First Steps in Developing High-Frequency Trading Models

Bruce Vanstone and Tobias Hahn

Bruce Vanstone
is an assistant professor at the Bond University Faculty of Business in Queensland, Australia.
bvanston@bond.edu.au

Tobias Hahn
is a senior teaching fellow at the Bond University Faculty of Business in Queensland, Australia.
thahn@bond.edu.au

The foreign exchange (FX) markets offer unusual opportunities to professional speculators. According to the Bank for International Settlements, the daily turnover of the FX markets is in excess of $4 trillion (Strong [1988]). FX markets are designed to assist international trade and investment and to allow participants to easily convert one currency into another at an agreed rate. These markets are also of particular interest to speculators not just for their size and correspondingly high liquidity but also for their low transaction costs and common use of leverage. This leverage and the ability to take both long and short positions with relatively low constraints are in stark contrast to many traditional markets.

Speculators, individuals, and trading firms frequently use off-the-shelf software and access the market through algorithmic brokers, in contrast to larger (and often utilitarian) traders. Algorithmic trading software and infrastructure provides access to high-frequency data and equally short, usually intraday, market positions.

Speculative, high-frequency market participants (speculators) face a number of questions that remain unaddressed by existing literature. High trading frequencies require system automation for at least the detection of trading opportunities and generally also for execution, without the typically more sophisticated infrastructure available to large investment firms. Equally important constraints often exist in regard to the breadth of skills in the information technology domain.

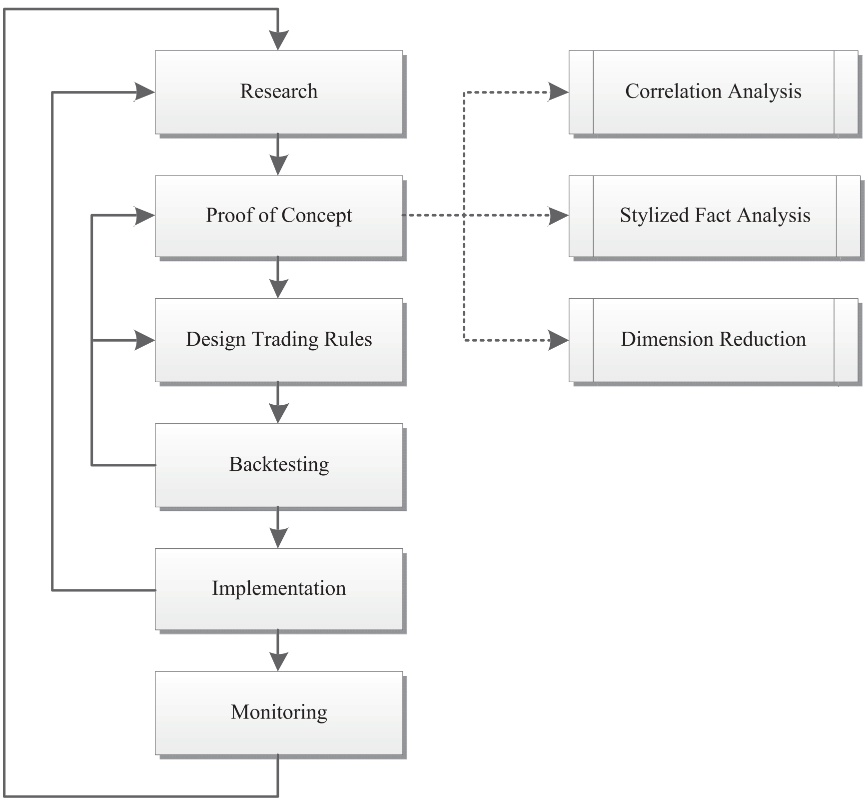
Modelling high-frequency time series is an area lacking in formal academic models. Speculators, along with many other professionals, often rely on stylized facts and data mining. Here, data mining is understood in its value-neutral sense, that is, as a technique to learn facts and relationships from data. Whether this results in a biased model is naturally an important question, but for the purposes of this article, a secondary one.

This article introduces a framework and trading rule development methodology based on both stylized facts and data-mining approaches. To this end, the methodology and results of a study for each of the two techniques introduced here are reported.

The two techniques provide an extension to an existing methodology for developing and testing trading rules as shown in Exhibit 1 (Vanstone and Finnie [2009]). They extend the correlation-based analysis of individual candidate variables, both fundamental and technical, to the more general case of dealing with a comprehensive set of individual candidates, not independent of one another, and in the case of stylized facts, statistical knowledge of the underlying time series rather than a candidate time series itself.

E X H I B I T 1

Trading Strategy Development Process



In regard to the stylized facts, a review and testing of commonly accepted high-frequency FX time series features is reported in the stylized fact analysis, focusing particularly on negative first-order autocorrelation. This study is based on one-minute EURUSD time series data. This currency pair is the most heavily traded currency in the FX markets and accounts for some 28% of the spot market; it should thus be considered the most efficient pair with respect to pricing. After confirming the stylized facts, a preliminary set of trading rules is postulated and the possibility of economic value is determined using benchmarking and robustness checks.

The dimension reduction study focuses on the second technique, the data-mining approach. In contrast to the stylized facts, no initial assumptions exist but instead a large amount of data is available. As in the case of stylized facts, the opportunity for economic value needs to be determined prior to proceeding with the trading system development. The large amount of data used in such cases is typically a combination of high-frequency price data and technical indicators.

Technical indicators are usually disregarded by academic research since the assumption of market efficiency suggests no economic profits are possible. Whether this assumption holds for a particular security at a particular time frame needs to be addressed by the speculator. There is, however, academic support for the profitability of technical analysis and their indicators. Park and Irwin [2007] provide a comprehensive review of the academic literature related to technical analysis and report "among a total of 92 modern studies, 58 found positive results regarding technical trading strategies, while 24 studies obtained negative results. Ten studies indicated mixed results." Park and Irwin conclude that although many modern studies report positive findings, they may be subject to problems in their testing procedures, such as data snooping, ex post selection of trading rules or search technologies, and difficulties in estimation of risk and transaction costs. In addition, there is substantial evidence that technical analysis and technical variables are heavily used in practice, particularly when trading in shorter time frames (Strong [1988]), and specifically in the FX markets (Taylor and Allen [1992]).

There is no pre-existing methodology that helps address these issues and still focuses on the ultimate goal of developing successful high-frequency trading systems. Methodological concerns are not the only issue practitioners and academics face, however. Because of the nature of the data-mining approach, there are a large number of variables (technical indicators) to be studied with an even larger number of possible rules to derive. It is unreasonable to assume that even considering varying measurement time frames (look-back periods), a large number of indicators based on a relatively limited number of time series would result in more meaningful indicators. Instead, any such transformation can only disperse the information content but not increase it.

It is thus necessary to limit the number of indicators to a set that is simultaneously smaller but retains any noise-reducing features present in the original set. The research objective in the case of data mining is thus not to test an assumed effect as in the case of stylized facts but to reduce the variable set—a case of dimension reduction.

Exacerbating the problem is the large number of variables to be studied, which in turn gives rise to a number of possibilities to use them in trading rules and in forming portfolios. Rather than attempting to evaluate each such combination, this article aims to derive the minimum set of technical variables using principal component analysis (PCA) as a dimension-reduction problem. This allows for future rule development based on a smaller set of key variables.

STYLIZED FACT ANALYSIS

Literature Review

Stylized facts of time series normally describe their dominant statistical features such as serial dependence, distributional characteristics, the presence of nonlinearity, and the presence and degree of scaling effects. It is often reported that those facts attributed to daily or weekly data do not hold for higher frequencies, that is, intraday data (Sewell [2011] and Dacoronga [2001]).

The prevailing opinion appears to be that high-frequency FX prices exhibit highly negative first-order autocorrelation at sampling frequencies of minutely or higher (Zhou [1996]). This effect is confirmed in other assets and sampling frequencies (see Sewell [2011], Dacoronga [2001], and Cont [2001]), such as Italian stock index futures for time periods smaller than 20 minutes (Bianco and Reno [2006]).

Although most financial time series returns are distributed approximately symmetrically with high kurtosis, this symmetry decreases and fat tails become more pronounced as the sampling frequency increases (Sewell [2011]).

Similarly, usual assumptions of linearity do not hold for many financial time series including FX data. Tests of five major FX rates find no evidence of linear correlation; however, substantial evidence of nonlinearity is found (Hsieh [1989]). Tests of 10 different GBP exchange rate pairs also document nonlinearity in many of the pairs (Evertsz [1995]). It is precisely this nonlinearity that leads many researchers to consider the use of machine learning techniques as a tool to learn rules from financial data. For these researchers, the technique of dimension reduction is of key importance.

Scaling effects refers to the absolute size of returns as a function of the time interval in which they are measured. Scaling laws provide a relationship between time interval and average volatility as a power of the absolute return of that interval. There is much evidence of the existence of power laws within high-frequency FX data. Scaling is reported in the mean absolute changes of logarithm prices (Muller [1990]), self-similarity is reported in USD/DEM (Evertsz [1995]) and multifractal scaling in DEM/USD (Fisher et al. [1997]). Scaling is reported in exchange rates (Galluccio [1997]) and exchange rate volatilities (Gencay et al. [2001]). Further support for FX scaling laws is provided by Andersen et al. [2001], Dacoronga [2001], and Gencay et al. [2001].

High-frequency time series in general, and in FX markets in particular, do not exhibit the character typically assumed in standard financial models.

Methodology

Building a trading system based on stylized facts requires meeting two preconditions: credibility and robustness. Credibility is either asserted along with the fact itself or requires the passing of particular tests, that is, absent additional information regarding the stylized facts, their validity needs to be confirmed by replicating the original research or parts thereof in the required time series.

Robustness requires at a minimum the confirmation of results outside the original data (time, place, instrument, and so on) as appropriate. In this article, both aspects are addressed simultaneously using an in-sample and out-of-sample backtesting approach.

The stylized facts are first confirmed in the in-sample data by testing for market efficiency and randomness. The general assumption in financial research is that markets are efficient and specifically that returns follow a random walk. Under those assumptions, no economically successful trading system is feasible. The stylized facts described here suggest that these assumptions do not hold at sufficiently high sampling frequencies.

Although market efficiency and randomness can be established in various ways, the Lo and MacKinlay variance ratio (Lo and MacKinlay [1988]) is frequently used for testing for market efficiency. The runs test, on the other hand, is a convenient way to test for randomness (of both values and direction, that is, sign of the return, respectively) along the return path (Aldridge [2010]). Autocorrelation, including partial autocorrelation (removing cumulative correlation at lower lags), is also used as a measure of serial dependence (or randomness).

Once the stylized facts and deviations from standard model assumptions are confirmed in-sample, rules suitable to exploiting the abnormal behavior of the time series can be postulated and tested in the out-of-sample data. These rules are not necessarily the same as would be used in the finished trading system. Rather, they are designed to extract the stylized fact-based feature from the time series for subsequent use. Aggregation of the data enables such tests at various sampling intervals (one minute versus one hour), also offering some insight into the presence of scaling effects.

Exhibit 2

Summary of Market Efficiency and Randomness Test Results (in-sample)

Period	Frequency	Variance Ratio Test, *p*-Value	Runs Test (+ve/–ve), *p*-Value
2008–In-sample partition 1	1 minute returns	Rejected, 1.11e–11	Rejected, 0.00e+00
2008–In-sample partition 1	1 hourly returns	5.34e–01	Rejected, 8.78e–05
2009–In-sample partition 2	1 minute returns	Rejected, 1.24e–133	Rejected, 0.00e+00
2009–In-sample partition 2	1 hourly returns	2.18e–01	Rejected, 2.60e–06

Exhibit 3

Summary of Autocorrelation Results

Period	Frequency	Autocorrelation Direction (Lag 1)	*p*-Value
2008–In-sample partition 1	1 minute returns	Negative	0.00e+00
2008–In-sample partition 1	1 hourly returns	Negative	1.04e–01
2009–In-sample partition 2	1 minute returns	Negative	0.00e+00
2009–In-sample partition 2	1 hourly returns	Negative	4.74e–01

The complete dataset covers one-minute price data from the EURUSD currency pair between January 1, 2008, and December 31, 2011 (inclusive), sourced from Thomson-Reuters TRTH database provided by SIRCA. This is split into in-sample (2008 and 2009) and out-of-sample data (remaining data).

Lacking credible benchmark portfolios for high-frequency FX trading, a naïve buy-and-hold strategy is used instead. Results from the postulated trading rules based on the stylized facts are thus compared to those based on the buy-and-hold strategy, using a number of metrics both in-sample and out-of-sample. In-sample results are used to confirm the stylized facts hold true for the data selected, whereas out-of-sample results are compared to in-sample outcomes as a robustness check. The models were implemented in MATLAB, whose statistical tests were also used to compute the results presented here. The backtesting metrics chosen here are typical of those presented by algorithmic trading software and were described in detail in Vanstone [2006].

Analysis of Data

In order to address concerns regarding the effect of the global financial crisis (GFC), the in-sample data were further subdivided into annual partitions (2008 and 2009). Exhibit 2 summarizes the results of the market efficiency and randomness tests by subdivision and sampling frequency. For each test, the table shows whether the null hypothesis was rejected including the test probability. The null hypotheses are that returns follow a random walk (variance ratio test) and that the observations sequence is random (runs test), respectively. As suggested by the stylized facts, the null hypothesis for the high-frequency data is rejected for one-minute data. More limited evidence can be found for the aggregated one-hour data, where the test for market efficiency cannot be rejected though the order of observations does not appear to be random (rejected runs tests in all but one case).

Similar results can be obtained using autocorrelation and partial autocorrelation for one-minute and hourly data and for the years 2008 and 2009. The results, along with the Ljung-Box Q-test statistics, are reported in Exhibit 3. There is significant negative autocorrelation at lag 1 for one-minute data, that is, the serial observation of residuals is not random. Although not significant, the autocorrelation term is also negative for hourly data.

This again confirms the existence of the previously documented stylized facts concerning negative autocorrelation in high-frequency FX data. All three tests—the variance ratio test, the runs test, and the Ljung-Box Q-test—indicate that the time series is suitable for development of a high-frequency trading strategy, particularly at the one-minute sampling frequency.

Based on these findings, a simple trading system is postulated exploiting the negative autocorrelation character of the time series. Speculative traders typically do not trade or even have the ability to trade every opportunity but instead select extreme cases. The following rules are a simple example, which reflects such typical behavior:

- Buy: Take a long position (one contract) when the one-minute closing price closes above the lowest price over the last week.
- Sell: Close the position after no new high has been reached within the last five minutes.

The specific parameters, that is, periods used in both the buying rule and the selling rule, were chosen arbitrarily. The results reported in Exhibits 4 and 5 summarize the in-sample and out-of-sample trade and portfolio statistics and include typical transaction cost with a large-scale retail broker (Interactive Brokers Commission [2011]). These metrics are typical of algorithmic trading system development software and were covered in detail in Vanstone [2006].

EXHIBIT 4

In-sample Trading Simulation Metrics vs. Naive Buy-and-Hold

Metric	Simulation	Buy-and-Hold
Raw Profit	$1,595.00	$577.50
Number of Trades	135	1
Average Minutes per Trade	25.98	740,652
Winning Percentage	64.44%	100.00%
Profit Factor	1.19	N/A
Recovery Factor	0.63	0.02
Payoff Ratio	0.65	0.00
Maximum Drawdown	–$2,535.00	–$37,060.00

EXHIBIT 5

Out-of-Sample Trading Simulation Metrics vs. Naive Buy-and-Hold

Metric	Simulation	Buy-and-Hold
Raw Profit	$1,710.00	–$13,682.50
Number of Trades	164	1
Average Minutes per Trade	29.16	748,702
Winning Percentage	57.93%	0.00%
Profit Factor	1.22	0.00
Recovery Factor	0.73	0.00
Payoff Ratio	0.89	0.00
Maximum Drawdown	–$2,330.00	–$27,010.00

The simulations yield similar results despite very different market conditions as illustrated by the buy-and-hold strategy, which even changes sign. The rules exploiting the stylized facts, on the other hand, exhibit similar and profitable behavior in both periods. This also demonstrates robustness across time of the resulting rules.

The time parameters used in the buying and selling rules were chosen arbitrarily. However, further testing shows this approach is robust to the selection of many different buying and selling periods. For the buying period rules, positive results exist for multiday time frames. For the selling period rules, positive results exist for small holding time frames, ranging from five minutes to one hour. The results for these further tests are not included in this article but are available from the authors on request.

This study described and demonstrated the technique used for assessing and confirming the existence and robustness of the stylized fact of highly negative first-order autocorrelation in high-frequency price data. A simple trading rule was then postulated to exploit this stylized fact, and the rule was back tested in both the in-sample and out-of-sample data. The result from the test confirms the stylized facts robustness and economic significance.

DIMENSION-REDUCTION STUDY

Overview

The alternative to working with stylized facts, that is, known abnormal but repeating effects, is the discovery of such behavior from the data itself. To this end, professional traders often rely on their intuition and experience in combination with a number of technical analysis indicators to develop rules. These are hardly ever formed explicitly.

The dimension-reduction process suggested in this section addresses both the development of new rules (pure data mining) and the simplification of existing rules (via filter rules). Such a process requires the decision maker to supply:

1. a list of (standard or proprietary) indicator values, and
2. parameters to control the dimension-reduction process.

This study is a demonstration of the proposed process applied to one-minutely data for the EURUSD exchange rates using a common set of indicators and reasonable parameter choices for the dimension-reduction process.

The dimension-reduction parameters are arbitrary albeit informed choices. Any decision maker needs to determine these parameters based on the specific problem domain (that is, asset class, sampling, and trading frequency, and so on), time series properties (especially the suspected signal/noise ratio), and decision makers' and investors' preferences, such as their level of uncertainty intolerance, risk aversion, and ability to cope with ambiguity and complexity. Equally, the indicator set may be enhanced by adding pricing or forecasting information beyond a commonly available set if the problem domain allows for this. The choices made here serve as a study into the application of the process to a selected instrument.

For the majority of speculative traders, indicator choices are often limited by their availability in the chosen trading software. As a representative example, Wealth-Lab [2013], an algorithmic trading software, presents 250 indicators and oscillators,[1] divided into three categories (104 community indicators, 71 TASC magazine indicators, and 75 Wealth-Lab standard indicators). Most indicators and oscillators require several parameters to be set by the user, resulting in an enormous number of possible combinations. Further, many of these variables can be applied to either underlying securities price data or overall market data (or even each other!). It is clearly not practical for an individual trader to consider all possibilities. There are a large number of markets, traded instruments, time frames, and approaches to trading. In addition to the large amount of work this would require, the marginal contribution of subsequent indicator choices quickly approaches zero as no additional information is added to the overall set after a number of time frames and classes of indicator have been included. Decision makers thus need a technique to reduce this set to a more manageable yet information-preserving set.

A commonly used technique in dimension reduction is principal component analysis (PCA), an orthogonal data transformation procedure mostly used in exploratory data analysis. PCA performs eigenvalue decomposition of a data co-variance (or correlation) matrix. PCA transforms a set of possibly correlated variables into a set of linearly uncorrelated variables called principal components. The first principal component accounts for the largest possible variance, and each subsequent principal component accounts for the largest amount of remaining variance uncorrelated to the prior principal components. The number of principal components will be less than or equal to the original number of variables. By design all variables are included in the final set, that is, PCA creates only an orthogonal representation of the indicators. However, it is common in many domains for the sum of the variances of the first few principal components to exceed 80% of the total variance of the original data.

The same discussion is relevant to the variables within each component as groups of variables often move together because they either measure the same effect, proximate or ultimate cause, or more simply they are similar mathematical representations or transformations of the same data source (that is, similar formulas), as prior studies have pointed out.

As a technique, PCA has widespread acceptance among the academic community, and has been used in prior research to create orthogonal variables for machine learning techniques from potential candidate variable sets (Cao et al. [2003]).

Since PCA guarantees that all principal components are orthogonal, it removes redundancy between components. Using the methodology in this study, the composition of each of the principal components can be determined in terms of the underlying original technical variables, leading to the discovery of a subset of variables that are of prime importance in describing the variability in the system that the data represent.

Economic relevance is not the focus in dimension reduction; instead, correlation analysis of individual variables should next be applied, as documented by Vanstone [2006], when economic significance is the goal. As noted there, rules should not be built using out-of-sample results or insights.

Out-of-sample PCA analysis is conducted in this article so that the set of technical variables that describe both the in-sample and out-of-sample data can be formed. This allows for determining which variables are common to both sets. These are the stable variables that describe the data variation over time. This approach can be used in a variety of ways. For example, the data could be initially divided into three samples, with dimension reduction being performed on two of those three samples. Variables common to both samples could then be used ex ante in the third sample for forming trading rules.

Methodology

Minutely data is obtained from SIRCA for the EURUSD contract for the period January 1, 2000, to December 31, 2013. The data is divided into two partitions. The in-sample partition covers the seven-year period January 1, 2000, to December 31, 2006. The out-of-sample partition covers the seven-year period January 1, 2007, to December 31, 2013. The PCA technique will be conducted on both the in-sample data and the out-of-sample data separately, and a determination can be made as to how many of the explanatory technical variables are common to each sample. As described here, typically, the dimension-reduction process is not conducted on out-of-sample data, which then also allows for the further ex ante investigation of technical trading rules based on any variables selected during in-sample PCA without the possibility of data snooping, a problem that Park and Irwin [2007] point out exists in many modern studies. Dimension reduction is performed in this article on out-of-sample data only for the purpose of understanding the time-varying character of the components of variability in the underlying time series data.

The technical variables considered as potential candidates to describe the high-frequency EURUSD data are all 75 of the Wealth-Lab standard indicators. For many variables, there are a number of parameters that need to be specified. Very often, little guidance is available to make sensible parameter choices. The Wealth-Lab trading platform itself provides defaults for these parameters if the trader does not specify them. For this reason, all parameters for the variables in the candidate set are provided at their default values. These default parameter values are also provided, along with a brief description of each variable, in Appendix A.

Technical Variables whose name is suffixed with an asterisk are variables whose values are expressed in terms of the price level of the specific underlying instrument. In this technique, the value of these variables is divided by the instrument's close price at the time they are calculated. As an example, a simple moving average reflects the underlying instrument's price level and is divided by the underlying instrument's current price to normalize it. This step addresses two related problems. Firstly, when applying the process to a wider set of instruments, the level of a price-sensitive variable is not relevant since prices (or more generally values) are largely arbitrary. A stock split, for example, has the effect of changing price levels but does not typically affect the economic position of the owner of the security. In the context of stock markets, in particular, any price level is thus achievable by *manipulating* the number of stocks outstanding.

The same principle applies when comparing different securities. It does not affect instruments that are unit-priced, such as foreign exchange and commodities. Secondly, and affecting all securities, the time series may not be mean-stationary. The typical remedy of first differencing is already implied in a number of variables, for example, ROC. It does not apply to others, those marked with an asterisk; here, rather than first differencing, the described adjustment is made to lead to a more appropriate metric. It should be noted that such indicators are typically used by comparing them to the price level or a similar indicator, possibly the same with only a different look-back period, when using them in specific rules. The division by price is thus not only meaningful in an analytical sense, it also replicates practical application.

A detailed description of each individual variable is not the objective here; their default values as used for this research are included in Exhibit A1. Readers interested in the formula and suggested application of specific technical variables will find this information easily available (Wealth-Lab [2013]). These technical variables form a fairly representative set across most of the major algorithmic trading platforms.

Wealth-Lab software is used to generate the time series values for the 75 indicators on a one-minutely basis. Four technical variables were immediately removed. The values of these variables are essentially array subscripts, and represent the array position in the time series where an event occurred. This leaves 71 technical variables on which to perform feature selection. Exhibit 6 lists the four array subscript variables that were removed prior to processing.

The remaining 71 technical variables are then imported into MATLAB, where the seven-step process described here is performed.

To implement this process, it is necessary to make judgements concerning the thresholds, that is, dimension-reduction parameters, to be applied. For example,

EXHIBIT 6

Technical Variables Whose Values Represent Array Subscripts

Technical Variable
HighestBar
LowestBar
PeakBar
TroughBar

the threshold used in this article to represent high correlation is 90%, but a different choice could also be considered appropriate by some users. The thresholds applied in this article are for illustrative purposes, and different underlying data and research objectives should be used to inform the decision maker's choices. In particular, lower thresholds would be more appropriate for higher frequencies of data due to the greater effects of noise in this kind of data.

To arrive at a minimum set of technical variables needed to describe the high-frequency EURUSD data, we propose these steps for both the in-sample and the out-of-sample data:

1. Eliminate any variables that are constant throughout the entire period. This step is necessary to remove a number of indicators that use volume in their construction and thus are absent in FX (over-the-counter market) data.
2. Standardize the data (by dividing each variable by its standard deviation, centering it around the mean, and clipping outliers).
3. Eliminate any variables that have a high correlation to another variable. We define *high correlation* to be correlated at the 0.9 level or higher. This is necessary since the PCA procedure would otherwise distribute coefficients among those highly correlated variables, leading to a large number of variables each with a marginal benefit but together contributing significantly to a component. Highly correlated variables thereby make no significant contribution to the overall value of the model. The net effect is a smaller set of variables without a significant loss of variability explained.
4. For the set of technical variables remaining after step 3, create a PCA model.
5. Select principal components such that a high percentage of the variability in the underlying data is explained by the model. In this article, the definition of *high percentage of the variability* in model selection is set to 60%.
6. Sort the underlying coefficient vectors for each component selected in step 4 by the absolute value of the coefficient for each technical variable.
7. Select technical variables from each component selected in step 4 such that a high percentage of each component is explained by the underlying technical variables. In this article, the definition of *high percentage of each component* is set to 60%.

As the coefficients for each of the principal components are expressed in terms of the loadings on the underlying technical variables used to construct the component, and as each component is orthogonal, this process allows for a general and customizable technique for selecting those technical variables that explain the greatest variability in the underlying data.

Principal-component analysis is used to determine variability within a set of data, not variability with respect to a given set of variables. To understand whether a set of variables is persistent in different time periods, it is necessary to conduct the PCA process on multiple subsamples and look for commonality between them. To illustrate this, we run the PCA process on both the in-sample and out-of-sample data and compare the results.

Analysis of Data

The underlying EURUSD in-sample dataset contains 2,484,168 observations (one for each minute), and the out-of-sample dataset contains 2,629,899 observations. Both datasets contain 68 columns: one for each technical variable described in Exhibit A1 with the four variables removed, as described in Exhibit 6, and with the date and time column added.

Step 1 removes all technical variables whose value is constant over the entire period. Exhibit 7 lists the eight technical variables that were removed. All were zero due to volume being specified as part of their construction.

Step 2 standardizes the data. PCA can be used on raw data when all variables are in the same units. Standardizing is preferable when variables are in different units or the variance is substantial, as in this case.

Step 3 removes all technical variables that are highly correlated to other technical variables. Exhibit 8 shows the technical variables removed with the definition of *high correlation* set at 90% or higher.

In the in-sample period, 24 variables were removed, compared to 28 variables removed during the out-of-sample period.

Step 4 creates the PCA model using the set of technical variables left after the constant exclusion (step 1), the variable standardization (step 2), and the correlation exclusion (step 3). Exhibit 9 shows the in-sample and out-of-sample pareto curves for the percentage of variance explained by the PCA models for the first 10 principal components.

EXHIBIT 7

Technical Variables Removed Due to Constant Values

Variables Removed
AccumDist
CADO
CMF
EMV
MFI
MoneyFlow
OBV
VMA

EXHIBIT 8

Technical Variables Removed Due to High Correlation to Other Technical Variables

In-Sample Data	Out-of-Sample Data
ADX	ADXR
AroonDown	ATRP
AroonUp	ATR*
AveragePriceC*	AroonDown
AveragePrice*	AroonUp
BBandLower*	AveragePriceC*
CumDown	AveragePrice*
CumUp	CumUp
DX	DIMinus
EMMinus	DIPlus
FIR	DX
Highest*	EMMinus
KeltnerLower*	FIR
Lowest*	KeltnerLower*
Parabolic*	Lowest*
Peak*	Parabolic*
QStick*	Peak*
ROC	QStick*
RSquared	ROC
RVI	RSquared
Trough*	RVI
TrueRange*	StochD
Vidya*	StochK
Volatility	Trough*
	TrueRange*
	UltimateOsc
	Vidya*
	Volatility*

EXHIBIT 9

PCA Model: Percent of Variance Explained: (A) In-Sample Data, (B) Out-of-Sample Data

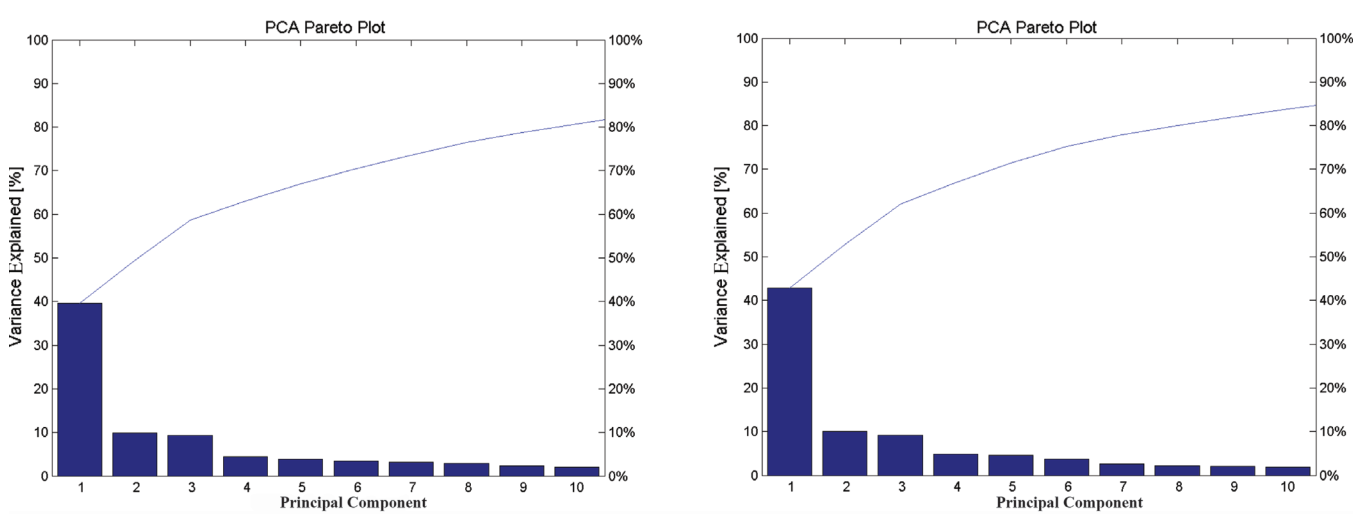


EXHIBIT 10
Minimum Sets of Technical Variables Required for the 60% Correlated Model

Technical Variables Required to Explain In-Sample Variance	Technical Variables Required to Explain Out-of-Sample Variance
ATR*	ADX
ATRP	BBandLower*
BBandUpper*	BBandUpper*
CCI	CCI
CMO	CMO
DIMinus	CumDown
DIPlus	Divergence*
Divergence*	DPO*
DPO*	DSS
DSS	EMA*
EMA*	EMPlus
EMPlus	FAMA*
FAMA*	Highest*
HV	HV
Kalman*	Kalman*
KAMA*	KAMA*
KeltnerUpper*	KeltnerUpper*
LinearReg*	LinearReg*
LinearRegSlope	LinearRegSlope
MACD	MACD
MAMA*	MAMA*
Median*	Median*
Momentum*	Momentum*
MomentumPct	MomentumPct
RSI	RSI
SMA*	SMA*
StdDev	StdDev
StdError	StdError
StochD	StochRSI
StochK	Sum*
StochRSI	TRIX
Sum*	VHF*
TRIX	WilderMA*
UltimateOsc	WilliamsR
VHF*	WMA*
WilderMA*	
WilliamsR	
WMA*	

EXHIBIT 11
Differences between the Sets of Explanatory Variables

Technical Variables Present in the In-Sample Set and Not Present in the Out-of-Sample Set	Technical Variables Present in the Out-of-Sample Set and Not Present in the In-Sample Set
ATR*	ADX
ATRP	BB and Lower*
DIMinus	CumDown
DIPlus	Highest
StochD	
StochK	
UltimateOsc	

Step 5 allows for determining how many principal components are required to explain a given percentage of the variability in the PCA model. Increasing (decreasing) the percentage of variability that must be explained directly increases (decreases) the number of technical variables that will be in the final feature set. For the in-sample data, the number of principal components that must be selected to explain 60% of the variability in the data is five. For the out-of-sample data, the number of principal components that must be selected to explain 60% of the variability in the data is four.

Step 7 allows for determining the amount of each required principal component that is to be explained, after sorting each required principal component coefficient by absolute value (step 6). Since the coefficients of each principal component are expressed in terms of the underlying technical variables' contribution to the component, this step effectively determines which technical variables are members of the final feature selection set for the model.

Exhibit 10 shows the minimum set of technical variables required to explain 60% of the variability of each component, where these components were chosen to explain 60% of total variability of the PCA model.

Exhibit 10 shows the minimum set of technical variables needed to explain 60% of the variance in the underlying EURUSD one-minute dataset over the in-sample and the out-of-sample periods.

In the in-sample period, there are 38 technical variables required to explain the variance in the underlying data from the original set of 75 Wealth-Lab standard indicators. In the out-of-sample period, there are 35 technical variables required to explain the same amount of variance.

Exhibit 11 shows the differences between the technical variables required to explain the variability of both the in-sample and out-of-sample data. Exhibit 12 shows those variables that are common and are required to explain the variability of both the in-sample and the out-of-sample data.

There are 31 technical variables in common between the sets of variables required to explain the in-sample variability and the out-of-sample variability. There are only seven variables that help explain the variability in the in-sample data that are not selected to explain the out-of-sample data, and there are only four variables that help explain the out-of-sample data that are not selected to explain the in-sample data.

EXHIBIT 12

Variables in Common between the In-Sample and Out-of-Sample Sets

Technical Variables in Common to Both Sets
BB and Upper*
CCI
CMO
Divergence*
DPO*
DSS
EMA*
EMPlus
FAMA*
HV
Kalman*
KAMA*
KeltnerUpper*
LinearReg*
LinearRegSlope
MACD
MAMA*
Median*
Momentum*
MomentumPct
RSI
SMA*
StdDev
StdError
StochRSI
Sum*
TRIX
VHF*
WilderMA*
WilliamsR
WMA*

It is particularly interesting that there is such a large overlap in explanatory variables between the in-sample and out-of-sample data, suggesting some persistence in the usefulness of these variables. It is equally interesting, although perhaps not unexpected, that a small number of variables that were useful in one phase of the market are not useful in another.

This study described and demonstrated a technique for dimension reduction in large financial datasets. The technique presented here is prescriptive and easily configured to allow the user to clearly state the required amount of variability that the final feature set must explain.

Variables retained during this process are the minimum set required to explain the chosen amount of variability among the data. The variables that are not chosen are either redundant or explain an extremely small amount of the residual variability. The main advantage of this approach is that it allows the user to decide which variables are likely to be useful in building trading models. Also, as documented earlier, FX high-frequency data contain substantial nonlinearity. Machine learning can be used to develop trading models under nonlinearity, and the input to a machine-learning process is a list of variables to be used. The variables returned by the PCA process are the ideal candidates for this purpose.

CONCLUSION

This article demonstrates the steps required to identify relevant algorithmic trading rule inputs based either on stylized facts or on available technical indicators. In each case, the appropriate technique was applied to EURUSD exchange rates.

In the case of the stylized fact of negative autocorrelation, the process required confirmation of the applicability of the knowledge to the chosen security and the robustness of the findings across time periods, both of which were confirmed for the EURUSD exchange rates. Furthermore, the relevance to building trading rules was established by deriving simple initial rules enabling back-testing simulation and evaluation of the rules from an economic rather than a purely statistical perspective.

In contrast, the dimension-reduction process focused on the discovery of a minimal representative set of technical indicators in the absence of stylized facts or explicit domain knowledge. Using a longer time period of EURUSD exchange rates, a large set of technical indicators readily available to currency traders was reduced to a small set of indicators, each contributing to the major principal components to a significant extent. This technique is particularly useful for traders looking to find technical variables of influence in their chosen trading domain, as well as academics looking for ways to determine which variables to use in filter rule tests and which variables to select as inputs to machine-learning techniques.

To the extent that there is a large overlap between the technical variables required to explain one-minutely EURUSD price movement over the in-sample and out-of-sample period, we conclude that some technical variables are indeed not only capable of helping explain the underlying variability, their explanatory power also appears to be robust over time.

FURTHER RESEARCH

Exhibit 1 introduced the basic methodology for developing high-frequency trading systems. This article has focused on the techniques used for assessing stylized facts and for performing dimension reduction in an orderly manner, both of which fall under the proof of concept stage. The first technique in this stage, correlation analysis, was described in detail in Vanstone [2006].

Next steps for this research are to focus on the design of high-frequency trading rules. Once variables have been correlated, assessed from stylized facts, or learned through dimension reduction, these same variables must be translated into prescriptive rules that can be implemented by algorithmic trading software. This is the purpose of the third step in Exhibit 1 and will be the focus of our future work.

APPENDIX A

TECHNICAL INDICATORS AND PARAMETERS

EXHIBIT A1

Candidate Set of Technical Variables Considered for Feature Selection

Name	Default Parameter Usage	Brief Description
AccumDist	N/A	Accumulation/distribution indicator. Created by Larry Williams (Williams [1986]). Attempts to measure whether accumulation or distribution is occurring in the market.
ADX	Period = 14	Average directional movement index. Created by Welles Wilder (Wilder [1978]). Used to measure the overall strength of a trend.
ADXR	Period = 14	Average directional movement index rating. Created by Welles Wilder (Wilder [1978]). Attempts to measure the strength of price movement in positive and negative directions, as well as the overall strength of the trend.

EXHIBIT A1 (*Continued*)

Name	Default Parameter Usage	Brief Description
AroonDown	Source = Close Period = 20	Created by Tushar Chande (Chande and Kroll [1994]). Measures how long it has been (as a percentage) since prices have recorded a new low within the specified period.
AroonUp	Source = Close Period = 20	Created by Tushar Chande (Chande and Kroll [1994]). Measures how long it has been (as a percentage) since prices have recorded a new high within the specified period.
ATR*	Period = 14	Average true range. Created by Welles Wilder (Wilder [1978]). Average of the true range of price movement, including gaps.
ATRP	Period = 14	Average true range percent. Created by Welles Wilder (Wilder [1978]). ATR expressed as a percentage of the closing price.
AveragePriceC*	N/A	Average of the high, low, and close price.
AveragePrice*	N/A	Average of the high and low price.
BbandLower*	Source: Close Period: 20 StdDevs: 2	Lower Bollinger band. Created by John Bollinger and later published (Bollinger [2001]). A Lower Bollinger band is calculated as a supplied standard deviation below a simple moving average of the price.
BbandUpper*	Source: Close Period: 20 StdDevs: 2	Upper Bollinger band. Created by John Bollinger and later published (Bollinger [2001]). An upper Bollinger band is calculated as a supplied standard deviation above a simple moving average of the price.
CADO	N/A	Chaikin oscillator. Created by Marc Chaikin (Chaikin [2013]). Attempts to measure whether accumulation or distribution is occurring in the market.
CCI	Period: 20	Commodity channel index. Created by Lambert (Lambert [1980]). CCI is designed to identify cyclical movements.
CMF	Period: 21	Chaikin money flow. Created by Marc Chaikin (Chaikin [2013]). An oscillator created from the accumulation/distribution line.
CMO	Source: Close Period: 20	Chande momentum oscillator. Created by Tushar Chande (Chande and Kroll [1994]). Calculates momentum using high and low series in its calculation.
CumDown	Source: Close Period: 4	The number of consecutive periods the underlying value was less than the value a certain number of periods ago.
CumUp	Source: Close Period: 4	The number of consecutive periods the underlying value was greater than the value a certain number of periods ago.
DIMinus	Period: 14	Created by Welles Wilder (Wilder [1978]). Used in conjunction with DIPlus to measure a market's directional movement.
DIPlus	Period: 14	Created by Welles Wilder (Wilder [1978]). Used in conjunction with DIMinus to measure a market's directional movement.
Divergence*	Source: Close MA Type: EMA Period: 9	Computes the divergence of a time series from a specified moving average.
DPO*	Source: Close Period: 20	Detrended price oscillator. Used to attempt to eliminate the trend in prices.

Exhibit A1 (*Continued*)

Name	Default Parameter Usage	Brief Description
DSS	Period 1: 10 Period 2: 20 Stochastic Period: 5	Created by William Blau (Blau [1995]). The double smoothed stochastic indicator applies exponential moving averages to a stochastic indicator.
DX	Period: 20	Developed by Welles Wilder (Wilder [1978]). Attempts to measure the overall and directional strength of a trend.
EMA*	Source: Close Period: 60	The exponential moving average gives greater weight to more recent data.
EMMinus	Source: Close Period: 40	The extreme motion index (minus) counts the number of new lows within the specified period.
EMPlus	Source: Close Period: 40	The extreme motion index (plus) counts the number of new highs within the specified period.
EMV	Period: 14	Created by Richard W. Arms Jr. (Arms [1971]). The ease of movement shows the relationship between volume and price change.
FAMA*	Source: Close FastLimit: 0.50 SlowLimit: 0.05	Following adaptive moving average. Created by John Ehlers (Ehlers [2000]).
FIR	Source: Close Filter: 1,2,2,1	The finite impulse response filter is used to smooth out historical time series data by using the parameters as weight values for the underlying data.
Highest*	Source: High Period: 50	The highest value in the specified time series within the period.
HighestBar	Source: High Period: 50	The actual array subscript to the highest value in the specified time series within the period.
HV	Source: Close Period: 20 Bars per Year Span: 252	Historical volatility computed as a standard deviation of the log of the changes in the underlying.
Kalman*	Source: Close	Created by R.E. Kalman (Kalman [1960]). Used to generate an estimate of future position based on current position and estimates of velocity, acceleration, and other uncertainties.*
KAMA*	Source: Close Period: 10	Created by Perry Kaufman (Kaufman [1995]). An adaptive moving average using noise level in the market to compute trend length.
KeltnerLower*	Period 1: 10 Period 2: 10	Created by Chester W. Keltner (Keltner [1960]). A price channel band plotted below a simple moving average.
KeltnerUpper*	Period 1: 10 Period 2: 10	Created by Chester W. Keltner (Keltner [1960]). A price channel band plotted above a simple moving average.
LinearReg*	Source: Close Period: 20	Least-squares trend line fitted to the specified number of periods for the time series.
LinearRegSlope	Source: Close Period: 20	Slope of the linear regression line.
Lowest*	Source: Low Period: 50	The lowest value in the specified time series within the period.
LowestBar	Source: Low Period: 50	The actual array subscript to the lowest value in the specified time series within the period.
MACD	Source: Close	Created by Gerald Appel (Appel [2005]). A momentum oscillator used to trade trends.

Exhibit A1 (*Continued*)

Name	Default Parameter Usage	Brief Description
MAMA*	Source: Close FastLimit: 0.50 SlowLimit: 0.05	Created by John Ehlers (Ehlers [2001]). The MESA adaptive exponential moving average.
Median*	Source: Close Period: 14	The middle value in the distribution of values during the specified period.
MFI	Period: 20	A momentum based measure of the money flowing in and out of a security.
Momentum*	Source: Close Period: 14	The difference between the current price and the price the specified period of bars prior.
MomentumPct	Source: Close Period: 14	The current price divided by the price the specified period of bars prior.
MoneyFlow*	N/A	Average price multiplied by volume.
OBV	N/A	Created by Joseph Granville (Granville [1975]). A momentum measure of positive and negative volume flow.
Parabolic*	AccelUp: 0.02 AccelDown: 0.02 AccelMax: 0.20	Created by Welles Wilder (Wilder [1978]). Parabolic sets a trailing stop level that moves up or down based on an acceleration factor.
Peak*	Source: Close Reversal Amount: 7 Peak/Trough Mode: Percent	The value of the highest peak during a specified period, subject to a reversal factor.
PeakBar	Source: Close Reversal Amount: 7 Peak/Trough Mode: Percent	The actual array subscript to the peak value in the specified time series within the period.
Qstick*	Period: 24	Created by Tushar Chande (Chande [2001]). The moving average of the difference between the open and close price.
ROC	Source: Close Period: 30	The percentage change in the security price over the specified period.
RSI	Source: Close Period: 20	Created by Welles Wilder (Wilder [1978]). Relative strength indicator is a measure of the market's internal strength.
RSquared	Source: Close Period: 30	The percentage of price action over the specified period explained by the regression line.
RVI	Period: 20	Created by John Ehlers (Ehlers [2001]). The relative vigor index measures the average difference between closing and opening prices normalized to the average trading range.
SMA*	Source: Close Period: 20	Simple moving average.
StdDev	Source: Close Period: 14	Standard deviation.
StdError	Source: Close Period: 20	Standard error of the estimate for a linear regression line over the specified period.
StochD	Period: 14 Smooth: 5	Created by George Lane (Lane [1984]). Compares the closing price to the price range over a specified period.
StochK	Period: 14	Created by George Lane (Lane [1984]). Measures the extent that price closes in the upper or lower area of the trading range.
StochRSI	Source: Close Period: 14	Created by Tushar Chande (Chande [2001]). A combination of the stochastic and relative strength indicators.
Sum*	Source: Close Period: 14	Returns the sum of values over a specified period.
TRIX	Source: Close Period: 10	ROC of a triple exponentially smoothed EMA over the specified period.

Exhibit A1 (*Continued*)

Name	Default Parameter Usage	Brief Description
Trough*	Source: Close Reversal Amount: 7 Peak/Trough Mode: Percent	The value of the lowest trough during a specified period, subject to a reversal factor.
TroughBar	Source: Close Reversal Amount: 7 Peak/Trough Mode: Percent	The actual array subscript to the trough value in the specified time series within the period.
TrueRange*	N/A	True range price has traveled between the current and the prior period.
UltimateOsc	N/A	Created by Larry Williams (Williams [1985]). The weighted sum of different oscillators.
VHF*	Source: Close Period: 24	Created by Adam White (White [1991]) to determine if prices are in a trending or congestion stage.
Vidya*	Source: Close StdDev Period: 10 Alpha: 0.10	Created by Tushar Chande (Chande [2001]). Similar to an exponential moving average using different periods determined by volatility.
VMA*	Source: Close Period: 60	Volume-weighted moving average.
Volatility*	EMA Period: 24 ROC Period: 12	Created by Marc Chaikin (Chaikin [2013]). The ROC of the EMA of the difference between high and low prices.
WilderMA*	Source: Close Period: 24	Created by Welles Wilder (Wilder [1978]). Moving average similar to EMA.
WilliamsR	Period: 30	Created by Larry Williams (Williams [1979]). A comparison of the close price to the high-low range over the specified period.
WMA*	Source: Close Period: 40	Weighted average of prices, with more recent points more heavily weighted.

ENDNOTES

The data used in this research were supplied by the Securities Industry Research Centre of Asia-Pacific (SIRCA) on behalf of Reuters and the ASX. The authors also gratefully acknowledge the financial assistance provided by Bond University through the Vice Chancellors Research Grant.

Furthermore, the research presented was greatly improved by the comments and suggestions given by Professor Tom Smith and several anonymous referees and participants at the 2012 Australasian Finance and Banking Conference, where Section 2 of the research outcomes was discussed.

[1]As of May 2013.

REFERENCES

Aldridge, I. *High-Frequency Trading.* Hoboken, NJ: Wiley, 2010.

Andersen, T.G., T. Bollerslev, F.X. Diebold, and P. Labys. "The Distribution of Realized Exchange Rate Volatility." *Journal of the American Statistical Association*, Vol. 96, No. 453 (2001), pp. 42-55.

Appel, G. *Technical Analysis: Power Tools for Active Investors.* New York, NY: Pearson Education, 2005.

Arms, R.J. *Profits in Volume.* Marketplace Books, 1971.

Bianco, S., and R. Reno. "Dynamics of Intraday Serial Correlation in the Italian Futures Market." *Journal of Futures Markets*, Vol. 26, No. 1 (2006), pp. 61-84.

Blau, W. *Momentum, Direction, and Divergence: Applying the Latest Momentum Indicators for Technical Analysis.* Wiley Traders Advantage Series. Hoboken, NJ: John Wiley & Sons, 1995.

Bollinger, J.A. *Bollinger on Bollinger Bands.* New York, NY: McGraw-Hill, 2001.

Cao, L.J., K.S. Chua, W.K. Chong, H.P. Lee, and Q.M. Gu. "A Comparison of PCA, KPCA and ICA for Dimensionality Reduction in Support Vector Machine." *Neurocomputing*, Vol. 55, No. 1-2 (2003), pp. 321-336.

Chaikin, M. *Chaikin Power Tools*. 2013, http://www.chaikinpowertools.com/.

Chande, T.S. *Beyond Technical Analysis: How to Develop and Implement a Winning Trading System, 2nd ed*. Hoboken, NJ: John Wiley & Sons, 2001.

Chande, T.S., and S. Kroll. *The New Technical Trader.* Wiley Finance. Hoboken, NJ: Wiley, 1994.

Cont, R. "Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues." *Quantitative Finance*, Vol. 1, No. 2 (2001), pp. 223-236.

Dacoronga, M.M. *An Introduction to High-Frequency Finance.* San Diego, CA: Academic Press, 2001.

Ehlers, J. *MESA Software*. Cited December 2000, www.mesasoftware.com.

——. "Technical Analysis of Stocks and Commodities." *MESA Adaptive Moving Averages*, Vol. 19, No. 9 (2001), pp. 30-35.

Evertsz, C.J. "Fractal Geometry of Financial Time Series." *Fractals*, Vol. 3, No. 3 (1995), pp. 609-616.

Fisher, A., L. Calvet, and B. Mandelbrot. "Multifractality of Deutschemark/U.S. Dollar Exchange Rates." Discussion paper 1166, Cowles Foundation, 1997.

Galluccio, S. "Scaling in Currency Exchange." *Physica A*, Vol. 245, No. 3-4 (1997), pp. 423-426.

Gencay, R., F. Selcuk, and B. Whitcher. "Scaling Properties of Foreign Exchange Volatility." *Physica A*, Vol. 289, No. 1-2 (2001), pp. 249-266.

Granville, J. *Granville's New Strategy of Daily Stock Market Timing for Maximum Profit*. 1975.

Hsieh, D.A. "Testing for Non-Linear Dependence in Daily Foreign Exchange Rates." *Journal of Business*, Vol. 62, No. 3 (1989), pp. 339-368.

Interactive Brokers Commissions. 2011, http://www.interactivebrokers.com/en/p.php?f=commission.

Kalman, R.E. "A New Approach to Linear Filtering and Prediction Problems." *Journal of Basic Engineering*, Vol. 82, No. 1 (1960), pp. 35-45.

Kaufman, P. *Smarter Trading: Improving Performance in Changing Markets.* R.R. Donnelley & Sons, 1995.

Keltner, C.W. *How to Make Money in Commodities*, 1960.

Lambert, D. "Commodity Channel Index." In *Futures*. Chicago, IL: Highline Media, 1980.

Lane, G. "Lane's Stochastics." In *Technical Analysis of Stocks and Commodities*, 1984.

Lo, A.W., and C. MacKinlay. "Stock Market Prices Do Not Follow Random Walks: Evidence from a Simple Specification Test." *Review of Financial Studies*, 1 (1988), pp. 41-66.

Muller, U.A. "Statistical Study of Foreign Exchange Rates, Empiricial Evidence of a Price Change Scaling Law, and Intraday Analysis." *Journal of Banking and Finance*, Vol. 14, No. 6 (1990), pp. 1189-1208.

Park, C.H., and S.H. Irwin. "What Do We Know About the Profitability of Technical Analysis?" *Journal of Economic Surveys*, Vol. 21, No. 4 (2007), pp. 786-826.

Sewell, M. "Characterization of Financial Time Series." UCL Research Note RN/11/01, 2011.

Strong, R.A. "A Behavioral Investigation of Three Paradigms in Finance." *Northeast Journal of Business and Economics*, 1988, pp. 1-28.

Taylor, M.P., and H. Allen. "The Use of Technical Analysis in the Foreign Exchange Market." *Journal of International Money and Finance*, 11 (1992), pp. 304-314.

Vanstone, B. *Trading in the Australian Stock Market Using Artificial Neural Networks*. Bond University, 2006.

Vanstone, B., and G. Finnie. "An Empirical Methodology for Developing Stock Market Trading Systems Using Artificial Neural Networks." *Expert Systems with Applications*, Vol. 36 (2009), pp. 6668-6680.

Wealth-Lab. 2013, www.wealth-lab.com.

White, A. "Vertical Horizontal Filter." In *Futures*, 1991.

Wilder, J.W. *New Concepts in Technical Trading*. Greensboro, NC: Trend Research, 1978.

Williams, L.R. *How I Made One Million Dollars Last Year Trading Commodities*. Windsor Books, 1979.

——. "The Ultimate Oscillator." In *Technical Analysis of Stocks and Commodities*, 1985.

——. *The Secret of Selecting Stocks for Immediate and Substantial Gains*, 2nd ed. Windsor Books, 1986.

Zhou, B. "High-Frequency Data and Volatility in Foreign-Exchange Rates." *Journal of Business and Economic Statistics*, Vol. 14, No. 1 (1996), pp. 45-52.