DISCRETE TIME MARTINGALE
and
A STATE SPACE APPROACH FOR MODELING FACTOR RETURNS

Isabella Lai | January. 24th 2024

Sisyphus' Way

This paper introduces a state-space approach in predicting asset returns. The content is structured as follows: Section One presents the background — it could have been slightly longer; Section Two describes the model — its premise, its mechanism, and its current base-form implementation; Section Three offers a heuristic demonstration of the model's martingale property, with the nice features that follows as well as the practical implications pending until another time; Section Four describes a toy application of our model using the realized returns of the five factors from the 2014 Fama and French 5-factor model, the results we have obtained, and the performance analysis on this model; the last section concludes with some comments and an outline of future work.

One

Our program is essentially a recursive state estimator that handles nonlinearities by deter ministically sampling a set of representative points for mean and covariance approximations (Julier and Uhlmann 1997). A thorough explanation is necessary to demonstrate that the performance it delivers is not mere coincidental, given the lesser popularity of state space models. The section consists of pieces that addresses some possible queries and comments.

Prediction vs. Prophecy

Prophecy is the prediction of a future moment, whose value you have no agency toward; prediction is the imputation of a missing value — what has been determined, although currently unknown. A prophet has a vague feeling about what they expect to happen in the future and seeks confirmatory evidence from all available sources. Sometimes, the prophet gets lucky. Sometimes the prophet gets lucky. At any point in time, there are prophets who

win, as the rules of the game dictate that the winner is determined not by correctness, but by simply outcompeting the losers.

Observed vs. Uncovered

Latent factor analysis, regression, and their combination are the primary methods in predicting asset returns (Fama and Stern 2016). The difference between latent factor analysis and regression with pre-specified factors lies in the assumptions behind the data, and the level of granularity that they assume to be knowable: the former is a composite, unsupervised approach has less interest in the causes recognizable by the naked eyes; the latter is never only confirmatory, but the causes must be verbalizable.

One advantage of the former approach is that it works well with raw, high dimensional data, rich of both structure and volume, but to be fully exploratory one has to consider all the possible features that have potential predictive power, and when it is almost the case that each observation becomes a feature on its own (Lo et.al 2010) and when the number of features exceeds the number of observations (Beyer et.al 1998), which, both are plausible when the prior is flat and uniform, the the p>>n problem arises when the number of variables (p) significantly exceeds the number of observations (n). In such cases, standard PCA faces challenges in possible overfitting and instability in the covariance matrix estimation. Overcoming p>>n would need regularization or additional dimensionality reduction (Dillon et.al 2011), but both introduces additional risks in possible distortions in scales, which PCAs are specifically sensitive to.

By contrast, the handful number of predefined factors do have an advantage that their presence has been verified consistently by people of conflicting perspectives. They are stable, they are interpolative, and therefore overfit isn't a matter of urgency. Occasionally, some can even be real.

State-Space vs. Regressions

A state space approach is preferable when autocorrelation needs to be explicitly modeled — deal with random walks, make short-term forecasts, or uncover the hidden processes that has the potential to hold significance in our analysis.

Regression-based models may censor the correlated errors, but oftentimes, the errors are present for legitimate reasons. In such cases, regressions fall short in capturing these aspects with the same level of accuracy — regression penalizes both randomness and chance, while state space models accept the system the way it is. A state space models is designed to capture stochasticity (see Note. 2).

In regressions, differencing for near unit root cases are difficult to assess as they result in substantial size distortions for tests for cointegration (Elliot 1995), who balances the instability of correlations during long-term forecasts — the task that regressions are supposed to excel at. The data is transformed in its entirety by a simple differencing and appears almost arbitrary, although nothing has been changed in the underlying.

Discrete vs. Continuous

State space models are formulated on the premise of discrete-time step, that time goes at one second per second — the dependency between the present state and the history is captured through, and only through the immediate state that precedes it. Common regression models typically assumes continuity — the value of the variable at the later state is treated as a natural subsequent of all previous state.

State space models run smoothly when changes are naturally discrete or occur at distinct intervals, such as digital signal processing or the market as it is actually implemented — its pulses can only be measured at each transactions. There are fewer analytical solutions available than the problems we face; and there are fewer causes conceivable than what we attempt to verbalize, and hence it comes the question of time on the x-axis, and hence the question of the price of an entire asset rather than each factor being priced on the y-axis, hence the problem of whether being perceived as a single entity actualizes an equity has an entitlement on its own.

Point vs. Distribution

State space models are used for sequential estimation. Very often, it models the full condi tional distributions of states given observations. Ideally, uncertainties, and measurable risks are incorporated at each steps of it yields another predictive distribution. In such a state space model, OLS is an intermediary in which all conceivable paths are at least considered.

Regression-based models treat least square as the intermediary of parameter estimation. It focuses on estimating the conditional mean or median of the response given predictors. The outcome is typically point predictions, while the entire predictive distribution is less of an emphasis. In the point-estimate approach, when an OLS estimate of an autoregressive parameter is close to 1, it tends to underestimate the true value (Quenouille 1949), and the solution was to adjust the estimate upwards (Andrews 1993), but like many other methods of debasing, this may introduce additional coloring of the results, whose risk may not be compensated but magnified when testing in realtime. Risk-adjusted return provides a less insufficient picture, but standard deviation is itself sensitive to scales, yet there is no probabilistic bankruptcy or risk-adjusted ruin— the losses cannot be recovered.

Closed-forms vs. Proxies

-

Two methods are available to us: the Extended Kalman Filter and the Sigma-Point Estima tor. The former is frequently used in econometrics (Javaheri et.al 2003; Kellerhals 2013). The Extended Kalman Filter linearizes a nonlinear system by computing the Jacobian matrices. It is time-intensive, and it's an additional opportunity to introduce errors, scale-sensitive biases, and uncertainties. As such, it is limited to a narrow range of nonlinearity and in cases when the distribution is unimodal.

The Unscented Estimator (Julier and Uhlmann 1997; Julier 1998) eliminates the necessity for explicit linearization. It selects just enough values of observations that enables the recovery of the probability distribution of state variables. This approach effectively captures both the mean and covariance of the state, thereby circumventing the Jacobian calculation. It is numerical. Multimodal distribution isn't a barrier. The answer is a proxy one but guaranteed to be the conditionally most optimal, as opposed to the closed-form approach in both the Extended Kalman Filter and in most of the regression-based methods. As we will see in Section Two, our points for recovering the PDF can capture up to the third moment when the distribution is arbitrary, and to the fourth when it is symmetric.

Diffusion vs. Filtration

Diffusion describes the movement of particles in response to concentration gradients until the concentration is uniform — for equilibrium is what the system aim for. Filtration involves a porous barrier, typically a membrane or a layered material, that allows certain parti-

cles to pass through while blocking others. Even within the same system, there are asymmetries between the particles in terms of the information they have access to. Unlike diffusion, which is a passive process driven by natural concentration gradients, filtration often requires a cause that propels the substance through the filter medium.

There are many ways to entertain on the metaphors. In filtration, the necessity of a first cause may suggest that what's driving the returns must be factors measurable, or that the factors are ultimately derived from men who imposes their will onto the particles. The presence of a porous barrier implies either the transaction costs aren't negligible and aren't evenly distributed, or that the time delay and asymmetry of information access means during the purification, there may be inefficiencies exploitable, but that is a self-defeating effort, since such opportunities exist only when the porous is still functioning, and taking the advantage undermines its ability to filter out the errors. Metaphors can only take us so far.

Errors and Mistakes

In a standard regression-based estimate, the error is the entire residual over time. There is usually no explicit attempt to dissect the error into different sources, although they can be divided into types. Still, it is not common to consider the error as a nonlinear accumulation over time.

One example of such consequence lies in the early stage of valuation. In discounted cash flow (DCF) models, incorrect calculations of terminal value, enterprise value, and equity value are common, and can materially change an analyst's target prices. These are errors of the measurements, but they are often censored once being discovered. A less solvable one is that the key aspects of DCF construction and execution may not be fully transparent to either the executors or the ones they report to, leading to cases such as overly-optimistic forecasts, especially in the terminal year (Green et.al 2016). This is an error of processing — it can sometimes be easier to identify, but an analytical solution might be impossible, and proxy ones don't age well.

-

A convenient method is mitigate both the errors and the mistakes is to systemize the treat ment of the errors, as in certain state space approaches. In state space models, errors are treated with higher degree of granularity and classified into two major categories: measurement errors, or generic errors, and process (or system) errors, or in the areas we are interested in, manmade mistakes. The errors are not to be minimized only after preliminary results are gathered, but used as information valuable for the adjustments in the next estimation.

A Recursive Least Square Filter, like a Kalman filter, is an algorithm that estimates the state of a process by minimizing the mean squared error (MSE). It's based on state space models, described by differential or difference equations. These models use state variables, represented in matrices, to capture system dynamics and control inputs. Observability, both a goal and a premise, ensures system state inference from output variables, which depend on the previous state. Originally designed for linear systems, the Kalman filter's optimality is limited in nonlinear contexts. Its effectiveness hinges on the accuracy of initial state estimates and covariance.

The Extended Kalman Filter (EKF) is an evolution of the standard form, tailored to handle nonlinear systems. It linearizes an inherently nonlinear system about the current estimate using a first-order Taylor series expansion — its usefulness is restricted to mildly nonlinear systems, but that was already a significant leap — the EKF has been used in econometric forecasting for more than two decades (Javaheri et.al 2003; Kellerhals 2013). The basic EKF process involves: state prediction, error covariance matrix updating, linearization of the processing model and measurement model around the current estimate with their Jacobians, which is a pain to calculate, then computing the Kalman gain to weigh the new measurement, combining the predicted state with new measurement data, and refining the error covariance estimate (see Note. 3 for details). Linearizing a highly nonlinear system using an EKF can lead to significant errors (Julier and Uhlmann 1997), but that is where we need EKF the most.

Building on its predecessors, the Unscented Kalman Estimator offers an improved solution in handling nonlinearities. It generates a just-enough symmetric set of sigma points from the state's distribution, satisfying a set of nonlinear constraints. These sigma points are adept at matching the mean, covariance, and higher moments, such as skew and kurtosis, of a distribution (Julier and Uhlmann 2004), particularly when the distribution is symmetrical. For what we are specifically interested is a system that wouldn't actually function if it were in equilibrium, our approach is essentially nothing more than adding a component that metabolizes the state variables by reducing the weights of observation over time and replacing them with weights added on the processing and observation errors, at a 6 for processing and 4 for observation ratio.

It is built upon the following premises, and it is applicable as long as the following premises are met:

1. System Dynamics and Noise Characteristics:

A nonlinear state transition and/or observation model. The model can be either deterministic or it can include stochastic components. The filter does not require Gaussian noise distributions; it is MMSE-optimal as long as the first and second moments are finite (Julier and Uhlmann 2022). The noise in both the process and measurement models is assumed to be known and characterized by its mean and covariance, with the covariance matrix representing an upper bound on the expected squared error.

2. Initial State Distribution:

The initial state x_0 is a random vector with a known mean μ_0 and covariance P_0. The initial state estimate doesn't have to be accurate, as long as the initial covariance could reasonably reflect the uncertainty.

3. Additive Noise for Basic Form:

The filter assumes an additive process and measurement noise. For non-additive noise, an augmented state approach is used (Grewal and Andrews 2014; for details see Note. 4)

4. Sigma Points and Weights:

Following convention, we chose $2n + 1$ points, where n is the dimension of the state space. The weights are subject to the condition that their sum equals 1: $\sum_{j=0}^{2n} W_j = 1$ (see Note. 5 for details)

5. Linear Regression Assumption:

A linear regression relationship can be established between the sigma points drawn from the prior distribution and their transformed values.

6. Discrete Time System and Differentiability:

The filter works at a discrete time scale; to effectively approximate the mean and covariance of the PDF, the nonlinear functions in the state-space model are typically assumed to be differentiable.

7. Observability and Controllability:

The system should be observable and controllable.

In our implementation, we adhere to industry conventions for sigma points selection (Julier and Uhlmann 2004). For our n-dimensional state space, where n represents the number of return contributing factors, we choose $2n + 1$ sigma points to ensure a comprehensive coverage of the distribution's characteristics. One point is the mean of the distribution, and the rest are symmetrically distributed around this mean (for the calculation of sigma points, see Note. 8).

We chose Cholesky decomposition to calculate the square root of the scaled covariance matrix $\sqrt{(n+\lambda)P}$, for the efficiency and numerical stability of this method (Trefethen and Bau 1997). P is the state covariance matrix, n is the dimensionality of the state, and λ is a scaling parameter. The Cholesky decomposition provides a lower triangular matrix U that satisfies the equation $(n+\lambda)P = UU^T$ (Trefethen and Bau 1997).

Our scaling factor λ is calculated as $\lambda = \alpha^2 \times (n+k) - n$, where alpha is a small positive value controlling the spread of sigma points, n is the state size (number of return contributing factors), and k is a scaling parameter that influences the distribution of sigma points. The choice of λ has an importance in the weighting of sigma points and hence the estimator's ability to capture the true mean and covariance of the state. The weighings ensure that the sigma points do represent the distribution's mean and covariance after transformation. The first weight has a greater influence, computed as $\frac{\lambda}{n+\lambda}$ for both mean and covariance. The remaining weights are uniformly $\frac{1}{2(n+\lambda)}$.

Our implementation updated standard nonlinear filters (Julier and Ulhmann 2004; Julier 1998) by adding two key elements: the fading memory approach based on Sorenson and Sacks (1971) and a hand-crafted time-dependent component method.

This method uses a time-dependent component to redistribute the weight on observation errors based on the rationale that it is easier to calibrate the observation model than the process model, whose default errors cannot be fully incorporated into the initial setup. Given this irreducible inaccuracy, we lessen the impact of estimations from the previous timestamps, and add weights to more recent observations by adjusting the coefficient of the state covariance in the update equation. The purpose is to manually correct the state estimation errors at a near spontaneous level.

The state predictor function adjusts the next state prediction by accounting for both momentum and mean-reversion influences. Momentum is captured as a time-dependent weight, so is mean-reversion. These weights are applied to the baseline long-term means of each factor, influencing the state prediction. The time dependent factor adjusts the model's sensitivity to recent changes, allowing it to be more responsive during periods of significant market movement, and relatively stable during quieter times. Without incorporating additional metrics and series with macro factors, our model mechanistically incorporate the basic shape for momentum and mean-reversion.

The pseudocode is provided below.

- *Given:*

{ State vector: $\bar{x}$; State dimension: n; Scaling factor: λ; Covariance matrix: P }

- Sigma points calculation:

{ First sigma point: $\sigma_0 = \bar{x}$; Subsequent sigma points for $i = 1$ to n: $\sigma_i = \bar{x} + U_i$;
Remaining sigma points for $i = n + 1$ to $2n$: $\sigma_i = \bar{x} - U_{i-n}$ }
Where U is the Cholesky decomposition of $(n + \lambda)P$.

- *Algorithm:*

1. Calculate the scaling factor λ.
2. Compute $(n + \lambda)P$ where P is the covariance matrix.
3. Cholesky decomposition on $(n + \lambda)P$ to obtain matrix U.
4. Initialize sigma points array with size $2n + 1$ by n.
5. Set the first sigma point σ_0 to the mean state $\bar{x}$.
6. For each column k in U (from 0 to $n - 1$):
   Calculate $\sigma_{k+1} = \bar{x} + U_k$.
   Calculate $\sigma_{n+k+1} = \bar{x} - U_k$.
7. Return the set of sigma points.

THREE

We attempt to demonstrates that, under certain premises, our filter's state predictions exhibit martingale properties in a general sense, implying no predictable trends or drifts.

Let $X_{t_{t=0}}^{\infty}$ be the stochastic process representing the state estimated by the Bayesian filter at discrete time steps t.

Let $\mathscr{F}tt = 0^{\infty}$ be the natural filtration representing the accumulation of all information up to time t, including all observable and unobservable factors that influence the state.

The state vector X_t is defined as a function of the current observation and the previous state estimate. We are guaranteed that X_t is $\mathscr{F}_t$-measurable, as it is composed of information solely contained within $\mathscr{F}_t$.

Given the bounded nature of the state space, the expected absolute value of X_t is finite: $E[\,|X_t|\,] < \infty$ for all t.

The filter updates the state vector according to $X_{t+1} = f(X_t, Y_{t+1})$, where Y_{t+1} is the new observation at time $t + 1$, and f is the update function defined by the filter, and as such, $E[X_{t+1} \,|\, \mathscr{F}t] = E[\,f(X_t, Yt + 1)\,|\,\mathscr{F}_t] = X_t$. By definition, the expected future value given all past and present information equals the current state.

For empirical validity, the filter underwent 10,000 timesteps, and examining the final state against the initial state, the state variables stayed close to their starting points. The means

were found to be relatively stable, indicating an absence of significant long-term trends in the state variables, and the standard deviations remained small and consistent, suggesting limited variability in the state estimates.

First Half of Metrics for results_df:

Index	Final Simulated State	Mean	Std Dev
R_MKT	0.00504446	0.02642352	0.03191402
R_ME	0.04194741	0.01755478	0.03741183
R_IA	0.02735818	0.02207541	0.01520889
R_ROE	0.02428835	0.03270153	0.02530338
R_EG	0.03716771	0.04444530	0.03240618

Second Half of Metrics for results_df:

Index	ADF P-Value	T-Test P-Value	Std Dev Slope	Std Dev Slope P-Value
R_MKT	0.00000000	0.00000000	-0.00000452	0.00000000
R_ME	0.00286732	0.00000000	-0.00000363	0.00000000
R_IA	0.00000000	0.00000000	-0.00000264	0.00000000
R_ROE	0.00000001	0.00000000	-0.00000792	0.00000000
R_EG	0.00000066	0.00000000	-0.00000854	0.00000000

We used an Augmented Dickey-Fuller (ADF) test to assess the stationarity of the state variables. The results showed absurdly low p-values — the process has no time-dependent structures or trends. Similarly, a one-sample t-test compares the mean of the state variables against their initial values, and the resulting p-values again suggest that the means of the state variables do not significantly differ from their initial values. The standard deviation of state variables over time analyzed using linear regression shows negative slopes and near 0 p-values — the variability of the state estimates does not increase over time.

We conclude that the model behaves like a martingale, but whether this property still holds after tailoring for one's favorite state functions is still contingent on the empirical tests (not necessarily so restrictive, see Note. 9).

Four

We tested our model on the historical returns of the five factors obtained from Fama and French (2014) (Fama and French 2023), which is an extension of the Fama-French three-factor model (Fama and French 1993) with two additional factors — operating profitability (RMW: Robust Minus Weak) and investment (CMA: Conservative Minus Aggressive), alongside the original factors. The 5-factor model has a nice feature: as long as the number of factors is five, the model's performance is consistent; how the factors are defined doesn't really matter. (Fama and French 2014). The dataset covers the daily returns spanning from July 1963 to November 2023 and annual returns from 1964 to 2022, offering a long-term perspective on raw, unportfolioed returns (see Note.10). The primary objective was to forecast these returns accurately, take them as they are — there is little connection with the factor

modeling framework as provided by Fama and French, who chose the initial three factor from an exploratory standpoint (Fama and Stern 2016). The implementation is exactly the same as the base model previously described, with two additional specifications introduced to mimic momentum and mean reversion.

For state prediction, long-term means for each factor were computed using rolling windows, assuming 252 days per year and a cycle indicative of a financial crisis every 7 years. These means serve as baselines in the state predictor function, in a very crude way anchoring the model's predictions to historically observed levels. Momentum weights are approximated using a sinusoidal function, as to model the cyclical nature — it peaks at the 6th month and phasing out after 12 months. The mean reversion weight linearly increases over a 5-year cycle, consistent with common sense.

In assessing the model's performance, we used 8 standard metrics. The Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) were calculated for each factor as indications of the model's precision. The MSE and RMSE for R_MKT were comparatively higher, suggesting relatively higher variance in the prediction errors for market returns. Mean Absolute Error (MAE) provided a linear scale of error magnitude, which is the average magnitude of the errors in the predictions. We used it in complementary with MSE and RMsE, as MAE is not influenced by the direction of the error and is a clear representation of average discrepancies. The MAE values across all factors were notably lower, suggesting that the model generally produced predictions that were closely aligned with the actual values, with R_EG showing the lowest MAE, indicating an accuracy in predicting the earnings growth factor.

For all factors, R-squared uniformly exceeding 0.82. we also considered the Symmetric Mean Absolute Percentage Error (SMAPE) in case MAE meets division by zero errors. The predictive accuracy resulted from SMAPE values was relatively moderate — its values spanning from 50 to 56 percent. The full result table is presented at the bottom. One minor aspect is that for the Max Error, the highest value is again associated R_MKT, most likely because we didn't incorporate in parallel the index and treasury data during the same time period.

Otherwise, the model demonstrates an adequate level of explanatory power and accuracy in predicting factor returns, and the consistency in its drawbacks offer the opportunities for directions of refinement. With sufficient tailoring, the model can be a potentially valuable tool for predicting the future state of returns if factors are reasonably specified.

First Half of Metrics:

Factor	MSE	RMSE	MAE	R-squared
R_MKT	0.14199214	0.37681844	0.24397735	0.83338050
R_ME	0.04618149	0.21489880	0.13898549	0.82769000
R_IA	0.02587032	0.16084254	0.10166101	0.82207180
R_ROE	0.02666343	0.16328941	0.10610510	0.82428360
R_EG	0.01914129	0.13835207	0.09191786	0.82999091

Second Half of Metrics:

Factor	SMAPE	Explained Variance	Mean Error	Max Error
R_MKT	50.39237355	0.83338879	-0.00265856	8.68710851
R_ME	51.19458131	0.82769799	-0.00146347	5.41985701
R_IA	53.21672660	0.82207208	0.00020183	3.34329173
R_ROE	54.47851511	0.82428368	0.00011389	1.93975721
R_EG	56.81220558	0.82999724	0.00084439	1.44818947

Reference

1. Andrews, Donald W.K. (1990). Tests for Parameter Instability and Structural Change With Un known Change Point. Cowles Foundation Discussion Papers. 1186.
2. Cheng, Y. & Liu, Z. Optimized Selection of Sigma Points in the Unscented Kalman Filter. 2011 International Conference on Electrical and Control Engineering, 2011, Pp. 3073-3075.
3. Dhillon, P. S., Foster, D. P., Kakade, S. M., & Ungar, L. H. (2011). A Risk Comparison of Ordi nary Least Squares vs Ridge Regression. arXiv Preprint arXiv:1105.0875.
4. Elliott, G. (1995). On the Robustness of Cointegration Methods When Regressors Almost Have Unit Roots. Econometrica, 66(1), 149–158.
5. Fama, E. F., & French, K. R. (1993). Common Risk Factors in the Returns on Stocks and Bonds. Journal of Financial Economics, 33(1), 3-56.
6. Fama, E. F., & French, K. R. (2014), a Five-Factor Asset Pricing Model. Fama-Miller Working Paper.
7. Fama, E. F., & French, K. R. (2023). Fama/French 5 Factors (2x3). Tuck School of Business at Dartmouth. From Https://Mba.Tuck.Dartmouth.Edu/Pages/Faculty/Ken.French/Data_Library/F-F_5_Factors_2X3.Html
8. Fama, E. F. and Stern, J. M. (2016), a Look Back at Modern Finance: Accomplishments and Limitations (Fall 2016). Journal of Applied Corporate Finance, Vol. 28, Issue 4, (Page 10-16).
9. Green, J., Hand, J. R. M. & Zhang, F. (2016). Errors and Questionable Judgments in Analysts' DCF Models. Review of Accounting Studies.
10. Grewal, M.S. & Andrews, A.P. (2014) Kalman Filtering Theory and Practice Using MATLAB. 4th Edition, John Wiley & Sons Inc., New York.
11. Javaheri, A., Lautier, D., & Galli, A. (2003). Filtering in Finance. Wilmott, 3, 67-83.
12. Julier, S. & Uhlmann, J. (1997). Consistent Debiased Method for Converting Between Polar and Cartesian Coordinate Systems. Proceedings of the 1997 SPIE Conference on Acquisition, Tracking, and Pointing. Vol. 3086. SPIE.
13. Julier, S. & Uhlmann, J. (2004) Unscented Filtering and Nonlinear Estimation. In Proceedings of the IEEE, Vol. 92, No. 3, Pp. 401-422.

14. Julier, S. & Uhlmann, J. (2022) Gaussianity and the Kalman Filter: a Simple yet Complicated Relationship, Jou. Cie. Ing., Vol. 14, No. 1, Pp. 21-26.
15. Julier, Simon (1998). "A Skewed Approach to Filtering". The Proceedings of the 12th Intl. Symp. On Aerospace/Defense Sensing, Simulation and Controls. Vol. 3373. SPIE.
16. Kabanov, Yuri. (2008). In Discrete Time a Local Martingale Is a Martingale Under an Equiva lent Probability Measure. Finance and Stochastics. 12. 293-297.
17. Kellerhals, B. P. (2013). Financial Pricing Models in Continuous Time and Kalman Filtering (Vol. 506). Springer Science & Business Media.
18. Khandani, Amir E, Adlar J Kim, & Andrew W Lo. (2010). Consumer Credit-Risk Models via Machine-Learning Algorithms. Journal of Banking & Finance 34 (2767–2787).
19. Quenouille, M. H. (1956). Notes on Bias in Estimation. Biometrika. 43 (3–4): 353–360.
20. Särkkä, S. (2013). Bayesian Filtering and Smoothing. Cambridge University Press. IMS Text book Series Vol. 3 (Page 51 - 53).
21. Sorenson, H. W., & Sacks, J. E. (1971). Recursive Fading Memory Filtering. Information Sci ences, 3(2), 101-119.
22. Trefethen, L.N. & Bau, D. (1997) Numerical Linear Algebra. (Page 172 - 177).
23. Wilmott, P., Howison, S., & Dewynne, J. (1995). The Mathematics of Financial Derivatives (Page 77; Page 103 -104).
24. Zdeborová, L. & Krzakala, F. (2016) Statistical Physics of Inference: Thresholds and Algo rithms, Advances in Physics, 65:5, 453-552.

Notes

1. Ridge and LASSO regression are designed to penalize overfitting, but this penalization indirectly impacts how randomness or stochasticity in the data is handled. Overfitting is when a model captures noise, or random fluctuations, in the data as if they were significant patterns. Randomness in data can lead to complex models that fit the noise rather than the underlying trend. L2 reg adds a penalty equal to the square of the magnitude of coefficients, reducing the impact of less important predictors, especially those that might be capturing noise instead of signal — by shrinking coefficients, ridge regression reduces the model's sensitivity to random fluctuations in the data. L1 imposes a penalty equal to the absolute value of the magnitude of coefficients, which can shrink coefficients of less relevant predictors to 0, indirectly reduces the model's tendency to fit the noise. In all three cases, the penalty on the coefficients doesn't just combat overfitting; it also moderates the extent to which the model fits to the random noise in the data. This results in a model that is not only less prone to overfitting but also less influenced by the stochastic variations in the dataset.

2. If a linear regression model has drift, time-dependent covariates t_k poses a significant problem. As the dataset grows, t_k increases indefinitely, leading to deteriorating problem conditioning over time. Such a model's dependency on absolute time complicates long-

term stability and accuracy. Transitioning to a state-space model offers a solution by focusing on the relative positioning of states and measurements in time, rather than their absolute values. It mitigates the unbounded growth of parameters like t_k, enhancing the model's conditioning and computational efficiency. State-space models handle dynamics and latent variables more effectively, offering adaptability to non-linear relationships and non-Gaussian noise. Such models maintain consistency in parameter estimation, regardless of the time series length.

3. We begin with linearizing the nonlinear state transition function $f(x, u)$ and measurement function $h(x)$, where x is the state vector and u is the control vector, around the current estimate ($\hat{x}_{k|k-1}$) using their Jacobians F_k and H_k. The state prediction equation is $\hat{x}_{k|k-1} = f(\hat{x}_{k-1|k-1}, u_k)$, with the error covariance prediction $P_{k|k-1} = F_k P_{k-1|k-1} F_k^T + Q_k$, where Q_k represents the process noise covariance. The Kalman gain is calculated as $K_k = P_{k|k-1} H_k^T (H_k P_{k|k-1} H_k^T + R_k)^{-1}$, where R_k is the measurement noise covariance. The state update formula is $\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k(y_k - h(\hat{x}_{k|k-1}))$, where y_k is the new measurement, and the error covariance update is given by $P_{k|k} = (I - K_k H_k) P_{k|k-1}$.

4. In its basic form, the filter assumes that both the process noise (affecting the state transition) and the measurement noise (affecting the observation model) are additive and as such, the noise terms are directly added to the state and observation equations. i.e. State Evolution with Process Noise:$x_k = f(x_{k-1}) + w_{k-1}$; Observation Model with Measurement Noise:$z_k = h(x_k) + v_k$, where w_{k-1} and v_k represent the process and measurement noise respectively. When dealing with non-additive noise, the state vector is expanded to include the noise components via a augmented state vector x_{aug} at time step k is given by: $x_{aug,k} = [x_k^T, w_{k-1}^T, v_k^T]^T$.

5. According to Cheng and Liu (2011) and Grewal and Andrews (2014), the sigma points $X_1, \ldots, X_m$ and their associated weights $w_1, \ldots, w_m$ must satisfy three conditions:

1. As always, the weights must sum to one.
2. The weighted sum of the sigma points must equal the mean of the distribution.

$$\sum_{i=1}^{m} w_i X_i = \mu$$

3. The weighted sum of the squared deviations of the sigma points from the mean must equal the covariance matrix of the distribution.

$$\sum_{i=1}^{m} w_i (X_i - \mu)(X_i - \mu)^T = \Sigma$$

To satisfy these conditions, a matrix X is constructed, where each column is one of the sigma points, and a matrix M can be constructed where each column is the mean μ. Additionally, a vector of ones, 1_m, is used. The following equivalences must then hold:

$(X - M)1_m = 0$

$(X - M)(X - M)^T = m\Sigma$

With only $n + 1$ sigma points while satisfying all the conditions above by setting the matrix X as: $X = M + \sqrt{m\Sigma}U$, where U is a matrix that satisfies $U1_m = 0$ and $UU^T = I$

6. This distinction is crucial for the proper functioning of the filter. In scenarios where noise is inherently additive, the basic form is applied. However, when the noise characteristics deviate from this additive nature, the augmented state approach provides a means to accommodate these complexities, ensuring the filter's applicability to a broader range of systems with different noise behaviors.

7. There is this fundamental difference in how regression models and recursive algorithms like a filter handle data. In standard regression analysis, especially in time series, the estimation of model parameters is typically done using a fixed dataset. When new data points are added to the dataset, the model parameters may need to be re-estimated to capture the updated information, and this re-estimation process often involves reprocessing the entire dataset, including both the old and new data, to recalibrate the model. This is especially true for models where the relationship between variables is assumed to be static, and any change in data might alter this relationship.

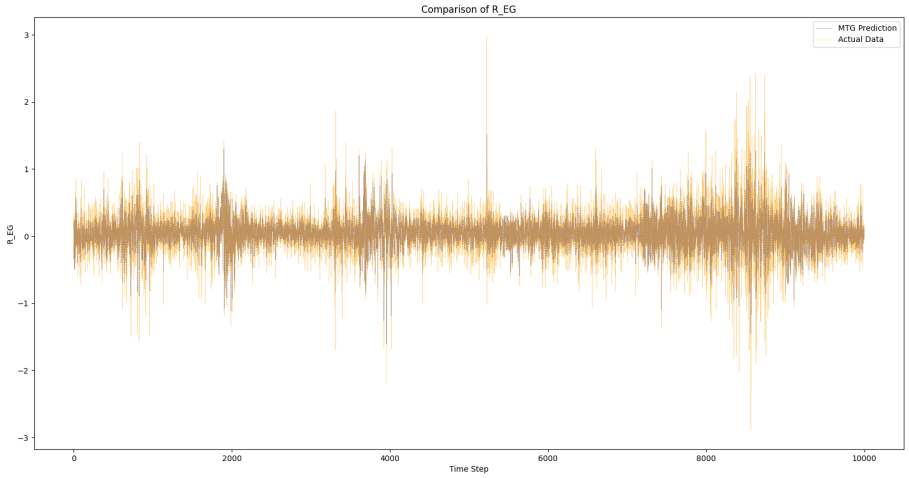
   For recursive algorithms like a filter are designed to update estimates *incrementally* as new data becomes available. These filters use the previous state estimate and the new data point to produce a new estimate, without the need to revisit the entire historical dataset. It is more efficient and suitable for real-time processing, especially in systems where data is continuously generated or almost continuously generated. Online learning algorithms can also update estimates incrementally, but traditional batch processing regression is still the industry standard (I assume with very little confidence), the need to reprocess the entire dataset for parameter re-estimation is a common practice.

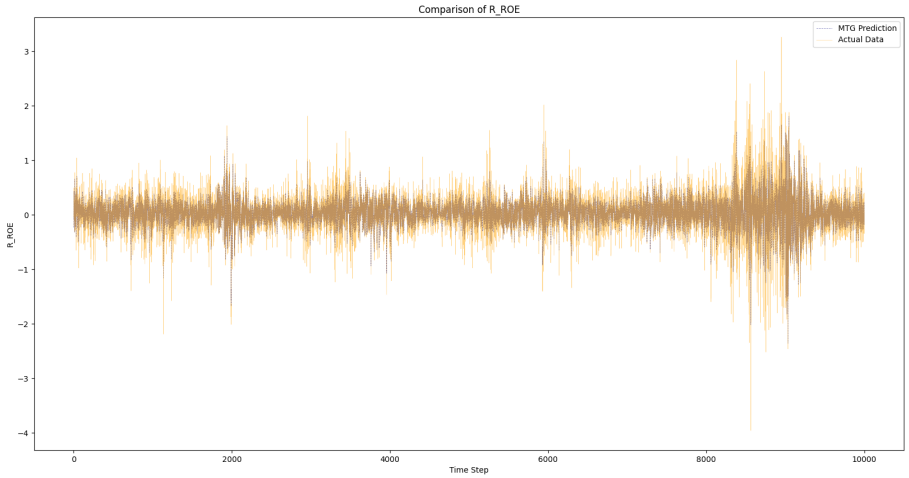
8. Sigma points are calculated by first scaling the covariance matrix by $n + \lambda$ and then using Cholesky decomposition to obtain its square root. The sigma points are generated around the mean (with adjustments made using the columns of the Cholesky factor) such that the sigma points contain the mean and covariance of the distribution.

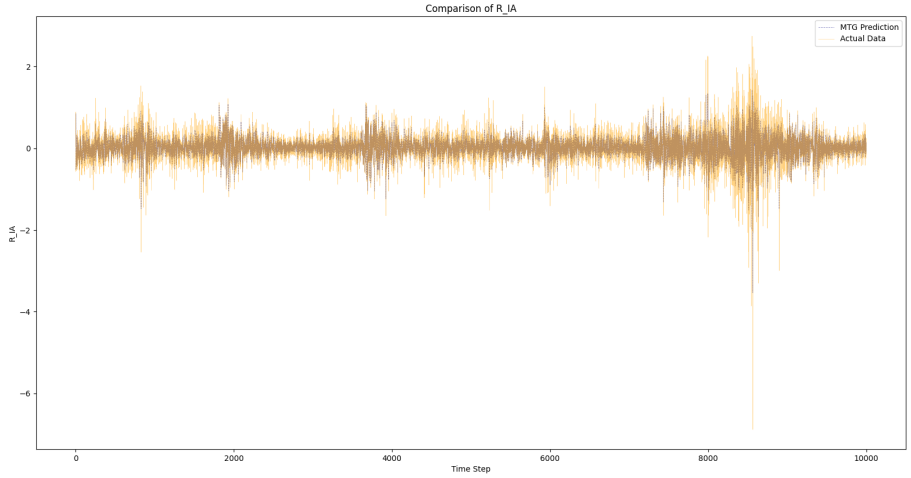
9. There is a proof in Kabanov (2008) that essentially says in a discrete-time infinite hori zon model, if a process is a local martingale, it can be treated as a global martingale under an equivalent probability measure, potentially generalizes the applicability of no-arbitrage in vacuums. There are oxygens in accounting books.

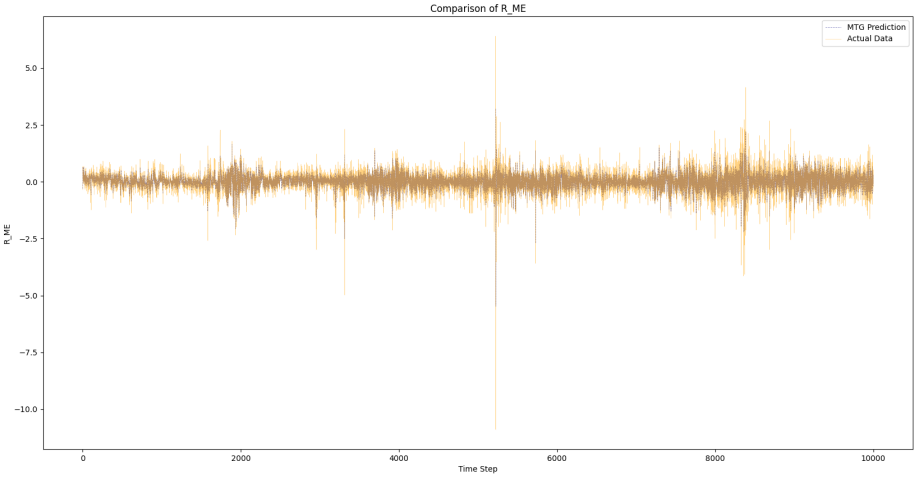
10. The dataset description says monthly returns, but it is in fact a daily one.

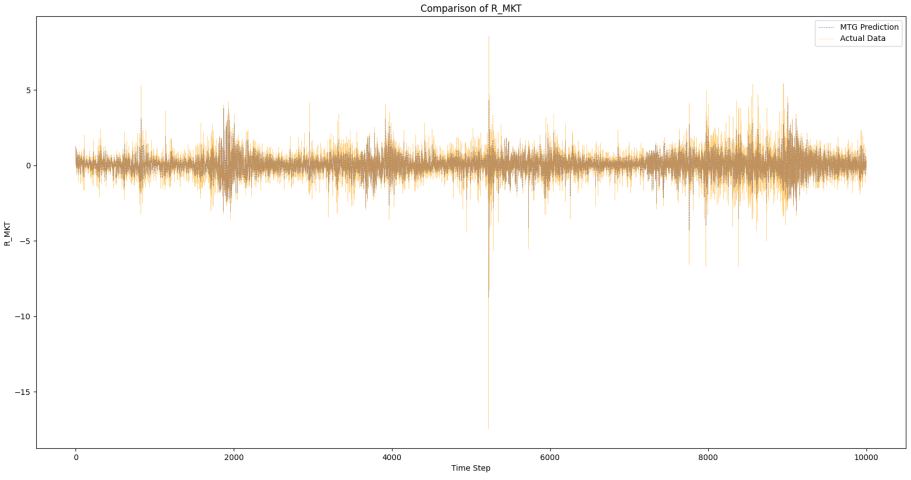
Figures:






Comparison of R_IA
MTG Prediction
Actual Data
R_IA
Time Step


Comparison of R_ME
MTG Prediction
Actual Data
R_ME
Time Step


Comparison of R_MKT
MTG Prediction
Actual Data
R_MKT
Time Step