Nearest Neighbor (KNN)[33] rule: Once we have located K nearest neighbors, we can average the output variable y for this subset and use that as our prediction.

- LOESS: Using data for the K nearest neighbors, for each new point we fit a linear regression based on the K nearest neighbors (subset of historical data), and predict the output using those coefficients. This is called LOESS or localized linear regression.

Non-parametric techniques offer a simple way to extrapolate analysis on past similar events. KNN is commonly used by financial analysts, who intuitively employ it without referring to it as such. The use of a historical sample makes the technique "supervised", while the absence of parameters or coefficient betas makes it "non-parametric". In many financial analyses, the output variable is often not linearly related to the inputs; this makes linear regression and its extensions like ridge and lasso unsuitable. In such cases, the K-nearest neighbor can capture those non-linear properties. A drawback of using the KNN rule lies in its extreme sensitivity to outliers.

One can view linear regression and k-nearest neighbor methods as two ends of the spectrum of classical Machine Learning[34]. On one hand, linear regression 'under-fits' the data and hence suffers from higher 'bias' to achieve lower 'variance'. On the other hand, locating the K-most similar inputs and averaging their output makes a weak structural assumption. In formal terms, we say that the K-nearest neighbor method can 'over-fit' and hence suffer from higher 'variance' (while having a low 'bias').

We can gain further insight into the problem of under/over-fitting by varying the value of K in the KNN. At one extreme, we can set K to be equal to the number of samples itself, in which case we simply predict the quantized sample average as the output for all inputs. As we reduce K, we obtain a decision boundary that splits the training set into a number of parts. In the limiting case of K=1, we form a tiny region around each training sample. In this case, we have over-fit the data, which is likely to lead to a high error rate on unseen samples, while yielding a deceptively low (in this case, zero) error rate on the training set. One can informally think of a K-nearest neighbor method as dividing the N samples in historical data into N/K sets and computing their respective means. This implies that the effective number of parameters in K-nearest neighbor method is N/K, which increases as K decreases. The notion of effective number of parameters or degrees of freedom is a

[33] K-nearest neighbor rules and extensions are covered in Dasarathy (1991), Ripley (1996), MacQueen (1967), Hastie and Tibshirani (1996a), Simard et al (1993). For learning vector quantization, see Kohonen (1989). For discussion on Bayesian non-parametrics including Dirichlet Process Models, see Dunson (2009, 2010b), Shen and Ghosal (2013), Tokdar (2011), Wang and Dunson (2011a), Carvalho et al (2010), Ohlssen, Sharples and Spiegelhalter (2007), Ray and Mullick (2006), Rodriguez, Dunson and Gelfand (2009), Bigelow and Dunson (2009).

[34] Splines (including B-splines, thin-plate splines and reproducing kernel Hilbert spaces) are covered in Green and Silverman (1994), Wahba (1990), Girosi et al (1995) and Evgenion et al (2000). For wavelets, consult Daubechies (1992), Chni (1992), Wickerhauser (1994), Donoho and Johnstone (1994), Vidakovic (1999), Bruce and Gao (1996). Local regression and kernel methods are covered in Loader (1999), Fan and Gijbels (1996), Hastie and Tibshirani (1990). Statistical treatment of parametric non-linear models is illustrated in Reilly and Zeringue (2004), Gelman, Chew and Shnaidman (2004), Denison et al (2002), Chipman, George and McCulloch (1998, 2002), DiMatteo et al (2001) and Zhao (2000). For basis function models, see Bishop (2006), Biller (2000), DiMatteo, Genovese and Kass (2001), Barbieri and Berger (2004), Park and Casella (2008), Seeger (2008), Ramsay and Silverman (2005), Neelon and Dunson (2004), Dunson (2005), Hazelton and Turlach (2011), Hannah and Dunson (2011), Patti and Dunson (2011).

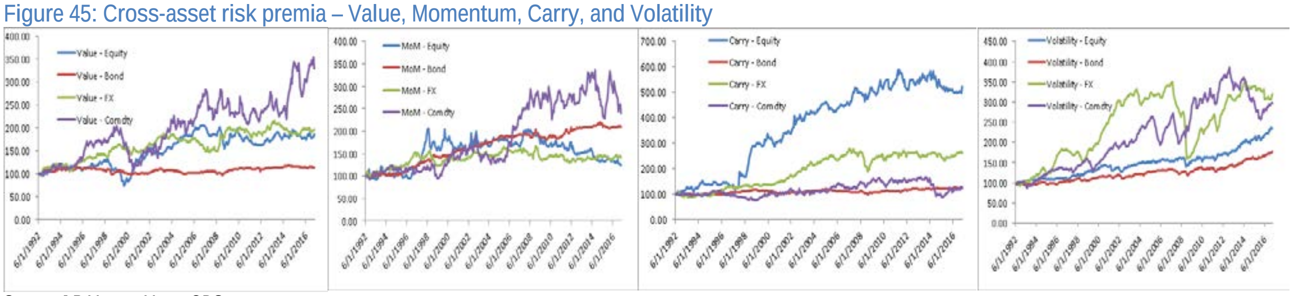
Marko Kolanovic, PhD
(1-212) 272-1438
marko.kolanovic@jpmorgan.com

way to characterize the model complexity. As K decreases, and the number of parameters increases, the model complexity essentially increases, thereby leading to over-fitting.[35]

Financial Example

In our previous work, we studied cross-asset risk premia investing.[36] In particular, we had constructed simplified risk premia factors of value, momentum, carry, and volatility (across equities, bonds, currencies and commodities). Our analysis had shown that risk premia strategies delivered a positive Sharpe ratio over an extended period of ~40 years and exhibited low and stable correlation to traditional asset classes. We will attempt to use the K-nearest neighbor algorithm to illustrate potential for timing of these cross-asset risk premia based on macro regimes.

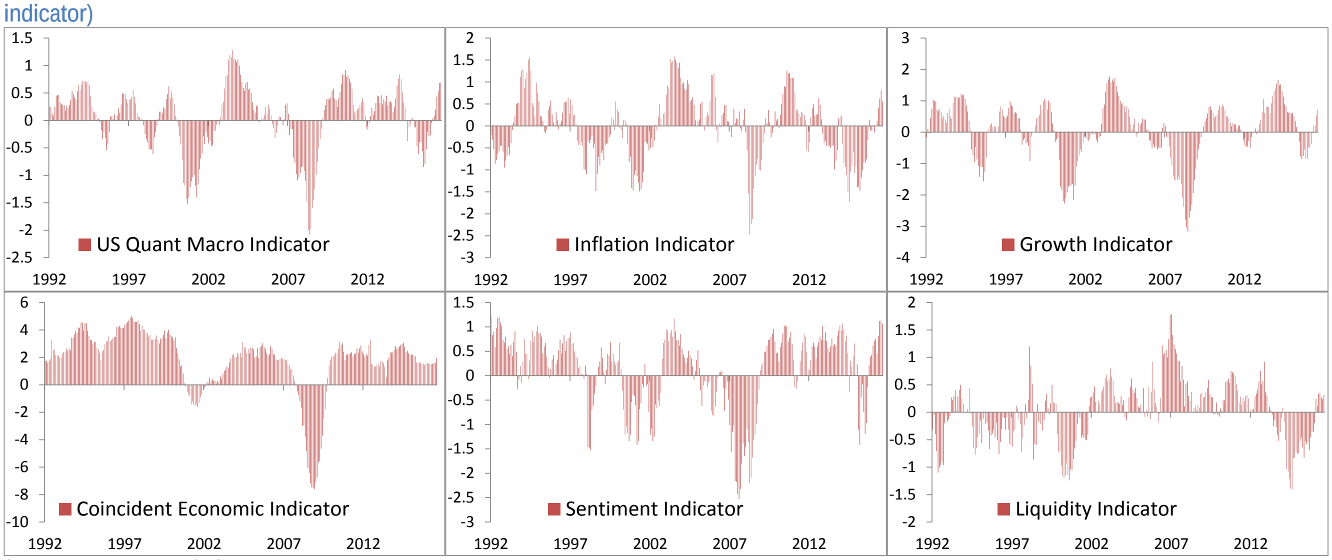
Figure 45: Cross-asset risk premia – Value, Momentum, Carry, and Volatility



Source: J.P.Morgan Macro QDS

To identify the current macro regime, we use 7 aggregated macro indicators: Quant Macro indicator, Growth, Liquidity, Sentiment, Inflation, Coincident economic indicator, as well as change in composite macro indicator. These aggregated indicators are based on 50 macro-economic signals; for details of construction, please see our earlier report titled "Framework for style investing: Style rotation and the business cycle".

Figure 46: Six aggregated macro indicators that are used to identify current economic regime (in addition, we use change in composite indicator)



Source: J.P.Morgan Macro QDS

[35] Formally, model complexity is characterized by Vapnik-Chervonenkis dimension of the set of classifiers or hypotheses. We describe that in a mathematical box in our discussion on model selection theory.

[36] Kolanovic, M and Wei, Z (2013), "Systematic strategies across asset classes: Risk factor approach to investing and portfolio management", J.P.Morgan Research Publication.

Marko Kolanovic, PhD
(1-212) 272-1438
marko.kolanovic@jpmorgan.com

Figure 47: List of 50 macro-economic signals used to construct the 7 macro indicators above (for more details see here)

ISM Manufacturing PMI	ISM Non-manufacturing PMI
Initial Jobless Claims	US Capacity Utilization
US Leading Indicator	Yield Curve – 10y less 2y
Global Leading Indicator	Railroads Freight Index, 12m change
Leading-Lagging Economic Indicator Spread	US relative to World Stock Index
Baltic Dry Index	Global Economic Momentum
Manufacturing New Orders ex. Transportation	Retail Sales ex. Transportation
New Housing Permits	Credit Spread Moody's AAA-BAA
Michigan Consumer Sentiment	New Company Revisions, 3M Avg
NAPM Percentage Surprise, 3M Wt. Avg.	JULI Inv Grade USD Spread over Tsy
CESI – Citigroup Economic Surprise Index	VIX
Dow Transportation/Utilities, 12M Chg.	Small – Large Cap Outperformance, 12M Chg.
Barclays Investment Grade Spread	Barclays High Yield Spread
Term Structure of Momentum, Diffusion Index	Stock Market, 3Y Change
Margin Debt Level	Margin Debt as Percentage of Market Cap
Loan Officers Survey	M2 Money Stock, 12M Percentage Change
10Y Bond Yield, Real Rate	Correlation of MZM (Money) and SPX
Term Structure of Fed Fund Futures	USD Trade Weighted Index
Avg Treasury Trading Volume % of Mkt Debt	US Credit Manager Index
Free Liquidity	Earnings Yield Dispersion
Real Loan Growth	ISM Prices Index, 12M Change
PPI, YOY Percentage Change	Unit Labor Cost, 12M Change
US Import Price Inflation	Wage Trend Index, 12M Change
Breakeven Inflation. 10Y	Oil Price, WTI
Commodity Price Index	Median Home Price, 12M Change

Source: J.P.Morgan Macro QDS

We construct a monthly rebalancing strategy, where every month we look at a rolling window of the prior 10 years. We characterize the macro regime in a given month by the value of the 7 aggregated indicators. Then we look for K nearest neighbors, i.e.

a. We compute the distance between the economic regime vector of the current month to the economic regime vector for all months in the past 10 year window.

b. We rank them in ascending order of distance and choose the first (nearest) K neighbors

For each of these K nearest neighbors, we find the average return over a succeeding month for all 20 cross-asset risk premia factors. We then rank these factors according to their average returns, and choose a subset S that performed the best.

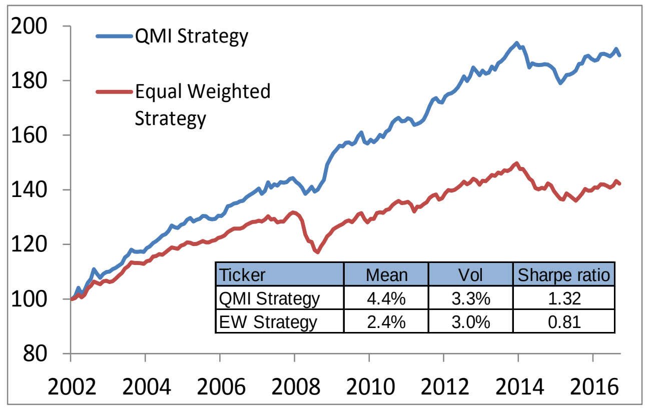
If we set S=1, it is equivalent to finding the best performing strategy historically in the same macro regime. If we set S=20, we will get an equal weighted strategy over the 20 risk premia. If we set K=1, we choose the nearest instance and replicate that for the current month. If we set K = 120, then we take all months in the input sample and average out the results, and invest in the best S strategies.

The Sharpe ratios of strategies based on K-nearest neighbors is tabulated below. We find that using K between 1 and 10, and number of risk factors between 10 and 19 generally outperforms simple equal weighted risk premia portfolio. For instance K=2 and S=14 yields the highest Sharpe ratio of 1.3, as compared to 0.8 shapre ratio of equal weighted strategy. We plot that case below alongside the equal weighted strategy that assigned equal weights to all 20 risk premia.

Marko Kolanovic, PhD
(1-212) 272-1438
marko.kolanovic@jpmorgan.com

Figure 48: Sharpe ratio of portfolio chosen using K-Nearest Neighbor rule over cross-asset systematic strategies (left, K=1 to 25); Performance by timing cross-asset risk premia using JPM's Quant Macro Indicators (right)

	Number of Risk Premia Strategy																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0.6	0.6	0.7	0.8	0.8	0.9	1.0	1.1	1.2	1.2	1.2	1.3	1.3	1.3	1.3	1.3	1.2	1.1	0.9	0.8
2	0.4	0.6	0.7	0.8	0.8	0.9	0.9	1.0	1.1	1.1	1.1	1.2	1.2	1.3	1.3	1.3	1.2	1.2	1.0	0.8
3	0.1	0.3	0.6	0.7	0.8	0.8	0.9	0.9	0.9	0.9	1.0	1.0	1.1	1.1	1.1	1.1	1.1	1.1	1.0	0.8
4	0.2	0.2	0.3	0.5	0.6	0.6	0.7	0.7	0.8	0.8	0.8	0.8	0.9	0.8	0.9	1.0	0.9	1.0	1.0	0.8
5	0.0	0.3	0.4	0.5	0.5	0.7	0.7	0.8	0.8	0.8	0.8	0.9	1.0	1.0	1.1	1.1	1.0	1.0	0.9	0.8
6	0.2	0.3	0.4	0.3	0.3	0.3	0.3	0.6	0.6	0.6	0.8	0.9	0.9	0.9	0.9	0.9	1.0	1.0	1.0	0.8
7	-0.1	0.3	0.4	0.4	0.4	0.3	0.4	0.5	0.5	0.6	0.7	0.8	0.9	0.9	0.9	1.0	1.0	0.9	0.9	0.8
8	0.2	0.2	0.2	0.4	0.6	0.5	0.6	0.7	0.7	0.8	0.8	0.9	1.0	0.9	0.9	1.0	1.0	1.0	0.9	0.8
9	0.1	0.2	0.2	0.5	0.6	0.7	0.7	0.7	0.8	0.9	1.0	0.9	1.0	1.0	1.0	0.9	0.9	0.9	0.9	0.8
10	0.0	0.1	0.2	0.3	0.5	0.5	0.6	0.7	0.7	0.8	0.7	0.8	0.8	0.9	0.8	0.9	0.9	0.9	0.9	0.8
11	-0.1	0.2	0.3	0.3	0.4	0.4	0.5	0.5	0.6	0.7	0.8	0.8	0.8	0.9	0.9	0.9	0.9	0.9	0.8	0.8
12	0.1	0.0	0.1	0.3	0.5	0.4	0.5	0.6	0.7	0.7	0.8	0.9	0.9	0.9	0.9	1.0	0.9	0.9	0.8	0.8
13	0.0	0.1	0.2	0.3	0.4	0.6	0.7	0.7	0.7	0.7	0.7	0.7	0.9	0.9	1.0	1.0	1.0	1.0	0.9	0.8
14	0.1	0.3	0.4	0.3	0.4	0.5	0.6	0.7	0.7	0.8	0.7	0.7	0.8	0.8	0.9	0.9	0.9	1.0	0.9	0.8
15	0.0	0.3	0.3	0.4	0.5	0.7	0.7	0.6	0.6	0.6	0.7	0.8	0.9	1.0	1.0	1.0	0.9	0.9	0.9	0.8
16	-0.1	0.1	0.3	0.4	0.6	0.6	0.7	0.7	0.7	0.7	0.7	0.7	0.8	0.9	0.9	0.9	0.9	0.9	0.9	0.8
17	0.1	0.2	0.2	0.5	0.3	0.4	0.6	0.6	0.6	0.7	0.7	0.8	0.8	0.9	1.0	0.9	0.9	0.9	0.9	0.8
18	0.2	0.4	0.3	0.5	0.6	0.5	0.6	0.7	0.7	0.8	0.8	0.8	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.8
19	0.2	0.2	0.1	0.4	0.5	0.6	0.7	0.7	0.7	0.6	0.7	0.8	0.9	0.8	0.8	0.9	0.9	0.9	0.9	0.8
20	0.3	0.1	0.3	0.4	0.6	0.6	0.6	0.7	0.7	0.7	0.7	0.8	0.9	0.9	0.8	0.8	0.9	0.9	1.0	0.8
21	0.1	0.2	0.4	0.6	0.7	0.6	0.6	0.6	0.7	0.7	0.7	0.7	0.7	0.8	0.8	0.8	0.9	0.9	0.9	0.8
22	0.3	0.1	0.3	0.5	0.6	0.5	0.5	0.5	0.6	0.6	0.6	0.7	0.7	0.7	0.8	0.8	0.9	0.8	0.8	0.8
23	0.3	0.4	0.3	0.6	0.7	0.5	0.6	0.6	0.7	0.7	0.7	0.7	0.8	0.7	0.8	0.8	0.8	0.9	0.9	0.8
24	-0.1	0.3	0.3	0.4	0.6	0.5	0.6	0.6	0.7	0.8	0.7	0.7	0.6	0.7	0.8	0.8	0.8	0.8	0.9	0.8
25	-0.1	0.3	0.4	0.5	0.6	0.6	0.6	0.7	0.7	0.7	0.6	0.7	0.7	0.7	0.8	0.8	0.8	0.8	0.8	0.8



Source: J.P.Morgan Macro QDS