Paper Summaries for Statistical Arbitrage

Steven K Xu

March 9, 2020

The Kalman Filter

Lacey: Tutorial: The Kalman Filter

This article is an instruction guide on what the Kalman filter is. The Kalman filter assumes a signal is of the form $y_k = a_k x_+ n_k$ and aims to estimate x_k. A Kalman filter uses state space techniques which allows the filter to be either a smoother, filter, or predictor. The ability of the filter to predict is much desired. Under the state space, we assume that the process evolves s.t.:

$$x_{k+1} = \Phi x_k + w_k$$

where Φ is a stationary transition matrix and w_k is process noise. Observations evolve under:

$$z_k = H x_k + v_k$$

where H is the connection between state and measurement, and v_k is measurement noise. We can estimate the next state using:

$$\hat{x_k} = \hat{x'_k} + K_k(z_k - H\hat{x'_k})$$

where $\hat{x'_k}$ is the prior estimate of $\hat{x_k}$. From this we can update our estimator, our believed covariance estimate, and then predict the next state.
The paper also describes the Kalman filter as a chi-square merit function, which means it has similarities to a least sqwuared fit but in a recursive form.

Sarkka et al.: Time Series Prediction by Kalman Smoother with Cross-Validated Noise Density

Sarkka et al. assume there are some time series with an underlying linear filtering model and a Gaussian noise. Thus, they use a Kalman filter to smooth noise time series data. They assume there is a long term linear dynamic model with a derivative obeying a Brownian noise process, and a short term process driven by an autoregressive process. They perform a long term prediction by running a Kalman filter over the data, and then running a Kalman smoother over the filtered result. They perform a short term prediction by filtering and smoothing residuals determined by their models.
The difference between a Kalman filter and a Kalman smoother is that a Kalman filter is conditioned on observations $y_{1:k}$, while a Kalman smoother is conditioned on all measurement data $y_{1:T}$.

In summary, this paper provides a way to infill missing data in-sample, but for the most part only captures the trend. It is probably not suitable for predicting the future because the smoother is conditioned on all the measurement data for a given period, and so may be pinned to 'known' data points, but we don't know the future.

Morrison & Pike: Kalman Filtering Applied to Statistical Forecasting

Morrison and Pike claim that the Kalman filter can be applied to statistical forecasting. They assume they are forecasting a time series with a varying mean and additive noise, driven by unknown variables and an unknown state vector. The Kalman Filter is used to estimate an optimal state vector that produces the observed result. The paper aims to create a general methodology for doing so. They perform a comparison of the Kalman filter with a Least Squares model to demonstrate that it may be superior to least squares. The advantage of the Kalman filter is that it is not necessary to assume that model coefficients are stationary.

Trend Detection

Yasar & Ray: Trend Detection and Data Mining via Wavelet and Hilbert-Huang Transforms

The paper describes the formulation of wavelet transforms and Hilbert-Huang transforms. A wavelet transform is essentially an adjustable windowed Fourier transform. It is defined as:

$$\Psi(s,t) = |s|^{-p} \int_{-\infty}^{\infty} f(\tau)\Psi^*(\frac{\tau - t}{s})d\tau$$

where Ψ^* is the 'mother wavelet', and t, s and τ are various time components. Wavelet transforms provide a way to reveal the superpositions of different structures on different time scales at different times, or on different spatial scales at different times. The Hilbert-Huang transform of a signal is its convolution with $1/t$.

$$H(t) = \frac{P}{\pi} \int_{-\infty}^{\infty} \frac{x(\tau)}{t - \tau} d\tau$$

It can estimate instantaneous frequencies in the data. It can demonstrate the Hilbert energy spectrum by representing amplitude and instantaneous frequency as functions of time. It can reveal the superposition of different structures at different frequencies at different times, or different energies at different locations, similar to the wavelet transform.

Bruder, Dao, & Roncalli: Trading Strategies with L1 Filtering

Bruder et al. demonstrate various filtering schemes upon the S&P500 from 2007 to 2011. These filtering schemes are meant to denoise the signals by applying a penalty to the numerical step difference in the signal. The scheme is such that a signal is composed of trend and noise:

$$y_t = x_t + \epsilon_t$$

and a filter works by seeking to minimize some objective function. For a filter under the assumption that the process is mean reverting, this objective function would be:

$$\frac{1}{2}\sum_{t=1}^{n}(y_t - x_t)^2 + \lambda \sum_{t=2}^{n} |x_t - x_{t-1}|$$

The paper covers the details of how to calibrate the regularization parameter via cross validation, and applies it to momentum strategies via backtesting and dynamic estimation of optimal filters.

Bruder, Dao, Richard, & Roncalli: Trend Filtering Methods for Momentum Strategies

This paper studies trend filtering methods, which are widely used in momentum strategies. It reviews various estimators for the trend of a time series, such as the trend cycle model, linear filtering, least squares filters, Kalman filters, non-parametric regression, L_1 filtering, and wavelet filtering. It also reviews multivariate filtering, as well as other practical considerations such as error correction, calibration of parameters, and measurement of efficiency. It also discusses the use of filtering as a predictive tool.

Roncalli & Weisang: Tracking Problems, Hedge Fund Replication, and Alternative Beta

The authors consider hedge fund replication as a general tracking problem and solve it using Bayesian filters. It is unclear whether indexes that can replicate the hedge fund industry. They apply particle filters to capture non-linearity and non-normality in hedge fund returns. They propose an investment strategy that invests both in highly liquid alternative beta and in less liquid investments.

The hedge fund industry has delivered higher Sharpe ratios than buy/hold strategies on traditional assets in the 15 years before 2008. Some investors want those returns without having to deal with the lack of transparency, poor liquidty, and management fees. Under asset based style factor model, there is some combination of factors that replicates a hedge fund's portfolio.

Global tactical asset allocation attempts to exploit short-term market inefficiencies by investing in various markets with the goal to profit across relative movements in those markets. The authors use Bayesian filters to capture tactical allocation. The authors use the Kalman filter to determine plausible weights for a replicating portfolio of a hedge fund index.

The authors define an alternative alpha and beta which is the difference not between the fund and the market but the replicating portfolio and the fund. The authors explain the success of hedge funds between 2000 and 2003 as because of high exposure to the directional equity market.

In conclusion, the authors formalize a framework for hedge fund replication as tracking problems with Bayesian filters.

d'Aspremont: Identifying Small Mean Reverting Portfolios

The author constructs the problem of forming mean reverting portfolios from multivariate time series as a sparse canonical correlation analysis, with a constrained number of included assets. Mean reversion is a good indicator of predictability in financial markets, and by constraining the number of included assets transaction costs are reduced and interpretability is increased.

Sparse portfolios are optimally mean reverting portfolios with few assets. Penalizing for sparsity makes a wider price range, highlighting more significant market inefficiencies. These arguments also apply symmetrically to momentum, but by maximizing predictability instead of minimizing. For some assets S, a portfolio can be constructed such that

$$dP_t = \lambda(\bar{P} - P_t)dt + \sigma dZ_t$$

$$P = \sum_{i=1}^{n} x_i S_{ti}$$

The goal is to maximize λ by adjusting x_i s.t. $||x|| = 1$ and there is restriction on the number of nonzero coefficients.

Canonical decompositions are used to do this. A sparse decomposition algorithm to do this is:

$$\lambda^{max}(A, B) = max\frac{x^T Ax}{x^T Bx}$$

They also implement greedy searches, a relaxation, and parameter estimation.

Backtesting

Bailey et al.: Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out of Sample Performance

The authors define backtest overfitting as when an investment strategy is tuned such that it performs well on the data it was designed on but not outside of that sample data. An overfit model has been tuned to target particular observations rather than general structure. In this paper, the authors show that relatively few trials are necessary to detect spuriously high backtested performance, as well as computing the minimum backtest length that an investor should require for a given number of trials. The more trials executed by a financial analyst, the higher the in sample an investor should require. The higher the in-sample Sharpe ratio, the more likely it is that its standard deviation is of the same order.

Bailey et al.: The Probability of Backtest Overfitting

This paper develops a framework that estimates the probability of backtest overfitting in the context of investment simulations. The authors reject the test set approach because it is inadequate for small samples and may be detrimental to the strategy's design. They also do not find approaches that model the underlying financial variable ideal because the underlying model may also overfit. Instead, they propose a model that requires only time series data of backtested performance to determine probability of backtest overfitting. They define a strategy selection process as overfitting if the expected performance of strategies selected in sample is less than the median performance rank out of sample of all strategies. They test their method, called CSCV, by generating large numbers of combinations of in sample subsets and determining the proportion that have overfitting. If there is a high probability of backtest overfitting, in-sample optimization weakens out of sample performance.

Machine Learning

Khaidem et al.: Predicting the direction of stock market prices using random forest

The authors propose treating forecasting as a classification problem using random forests. Their motivation is that it is easier to forecast whether the stock will move up or down instead of stock price itself. Their model incorporates the relative strength index, the stochastic oscillator, Williams %R, moving average convergence divergence, price rate of change, and on balance volume as features. Their method is a standard implementation of random forest, which generates a large number of random decision trees with random features and uses them as a voting classifier. The authors claim they have good out of sample performance with out of sample error rate converging towards 5% for a 90 day forecast at around 40 estimators.

In practice, even with the given features, the random forest is fundamentally a black box and can only be evaluated on historical performance.

Szklarz et al.: Application of Support Vector Machine on Algorithmic Trading

The authors use a support vector machine to predict whether the market will move up or down. They use relative strength indicator, moving average convergence divergence, momentum, and Bollinger Bands as their technical inputs. Using a very niely hand selected set of training and test data, the authors feed the data to a SVM-KM, which combines the SVM with a k-means clustering process. In conclusion, the authors find that the SVM doesn't work well in an up-trending market and doesn't beat buy and hold, but does well in down-trending markets, and that this strategy should be used in down trending markets.

Overall, it seems that this paper has a number of problems in that they do not sufficiently motivate the time periods which they are using as their test and training data, they do not explain how exactly their SVM-km algorithm works, even though they describe an SVM, and in many ways their paper simply describes their result.

Krauss et al.: Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P500

This paper implements and analyzes the effectiveness of deep neural networks, gradient-boosted-trees, random forests, and a combination of these methods on the S&P500. The models are trained on lagged returns of all stocks in the S&P500 and generate one day ahead trading signals of the probabilities of stocks to outperform the market. High probabilities become long signals and the lowest probabilities become short signals. They find out-of-sample returns exceeding 0.45% per day, posing a challenge to semi-strong market efficiency.

Deep neural networks consist of an input layer, several hidden layers, and an output layer. Each is composed of neurons. Their network has 31 inputs and 2746 parameters, which is small in the context of deep learning.

Gradient boosting trees sequentially apply weak learners to re-weighted versions of the training data such that each classifier focuses on hard-to-classify examples. Then, the series are put together in a voting classifier to determine the overall classification.

Boosting consists of successively fit shallow decision trees.
Random forests consist of deep but decorrelated trees on different samples of data.
Finally, they also use an equally weighted ensemble that takes the probabilities forecasted by each of their models and averages them.

Visser: depmixS4: An R Package for Hidden Markov Models

This is a specification for a specific implementation of Hidden Markov models. It can fit dependent mixture models, latent class regression, and finite mixture models. It is intended for modeling time series data with the aim of characterizing underlying transition processes.
A dependent mixture model assumes that at any given time, observations are distributed as a mixture with some number of states, and that time dependencies between observations arise from time dependencies between states. These dependencies between states are assumed to follow a first order Markov process.

Nguyen: Hidden Markov Model for Stock Selection

Hidden Markov Models can predict hidden regimes of observation data. The authors use HMM to make monthly regime predictions of inflation, industrial production index, the S&P500, and the VIX. They calibrate the model monthly and predict the next month's regimes. They look at the historical data for time periods similar to the forecasted regimes, and select stocks which performed well in those regimes and compare to a benchmark of the S&P500. According to them, they had a 14.9% return over 1999 to 2014.
An HMM's parameters are a transition matrix A, an observation probability matrix B, and a probability vector p of the odds of being in state S at time t. An HMM needs to determine probability of observation given model parameters, the best sequence given observations under model parameters, and how to best calibrate HMM parameters to maximize the probability of observation data given parameters.

Arbitrage

AQR: A Century of Evidence in Trend-Following Investing

Trend-following has strong positive returns and realized a low correlation to traditional asset classes for each decade for more than a century. This is true for the past 100 years. A basic trend following strategy is time series momentum, longing positive returns and shorting negative returns. They construct a simple equal weighted combination of 1, 3, and 12 month time series across 59 markets for 4 asset classes. The strategy is always long or short every market, but targets the same amount of volatility to avoid too much risk from one market. Across the full sample the strategy has a Sharpe Ratio of 1.0 across a century from 19093 to 2012, which beats the S&P and is anticorrelated with it.

Kanamura & Fabozzi: A Profit Model for Spread Trading with an Application to Energy Futures

The paper proposes a profit model for spread trading, general enough for any financial instrument, that focuses on the first hitting time probability density of a price spread. Profit is easily calculated if first hitting time probability density is known. It then specifies the model for the energy futures market. Energy futures spreads are shown to follow a mean reverting process. Mean reverting processes have approximately known first hitting time probability density. The expected return of a spread convergence is given by:

$$r_{p,c} = x \int_0^T f_{\tau_{x \to 0}}(t)dt$$

where the function term is the first hitting time density for a price spread process. The total profit for spread trading is the sum of the expectations of profit with convergence and profit without convergence. They assume a spread trade is a mean reverting process and perform empirical trades on the energy futures markets. They find that a backtested profit could be found to be relatively stable. Profits from heating oil and natural gas were positively affected by seasonality. High volatility and long term mean reversion may caues high total profits from pairs trading.

Boguslavsky(a): Arbitrage under Power

The authors present their work on determining when to open and close a position on an asset diverging from its fair price in order to avoid untenable losses. They assume an Ornstein-Uhlenbeck process for the price, and power utility for the wealth of a finite horizon agent. Using the Hamilton-Jacobi-Bellman equation, they solve to find that the optimal strategy is given by:

$$\alpha_t^* = -wxD(\tau)$$

where:

$$\tau = T - t$$

$$v = \frac{1}{\sqrt{1-\gamma}}$$

$$C(\tau) = \cosh v\tau + v \sinh v\tau$$

$$D(\tau) = \frac{C'(\tau)}{C(\tau)}$$

The optimal position is linear both in wealth W and asset price X.
Thus the authors conclude that the position should be opened continuously as spread deviates from 0, that a losing position should be cut as soon as the position spread exceeds current wealth, that under power utility spread widening is always bad, that faster mean reversion and lower noise make traders more aggressive, that averse traders become less aggressive as a finite time horizon approaches, and that is safer to underestimate mean reversion speed and overestimate noise than vice versa. Of course this model assumes perfect liquidity, no market friction, and other highly unrealistic assumptions.

Lubnau: Spread trading strategies in the crude oil futures market

The article tests whether Bollinger Bands can be employed profitably in markets for WTI and Brent crude oil. The Bands are based on a mean-reverting hedge portfolio of WTI and Brent. The author finds that some setups of the system are profitable over every five year period tested. This is a fairly unconvincing result given that backtest overfitting is a common phenomenon. The best Sharpe ratios end up being in excess of three, and are found when a dynamic linear model with Kalman filtering and maximum likelihood estimates of the unknown variance of the state equation are used to constantly update the hedge ratio of the portfolio.
A Bollinger Band for a portfolio follows the form:

$$z_t = \frac{p_t - MA_t}{\sigma_t}$$

where p is the value of the portfolio at the time, and MA and t are rolling mean and standard deviation. Traditionally t = 20, though this paper uses 50, 100, and 200. Signals are generated when this indicator reaches -2 or 2. Exit values are 0, .2.
The paper uses a Kalman filter to dynamically update hedge ratio daily.
Because the Bollinger Band, which does not provide any predictive power, generates a substantial profit and higher Sharpe ratio than a random entry strategy, this suggests that the WTI-Brent market is weakly efficient. However, the strategy is weak to global macro factors.

Dare: Statistical arbitrage in the U.S. treasury futures market

The paper argues that the U.S. treasury futures market is informationally inefficient because a strategy that assumes cointegrated treasury futures prices earns a significant excess over an equally weighted portfolio.
The paper assumes that the yield curve can be treated with a factor model, and uses PCA to extract principal components of futures time series. The paper argues that a factor structure description of the yield curve implies cointegration and that cointegration implies a factor structure. If yield can be modeled by:

$$y_t = Af_t + u$$

Then yield can be written as:

$$y_t = Bh_t + v_t$$

If factors are a multivariate random walk:

$$f_t = f_{t-1} + \phi(L)\epsilon_t$$

Then yield is the sum of a stationary process and a unit root process:

$$y_t = w_t + z_t$$

Thus, cointegration. The underlying assumption is that interest rates and bond futures admit a factor structure, and from there cointegration is naturally. However, this hypothesis cannot be directly verified because price is likely not stationary.
In some ways, this is just an elaborate exploration of what the 3 principal components of the yield curve is. However, somehow, this generates an arbitrage, which means that the market is not informationally efficient.

Leung & Li: Optimal Mean Reversion Trading with Transaction Costs and Stop-Loss Exit

This paper considers optimal timing strategies for trading a mean-reverting price spread and extends the problem with a stop-loss constraint to limit the maximum loss. The paper finds that the entry level for a trade strictly lies above the stop-loss constraint, and that a higher stop-loss always implies a lower optimal take-profit level.
The problem is defined as a double optimal stopping problem of both entry and exit. An investor's timing strategy can be seem in terms of a liquidation region $[b_L^*, +\inf)$, a delay region (L, b_L^*), and a stop-loss region $(-\infty, L]$. An investor's value function is negative in the liquidation and stop-loss region, and so they will immediately close their positions. If stop loss is too high, the investor will liquidate every trade. The problem is formulated such that a higher-stop loss will induce the investor to voluntarily liquidate earlier at a lower take-profit level, but it is unclear whether this is an artifact of the math or an actual observation on how risk-averse traders function.

Misc

Greenwood & Tymerski: A Game-Theoretical Approach for Designing Market Trading Strategies

The authors adopt an approach that uses fuzzy rules instead of hard and fast rules in determining whether they should buy or sell data. They define three metrics: DOJI, which indicates that open and close for the day are within a small percentage of each other, indicating a market reversal or price indecision; NRk, that the range between high and low of the day is less than the previous k-1 days, suggesting volatility contraction; Hook days, when the price opens outside the previous day's range and then reverses, a reaction to overbought or oversold conditions. The authors fuzzify these conditions to map them to [0,1] instead of being the binary condition of 0,1. They also use an evolutionary strategy to determine the best set of weights to give to each of their indicators. Finally, they find that the fuzzy rule base does not make excessive returns but does not lose during bad periods, and can be used to protect capital.

Hamdan et al.: A Primer on Alternative Risk Premia

Alternative risk premia denote all systemic risk factors that have resulted in positive performance in the past. They are extremely complicated. Related concepts are risk factors, skewness premia, market anomalies, and bad times. The authors present a process for identifying and assessing alternative risk premia, their inclusion in investment strategies, and their use in estimating hedge fund performance.
The CAPM is

$$E[R_i] - R_f = \beta_i(E[R_{mkt}] - R_f$$

Its related risk premium is how much the asset outperforms the market by wrt the risk free asset.
The primary lesson of their research is such that alternative risk premia extend the universe of risk premia that cannot be capitalized on, and that the question of whether to invest in certain risk premia becomes more difficult. They also suggest

that the existence of alternative risk premia could explain why active management may perform better than passive management, and expect their work to renew the benchmarking issue.

Andersen & Benzoni: Realized Volatility

Realized volatility is important because it's an estimator of return volatility after the fact without any measurement error. It is also conceptually related to the cumulative expected variability. Asset prices reveal a lot about expected volatility but not expected return. Vol is calculated with an estimator, and as number of observations $n \to \infty$

$$\sqrt{nK}(\hat{\sigma_n}^2 - \sigma^2) \to N(0, 2\sigma^4$$

so vol of vol is known. Quadratic variation of return and realized volatility are related in such a way that:

$$\sqrt{nk}(\frac{RV(t,k;n) - QV(t,k)}{\sqrt{2IQ(t,k)}}) \to N(0,1)$$

where $IQ = \int_{t-k}^{t} \sigma^4(\tau) d\tau$. Further extension of these results can be applied to volatility forecasting, model specification, and violations of the no-arbitrage condition.

Gueant et al.: Dealing with the Inventory Risk: A solution to the market making problem

A market maker is a participant that provides liquidity and maintains inventory. Market makers need to determine optimal quotes that they provide to other market participants. For market makers, setting a bid/ask spread drives buying and selling. The optimal bid/ask spread is given by:

$$\psi^*(t,q) = -\frac{1}{k} ln(\frac{v_{q+1}(t)v_{q-1}(t)}{v_q(t)^2} + \frac{2}{\gamma} ln(1 + \frac{\gamma}{k})$$

Unfortunately, since this depends on utility, which is observable, approximations are necessary.

Optimal quotes are independent of almost time. The bid-ask spread widens with variance because the market maker wants to earn more per trade to compensate for increased inventory risk. Optimal quotes drift in the same direction as expected price drift because market makers want a fair price. As liquidity increases, spread tightens because inventory risk decreases. Dependence on the utility parameter is ambiguous.

Goshal & Roberts: Optimal Market Making under Inventory Risk and Adverse Selection Constraints

Market makers need to optimize profit while minimizing inventory risk, but avoiding inventory is most easily done by setting a wide bid/offer spread, which reduces profitability. This paper aims to address adverse selection in OTC markets with counterparty tuned prices via a model that uses the entropy of a high dimensional, probabilistic representation of connection behavior.

Inventory risk stems from variations in an asset held. Adverse selection risk is client-specific, and may not fit in the Hamilton-Jacobi-Bellman approach. Adverse selection arises from the risks involved with information asymmetry. The innovation of this paper is quantifying adverse selection via entropy in a way that will fit within differential equations governing dynamic inventory risk management. Differential entropy is given by:

$$H(i) = \frac{1}{2}log((2\pi e)^D|K_i|$$

where K_i is the covariance matrix of connection/counterparty i. Adverse selection corresponds to a highly entropic probability distribution. Low entropy counterparties have predictable trading activity which makes managing inventory risk easy.

Martin: Optimal multifactor trading under proportional transaction costs

This paper considers how proportional, not fixed, transaction costs affect the trading of a synthetic asset following a diffusion process. The no-trade zone is found to be proportional to the cube root of the transaction cost, and also proportional to the 2/3 power of the volatility of the target position. Thus, a faster trading strategy is more buffered than a slower one. This theory is set in continuous time.

This is important because strategies need to make money outside of costs. Properly buffering a strategy is necessary to deal with high transaction costs. The buffer is the area around the target position in which the position must be held, presumably because transaction costs would exceed profit, called a no-trade (NT) zone. There is a discrete-trade (DT) zone in which one will trade to reach the NT zone. The point of the buffer is to prevent trading backwards and forwards on small amounts. If it is too narrow, there will be too much overtrading and too many costs, but too wide and trading is performed far too rarely.The paper finds that he buffer is wider when trading speed is higher. The buffer is thinner when volatility of the asset is higher, but so is the target position, so the buffer width remains about the same proportion. The rule for optimal buffer seems to overestimate optimal width for low costs.

Almgren: Optimal Trading with Stochastic Liquidity and Volatility

This paper deals with optimal trade scheduling. Under high-frequency trading schemes, there will be a price difference between the price that arrives into the trading system and the price an order is executed for. Thus, this trade cost is a random variable.

The paper formulates this problem into the HJB equation and solves it numerically instead of analytically. The paper demonstrates that trading speed depends on market state ξ and time to expiration $\tau = (T - t)/\delta$ as a multiple of market relaxation time δ. The paper finds that when τ is small, trade rate increases. When ξ is large positive, market impact is high and volatility is low, so an optimal strategy trades slowly. When ξ is large negative, market impact is low and volatility is high, so an optimal strategy trades rapidly. These are in the parameters of ξ, but they also appear in the numerical solution of the HJB equation.