# Analytic Methods for Asset Allocation with Illiquid Investments and Low-Frequency Data

by Adam R. Slakter

*S.B. Computer Science M.I.T., 2017*
*S.B. Management Science M.I.T., 2017*

Submitted to the
Department of Electrical Engineering and Computer Science
In Partial Fulfillment of the Requirements for the Degree of

Master of Engineering in Electrical Engineering and Computer Science

at the Massachusetts Institute of Technology

June, 2018



Author: ______________________________

Department of Electrical Engineering and Computer Science
June 1, 2018

Certified by: ______________________________

Peter Kempthorne, Ph.D. Department of Mathematics, Thesis Supervisor
June 1, 2018

Certified by: ______________________________

Ray Ming Yeh. Two Sigma Investments, Thesis Co-Supervisor
June 1, 2018

Accepted by: ______________________________

Katrina LaCurts, Ph.D. Chair, Masters of Engineering Thesis Committee

# Analytic Methods for Asset Allocation with Illiquid Investments and Low-Frequency Data

by Adam R. Slakter

*S.B. Computer Science M.I.T., 2017*
*S.B. Management Science M.I.T., 2017*

Submitted to the
Department of Electrical Engineering and Computer Science
In Partial Fulfillment of the Requirements for the Degree of

Master of Engineering in Electrical Engineering and Computer Science

at the Massachusetts Institute of Technology

June, 2018



**Abstract**

Investing in illiquid assets poses a challenge to investors, as the low-frequency data makes it difficult to quantify the risks across portfolios and make asset allocation decisions. This work reviews several principal methods to infer missing data and tests their implications for asset allocation. It compares these methods by applying them to hypothetical portfolios in a realistic simulation environment, helping allocators decide which methodology to use and when. Proxy-based methods, which utilize a related series of higher-frequency observations, outperform non proxy-based inference techniques when the correlation of the available proxy is above 0.3. If data autocorrelation is high, models such as Kalman filters, which are capable of explicitly modeling the autocorrelation outperform other proxy-based methods. In normal market conditions, the CL Method gives the best overall performance of methods tested, indicated by low RMSEs and reliable forecasts for mean return, volatility, Sharpe Ratio, and drawdown.

**Keywords: Illiquid Investments, Low-Frequency Data, Missing Data**

[1]The views and opinions expressed in this thesis are those of the author and do not necessarily reflect those of Two Sigma Investments. Send correspondence to Adam R. Slakter, Two Sigma Advisers, LP, 100 Avenue of the Americas, New York, NY. 10013. E-mail: Adam.Slakter@TwoSigma.com

# Contents

# Acknowledgements

Without the guidance and inspiration of several individuals, this work would not have been possible. I would first like to thank my MIT advisor, Dr. Peter Kempthorne, for his continued feedback and support throughout the thesis writing process. His careful eye, thoughtful commentary, and persistent dedication were immensely helpful for increasing the rigor of the mathematics and elevating the project to its current form.

Special thanks to Jeffrey Saret for the inspiration for the thesis idea and contributions throughout the early stages of the project. I would also like to thank Gerardo Manzo for his many hours of instruction on filtering and stochastic processes that helped make the work a reality. Also, Ray Ming Yeh and the entire Asset Management Modeling Team have been invaluable in supporting my efforts every step of the way as I transitioned to full-time life at Two Sigma.

I would also like to thank Tomas Palacios and the 6-A Program staff for their flexibility in helping make this joint research between MIT and Two Sigma a reality. I feel honored that I was able to work on a project sponsored by two amazing institutions.

Most importantly, I would like to thank my parents, sister, and relatives for their endless love, support, and encouragement. Without them, none of this would have been possible.

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Lack of good data makes informed asset allocation decisions with illiquid assets difficult. There are two main issues. The first is the fidelity of the data, manifested by reported returns of private equity firms that have lower volatility than respective cash flows, suggesting that firms mark companies to smooth out returns [Jenkinson et al., 2013]. The second is low volume or frequency of available data for illiquid assets. This research addresses the second issue.

Consider an institutional investor with wealth held in liquid assets such as stocks and bonds, observed daily or monthly, and illiquid assets such as real estate, observed quarterly. To measure and modify her asset allocation within the quarter, she has up-to-date information (within the month) on her liquid assets but often must make assumptions for the recent risk and performance of her illiquid assets. In order to apply many conventional portfolio allocation tools like risk decomposition and mean-variance optimization that require data sampled at the same rate, she must transform the data. Two approaches are aggregating to the lowest common frequency or inferring higher-frequency data.

Aggregation, though simple and convenient, throws away valuable information from the dataset. In the case of this investor, the number of relevant data points per asset is reduced by a factor of three, from twelve to four each year. Insights gleaned in this low-frequency setting like regression betas or factor exposures might not be relevant when applied at the higher-frequency [Mallinger-Dogan and Szigety, 2014]. In contrast, inferring higher-frequency data preserves all original data points but runs the risk of introducing noise. If this technique adds more information than noise to a dataset, then the investor should prefer it over aggregation. Thus, a goal for an investor of illiquid assets looking to apply conventional portfolio allocation tools is to augment the original dataset with meaningful higher-frequency information. Often, the choice of which augmentation method to use can yield different asset allocation decisions, and Chapter 2 presents a real-world example highlighting the practical issues in dealing with illiquid data.

The research explores a number of common solutions for this data inference problem, falling into two categories: proxy-based and non proxy-based. Chapter 3 discusses these methods in greater detail. Proxy-based methods utilize a related series to help forecast the fair value of the illiquid asset within the quarter, while non-proxy-based methods make assumptions on the underlying process driving asset returns without the use of a related series. Proxy-based models inject outside information into the dataset and can provide meaningful augmentation if a useful related series is available. Otherwise, non-proxy-based methods are the best an asset allocator can do. The research explores non-proxy-based methods such as simple interpolations as well as important proxy-based methods: Chow Lin method [Chow and Lin, 1971], Mixed Data Sampling [Ghysels et al., 2004, Bai et al., 2013], and Kalman Filtering [Kalman, 1960, Bai et al., 2013]. This work builds on that of Dogan and Szigety [Mallinger-Dogan and Szigety, 2014] and provides an overview of data inference methods applicable to investing and compares these methods by applying them to hypothetical portfolios in a realistic simulation environment. The research objective is to help allocators decide which methodology to use and when.

To assess the performance of the inference models, a robust simulation framework capable of generating asset return data under various market conditions is needed. A significant part of the research is the development of the Multi-Factor Market Simulator, which applies Monte Carlo methods to simulate baskets of asset returns (with corresponding proxies for the illiquid assets) according to a specified covariance structure and market parameters. These market parameters control the level of autocorrelation in the illiquid returns as well as the degree of tail risk affecting each asset class. For instance, if an allocator believes there are major geopolitical risks affecting a particular illiquid market, she should seek data inference methods that perform well in conditions of high market tail risk. The research does not specify an answer on how market parameters should be set, and the allocator should have her own assumptions on what these should be. The research does, however, prescribe the best inference method to use according to common portfolio performance metrics, given a particular state of the world decided upon by the allocator.

The Multi-Factor Market Simulator generates asset returns that vary according to three factors that impact the performance of data inference tools: market tail risk, data autocorrelation, and availability of a good proxy. For instance, when large price dislocations in markets are common, methods that smooth out returns tend to understate volatility. Since illiquid assets are often marked to smooth volatility, illiquid returns often exhibit autocorrelation [Getmansky et al., 2004, Khandani and Lo, 2011]. Models which are able to incorporate this tendency will likely give better forecasts. Also, since proxies supply information for augmentation, the availability of a particularly good proxy can help better signal the fair value of the illiquid asset within the period. For example, a go-anywhere private equity firm might not have a good proxy available whereas a venture capital firm investing in Internet companies might have a small-cap tech proxy that is useful.

By parameterizing these three dimensions – market tail risk, data autocorrelation, and availability of a good proxy — the framework can simulate different market environments and help allocators decide the best inference methods to use, given their expectations about these parameters. Chapter 4 details the data generation methodology. Merton's jump-diffusion process is a model for asset returns that incorporates small day-to-day "diffusion-type" movements with larger randomly occurring "jumps" [Merton, 1976]. By controlling the frequency of these "jumps", the simulator controls how often market dislocations (e.g. crashes) occur. The research defines market tail risk as the likelihood of a market crash. Fractional Brownian motion (fBm) is a generalization of classical Brownian motion in which the increments of the process have a user-prescribed level of autocorrelation. The Multi-Factor Market Simulator utilizes this process along with Merton's jump-diffusion model, so that the small day-to-day movements (and not market crashes) are autocorrelated. By assuming that asset returns are driven by Merton's jump-diffusion in conjunction with fBm, the simulator controls the levels of tail risk and return autocorrelation in the market. The research defines the goodness of a proxy to an illiquid asset as the proxy's similarity to the higher-frequency representation of the illiquid asset. This definition allows good proxies to provide meaningful higher-frequency information about the illiquid asset. Proxies are generated by perturbing the true monthly illiquid asset returns with various amounts of noise. Superior proxies are generated with less noise.

Asset allocators are ultimately concerned with portfolio performance via metrics such as mean return, Sharpe Ratio, and drawdown. In order to use these common portfolio performance measures as criteria for data augmentation models, the research allocates and compares two portfolios. The first portfolio serves as a baseline and is allocated from perfect data observed monthly. The second "experimental" portfolio, is allocated from the same data observed quarterly with inference methods applied to convert it to monthly frequency. Chapter 5 discusses the simulation framework's design in greater detail. The simulation framework allocates portfolios via Markowitz mean-variance optimization, as is common in industry, using a rolling window of three years. This assumption is in line with a representative investor who re-balances the portfolio every three months and keeps the target allocation constant within the quarter. Allocations are out-of-sample, using weights fit on the previous three years of data to predict performance over the next quarter.

Under these assumptions of portfolio performance and market conditions, the data imply that the recommendation of which inference model to use changes with the goodness of available proxy and data autocorrelation but is not significantly affected by market tail risk. Proxy-based methods, which utilize a related series of higher-frequency observations, outperform non proxy-based inference techniques when the correlation of the available proxy is above 0.3. If data autocorrelation is high, models such as Kalman filters, which are capable of explicitly modeling the autocorrelation outperform other proxy-based methods. In normal market conditions, the CL Method gives the best overall performance of methods tested, indicated by low RMSEs and reliable forecasts for mean return, volatility, Sharpe Ratio, and drawdown.

Chapter 6 discusses these results in detail. Chapter 7 gives a practical application of the research to asset allocation, and Chapter 8 concludes with suggestions for future work.

# Chapter 2

# Allocating with Illiquid Investments: Practical Issues

As a real-world example of the practical difficulties of illiquid investments that can influence asset allocation decisions, consider an institution that holds four liquid index positions observed monthly in stocks (MSCI World Investable Markets), bonds (Barclays U.S. Aggregate Bond Total Return), hedge funds (HFRI Fund Weighted Composite), and commodities (S&P Goldman Sachs Commodities Total Return). The institution's asset allocator is considering adding an index position in private equity (Cambridge Associates US Private Equity), an illiquid asset observed quarterly, to the portfolio. For further descriptions of these indices, see Table 4.2 in Section 4.4. Her historical data for each index extends from June 1994 to March 2017. To evaluate whether or not to incorporate private equity in the portfolio, the allocator seeks to assess the historical performance of her portfolio with private equity added to the allocation and compare it to her realized performance.

However, she faces a problem: how to approximate historical performance with private equity added to the allocation. She wishes to use mean-variance optimization[1], which is commonly used in the asset management industry, as her portfolio allocation tool, which requires all asset returns to be sampled at the same frequency (i.e. all quarterly or all monthly). Therefore, she must figure out how to match the sampling frequencies of her time series prior to allocation. She sees two options: (1) aggregating all of the asset returns to the lowest common frequency, which is quarterly, or (2) attempting to infer higher-frequency

[1]Section 5.3 provides details on mean-variance (Markowitz) optimization.

private equity returns, which are monthly.

This example will show that these two options often give radically different results, which makes allocations with different asset return sampling frequencies difficult. For inferring higher-frequency returns, the institution uses back fill[2], which assigns trailing monthly returns to the most recent quarterly return divided by three. Since the illiquid returns are available quarterly, portfolios allocated on monthly inferred returns approximate historical performance with accuracy proportional to how closely the inference method recovers the true monthly returns of the illiquid asset (which the allocator will never know). This analysis is entirely in-sample; since the allocator does not know the true monthly returns, she must use her estimated returns to approximate historical portfolio performance. In this example, portfolios are constructed via mean-variance optimization using a rolling window of three years, an assumption in line with a representative investor who re-balances the portfolio every three months and keeps the target allocation constant within the quarter. Allocations are long-only and fully-invested, with an annualized target volatility of 8%. The allocator compares the performance of three portfolios, detailed below:

1. **AGGREGATION**: the allocator aggregates her four liquid asset returns to quarterly frequency, to match the frequency of the private equity returns. Her portfolio consists of her four liquid assets plus private equity, which now are each observed quarterly. She determines rolling portfolio weights using mean-variance optimization over a window of 12 quarters and applies these weights to her quarterly returns.

2. **BACK FILL**: the allocator uses back fill to convert her quarterly series of private equity returns to monthly frequency. Her portfolio consists of her four liquid assets plus her inferred private equity returns, which are each observed monthly. She determines rolling portfolio weights using mean-variance optimization over a window of 36 months and applies these weights to her monthly returns.

3. **BASELINE LIQUID**: this is her realized historical performance without any illiquid asset. Her portfolio consists of four liquid assets observed monthly with allocations determined via mean-variance optimization.

To compare the performance of these three portfolios, the allocator computes mean return, volatility, Sharpe ratio, and maximum drawdown[3]. Mean return, volatility, and Sharpe ratio

[2] Back fill is detailed in Section 3.

[3] Section 5.4 explains each of these performance measures in detail.

are each annualized. For her Sharpe ratio calculation, she assumes the risk-free rate is 3% annualized. Her results, including a plot of cumulative return, are shown in Figure 2.1.

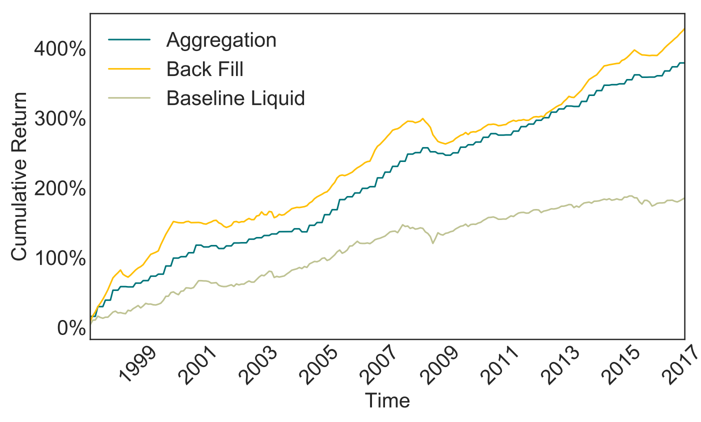
**Figure 2.1** – Aggregated vs. Inferred Mean-Variance Portfolios Performance

The performance and cumulative return of three mean-variance portfolios when (a) higher-frequency series are aggregated to the lowest common frequency of the illiquid asset (Aggregation), (b) the illiquid series is interpolated to higher frequency via back fill (Back Fill), and (c) no illiquid assets are contained in the allocation (Baseline Liquid). The data source is four liquid assets – stocks (MSCI World Investable Markets), bonds (Barclays U.S. Aggregate Bond Total Return), hedge funds (HFRI Fund Weighted Composite), and commodities (S&P Goldman Sachs Commodities Total Return) – observed monthly and a single illiquid asset, private equity (Cambridge Associates US Private Equity), observed quarterly. The data extends from June 1994 to March 2017 and performance covers June 1997 to March 2017. The performance measures mean return, volatility, and Sharpe ratio (which assumes a 3% annualized risk-free rate) are each annualized.

*(a) Performance Metrics*

| | Aggregation | Back Fill | Baseline Liquid |
|---|---|---|---|
| Mean Return | 19.18% | 21.70% | 9.41% |
| Volatility | 9.00% | 9.99% | 8.63% |
| Sharpe Ratio | 1.80 | 1.89 | 0.74 |
| Max Drawdown | -2.95% | -9.12% | -10.84% |
| Contains Illiquids | Yes | Yes | No |
| #Observations | 79 | 237 | 237 |
| Frequency | Q | M | M |

*(b) Cumulative Return*




For both the aggregation and back fill portfolios, overall performance increases over her current portfolio. Annualized mean return increases by 3.43% and 4.20% respectively for the two portfolios while volatility increases by 0.37% and 1.36% respectively. These effects drive a Sharpe ratio increase of 1.06 and 1.15 respectively for the aggregation and back fill portfolios. This large Sharpe ratio increase suggests that the allocator would likely see benefit from adding private equity to her allocation. However, with respect to maximum drawdown, a statistic widely used in industry to measure portfolio downside risk, aggregation and back fill give dramatically different results, -2.95% and -9.12% respectively. Maximum drawdown corresponds to the largest peak-to-trough decline in portfolio value in percentage terms. Since aggregation smooths out the return series and private equity has done very well historically, trending upwards over time, aggregation produces an unrealistically low maximum drawdown value. The drawdown, which takes place during the onset of the financial crisis in late 2008 for all three portfolios can be seen visually in Figure 2.1. Note that the cumulative return of the "Aggregation" portfolio appears as a step function because it is measured quarterly, while the other two portfolios have performance measured monthly.

Furthermore, the estimates of mean return differ by 0.77% between the two techniques, causing a Sharpe ratio difference of 0.09. The allocator faces an issue: since true monthly returns for her illiquid asset will never exist, there is no way to determine whether aggregation or back fill produces a more accurate estimate of monthly historical performance. Her goal should be to find an inference method which produces an accurate estimate of historical performance, so that she can make asset allocation decisions with illiquid investments with confidence in any market condition. Furthermore, she should be aware of the caveats and pitfalls of her inference method of choice.

It is clear that making asset allocation decisions with low-frequency data may be difficult for an investor. Without a clear understanding of how market conditions affect the performance of data inference methods, the investor cannot authenticate estimates of historical portfolio risk and performance. The research objective is to arm the investor with a deeper understanding of inference methods for illiquid data by assessing their impact on portfolio allocation across a variety of market conditions.

# Chapter 3

# Inference Methodologies

The following chapter gives an overview of the inference models explored in the research to convert quarterly data to monthly frequency. Inference methodology selection as a function of the allocator's situation and beliefs is the main purpose of the research. For instance, if the allocator believes that the global political climate is unstable, she might think there is higher tail risk in that market. To allocate a high-performing portfolio, she would favor a data inference methodology that handles tail risk well. She might also be allocating a portfolio of illiquid assets that have no suitable proxy. For instance, in an emerging African or East Asian market, there might not be a reliable high-frequency index to use as a proxy for a tradable illiquid asset. In that case, she would favor an inference method that does not require the use of related series for inference.

The research explores inference methodologies that fall into two main categories: proxy-based and non proxy-based. Proxy-based methods utilize one or more related series to incorporate useful external information into the data while non-proxy methods do not. In general, proxy-based methods are more complex and therefore more computationally demanding, though they have the potential for superior inference if a good proxy is available. Section 3.1 gives the motivation and mathematical specification for each non proxy-based method while Section 3.2 explains the proxy-based methods. A list of inference methodologies explored in the research along with summarizing details is found in Table 3.1.

| Inference Methodology | Uses Proxy | Type | Ease of Use |
|---|---|---|---|
| Back Fill | Non-Proxy | Interpolation | Easy |
| Chow Lin Method | Proxy | Regression | Moderate |
| Cubic Spline Interpolation | Non-Proxy | Interpolation | Easy |
| Kalman Filtering | Proxy & Non-Proxy | State-Space | Difficult |
| Linear Interpolation | Non-Proxy | Interpolation | Easy |
| MIDAS | Proxy | Regression | Moderate |

**Table 3.1** – Methodologies Summary

A summary of the data inference methodologies used in the research. Displays whether each model is proxy or non-proxy-based, the model's general type, and general ease of use. An ease of use of "easy" is simple to perform in Microsoft Excel while a labeling of "moderate" requires optimization routines. A rating of "difficult" has greater computational complexity and requires significant parameter estimation.

## 3.1 Non Proxy-Based Methodologies

The four non-proxy-based methods the research explores are back filling, linear interpolation, cubic spline interpolation, and a non proxy-based configuration of the Kalman filter. The first three methods are discussed in this section, while the Kalman filter is introduced with the proxy-based methods in Section 3.2.3.

In the research, log returns are used. When applying any of the non-proxy-based inference methods to infer quarterly log returns, quarterly returns are first divided[1] by 3. For back filling, this division step ensures the inferred monthly returns sum to the observed quarterly values. For linear and cubic spline interpolations, the division ensures that monthly returns are not produced with an unreasonably high volatility, at the expense of being potentially volatility reducing overall.

### 3.1.1 Back Filling

Back filling, assuming that return is constant for each month within the quarter, is a quick and convenient way of inferring higher frequency data. It is included in the research as a benchmark for more sophisticated methods. To back fill, one simply assigns trailing monthly returns to the most recent quarterly return divided by three. For instance, say the true quarterly log return for an illiquid asset for the 1st fiscal quarter (from October 1st to December 31st) is 3%. Then, back filling would assign the monthly returns in October, November, and December each to 1%. Another common technique is forward filling [Pandas,

[1]Log returns are additive, so simple division rather than geometric mean suffices.

2017], which infers the leading monthly returns instead of trailing. The research opted for back filling rather than forward filling to prevent the introduction of a lag effect that would corrupt the covariance structure of assets in the portfolio.

### 3.1.2 Spline Interpolation

Both linear interpolation and cubic spline interpolation are types of spline interpolation. The idea of a spline is simple: on each interval between data points, represent the graph with a function. Linear interpolation is the simplest form of a spline, which is formed by connecting data with lines. Let $(x_0, y_0)$ and $(x_1, y_1)$ be adjacent points. Then, the linear interpolant is the straight line between these points. For a value $x$ within the interval $[x_0, x_1]$, the value $y$ along the straight line is given from the equation of slopes

$$\frac{y - y_0}{x - x_0} = \frac{y_1 - y_0}{x_1 - x_0} \tag{3.1}$$

After linear interpolation, the next simplest function is quadratic. As the order of the function increases, the resultant graph becomes smoother. Cubic is the smallest degree that allows for the slope to be continuous at each point (e.g. no "corners"), which is a desirable property for modeling curves or motion trajectories. To enforce this property, at each point, one must either specify the first derivative or require that the second derivative be zero, called the "natural spline" condition. Due to their smoothness and simplicity, cubic splines, also referred to as Hermite splines or csplines, are used frequently in computer graphics to help render smooth images [Treuille, 2010]. Given a set of $n + 1$ data points $(x_i, y_i)$, where no two $x_i$ are the same and $a = x_0 < x_1 < \cdots < x_n = b$, a cubic spline $S(x)$ is a function satisfying three conditions [Young and Mohlenkamp, 2009]:

1. $S(x) \in C^2[a, b]$, meaning that $S(x) : [a, b] \to \mathbb{R}$ is a function for which both $S'(x)$ and $S''(x)$ exist and are continuous
2. On each sub-interval $[x_{i-1}, x_i]$, $S(x)$ is a polynomial of degree 3, where $i = 1, \ldots, n$
3. $S(x_i) = y_i$, $\forall i = 0, 1, \ldots, n$

Thus, $S(x)$ takes the form

$$S(x) = \begin{cases} C_1(x), & x_0 \leq x \leq x_1 \\ \dots & \\ C_i(x), & x_{i-1} \leq x \leq x_i \\ \dots & \\ C_n(x), & x_{n-1} \leq x \leq x_n \end{cases} \tag{3.2}$$

where for each $i = 1, \dots, n$, $C_i = a_i + b_i x + c_i x^2 + d_i x^3$ $(d_i \neq 0)$ is a cubic function. To determine the cubic spline $S(x)$, coefficients $a_i$, $b_i$, $c_i$, and $d_i$ must be found for each $i$ via the constraints

- $C_i(x_{i-1}) = y_{i-1}$ and $C_i(x_i) = y_i$, $\forall i = 1, \dots, n$
- $C_i^{'}(x_i) = C_{i+1}^{'}(x_i)$, $\forall i = 1, \dots, n-1$
- $C_i^{''}(x_i) = C_{i+1}^{''}(x_i)$, $\forall i = 1, \dots, n-1$

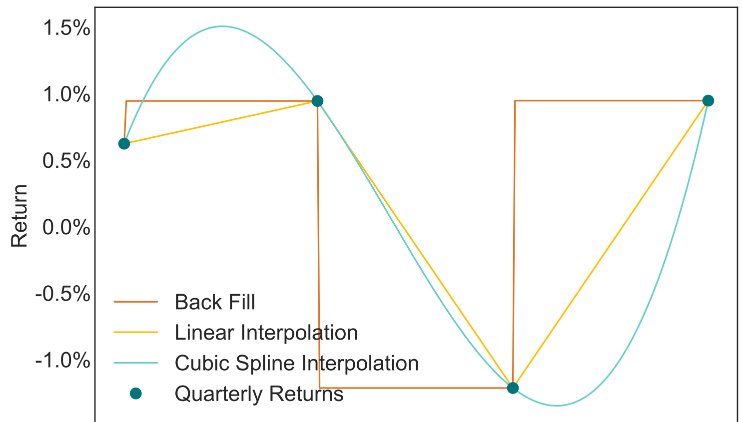
There are $n + n + (n-1) + (n-1) = 4n - 2$ conditions, but $4n$ coefficients are needed, so two boundary [2] conditions on the second derivative: $C_1^{''}(x_0) = C_n^{''}(x_n) = 0$ are added. Plugging in known values $x_0, \dots, x_n$ results in a system of $4n$ equations with $4n$ unknowns, which commercial linear algebra packages can solve efficiently.

Figure 3.1 shows an example application of these non-proxy-based inference methods to a simulated return series. Quarterly returns divided by 3 are plotted as blue dots. The orange line, indicating back fill, is horizontal between quarters, indicating that this method infers constant monthly returns that sum to the next quarterly value. Linear interpolation, indicated by the yellow line, joins the quarterly returns with straight lines while cubic spline interpolation, indicated by the light blue line, fits a 3rd order polynomial between adjacent quarterly returns. All three inferences have a smoothing effect compared to the original monthly return series, which understates volatility, and impose highly unrealistic assumptions on the data. For instance, all of these methods are incapable of inferring intra-quarter spikes in monthly returns. Linear interpolation also assumes that return never reverses direction within the quarter, which is certainly not always the case.

[2] These conditions are called "natural" or "simple" boundary conditions

**Figure 3.1** – Non Proxy-Based Inference

Inference performed by non proxy-based models on one year of simulated quarterly log returns. The plotted quarterly returns are observed quarterly returns divided by 3, which yields monthly inferred returns that roughly sum to the observed quarterly values.




| | Q1 | | | Q2 | | | Q3 | | | Q4 |
|---|---|---|---|---|---|---|---|---|---|---|
| Quarterly Returns | 0.6% | NaN | NaN | 0.9% | NaN | NaN | -1.2% | NaN | NaN | 0.9% |
| Back Fill | 0.6% | 0.9% | 0.9% | 0.9% | -1.2% | -1.2% | -1.2% | 0.9% | 0.9% | 0.9% |
| Linear Interpolation | 0.6% | 0.7% | 0.8% | 0.9% | 0.2% | -0.5% | -1.2% | -0.5% | 0.2% | 0.9% |
| Cubic Spline Interpolation | 0.6% | 1.4% | 1.4% | 0.9% | 0.2% | -0.7% | -1.2% | -1.3% | -0.7% | 0.9% |

## 3.2 Proxy-Based Methodologies

Proxy-based inference models utilize one or more higher-frequency proxies for inference of a lower-frequency series. The idea is that higher-frequency information contained in proxies can be useful for estimating where the lower-frequency series lies in the higher-frequency space. In general, the greater correlation the proxy has to the true higher-frequency representation of a lower-frequency series, the superior inference it should provide. Conversely, if the proxy has low correlation, then it likely will not be useful. The research explores proxy-based models that fall into two main categories: regressions and state-space approaches.

The basic logic for regression methods is to attempt to find the best fit with a proxy to make a prediction. They expand on the simple concept of linear regression in the case that the proxy, the regressor, is sampled at a higher frequency than the response series to be interpolated. The research explores two proxy-based regression models: the Chow Lin Method and Mixed Data Sampling.

State-space models assume that the low-frequency series has a higher-frequency representation that lies outside of one's ability to observe [Mallinger-Dogan and Szigety, 2014]. An analogy is seeing pictures of a sporting event when there is no available video. In the research, these “pictures” of the high-frequency series are taken once a quarter, when the illiquid returns are observed. Monthly observations of the proxy help influence one’s understanding of the events within the quarter to infer frames of the video. Kalman Filtering, introduced by R.E. Kalman in 1960, is a convenient algorithm for inferring unobserved data [Kalman, 1960]. It handles complex residual structures easily and incorporates different priors on the process driving underlying data. The research explores four configurations of the Kalman filter of varying complexity, including one that is non proxy-based.

This section introduces the research’s three proxy-based models – Chow Lin Method, Mixed Data Sampling, and Kalman Filtering.

### 3.2.1 Chow Lin Method

The first widely used model that attempted to solve the proxy-based low-frequency inference problem was proposed by Chow and Lin in 1971, hereafter the CL Method [Chow and Lin, 1971]. The methodology has two main components. The first component applies low-frequency linear regression betas fit over the entire sample to the higher-frequency observations of the proxy. The second component is an estimate of the high-frequency residuals, which are deviations from what might be expected from using the low-frequency betas, based on a user-specified error model [Mallinger-Dogan and Szigety, 2014]. The research utilizes an autoregressive order one process as the error model, as was done by Chow and Lin. Though the Chow Lin framework generalizes for inference of any evenly spaced data, this discussion assumes that quarterly data is being inferred to monthly frequency.

The methodology is specified as follows. Let the subscripts $M$ and $Q$ represent quarterly and monthly data respectively, $n$ be the number of quarterly data points, and $d$ be the number of proxies being used. Let $\mathbf{P}_Q \in \mathbb{R}^{n\times 1}$ represent the observed quarterly log returns and $\mathbf{X}_M \in \mathbb{R}^{3n\times d}$ represent the corresponding monthly proxies in stacked matrix form,

$$\mathbf{X}_M = \begin{bmatrix} \uparrow & & \uparrow \\ \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_d \\ \downarrow & & \downarrow \end{bmatrix}$$

The low-frequency regression model [Chow and Lin, 1971, Mallinger-Dogan and Szigety, 2014] is given by

$$
\begin{aligned}
\mathbf{P}_Q &= \mathbf{C}\mathbf{P}_M \\
&= \mathbf{C}\mathbf{X}_M\beta + \mathbf{C}\mathbf{e}_M \\
&= \mathbf{X}_Q\beta + \mathbf{e}_Q
\end{aligned}
\tag{3.3}
$$

The residuals are normally distributed according to $\mathbf{e}_Q \sim \mathcal{N}\left(\mathbf{0}, \mathbf{V}_Q\right)$.The matrix $\boldsymbol{C} \in \mathbb{R}^{n\times 3n}$ links the monthly log returns to the quarterly log returns, assuming a 3:1 ratio, and is given by

$$
\begin{aligned}
\mathbf{C} = \boldsymbol{I}_n \otimes \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \\
= \begin{bmatrix} 1110000 & ... & 000 \\ 0001110 & ... & 000 \\ ... & & ... \\ 0000000 & ... & 111 \end{bmatrix}
\end{aligned}
$$

where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix and $\otimes$ indicates Kronecker's product. It can be shown [Chow and Lin, 1971] that the best linear unbiased estimator (BLUE) of higher-frequency is given by

$$
\hat{\mathbf{P}}_M = \mathbf{X}_M\hat{\beta} + \left(\mathbf{V}_{\mathbf{P}_Q}\mathbf{V}_Q^{-1}\right)\hat{\mathbf{e}}_Q \tag{3.4}
$$

where $\hat{\beta} = \left(\mathbf{X}_Q^T\mathbf{V}_Q^{-1}\mathbf{X}_Q\right)^{-1}\mathbf{X}_Q^T\mathbf{V}_Q^{-1}\mathbf{P}_Q$ is the least-square estimate of the low-frequency regression coefficients and $\hat{\mathbf{e}}_Q = \mathbf{P}_Q - \mathbf{X}_Q\hat{\beta}$ is the vector of residuals from the low-frequency regression, and $\mathbf{V}_{\mathbf{P}_Q} = \mathbb{E}\left[\mathbf{e}_M\mathbf{e}_Q^T\right]$.

The term $\mathbf{V}_{\mathbf{P}_Q}\mathbf{V}_Q^{-1}$ in (3.4) dictates how the quarterly residuals are distributed at the monthly frequency. Assuming that the monthly residuals follow a zero-mean $AR(1)$ process

$$
e_t = ae_{t-1} + \epsilon_t \ , \ e \sim \mathcal{N}\left(0, \sigma^2\right)
$$

where $a$ can be estimated from its low-frequency counterpart, one can derive

$$
\mathbf{V}_{\mathbf{P}_Q}\mathbf{V}_Q^{-1} = \mathbf{V}_M\mathbf{C}^T\left(\mathbf{C}\mathbf{V}_M\mathbf{C}^T\right)^{-1} = \mathbf{A}\mathbf{C}^T\left(\mathbf{C}\mathbf{A}\mathbf{C}^T\right)^{-1} \tag{3.5}
$$

where $\mathbf{V}_M$ is the covariance of $\mathbf{e}_t$,

$$\mathbf{A} = \begin{bmatrix} 1 & a & a^2 & \cdots & a^{3n-1} \\ a & 1 & a & \cdots & a^{3n-2} \\ a^2 & a & 1 & \cdots & a^{3n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a^{3n-1} & a^{3n-2} & a^{3n-3} & \cdots & 1 \end{bmatrix}$$

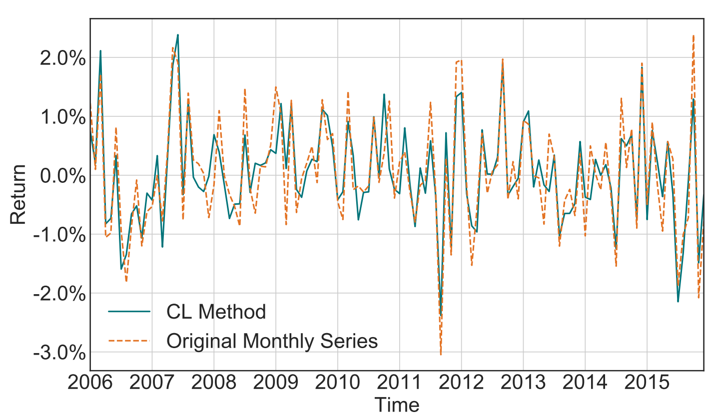
Finally, through substitution, (3.4) becomes

$$\hat{\mathbf{P}}_M = \mathbf{X}_M \hat{\beta} + \mathbf{A}\mathbf{C}^T \left(\mathbf{C}\mathbf{A}\mathbf{C}^T\right)^{-1} \hat{\mathbf{e}}_Q \tag{3.6}$$

The main benefits of this methodology are that it is fast and intuitive. Some cons include a potential bias from using low-frequency betas in a high-frequency setting and that complex residual structures are hard to implement [Mallinger-Dogan and Szigety, 2014]. For instance, one potential source of this bias is that quarterly returns are often much smoother than monthly returns, causing quarterly betas to underestimate monthly volatility. Example inference using the CL Method can be seen in Figure 3.2. Inference is performed on 10 years of simulated monthly data aggregated to quarterly frequency via a monthly proxy with 0.8 correlation. Higher-frequency residuals are assumed to follow an $AR(1)$ process. When proxy goodness is high, the CL Method provides inference that closely matches the shape of the original monthly series.

**Figure 3.2** – CL Method Inference

Example inference using the CL Method.




.

### 3.2.2 Mixed Data Sampling

In 2004, Ghysels, Santa-Clara, and Valkanov proposed the Mixed Data Sampling regression model, hereafter MIDAS, which accommodates independent variables that appear at a higher frequency than the dependent variable [Ghysels et al., 2004]. MIDAS's novelty is its introduction of a parsimonious weighting function that captures the lag structure between the high and low-frequency variables [Ghysels et al., 2004]. In the research, the weighting function specifies the relationship within the quarter between monthly proxies and quarterly observations of the illiquid asset using a small number of parameters. These parameters are estimated via nonlinear least squares.

Let $y_t$ be illiquid data and $x_t^{(m)}$ be the liquid proxy, where $x_t^{(m)}$ is sampled $m$ times faster than $y_t$. In the research, $x_t^{(m)}$ is monthly while $y_t$ is quarterly, so $m = 3$. The goal is to determine the dynamic relationship between $y_t$ and $x_t^{(m)}$. Specifically, $y_t$ is to be projected onto a history of lagged observations of $x_{t-1/m}^{(m)}, x_{t-2/m}^{(m)}, \ldots, x_{t-K/m}^{(m)}$. A single-variable MIDAS regression model looks like [Ghysels et al., 2004, Eric Gysels, 2006]:

$$y_t = \beta_0 + \beta_1 B\left(L^{1/m}; \theta\right) x_t^{(m)} + \epsilon_t^{(m)} \tag{3.7}$$

where:

- $L^{1/m}$ is the lag operator such that $L^{1/m} x_t^{(m)} = x_{t-1/m}^{(m)}$
- $B\left(L^{1/m}; \theta\right) = \sum_{k=0}^{K} B\left(k; \theta\right) L^{k/m}$, $K$ is the max number of lag periods to consider, and $B\left(k; \theta\right)$ is a user-specified function modeling the weight of lag values, parameterized by a small vector of parameters $\theta$.
- Parameters $\beta_0, \beta_1$ capture the overall impact of $x_t^{(m)}$ on $y_t$,
- $\epsilon_t^{(m)}$ is a white noise process.

To attain monthly values $y_t^{(m)}$, the fitted parameters $\beta_0$, $\beta_1$, and $\theta$ are applied[3] to the rolling $K$ monthly proxy observations $x_t^{(m)}$ using (3.7). The max lag length parameter is set to $K = 12$, representing that the trailing year of monthly observations affects the estimates.

In a MIDAS regression, the lag structure is captured by a user-specified function of a few parameters summarized in vector $\theta$ [Eric Gysels, 2006]. For different values of $\theta$, the

[3] MIDAS was not designed for the illiquid inference application in the research. While applications of MIDAS to forecast the next quarterly data point are common [Armesto et al., 2010, Bai et al., 2013, Eric Gysels, 2006], the framework is seldom used to infer monthly data points within the quarter in this manner.

function can take different shapes. Also, the parameterized weights can decrease at different rates as the number of lags increases. Once the functional form of $B(k;\theta)$ is specified by the user, the shape of this function, and therefore the number of lags that are included in the regression, is purely data driven by optimizing for $\theta$ via nonlinear least squares.

MIDAS prevents parameter proliferation by parameterizing $B(k)$ in terms of a small vector of parameters $\theta$, rather than fitting $K$ individual lag weights. [Eric Gysels, 2006]. If instead the parameters of the lagged polynomial were left unrestricted, such that $B(k)$ did not depend on $\theta$, the number of lags would quickly grow large as a function of $K$. For instance, if quarterly observations of $y_t$ were influenced by six quarters of lagged monthly $x_t^{(m)}$'s ($K = 18$), then one would need to fit 18 lags. If instead $x_t^{(m)}$'s were sampled daily, one would need over one hundred parameters to model the same level of dependence.

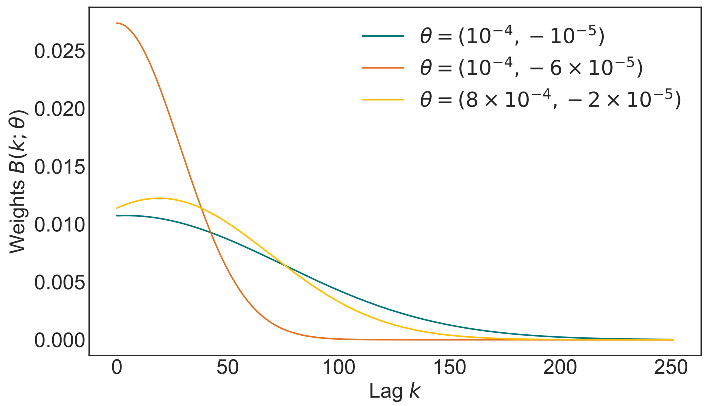
There are several possible polynomial specifications of $B(k;\theta)$, and the two most commonly used are exponential Almon lag and beta lag [Ghysels et al., 2004]. These two families provide positive coefficients that sum up to unity. These characteristics allow for the identification of the scale parameter $\beta_1$ found in the single variable MIDAS regression. Exponential Almon lag relates to the "Almon Lags" that are popular in distributed lag literature [Almon, 1965]. Exponential Almon lag weighting is expressed as

$$B(k;\theta) = \frac{\exp\left(\theta_1 k + \theta_2 k^2 + \cdots + \theta_p k^p\right)}{\sum_{j=1}^{K} \exp\left(\theta_1 j + \theta_2 j^2 + \cdots + \theta_p j^p\right)} \tag{3.8}$$

This parameterization is flexible and can take on various shapes with only a few parameters. The research uses the same functional form with two parameters utilized by Ghysels, Santa-Clara, and Valkanov [Ghysels et al., 2004]. Figure 3.3 plots various shapes of (3.8) for different parameterizations of $\theta$ that are relevant. It is assumed that more recent observations of the proxy should be given more influence than older ones, so cases for which weights are increasing are not considered. For $\theta_1 = \theta_2 = 0$, weights are equal, and this case is not plotted. Weights can decline slowly or fast with the lag. The exponential function can also produce hump shapes. Declining weights are guaranteed as long as $\theta_2 \leq 0$ [Eric Gysels, 2006].

**Figure 3.3** – Exponential Almon Lag MIDAS Weights

The various shapes of the Exponential Almon specification. Weights are plotted on the first 252 lags (corresponding to a year's worth of daily lags) to highlight the flexibility of the specification. Shapes are determined by the values of the parameters $\theta$.




Beta lag is based on the beta distribution, which is known in Bayesian econometrics for flexible, yet parsimonious, prior distributions [Eric Gysels, 2006]. It is specified as

$$B(k;\theta) = \frac{f\left(\frac{k}{K}, \theta_1, \theta_2\right)}{\sum_{j=1}^{K} f\left(\frac{j}{K}, \theta_1, \theta_2\right)} \tag{3.9}$$

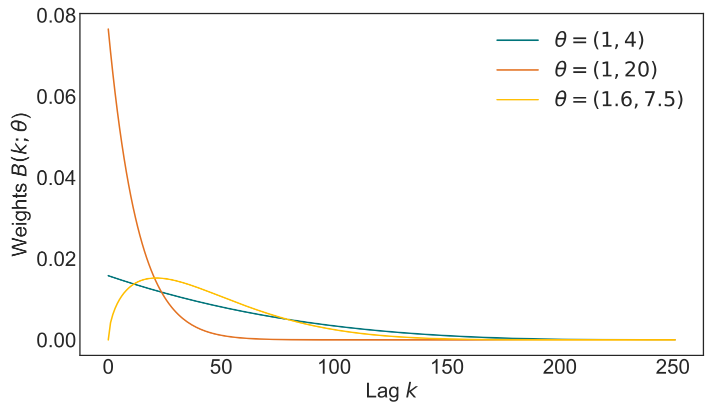
where:

$$\begin{aligned} f(x, c_1, c_2) &= \frac{x^{c_1-1}(1-x)^{c_2-1}}{Beta(c_1, c_2)} \\ Beta(x, y) &= \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} \\ \Gamma(x) &= \int_0^\infty e^{-y} y^{x-1} dy \end{aligned} \tag{3.10}$$

The beta lag parameterization is also quite flexible. Figure 3.4 plots various shapes of (3.9) for different parameterizations of $\theta$. The case of slowly declining weights corresponds to $\theta_1 = 1$ and $\theta_2 > 1$, and as $\theta_2$ increases weights decline faster [Eric Gysels, 2006]. Like the Almon specification, beta lag is also capable of producing hump-shaped patterns.

**Figure 3.4** – Beta Lag MIDAS Weights

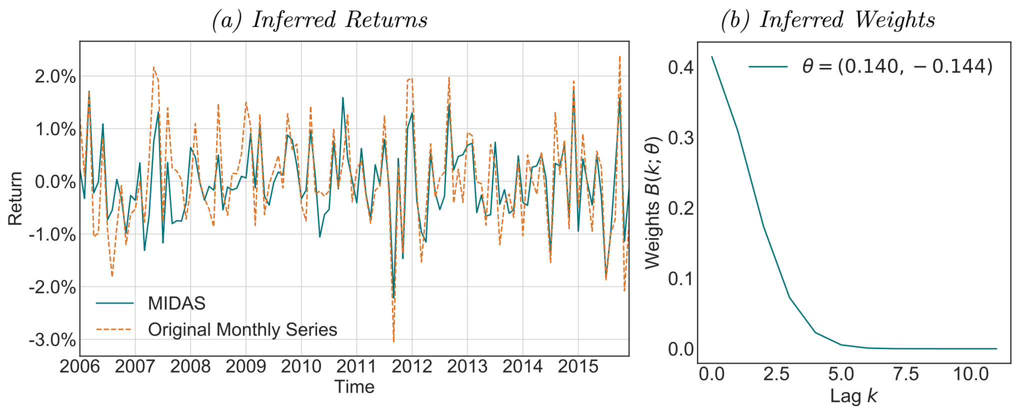
The various shapes of the Beta specification. Weights are plotted on the first 252 lags (corresponding to a year's worth of daily lags) to highlight the flexibility of the specification. Shapes are determined by the values of the parameters $\theta$.




Example inference using MIDAS can be seen in Figure 3.5. The lag length is set $K = 12$ months and two-parameter exponential Almon lag weighting is used. Inference is performed on 10 years of simulated monthly data aggregated to quarterly frequency via a monthly proxy with 0.8 correlation. To attain monthly values $y_t^{(m)}$, the fitted parameters $\beta_0$, $\beta_1$, and $\theta$ are applied to the rolling $K$ monthly proxy observations $x_t^{(m)}$ using (3.7). With a good proxy, the MIDAS framework is powerful enough to provide inference that closely matches the original monthly series, shown by the overall similarity of shape and variance in Panel *(a)*. The fitted parameter values are $\hat{\beta}_0 = -1.224 \times 10^{-3}$, $\hat{\beta}_1 = 0.761$, and $\hat{\theta} = \begin{bmatrix} 0.140 & -0.144 \end{bmatrix}^T$. As seen in Panel *(b)*, this value of $\theta$ corresponds to steeply declining weights, showing that the most recent proxy observation is given the most weight. The trailing 6 months in the year have little to no effect on the estimates. The extremely small value of $\beta_0$ indicates that the proxy has a similar mean return to the illiquid series. A scalar value of $\beta_1 \approx 0.8$ makes intuitive sense, as one would expect this parameter to capture the correlation between the asset and its proxy.

**Figure 3.5** – MIDAS Inference

Example inference using MIDAS, specified in (3.7), with exponential Almon lag weighting and $K = 12$. Inference is performed on 10 years of simulated monthly data aggregated to quarterly frequency via a monthly proxy with 0.8 correlation. Panel *(a)* shows the inferred returns while Panel *(b)* shows the fitted exponential Almon lag weights.




### 3.2.3 Kalman Filtering

Kalman filtering uses a series of measurements over time, sometimes containing noise, and produces estimates of unknown variables driving those measurements. Kalman filtering is a state-space model. state-space models assume that the data is driven by an underlying state process which is observed infrequently. The research assumes that the observed quarterly data is driven by a monthly state process. Kalman filters produce estimates of the underlying state variables by estimating their mean and covariance at each time-step. This section discusses the Kalman Filter models explored in the research and how the filter parameters are estimated.

#### 3.2.3.1 Kalman Filter Model

To apply a Kalman filter, one must first specify a state equation and a measurement equation. The state equation describes the dynamics of the state variables while the measurement equation links the observed series to the hidden state process. The state at the next time-step is influenced by the previous state, observations of the proxy, and potentially noise. The observed data is assumed to be a function of the state and potentially noise. The general form of a Kalman filter is

$$\boldsymbol{\xi}_t = \boldsymbol{F}_t\boldsymbol{\xi}_{t-1} + \boldsymbol{C}_t\boldsymbol{x}_t + \boldsymbol{\alpha}_t + \boldsymbol{u}_t \qquad \text{State Equation} \qquad (3.11)$$

$$\boldsymbol{y}_t = \boldsymbol{H}_t\boldsymbol{\xi}_t + \boldsymbol{v}_t \qquad \text{Measurement Equation}$$

where:

- $\boldsymbol{\xi}_t \in \mathbb{R}^n$ is the state estimate at time $t$
- $\boldsymbol{y}_t \in \mathbb{R}^m$ is the measurement at time $t$, an observed data point
- $\boldsymbol{x}_t \in \mathbb{R}^\gamma$ are values of the control variables (proxies) at time $t$. The parameter $\gamma$ represents the number of control variables that are used.
- $\boldsymbol{F}_t \in \mathbb{R}^{n\times n}$ is the state transition matrix, which relates the state at time $t-1$ to time $t$ in the absence of a driving function or process noise
- $\boldsymbol{C}_t \in \mathbb{R}^{n\times\gamma}$ is the state control matrix, which determines the control variables' effect on the state
- $\boldsymbol{\alpha}_t \in \mathbb{R}^n$ is the state intercept matrix, which controls the drift of the state
- $\boldsymbol{H}_t \in \mathbb{R}^{m\times n}$ is the observation matrix, which specifies how measurements at time $t$ depend on the state at time $t$ in the absence of measurement noise
- $\boldsymbol{u}_t \in \mathbb{R}^n$ is the process noise, assumed to be mean-zero Gaussian such that $\boldsymbol{u}_t \sim \mathcal{N}(0, \boldsymbol{Q}_t)$ where $\boldsymbol{Q}_t \in \mathbb{R}^{n\times n}$ is the process noise covariance
- $\boldsymbol{v}_t \in \mathbb{R}^m$ is the measurement noise, assumed to be mean-zero Gaussian such that $\boldsymbol{v}_t \sim \mathcal{N}(0, \boldsymbol{R}_t)$ where $\boldsymbol{R}_t \in \mathbb{R}^{m\times m}$ is the measurement noise covariance
- The process noise $\boldsymbol{u}_t$ and measurement noise $\boldsymbol{v}_t$ are assumed independent

There is vast literature on the Kalman filter, and this research follows the work of [Mariano and Murasawa, 2010] that uses the filter to estimate the monthly GDP of Switzerland. The research keeps a state of size $n = 3$, since there are three months within a quarter. To simplify the data generation framework, a single proxy is generated for each illiquid asset, so $\gamma = 1$. A single return value is observed quarterly, so $m = 1$. While the Kalman filter is capable of accommodating noise in the measurement equation, the observed quarterly returns are true values rather than noisy values. Thus, $\boldsymbol{v}_t = 0$ for all $t$.

The transition matrix captures underlying state behavior, which is assumed to be auto-regressive order $p$. Though more complex specifications are possible, taking the form ARIMA$(p, d, q)$, assuming AR$(p)$ reduces the number of parameters to estimate. The observation matrix introduces a sum constraint: the three monthly predictions must sum to the observed quarterly value. No other proxy-based inference methodology explored in the research has this feature. The control matrix scales the proxy observation by a constant to account for differences in volatility between the illiquid asset and its proxy.

The research explores four different state equation specifications, with varying degrees of complexity. As the number of parameters increases, estimating the filter becomes much more computationally demanding, so the maximum number of estimated parameters is limited to five. The research's four state equation specifications are outlined in Table 3.2.

| Name | #Params | State Process | Intercept | Proxies | Specification |
|---|---|---|---|---|---|
| Non-AR Proxy-Based | 3 | None | Yes | Yes | $y_t = cx_t + \alpha + qu_t$ |
| Simple AR Proxy-Based | 3 | $AR\,(1)$ | No | Yes | $y_t = \phi_1 y_{t-1} + cx_t + qu_t$ |
| Complex AR Proxy-Based | 5 | $AR\,(2)$ | Yes | Yes | $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + cx_t + \alpha + qu_t$ |
| Non Proxy-Based | 4 | $AR\,(2)$ | Yes | No | $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \alpha + qu_t$ |

**Table 3.2** – Kalman Filter Configurations

Names and descriptions for the Kalman filter configurations explored in the research. #Params represents the size of the estimated parameter space. $y_t$, $x_t$, and $u_t$ are the state, measurement of the proxy, and residual respectively at time $t$. The parameters $\phi_1$, $\phi_2$, $c$, $\alpha$, and $q$ are estimated via maximum likelihood. The notation $AR\,(p)$ represents an auto-regressive order $p$ process.

The complex AR proxy-based model is as follows[4]:

$$\underbrace{\begin{bmatrix} z_{t+1} \\ z_t \\ z_{t-1} \end{bmatrix}}_{\boldsymbol{\xi}_t} = \underbrace{\begin{bmatrix} \phi_1 & \phi_2 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}}_{\boldsymbol{F}} \underbrace{\begin{bmatrix} z_t \\ z_{t-1} \\ z_{t-2} \end{bmatrix}}_{\boldsymbol{\xi}_{t-1}} + \underbrace{\begin{bmatrix} c \\ 0 \\ 0 \end{bmatrix}}_{\boldsymbol{C}} \boldsymbol{x}_{t+1} + \underbrace{\begin{bmatrix} \alpha \\ 0 \\ 0 \end{bmatrix}}_{\boldsymbol{\alpha}} + \underbrace{\begin{bmatrix} q & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}}_{\boldsymbol{U}} \underbrace{\begin{bmatrix} u_{t+1} \\ u_t \\ u_{t-1} \end{bmatrix}}_{\boldsymbol{u}_t}$$

$$\boldsymbol{y}_t = \boldsymbol{h}_t \boldsymbol{\xi}_t \tag{3.12}$$

[4]The non-AR proxy-based model forces $\phi_1 = \phi_2 = 0$, the non proxy-based model forces $c = 0$, and the simple proxy-based model forces $\phi_2 = \alpha = 0$.

where:

- $\boldsymbol{\xi}_t = \begin{bmatrix} z_{t+1} & z_t & z_{t-1} \end{bmatrix}^T$ is the state at time $t$. It is the stacked vector dynamics of the unobserved monthly returns.

- The transition matrix $\boldsymbol{F}$ assumes an $AR(2)$ process drives the state according to $z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \text{other terms}$

- The observation matrix enforces the sum constraint, that monthly predictions should sum to the observed quarterly values:

$$\boldsymbol{h}_t = \begin{cases} \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T & \text{for } t = 1, 2, 4, 5, 7, \ldots \\ \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T & \text{for } t = 3, 6, 9, \ldots \end{cases} \tag{3.13}$$

- A single proxy $\boldsymbol{x}$ impacts the state through the control matrix[5]

- The time-invariant parameters $c$, $\phi_1$, $\phi_2$, and $\alpha$, and $q$ are estimated via maximum likelihood.

Example inference using two configurations of the Kalman filter can be seen in Figure 3.6. Inference is performed on 10 years of simulated monthly data aggregated to quarterly frequency via a monthly proxy with correlation 0.8. Data were simulated using a simple mean-zero Brownian motion process. One filter configuration uses a proxy for inference while the other is non proxy-based. The proxy-based configuration captures the overall shape of the monthly series with similar variance to the true monthly returns. Its fitted state equation is given by

$$\underbrace{\begin{bmatrix} z_{t+1} \\ z_t \\ z_{t-1} \end{bmatrix}}_{\boldsymbol{\xi}_t} = \underbrace{\begin{bmatrix} -0.13 & -0.08 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}}_{\boldsymbol{F}} \underbrace{\begin{bmatrix} z_t \\ z_{t-1} \\ z_{t-2} \end{bmatrix}}_{\boldsymbol{\xi}_{t-1}} + \underbrace{\begin{bmatrix} 0.87 \\ 0 \\ 0 \end{bmatrix}}_{\boldsymbol{C}} \boldsymbol{x}_{t+1} + \underbrace{\begin{bmatrix} \approx 0 \\ 0 \\ 0 \end{bmatrix}}_{\boldsymbol{\alpha}} + \underbrace{\begin{bmatrix} 0.51 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}}_{\boldsymbol{U}} \underbrace{\begin{bmatrix} u_{t+1} \\ u_t \\ u_{t-1} \end{bmatrix}}_{\boldsymbol{u}_t}$$

The parameter value $c \approx 0.8$ makes sense as the proxy has a 0.8 correlation and approximately the same variance as the true monthly series. A value of $\alpha \approx 0$ indicates that there is little to no drift in the data, which is true, since the driving process in this case is mean-zero Brownian

[5] For a discussion of this design choice, see Section 5.5

motion. Values of $\phi_1$ and $\phi_2$ close to 0 indicate that there is insignificant autocorrelation in the data, which is also the case.

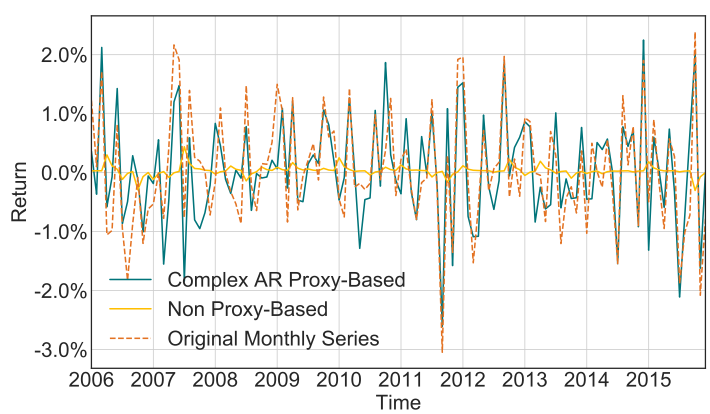
Since the non proxy-based configuration has no monthly proxy, it observes only quarterly returns, which have lower variance than the true monthly returns. Therefore, the non-proxy configuration has a denoising effect and dramatically reduces the variance of the inferred monthly series. In the non-proxy configuration, the inferred series is not the Kalman filter mean estimate. Rather, the inferred series is a distribution with mean given by the Kalman filter mean estimate and variance computable from the Kalman filter. The non proxy-based configuration's fitted state equation is given by

$$\underbrace{\begin{bmatrix} z_{t+1} \\ z_t \\ z_{t-1} \end{bmatrix}}_{\boldsymbol{\xi}_t} = \underbrace{\begin{bmatrix} 0.27 & 0.01 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}}_{\boldsymbol{F}} \underbrace{\begin{bmatrix} z_t \\ z_{t-1} \\ z_{t-2} \end{bmatrix}}_{\boldsymbol{\xi}_{t-1}} + \underbrace{\begin{bmatrix} 0.55 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}}_{\boldsymbol{U}} \underbrace{\begin{bmatrix} u_{t+1} \\ u_t \\ u_{t-1} \end{bmatrix}}_{\boldsymbol{u}_t}$$

Since $\phi_1$ and $\phi_2$ are again small, the filter's mean estimate is close to zero, causing the aforementioned smoothing effect.

**Figure 3.6** – Kalman Filter Inference

Example inference using two Kalman Filter configurations.




#### 3.2.3.2 Estimating the Kalman Filter

In order to estimate the system in (3.12), Kalman filters use a form of feedback control: the filter predicts using the state at each time-step and then obtains feedback in the form of measurements. The state equation is used to project forward (in time) the current state and error covariance estimates to obtain the *a priori* estimates for the next time-step. The measurement equation is responsible for providing the feedback to update the *a priori* prediction to an improved *a posteriori* estimate incorporating the measurement [Welch and Bishop, 2006]. The following discussion derives this estimation algorithm.

Let $\hat{\boldsymbol{\xi}}_t^- \in \mathbb{R}^n$ be the *a priori* state estimate at time $t$ given knowledge of the process prior to step $t$ and $\hat{\boldsymbol{\xi}}_t \in \mathbb{R}^n$ be the *a posteriori* state estimate at time $t$ given measurement $\boldsymbol{y}_t$. The *a priori* and *a posteriori* estimate errors are therefore defined as

$$\begin{aligned}\boldsymbol{e}_t^- &\equiv \boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_t^-, \text{ and}\\ \boldsymbol{e}_t &\equiv \boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_t\end{aligned}$$

The *a priori* estimate error covariance is then

$$\boldsymbol{P}_t^- = \mathbb{E}\left[\boldsymbol{e}_t^- \boldsymbol{e}_t^{-T}\right] \tag{3.14}$$

and the *a posteriori* estimate error covariance is

$$\boldsymbol{P}_t = \mathbb{E}\left[\boldsymbol{e}_t \boldsymbol{e}_t^T\right] \tag{3.15}$$

Deriving the equations for estimating the Kalman filter begins with the goal of finding an equation that computes the *a posteriori* state estimate $\hat{\boldsymbol{\xi}}_t$ as a linear combination of an *a priori* estimate $\hat{\boldsymbol{\xi}}_t^-$ and a weighted difference between an actual measurement $\boldsymbol{y}_t$ and a measurement prediction $\boldsymbol{H}\hat{\boldsymbol{\xi}}_t^-$ as shown below [Welch and Bishop, 2006]:

$$\hat{\boldsymbol{\xi}}_t = \hat{\boldsymbol{\xi}}_t^- + \boldsymbol{K}\left(\boldsymbol{y}_t - \boldsymbol{H}\hat{\boldsymbol{\xi}}_t^-\right) \tag{3.16}$$

This difference $\left(\boldsymbol{y}_t - \boldsymbol{H}\hat{\boldsymbol{\xi}}_t^-\right)$ is called the measurement *innovation* or the *residual* and reflects the discrepancy between the predicted measurement $\boldsymbol{H}\hat{\boldsymbol{\xi}}_t^-$ and the actual measurement $\boldsymbol{y}_t$. A residual of zero means the two are in complete agreement. The matrix $\boldsymbol{K} \in \mathbb{R}^{n\times n}$, called the Kalman gain, is chosen to be the blending factor that minimizes the *a posteriori* error covariance given in (3.15). This minimization can be accomplished by substituting (3.16) into (3.15), performing the necessary expectations, taking the derivative of the trace of the result with respect to $\boldsymbol{K}$, setting that result equal to zero, and then solving for $\boldsymbol{K}$ [Welch and Bishop, 2006].

One form of the resulting $\boldsymbol{K}$ that minimizes (3.16) is given by

$$\boldsymbol{K}_t = \boldsymbol{P}_t^- \boldsymbol{H}^T \left( \boldsymbol{H} \boldsymbol{P}_t^- \boldsymbol{H}^T + \boldsymbol{R} \right)^{-1} \tag{3.17}$$

Weighting by $\boldsymbol{K}$ establishes a clear trade off between the measurement error covariance and estimate error covariance. Looking at (3.17), it is clear that as the measurement error covariance $\boldsymbol{R}$ approaches zero, the gain $\boldsymbol{K}$ weights the residual more strongly. Specifically,

$$\lim_{\boldsymbol{R}_t \to \boldsymbol{0}} \boldsymbol{K}_t = \boldsymbol{H}^{-1}$$

Conversely, as the a priori estimate covariance $\boldsymbol{P}_t^-$ approaches zero, the gain $\boldsymbol{K}$ weights the residual less heavily. Specifically,

$$\lim_{\boldsymbol{P}_t^- \to \boldsymbol{0}} \boldsymbol{K}_t = \boldsymbol{0}$$

Another way to think about the weighting by $\boldsymbol{K}$ is that as the measurement error covariance $\boldsymbol{R}$ approaches zero, the actual measurement $\boldsymbol{y}_t$ is trusted more and more, while the predicted measurement $\boldsymbol{H}\hat{\boldsymbol{\xi}}_t^-$ is trusted less and less. On the other hand, as the a priori estimate error covariance $\boldsymbol{P}_t^-$ approaches zero, the actual measurement $\boldsymbol{y}_t$ is trusted less and less, while the predicted measurement $\boldsymbol{H}\hat{\boldsymbol{\xi}}_t^-$ is trusted more and more. In the research, since true quarterly return values are observed, the measurement error $\boldsymbol{R} \approx \boldsymbol{0}$, so the actual measurements $\boldsymbol{y}_t$ are trusted completely.

Using the above equations, the Kalman filter estimation algorithm follows. The algorithm is recursive, alternating between a "predict" and "update" phase on each time-step. In prediction, the filter uses the previous *a posteriori* state estimate $\hat{\boldsymbol{\xi}}_{t-1}$ and error covariance $\boldsymbol{P}_{t-1}$ to predict current *a priori* estimates. In the update phase, the filter computes the current *a posteriori* state estimate $\hat{\boldsymbol{\xi}}_t$ and error covariance $\boldsymbol{P}_t$ given the measurement $\boldsymbol{y}_t$. The algorithm is as follows [Reid and Term, 2001]:

- **Predict:**

$$\begin{aligned} \hat{\boldsymbol{\xi}}_t^- &= \boldsymbol{F}\hat{\boldsymbol{\xi}}_{t-1} + \boldsymbol{C}\boldsymbol{x}_{t-1} + \boldsymbol{\alpha} && \text{Project the state ahead} \\ \boldsymbol{P}_t^- &= \boldsymbol{F}\boldsymbol{P}_{t-1}\boldsymbol{F}^T + \boldsymbol{Q}_{t-1} && \text{Project the error covariance ahead} \end{aligned} \tag{3.18}$$

- **Update:**

$$\boldsymbol{K}_t = \boldsymbol{P}_t^- \boldsymbol{H}^T \left( \boldsymbol{H} \boldsymbol{P}_t^- \boldsymbol{H}^T + \boldsymbol{R}_t \right)^{-1} \quad \text{Compute the Kalman gain} \tag{3.19}$$

$$\hat{\boldsymbol{\xi}}_t = \hat{\boldsymbol{\xi}}_t^- + \boldsymbol{K}_t \left( \boldsymbol{y}_t - \boldsymbol{H} \hat{\boldsymbol{\xi}}_t^- \right) \quad \text{Update estimate with measurement } \boldsymbol{y}_t$$

$$\boldsymbol{P}_t = \left( \boldsymbol{I} - \boldsymbol{K}_t \boldsymbol{H} \right) \boldsymbol{P}_t^- \quad \text{Update the error covariance}$$

The parameters of the filter are estimated via maximum likelihood.

#### 3.2.3.3 Deriving the Likelihood Function

Many signals can be described as

$$\boldsymbol{y}_t = \boldsymbol{a}_t \boldsymbol{x}_t + \boldsymbol{n}_t$$

where $\boldsymbol{y}_t$ is the observed signal, $\boldsymbol{a}_t$ is a gain term, $\boldsymbol{x}_t$ is the information-bearing signal, and $\boldsymbol{n}_t$ is additive noise [Thacker and Lacey, 2006]. The overall objective is to estimate $\boldsymbol{x}_t$. The difference between the estimate of $\hat{\boldsymbol{x}}_t$ and $\boldsymbol{x}_t$ itself is termed the error

$$f\left(\boldsymbol{e}_t\right) = f\left(\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t\right)$$

The particular shape of $f\left(\boldsymbol{e}_t\right)$ depends on the application, but it is clear that the function should be both positive and monotonically increasing [Thacker and Lacey, 2006]. One error function with these properties is the squared error function

$$f\left(\boldsymbol{e}_t\right) = \left(\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t\right)^2$$

Since the filter is applied over many data points over a period of time, a more meaningful metric is the expected value of the error function. This results in the mean squared error (MSE) loss function

$$\epsilon\left(t\right) = \mathbb{E}\left[\left(\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t\right)^2\right] \tag{3.20}$$

This derivation for MSE, though intuitive, is a heuristic. A more rigorous derivation follows using maximum likelihood. This is achieved by restating the goal of the filter to finding the $\hat{\boldsymbol{x}}$ that maximizes the probability of observing the data $\boldsymbol{y}$

$$\hat{\boldsymbol{x}}_{\text{best}} = \arg\max_{\hat{\boldsymbol{x}}} P\left(\boldsymbol{y} \mid \hat{\boldsymbol{x}}\right)$$

Assuming that the additive random noise is Gaussian distributed with a standard deviation $\sigma_t$ gives

$$P\left(\boldsymbol{y}_t \mid \hat{\boldsymbol{x}}_t\right)=\frac{1}{\sqrt{2\pi\sigma_t^2}}\exp\left[-\frac{\left(\boldsymbol{y}_t-\boldsymbol{a}_t\hat{\boldsymbol{x}}_t\right)^2}{2\sigma_t^2}\right]$$

Assuming mutual independence of the error terms gives the maximum likelihood function of the data

$$P\left(\boldsymbol{y} \mid \hat{\boldsymbol{x}}\right)=\prod_{t=1}^{T}\frac{1}{\sqrt{2\pi\sigma_t^2}}\exp\left[-\frac{\left(\boldsymbol{y}_t-\boldsymbol{a}_t\hat{\boldsymbol{x}}_t\right)^2}{2\sigma_t^2}\right]$$

Taking the log of both sides gives the log-likelihood function

$$\mathcal{L}\left(\boldsymbol{y} \mid \hat{\boldsymbol{x}}\right)=-\frac{T}{2}\sum_{t=1}^{T}\ln\left(2\pi\right)-\frac{1}{2}\sum_{t=1}^{T}\ln\sigma_t^2-\frac{1}{2}\sum_{t=1}^{T}\frac{\left(\boldsymbol{y}_t-\boldsymbol{a}_t\hat{\boldsymbol{x}}_t\right)^2}{2\sigma_t^2} \tag{3.21}$$

This function (3.21) is driven by sum of squared error terms in the same way as the MSE loss function (3.20). The measurement function of the Kalman filter can be estimated as

$$\boldsymbol{y}_t=\boldsymbol{H}\hat{\boldsymbol{\xi}}_t^-+\boldsymbol{n}_t$$

where $\boldsymbol{y}_t$ is the observed series, $\boldsymbol{H}\hat{\boldsymbol{\xi}}_t^-$ is the prediction, and $\boldsymbol{n}_t$ is added noise. The variance of this prediction is given by

$$\operatorname{var}\left(\boldsymbol{H}\hat{\boldsymbol{\xi}}_t^-\right)=\boldsymbol{H}^T\boldsymbol{P}_t^-\boldsymbol{H}$$

where $\boldsymbol{P}_t^-$ is the *a priori* estimate error covariance of the *a priori* state estimate $\hat{\boldsymbol{\xi}}_t^-$. Therefore, assuming normally distributed observations of the illiquid series, given previous observations and observations of the proxy,

$$\forall t=1,\ldots,T:\ \boldsymbol{y}_t \mid \left(\boldsymbol{y}_0,...,\boldsymbol{y}_{t-1},\boldsymbol{x}_1,...,\boldsymbol{x}_t\right)\sim\mathcal{N}\left(\boldsymbol{H}\hat{\boldsymbol{\xi}}_t^-,\boldsymbol{H}^T\boldsymbol{P}_t^-\boldsymbol{H}\right)$$

and substituting into (3.21), the log-likelihood of the filter is given by [Cuche and Hess, 2000]

$$\mathcal{L}\left(\boldsymbol{y} \mid \hat{\boldsymbol{x}}\right)=-\frac{T}{2}\ln\left(2\pi\right)-\frac{1}{2}\sum_{t=1}^{T}\ln\left|\boldsymbol{H}^T\boldsymbol{P}_t^-\boldsymbol{H}\right|-\frac{1}{2}\sum_{t=1}^{T}\frac{\left(\boldsymbol{y}_t-\boldsymbol{H}\hat{\boldsymbol{\xi}}_t\right)^2}{\boldsymbol{H}^T\boldsymbol{P}_t^-\boldsymbol{H}} \tag{3.22}$$

# Chapter 4

# Multi-Factor Market Simulator

The Multi-Factor Market Simulator generates ten years of monthly asset returns (and corresponding monthly proxies) according to specified parameters with a covariance structure mirroring realized asset class index returns. These parameters vary along three dimensions: market tail risk, data autocorrelation, and availability of a good proxy. To remove the effects of initial conditions, the framework generates several thousand monthly observations and keeps only the last ten years for each asset. Returns are generated for four liquid asset classes, observed monthly, and three illiquid asset classes, observed quarterly. A single monthly proxy is generated for each illiquid asset.

This chapter first discusses how the Multi-Factor Market Simulator models each of the three parameter dimensions — market tail risk, data autocorrelation, and availability of a good proxy — individually and then concludes with the data simulation algorithm.

## 4.1 Market Tail Risk

To generate asset returns with varying levels of market tail risk, the Multi-Factor Market Simulator employs a Merton jump-diffusion process ([Carr and Wu, 2007, Kelly et al., 2016], among others). Merton's jump-diffusion process is a model for asset returns that incorporates small day-to-day "diffusion-type" movements with larger randomly occurring "jumps" [Merton, 1976]. By controlling the frequency of these "jumps", the simulator controls how often market crashes occur. The research defines market tail risk as the likelihood of a market crash.

The risk-neutral jump-diffusion process for the stock price is [Merton, 1976, Lyuu, 2015]

$$\frac{d\boldsymbol{S}_t}{\boldsymbol{S}_t} = \left(\mu - \lambda\bar{k}\right) dt + \sigma d\boldsymbol{W}_t + k dq_t \tag{4.1}$$

where $\mu$ and $\sigma$ are the instantaneous expected return and volatility of the stock and $d\boldsymbol{W}_t$ is an increment of Brownian motion. The jump event is governed by a compound Poisson process $q_t$ with intensity $\lambda$, where $k$ represents the magnitude of the *random* jump [1]. The distribution of $k$ obeys $\ln(1+k) \sim \mathcal{N}\left(\mu_q, \sigma_q^2\right)$ with mean $\bar{k} \equiv \mathbb{E}[k] = e^{\mu_q + \sigma_q^2/2} - 1$. The jump-diffusion model is flexible in that it can model jump risks that are symmetrical about zero or concentrate more on large negative returns, by choice of $\mu_q$ and $\sigma_q$. The research assumes that market tail risks are symmetric and sets $\mu_q = 0$, since the research's inference models are unbiased to whether large dislocations are positive or negative. The solution to 4.1 is [Merton, 1976, Lyuu, 2015]

$$\boldsymbol{S}_t = \boldsymbol{S}_0 e^{\left(r - \lambda\bar{k} - \sigma^2/2\right)t + \sigma\boldsymbol{W}_t} U\left(n\left(t\right)\right) \tag{4.2}$$

where:

$$U\left(n\left(t\right)\right) = \prod_{i=0}^{n(t)} \left(1 + k_i\right) \tag{4.3}$$

- $k_i$ is the magnitude of the $i$th jump with $\ln(1+k_i) \sim \mathcal{N}\left(\mu_q, \sigma_q^2\right)$
- $k_0 = 0$
- $n(t)$ is the number of jumps up to time $t$ for the Poisson process with intensity $\lambda$

To generate asset price paths in a Monte Carlo fashion, one can evolve

$$\boldsymbol{S}_{t+\Delta t} = \boldsymbol{S}_t e^{\left(\mu - \lambda\bar{k} - \sigma^2/2\right)\Delta t + \sigma d\boldsymbol{W}_t} U_t \tag{4.4}$$

where:

$$U_t = e^{p \times \mu_q + \sqrt{p} \times \sigma_q \times \mathcal{N}(0,1)} \tag{4.5}$$

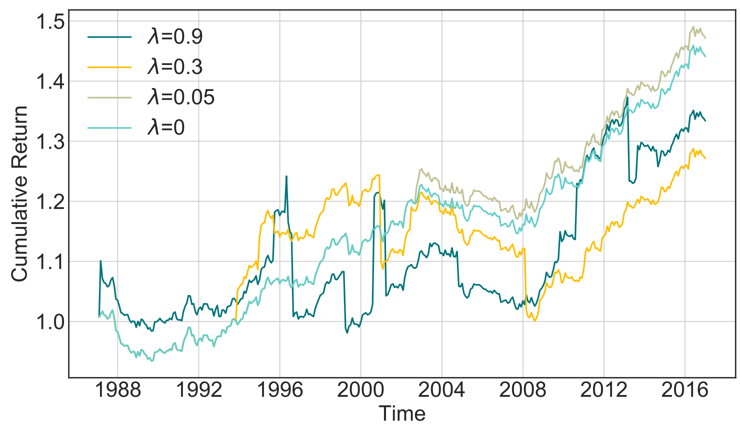
and $p \sim \text{Poisson}\left(\lambda \times \Delta t\right)$.

[1] When $\lambda = 0$, the model reduces to the Black-Scholes model.

Figure 4.1 shows data simulated by varying the jump risk intensity $\lambda$. As $\lambda$ increases large price dislocations become more frequent. The jumps appear to be symmetrically distributed about 0, as expected given the parameters.

**Figure 4.1** – Generating Data as a Function of Jump Risk

An example simulation of a series where the jump-risk parameter $\lambda$ varies in the set $\{0, 0.05, 0.3, 0.9\}$, representing zero, low, medium, and high frequency of jumps. Correspondingly, a jump is expected to occur every$\{\text{never}, 20, \approx 3, \approx 1\}$ years, since $1/\lambda$ is the expected time between jumps. Expected jump size $\mu_q = 0$, and for the purpose of easy visualization, the volatility of jump size is set to an unrealistically large value $\sigma_q = 0.2$. In actual simulations, the default value of this parameter is $0.2/4 = 0.05$.




## 4.2 Data Autocorrelation

To generate asset returns with varying levels of autocorrelation, the Multi-Factor Market Simulator employs a fractional Brownian motion (fBm) process. Fractional Brownian motion is a generalization of Brownian motion in which the increments are correlated according to a user-specified parameter. The research utilizes fBm to govern the day-to-day diffusive movements, replacing the Brownian motion component in Merton's jump-diffusion process. By using fBm in conjunction with Merton's jump-diffusion, the research parameterizes data autocorrelation and market tail risk.

Fractional Brownian motion starts at zero, has expectation zero for all times $t \in [0, T]$, and has the following covariance function:

$$\mathbb{E}\left[\boldsymbol{W}^H(t)\,\boldsymbol{W}^H(s)\right] = \frac{1}{2}\left(|t|^{2H} + |s|^{2H} - |t-s|^{2H}\right) \tag{4.6}$$

where $H$ is a real number in $(0, 1)$ called the Hurst index associated with the fBm process [Mandelbrot and Ness, 1968]. $H$ describes the raggedness of the resultant motion, with a higher value leading to a smoother motion. Specifically,

- if $H = 1/2$, the process is a Brownian motion process. The covariance evaluates to $\mathbb{E}\left[\boldsymbol{W}^{1/2}(t)\,\boldsymbol{W}^{1/2}(s)\right] = \frac{1}{2}(|t| + |s| - |t - s|) = \min(s, t)$, the same as Brownian motion.

- if $H > 1/2$, the covariance is positive, so increments of the process are positively correlated

- if $H < 1/2$, the covariance is negative, so increments of the process are negatively correlated

To simulate values of the fBm at times $t_1, \ldots, t_n$, one can perform the following steps [Dieker, 2004, Asmussen et al., 1998]:

1. Compute the variance-covariance matrix

$$\Gamma = \begin{bmatrix} R(t_1,t_1) & R(t_1,t_2) & \cdots & R(t_1,t_n) \\ R(t_2,t_1) & R(t_2,t_2) & \cdots & R(t_2,t_n) \\ \vdots & \vdots & \ddots & \vdots \\ R(t_n,t_1) & R(t_n,t_2) & \cdots & R(t_n,t_n) \end{bmatrix} \tag{4.7}$$

   where $R(s,t) = \frac{1}{2}\left(s^{2H} + t^{2H} - |t - s|^{2H}\right)$

2. Compute $\Lambda$, the square root matrix of $\Gamma$, given by $\Lambda^2 = \Gamma$ using the Cholesky decomposition method. Loosely speaking, $\Lambda$ is the "standard deviation" matrix associated with the variance-covariance matrix $\Gamma$.

3. Generate a Gaussian random vector $v = (v_1, \ldots, v_n)^T$, such that $v_i \sim \mathcal{N}(0, 1)$. Applying the Cholesky matrix correlates the increments of the fBm process, producing the correlated random vector $d\boldsymbol{W}^H = \Lambda v$.

Given a weakly stationary time series $\boldsymbol{x}$, which is true of fBm increments, the sample autocorrelation can be computed as a function of the time lag. The lag $\tau$ autocorrelation of $\boldsymbol{x}$ is

$$R(\tau, \boldsymbol{x}) = \frac{\mathbb{E}\left[(\boldsymbol{x}_t - \mu)(\boldsymbol{x}_{t+\tau} - \mu)\right]}{\sigma^2} \tag{4.8}$$

where $\mu = \bar{\boldsymbol{x}}$ and $\sigma^2 = \mathrm{var}(\boldsymbol{x})$. The autocorrelation function of fBm can be computed in closed from from the fBm autocovariance function. The autocorrelation function of fBm as a function of lag length $\tau$ follows from (4.6) and is given by [Tarnopolski, 2006]

$$\rho(\tau, H) = \frac{1}{2}\left(|\tau + 1|^{2H} - 2|\tau|^{2H} + |\tau - 1|^{2H}\right) \tag{4.9}$$

As expected, $\rho(\tau, H) \in [0, 1]$ and takes on negative values for $H < 1/2$, a value of 0 for $H = 1/2$ and positive values for $H > 1/2$. A Taylor expansion of $\rho(\tau)$ from (4.9) gives

$$\frac{\rho(\tau)}{H(2H-1)|\tau|^{2H-2}} \to 1$$

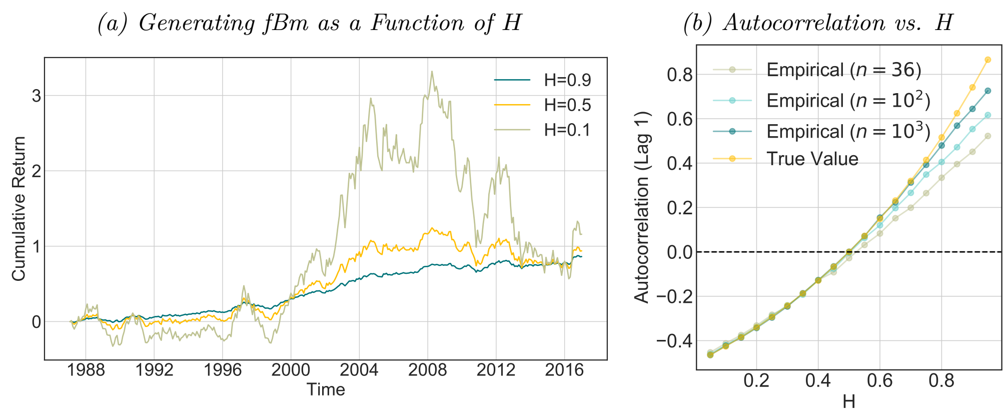
as $\tau \to \infty$. It follows that for $H > 1/2$, the autocorrelation $\rho(\tau)$ behaves like $|\tau|^{2H-2}$ for large $\tau$.

Figure 4.2 visualizes fBm processes simulated by varying the parameters $H$ and number of data points $n$. In the left panel, the figure shows three fBm processes simulated by varying $H \in \{0.1, 0.5, 0.9\}$where $n = 360$, corresponding to 30 years of monthly observations. The left panel shows visually that increasing the Hurst index $H$ increases autocorrelation, smoothing the resultant series. A value of $H = 0.5$ means there is no autocorrelation. $H = 0.1$ corresponds to strong negative autocorrelation while $H = 0.9$ corresponds to strong positive autocorrelation. There is no drift in these series, since fBm increments have expectation 0.

The right panel shows the effect of $H$ and $n$ on autocorrelation empirically, by plotting the mean lag-1 autocorrelation coefficients over 1000 trials of simulated fBm increments as a function of $H$ and $n$ ("Empirical $n =$..."). It juxtaposes these empirical values with true values ("True Value") computed using (4.9). $H$ varies in the range $H \in [0, 0.95]$ in increments of 0.05 while $n$ varies in $n \in \{36, 100, 1000\}$. For extremely low values of $H$ both empirical and true autocorrelations are strongly negative and take on similar values. As $H$ increases, both empirical autocorrelations and true autocorrelation increase monotonically, but the empirical autocorrelations diverge from the true autocorrelation. This effect is due to the fact that as $H$ becomes very large with small $n$, the fBm process becomes more difficult to simulate. When $n = 36$, the empirical value for $H = 0.95$ of 0.52 disagrees with the exact value indicated by the function of $\rho(\tau = 1, H = 0.95) \approx 0.87$. When repeating the experiment with $n = 1000$ increments instead of $n = 36$, the empirical autocorrelation of 0.72 much more closely matches this value. In general, as $n$ increases, the empirical autocorrelations approach true values. Theoretically, as $n \to \infty$, the empirical mean autocorrelations should converge to true values.

**Figure 4.2** – Generating Data as a Function of Autocorrelation

Visualization and autocorrelation coefficients for simulated fBm processes as a function of $H$. Panel *(a)* shows single simulations of fBm processes where the autocorrelation parameter $H$ varies in the set $\{0.1, 0.5, 0.9\}$ and $n = 360$, corresponding to 30 years of monthly data. Panel *(b)* shows mean lag 1 autocorrelation coefficients over 1000 simulations of fBm increments ("Empirical $n =$...") where $n$ varies in $n \in \{36, 100, 1000\}$ as well as true autocorrelation values ("True Value") computed directly from the fBm autocorrelation function for each value of $H \in [0.05, 0.95]$ .




## 4.3 Proxy Goodness

To generate proxies in the framework true monthly returns are perturbed with varying levels of noise. Proxy goodness is defined as the inverse of the level of noise used to perturb true returns. The smaller the amount of noise is used, the better the proxy is. The research's methodology for generating proxies is inspired by the Cholesky decomposition method, from which one can derive that if $x, y$ are independent and identically distributed standard Gaussians, then $\text{corr}\left(x, \rho x + \sqrt{1-\rho^2}y\right) = \rho \in [0, 1]$. In the research, increments are neither independent (since fractional Brownian motion is used) nor identically distributed (since independent jump-diffusion processes affect all assets). Therefore, references to the correlation coefficient $\rho$ mean "approximate correlation" rather than true correlation. Loosely speaking, the higher the value of $\rho$, the less noise is being used to perturb the asset returns. In the case of $\rho = 1$, the proxy is the exact same as the monthly series.

Proxies are generated according to the following methodology:

1. Generate a monthly series using Merton's jump-diffusion for market dislocations in

conjunction with fBm for the day-to-day increments, as explained in the next section. Let this series be $d\boldsymbol{S}_t$ with corresponding jump-diffusion component $d\boldsymbol{U}_t$.

2. Generate an independent Brownian motion process $d\boldsymbol{W}_t$, with the same drift and variance as the asset's fBm process. Apply $d\boldsymbol{U}_t$ to $d\boldsymbol{W}_t$ to produce a new series $d\boldsymbol{S}_t^{'}$. The framework assumes that an asset and its proxy are exposed to the same tail risks, so the simulator uses the same jump-diffusion process to generate the original series and its corresponding proxy. This assumption allows proxies to supply higher-frequency information about the timing of dislocation events to improve inference of the proxy-based models. Under conditions of market stress, asset classes can have increased correlation and lower diversification benefits, and this effect is captured by assuming assets and their corresponding proxies are exposed to the same tail risks. Formally, this step uses (4.1) and looks like

$$\begin{aligned} \frac{d\boldsymbol{S}_t^{'}}{\boldsymbol{S}_t^{'}} &= d\boldsymbol{U}_t + \sigma d\boldsymbol{W}_t \\ d\boldsymbol{U}_t &= \left(\mu - \lambda \bar{k}\right) dt + k dq_t \end{aligned} \tag{4.10}$$

One would not expect high-frequency returns to be smoothed to the same extent as illiquid returns. As a result, this step includes the design choice to generate proxies from an independent geometric Brownian motion process, where increments have zero autocorrelation, rather than a fractional Brownian motion process with positive autocorrelation.

3. Perturb the original monthly series with noise, represented by this new process. The proxy becomes

$$d\boldsymbol{P}_t = \rho d\boldsymbol{S}_t + \sqrt{1-\rho^2} d\boldsymbol{S}_t^{'} \tag{4.11}$$

Empirically, it is observed that $\operatorname{corr}(d\boldsymbol{P}_t, d\boldsymbol{S}_t) \approx \rho$ for the majority of cases. In the cases where $\operatorname{corr}(d\boldsymbol{P}_t, d\boldsymbol{S}_t)$ is meaningfully different from $\rho$, then step (2) is repeated until a satisfactory proxy is generated. Meaningfully different is defined as $\rho \notin [r-\delta, r+\delta]$, where $r$ is the empirically observed Pearson correlation coefficient between the asset returns and the proxy and $\delta = 0.1$. Intuitively, this parameter $\delta$ should depend on the lengths of the series that are generated, since the significance of correlations depends on sample size. However, in the research, the lengths of series are fixed to 120 monthly data points. Empirically, a parameter value of $\delta = 0.1$ allows for proxies with correlation close to $\rho$ to be generated using reasonable computational resources.

Figure 4.3 shows three proxies generated with varying levels of goodness, with $\rho \in \{0, 0.5, 0.9\}$. This figure depicts visually that as $\rho$ increases, proxies increase in similarity to the original series.

**Figure 4.3** – Proxy Generation of Varying Goodness

Example generation of three monthly proxies spanning three years with varying levels of goodness. Proxy goodness ranges in $\rho \in \{0, 0.5, 0.9\}$. For simplicity, these series were generated with zero jump risk ($\lambda = 0$).

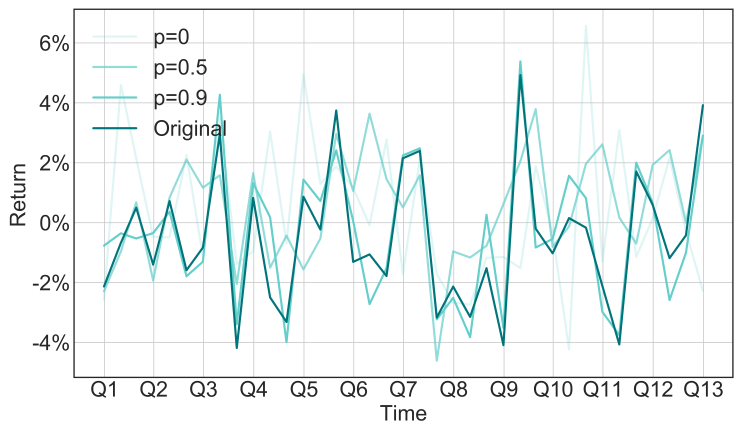


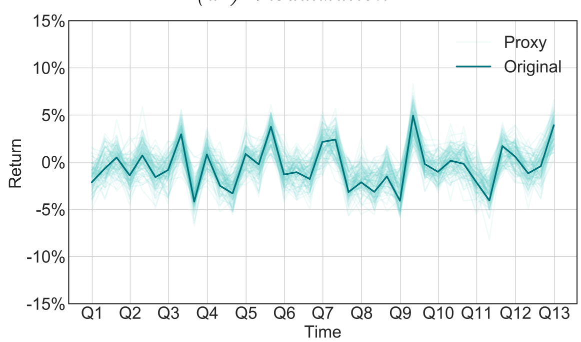

Figure 4.4 shows the simulation of 100 monthly proxies spanning three years with approximate correlation $\rho = 0.8$ to the original series in three scenarios: zero jump risk ($\lambda = 0$), moderate jump risk ($\lambda = 0.5$), and high jump risk ($\lambda = 1$). In the absence of jump risk, the distribution of sample correlations is centered near $\rho$, with a mean of 0.802. However, as jump risk increases, the distribution of sample correlations skews right. With moderate jump risk, the distribution of sample correlations has a mean of 0.828 while under high jump risk, the mean is even higher, at 0.879. This effect is due to the design choice to use the same jump-diffusion process to generate the original series and its proxy, which was motivated by the assumption that an asset and its proxy should be exposed to the same market tail risks. Section 5.5 provides further motivation for this design choice.

**Figure 4.4** – Sample Correlations and Proxy Generation

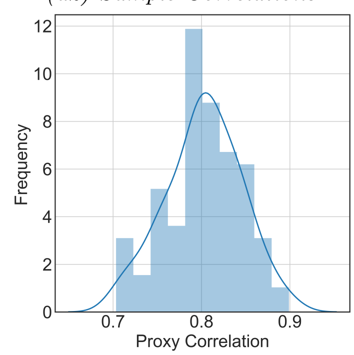
Example generation of 100 monthly proxies spanning three years with approximate correlation $\rho = 0.8$ to the original series. In the top two panels, the original series was generated with no jump risk ($\lambda = 0$). In the middle two panels, the original series was generated with moderate jump risk ($\lambda = 0.5$), corresponding to an expected jump every two years. In the bottom two panels, the original series was generated with high jump risk ($\lambda = 3$), corresponding to an expected three jumps a year. The left panels visualize the relationship between the generated proxies and the original series while the right panels show the distribution of the sample correlations of generated proxies to the original series. The distribution of possible jump sizes has a mean of zero ($\mu_q = 0$) and and a standard deviation of 5% ($\sigma_q = 0.05$).

No Jump Risk

*(a1) Visualization*

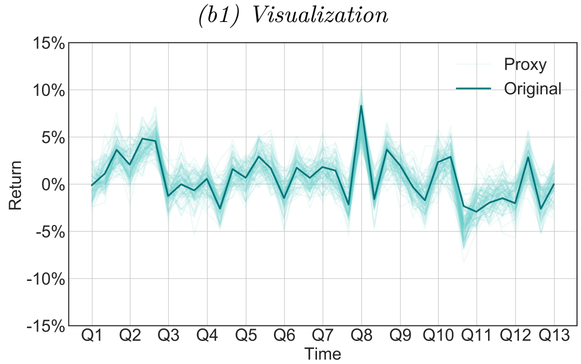



*(a2) Sample Correlations*

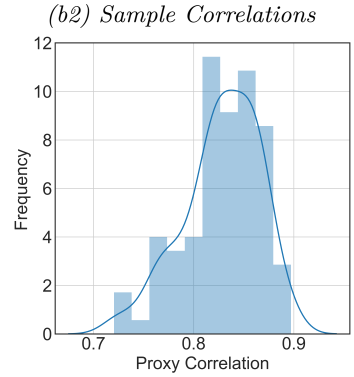



Moderate Jump Risk

*(b1) Visualization*

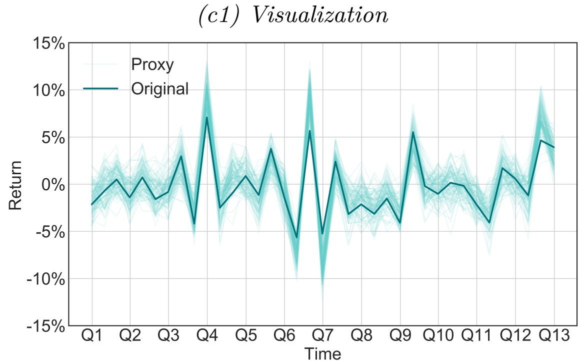



*(b2) Sample Correlations*

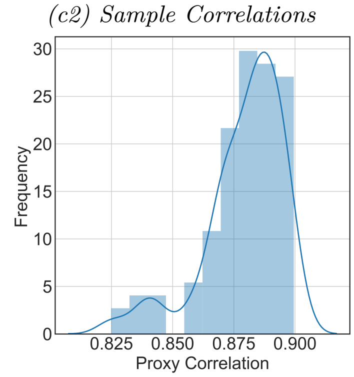



High Jump Risk

*(c1) Visualization*




*(c2) Sample Correlations*



## 4.4 The Complete Data Generation Algorithm

The Multi-Factor Market Simulator's complete data generation algorithm makes use of the parameters summarized in Table 4.1. Further intuition behind these parameters is explained in Section 5.2. The methodology produces cross-correlated and longitudinally-correlated assets with an added jump-diffusion process to capture market tail risk. The asset returns are cross-correlated to replicate the historical covariance structure of seven asset class indices. Cross-correlation is necessary to ensure that the methodology produces realistic returns that mirror historical realized performance and covariance of assets typical of institutional portfolios. The simulated asset returns have longitudinal correlation, or autocorrelation, to mirror the smoothing effect that often occurs in illiquid return reporting [Jenkinson et al., 2013], which is controlled by an fBm process.

| Symbol | Description |
|---|---|
| $H$ | autocorrelation Hurst Index |
| $\lambda$ | jump process intensity |
| $\mu_q$ | mean jump size |
| $\sigma_q$ | jump size volatility |
| $\rho$ | proxy goodness (correlation) |

**Table 4.1** – Data Generation Parameters Summary
Symbols and descriptions for each of the parameters used to generate data.

The data generation algorithm starts by estimating an intended covariance structure for the simulated returns via historical asset class indices. Next, it generates a basket of independent autocorrelated factor returns using fBm processes. Cross-correlating those factor returns using the Cholesky decomposition of the estimated covariance matrix produces a basket of returns with intended covariance that maintains proper longitudinal autocorrelation structure. An independent jump-diffusion process is then added to each asset class, and proxies are generated for the illiquid assets. This section first formalizes the data generation algorithm and then provides empirical evidence to validate the framework.

### 4.4.1 Data Generation Algorithm Specification

The following steps generate a ten-year basket of monthly historical asset returns (step (1) of the research design) and corresponding proxies:

1. Estimate the mean vector $\vec{\boldsymbol{\mu}}$, volatility vector $\vec{\boldsymbol{\sigma}}$, and variance-covariance matrix

$\mathbf{\Sigma}$ from historical data. The research uses quarterly returns of seven asset classes – commodities, equity, fixed income, hedge funds, private equity, real estate, and venture capital – with observation dates and descriptions given in Table 4.2. The following annualized estimates are found:.

$$\vec{\boldsymbol{\mu}} = \begin{bmatrix} \text{Commodities} & 2.3\% \\ \text{Equities} & 5.9\% \\ \text{Fixed Income} & 5.5\% \\ \text{Hedge Funds} & 8.0\% \\ \text{Private Equity} & 14.3\% \\ \text{Real Estate} & 9.8\% \\ \text{Venture Capital} & 16.7\% \end{bmatrix}, \qquad \vec{\boldsymbol{\sigma}} = \begin{bmatrix} 25.0\% \\ 16.8\% \\ 3.7\% \\ 8.0\% \\ 10.3\% \\ 9.1\% \\ 24.1\% \end{bmatrix} \tag{4.12}$$

$$\Sigma = \begin{bmatrix} 6.23\% & & & & & & \\ 1.33\% & 2.84\% & & & & & \\ -0.14\% & -0.18\% & 0.13\% & & & & \\ 0.80\% & 1.16\% & -0.07\% & 0.64\% & & & \\ 0.88\% & 1.33\% & -0.10\% & 0.64\% & 1.05\% & & \\ 0.60\% & 0.53\% & -0.03\% & 0.21\% & 0.53\% & 0.83\% & \\ 1.03\% & 1.74\% & -0.17\% & 1.13\% & 1.64\% & 0.35\% & 5.83\% \end{bmatrix}$$

2. Generate monthly increments of a fractional Brownian motion process $\overrightarrow{d\boldsymbol{W}_t^H}$ for each factor, as a function of the Hurst Index parameter $H$, according to the methodology explained in Section 4.2. The fBm processes across factors are pairwise independent and use the same value of $H$.

3. Cross-correlate the factor returns by multiplying $\overrightarrow{d\boldsymbol{W}_t^H}$ by $\text{Chol}\,(\mathbf{\Sigma})$, where Chol indicates the Cholesky decomposition of the estimated covariance matrix. This step takes a linear combination of autocorrelated processes, producing a basket of assets that remains longitudinally autocorrelated but also has a cross-sectional covariance structure according to the historical indices in Table 4.2. Intuitively, one could think of this step as producing asset class returns from orthogonal factor returns. Define a new series $\overrightarrow{d\boldsymbol{W}_t^{\text{assets}}} = \text{Chol}\,(\mathbf{\Sigma}) \times \overrightarrow{d\boldsymbol{W}_t^H}$.

4. Generate an independent Merton jump-diffusion process for each asset class $\overrightarrow{d\mathbf{U}_t}$, using $\overrightarrow{d\boldsymbol{W}_t^{\text{assets}}}$ as the underlying motion process via the jump intensity $\lambda$, mean jump size $\mu_q$, and jump volatility $\sigma_q$ parameters. The simulator assumes these same jump parameters are common to each asset class. Specifically, the simulator extends (4.4) and (4.5) and evolves the following

$$\overrightarrow{\boldsymbol{S}_{t+\Delta t}} = \overrightarrow{\boldsymbol{S}_t} e^{\left(\overrightarrow{\boldsymbol{\mu}} - \lambda \bar{k} - \overrightarrow{\boldsymbol{\sigma}}^2/2\right)\Delta t + \overrightarrow{d\boldsymbol{W}_t^{\text{assets}}}} \odot \overrightarrow{\mathbf{U}_t} \tag{4.13}$$

where:

$$\overrightarrow{\mathbf{U}_t} = e^{\overrightarrow{p} \times \mu_q + \sqrt{\overrightarrow{p}} \times \sigma_q \times \mathcal{N}(0,1)}$$
$$\bar{k} = e^{\mu_q + \sigma_q^2/2} - 1$$
$$\overrightarrow{p} \sim \text{Poisson}\left(\lambda \times \Delta t\right)$$

and $\overrightarrow{\boldsymbol{\mu}}$, $\overrightarrow{\boldsymbol{\sigma}}$, and $\boldsymbol{\Sigma}$ are given by values in (4.12) and $\odot$ represents the element-wise product.

5. Generate proxies for each illiquid asset class $\overrightarrow{d\boldsymbol{P}_t}$ according to the proxy goodness parameter $\rho$ and using the respective jump processes $\overrightarrow{d\mathbf{U}_t}$ via the methodology explained in Section 4.3

| Asset Class | Index Name | Description | Obs. Period | Liquidity |
|---|---|---|---|---|
| Commodities | S&P Goldman Sachs Commodities Total Return CME | A leading measure of general commodity price movements and inflation in the world economy, calculated primarily on a world production-weighted basis of physical commodities futures contracts. | Feb 1970 - Nov 2017 | Liquid |
| Equities | MSCI World Investable Markets | World investable market equities index, capturing large, mid, and small cap equities across 23 developed markets and 23 emerging markets. | Jun 1994 - Nov 2017 | Liquid |
| Fixed Income | Barclay's US Aggregate Bonds Total Return | Broad-based benchmark for investment grade, US dollar-denominated, fixed-rate taxable bond market. Index includes treasuries, government-related and corporate securities, MBS, and others. | Feb 1976 - Nov 2017 | Liquid |
| Hedge Funds | HFRI Fund Weighted Composite Index | Global equal-weighted index of over 2,000 single-manager funds with at least $50m AUM that report monthly net of all fees performance. | Jan 1990 - Nov 2017 | Liquid |
| Private Equity | Cambridge Associates US Private Equity | Index based on data compiled from more than 1,400 institutional-quality buyout, growth equity, private equity energy, and subordinated capital funds. | Jun 1986 - Jun 2017 | Illiquid |
| Real Estate | Cambridge Associates Real Estate | Index based on data compiled from more than 900 institutional-quality real estate funds. | Jun 1986 - Jun 2017 | Illiquid |
| Venture Capital | Cambridge Associates Venture Capital | Index based on data compiled from over 1,700 institutional-quality venture capital funds. | Jun 1986 - Jun 2017 | Illiquid |

**Table 4.2** – Asset Class Indices Descriptions

Names and descriptions for historical asset return indices used to estimate the mean vector $\vec{\mu}$ and variance-covariance matrix $\mathbf{\Sigma}$. Indices that are liquid are observed monthly while indices that are illiquid are observed quarterly.

### 4.4.2 Multi-Factor Market Simulator Validation

This section provides empirical evidence to validate the data generation algorithm. It is not obvious how cross-correlating fBm increments in step (3) of the algorithm produces asset returns with intended autocorrelation that mirror the historical covariance structure estimated in step (1). To validate this property of the algorithm, this section defines two test statistics. The first is the mean autocorrelation of the returns. As $H$ increases, one would expect this quantity to go up. The second statistic measures the error between the simulated returns' covariance and the historical covariance $\mathbf{\Sigma}$. One would expect this error to be low and independent of the Hurst index $H$.

Empirically, it is found that the error between the simulated returns' covariance and $\mathbf{\Sigma}$ is low and independent of the fBm Hurst parameter $H$, when jump intensity $\lambda = 0$. Furthermore, as $H$ increases, longitudinal autocorrelation increases, as expected. These two results validate the framework. The following discussion more rigorously defines these test statistics and provides supporting evidence. The statistics are defined as:

1. *Mean Autocorrelation*: the mean lag-1 autocorrelation correlation coefficient across the simulated returns. To validate the framework, this quantity should vary directly with $H$. Since the simulated returns have mean and variance that are static throughout time, they are weakly stationary, so autocorrelation can be computed as a function of the time lag. From (4.8), the lag $\tau$ autocorrelation of a weakly stationary time series $\boldsymbol{x}$ is

$$R\left(\tau, \boldsymbol{x}\right) = \frac{\mathbb{E}\left[\left(\boldsymbol{x}_t - \mu\right)\left(\boldsymbol{x}_{t+\tau} - \mu\right)\right]}{\sigma^2}$$

   where $\mu = \bar{\boldsymbol{x}}$ and $\sigma^2 = \mathrm{var}\left(\boldsymbol{x}\right)$. Thus a measure for the overall autocorrelation in the simulated data is the mean autocorrelation over asset classes. Let $\boldsymbol{X}$ be the simulated quarterly asset returns in stacked matrix form. Then, mean autocorrelation is specified as

$$\text{Mean Autocorrelation} = \frac{1}{d}\sum_{i=1}^{d} R\left(1, \boldsymbol{X}_i\right)$$

   where $d$(=7) is the number of asset classes.

2. *Covariance Error*: the overall error between the empirical covariance matrix and the historical covariance matrix. To validate the framework, this quantity should be small and not change with $H$. Because it is expected that increasing $H$ will decrease the variance of simulated assets, these two matrices are each normalized by variance.

Correlation matrices are computed according to:

$$\boldsymbol{\varrho} = \left(\operatorname{diag}\left(\boldsymbol{\Sigma}^{(\boldsymbol{X})}\right)\right)^{-1/2} \boldsymbol{\Sigma}^{(\boldsymbol{X})} \left(\operatorname{diag}\left(\boldsymbol{\Sigma}^{(\boldsymbol{X})}\right)\right)^{-1/2}$$

where $\Sigma^{(\boldsymbol{X})}$ is the covariance matrix of $\boldsymbol{X}$ and $\operatorname{diag}\left(\boldsymbol{\Sigma}^{(\boldsymbol{X})}\right)$ is the matrix of the diagonal elements of $\boldsymbol{\Sigma}^{(\boldsymbol{X})}$ (i.e. the variances of $\boldsymbol{X}_i$ for $i = 1, \ldots, n$). Let $\boldsymbol{\varrho}^{(\text{hist})}$ be the correlation matrix of quarterly asset returns found in Table 4.2 and $\boldsymbol{\varrho}^{(\text{sim})}$ be the correlation matrix of quarterly simulated asset returns. Thus, the error between $\boldsymbol{\varrho}^{(\text{hist})}$ and $\boldsymbol{\varrho}^{(\text{sim})}$ is given by

$$\text{Covariance Error} = \frac{1}{d^2} \left\| \boldsymbol{\varrho}^{(\text{hist})} - \boldsymbol{\varrho}^{(\text{sim})} \right\|_F$$

where $\|\boldsymbol{A}\|_F$ indicates the Frobenius norm of $\boldsymbol{A}$ given by $\|\boldsymbol{A}\|_F = \sqrt{\sum_{i,j} \boldsymbol{A}_{ij}^2} = \sqrt{\operatorname{trace}\left(\boldsymbol{A}^T \boldsymbol{A}\right)}$ and $d(=7)$ is the number of asset classes. Dividing by $d^2$ scales the result to give the mean error per element, since $\boldsymbol{\varrho} \in \mathbb{R}^{d \times d}$.

Figure 4.5 shows the sample means of these two statistics computed over 1000 trials by varying $H \in [0.5, 0.95]$ in increments of 0.05. Each basket of returns consists of 7 assets, each with 10 years of monthly data points ($n = 120$). Covariance error is roughly constant across values of $H$ while mean autocorrelation increases with $H$, as expected. As $H$ becomes large (>0.85), the fBm processes become more difficult to simulate with small $n$, so the slope of the mean autocorrelation line slightly decreases and error slightly increases. However, this result is not an issue, since if $n$ were sufficiently large, one would not expect to see this effect. The final step in validating the framework consists of showing that the baseline covariance error is low. Due to sampling variability, there is a null distribution in this test statistic; even if the two baskets of asset returns are generated from the same process, one would expect this error to be nonzero for any given trial. One would instead expect the mean empirical covariance (over a large number of trials) to much more closely match the historical covariance. As a final check, the following quantity is computed

$$\text{Mean Covariance Error} = \frac{1}{d^2} \left\| \boldsymbol{\varrho}^{(\text{hist})} - \overline{\boldsymbol{\varrho}}^{(\text{sim})} \right\|_F$$

where $\overline{\boldsymbol{\varrho}}^{(\text{sim})} = \frac{1}{T} \sum_{t<T} \boldsymbol{\varrho}_t^{(\text{sim})}$, $\boldsymbol{\varrho}_t^{(\text{sim})}$ indicates the correlation matrix of simulated returns on trial $t$, and $T(=1000)$ is the number of trials. To validate the simulator, this quantity

should be close to 0, since it averages out the effects of sampling variability. When $H = 0.85$, Covariance Error $= 0.0236 \gg$ Mean Covariance Error $= 0.0011 \approx 0$, as expected. Thus, even when autocorrelation is high, the expected covariance structure matches the historical covariance. Figure 4.6 gives a side-by-side visualization of the historical correlation matrix $\boldsymbol{\varrho}^{(\text{hist})}$ with the mean correlation matrix $\overline{\boldsymbol{\varrho}}^{(\text{sim})}$ over $\boldsymbol{T}$(=1000) trials. There is not a single cell where the matrices differ by more than 0.01. These results imply (1) longitudinal autocorrelation increases with $H$ and (2) the asset returns have similar covariance structure as the historical data for all values of $H$, validating the data generation methodology.

**Figure 4.5** – Cross Correlation and Longitudinal Correlation

The overall error between the empirical covariance matrix and the historical covariance matrix (Covariance Error) and empirical lag-1 autocorrelation (Mean Autocorrelation) of simulated asset returns as a function of Hurst index $H$. Returns are 120 monthly data points. All reported statistics are the sample means over 1000 trials.

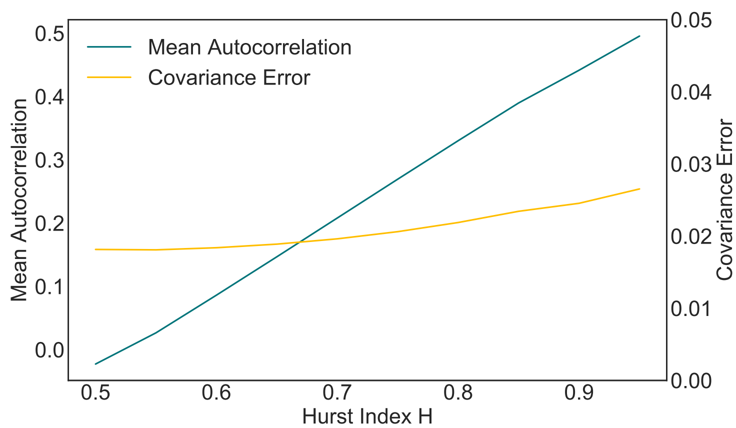
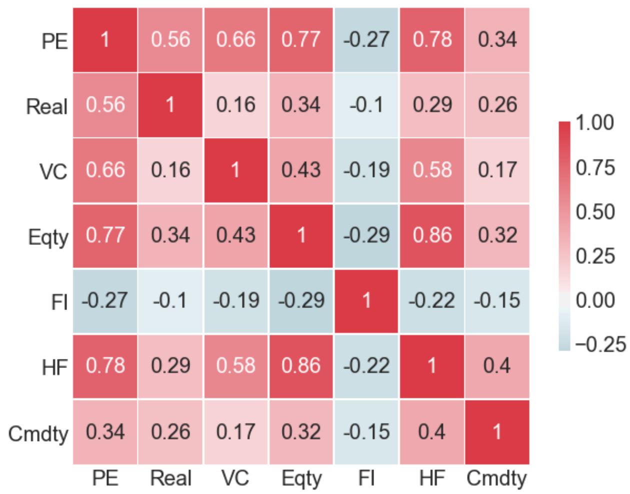


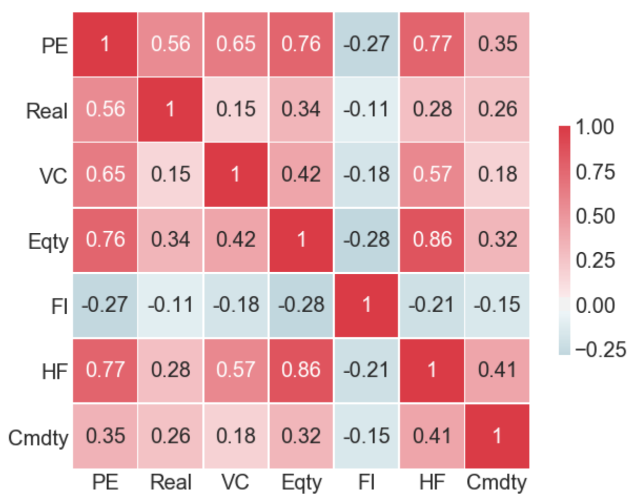
**Figure 4.6** – Correlation Matrix Comparison

A side-by-side comparison of the historical correlation matrix $\boldsymbol{\varrho}^{(\mathrm{hist})}$ with the mean simulated correlation matrix $\overline{\boldsymbol{\varrho}}^{(\mathrm{sim})}$ over 1000 trials, for $H = 0.85$. Panel *(a)* shows $\boldsymbol{\varrho}^{(\mathrm{hist})}$ while Panel *(b)* shows $\overline{\boldsymbol{\varrho}}^{(\mathrm{sim})}$. Simulated returns are 120 monthly data points.

*(a) Historical Correlation Matrix* $\boldsymbol{\varrho}^{(hist)}$




*(b) Mean Simulated Correlation Matrix* $\overline{\boldsymbol{\varrho}}^{(sim)}$



# Chapter 5

# Simulation Framework

The simulation framework was designed to apply the data inference methods to portfolios and test them in a realistic environment. By generating asset returns that vary according to market tail risk, autocorrelation, and proxy availability, the Multi-Factor Market Simulator generates data in several different realistic states of the world. The research outputs the preferred inference model to maximize the performance of the allocated portfolio in each of these states of the world via investor-focused metrics like Sharpe Ratio and drawdown. In applying the research, it is the allocator's responsibility to decide which state of the world is appropriate given their current observations about the market. For instance, if the allocator believes that illiquid returns are smoothed, as is often the case with reported private equity returns, then they would favor an inference method that is capable of handling data autocorrelation. Similarly, if an allocator possesses a good high-frequency proxy, for instance the NASDAQ Index for venture capital firms investing in technology companies, then they would likely favor a proxy-based inference method, which is capable of making use of this higher-frequency information, over a non proxy-based one. The simulation framework is inspired by that of Dogan and Szigety [Mallinger-Dogan and Szigety, 2014] and involves five main steps: 1) generating monthly asset returns, 2) aggregating to quarterly frequency, 3) inferring monthly returns from quarterly observations, 4) allocating portfolios, and 5) assessing performance. This section first discusses the high-level research design. Since the Multi-Factor Market Simulator was discussed previously, it then discusses research tests, method for portfolio allocation, and performance criteria. This section finishes with assumptions surrounding the design choices made in the framework.

## 5.1 Research Design

To compare the performance of the various data inference models, the research allocates two portfolios: a baseline and an experimental portfolio. The baseline portfolio is allocated from original monthly returns. The experimental portfolio is allocated based on the original monthly returns aggregated to quarterly frequency with inference methods applied to convert the returns back to monthly frequency. Performance is measured from portfolio weights applied out-of-sample to the original monthly data. The baseline portfolio serves as a benchmark, since it is allocated from complete data. Superior experimental portfolios have performance that closely matches the benchmark, and better inference leads to better experimental portfolios.

All reported statistics are the sample means across 1000 trials, which converge towards true values by the weak law of large numbers. The simulation design has five main steps, explained below and depicted visually in Figure 5.1:

1. **GENERATION**: generate ten years of monthly asset returns (and corresponding monthly proxies) according to specified parameters and historical covariance structure, as described in Section 4, using the Multi-Factor Market Simulator. Monthly returns are generated for seven asset classes, four liquid and three illiquid. The monthly returns for illiquid assets will be aggregated to quarterly in the next step. A single monthly proxy is generated for each illiquid asset with correlation to the true monthly returns.

2. **AGGREGATION**: aggregate returns of illiquid assets from monthly to quarterly frequency to create a basket of mixed-frequency assets. Since log returns are used in the analysis, quarterly return is simply the sum of the three corresponding monthly returns.

3. **INFERENCE**: perform inference to convert quarterly time series to monthly frequency. This step applies each of the inference methodologies explained in Sections 3 and 3.2 to the illiquid returns with applicable liquid proxies.

4. **ALLOCATION**: allocate portfolios via Markowitz mean-variance optimization using a rolling window of three years. This assumption is in line with a representative investor who re-balances the portfolio every three months and keeps the target allocation constant within the quarter. The baseline portfolio is allocated from the original monthly data while the experimental portfolio is allocated using partially-inferred data.

Allocations are out-of-sample, using weights fit on the previous three years of data to predict performance over the next quarter. Section 5.3 provides details on allocation.

5. **ASSESSMENT**: assess relative portfolio performance via metrics explained in Section 5.4. The experimental portfolio is expected to underperform the baseline portfolio, since allocations with incomplete data are expected to be worse than allocations with full data. Otherwise, there would be alpha in purposely adding noise to data.

## 5.2 Research Tests

The study consists of three main tests – a proxy goodness test, market tail risk test, and data autocorrelation test. In each of these tests, a single parameter is varied to isolate how the respective factor impacts the performance of the data inference models. Before discussing each of these tests in detail, this section first defines default values for the parameters used to simulate data and discusses the assumptions behind these choices. Default values for simulation parameters correspond to what should be expected under normal market conditions. The definition of "normal" is based on intuition and relevant literature. Fitting these parameters to historical data is a difficult problem in and of itself, which is out of the scope of this thesis. Default parameter values and assumptions are shown in Figure 5.1.

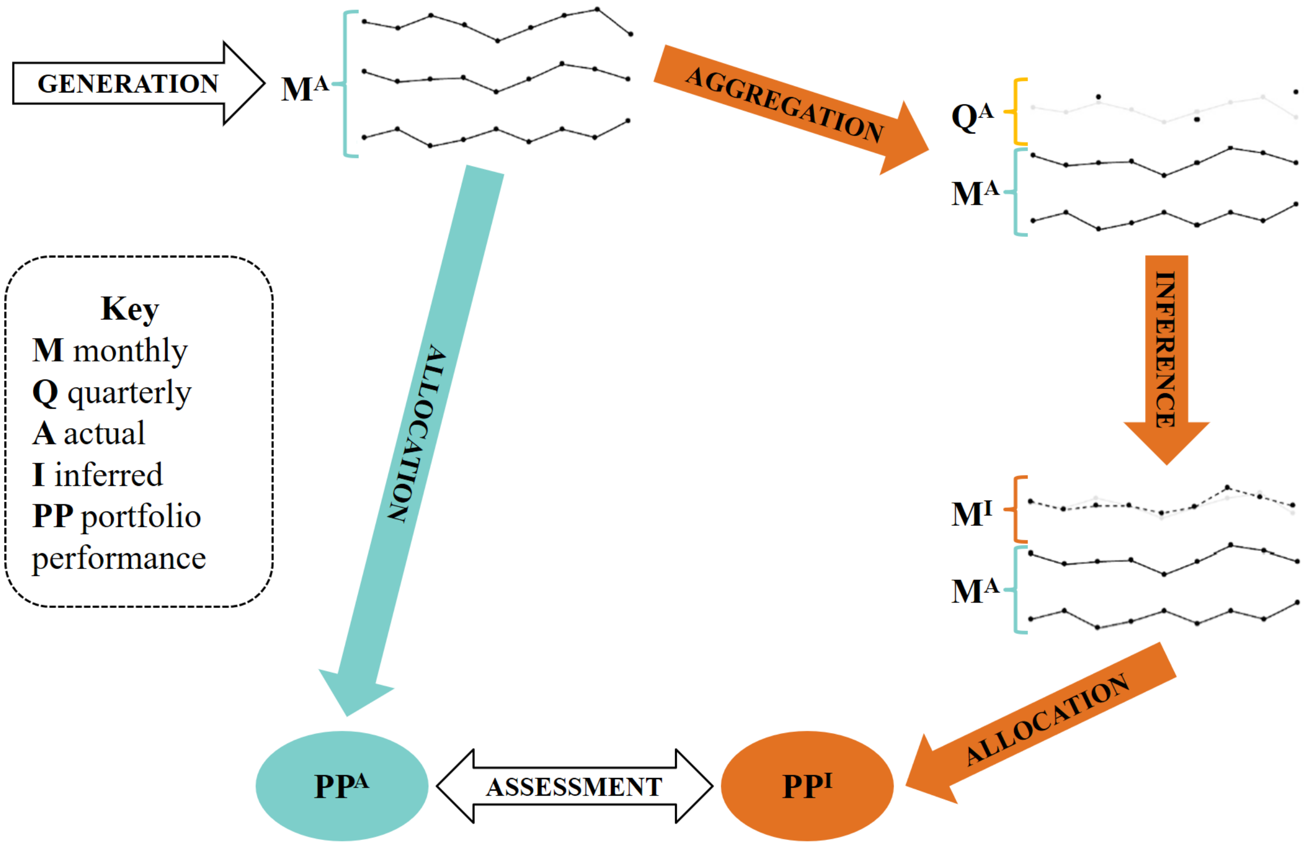
| Parameter | Default Value | Assumption |
|---|---|---|
| $H$ | 0.6 | There exists a small degree of illiquid return autocorrelation under normal market conditions. This choice is based on research studying illiquid returns [Getmansky et al., 2004, Khandani and Lo, 2011]. |
| $\lambda$ | 0.2 | Market tail risk is low under normal conditions. A major market dislocation event is expected to happen every 5 ($1/\lambda$) years for each asset class. |
| $\mu_q$ | 0 | Upside and downside tail risk are equal in normal market conditions. Therefore, the expected jump size is 0. |
| $\sigma_q$ | 0.05 | The distribution of market dislocation events on asset returns has a standard deviation of 5%. |
| $\rho$ | 0.6 | Under normal conditions, there exists a reasonably good proxy available, with about 0.6 correlation. |

**Table 5.1** – Default Values and Corresponding Assumptions for Simulation Parameters
Assumptions for the default value choices for each of the parameters used to generate data.

In each of the simulation tests, the default parameters values are used except for the

**Figure 5.1** – Simulation Framework

A flow-diagram of the Simulation Framework. Each cycle (from top left to bottom center) is repeated 1000 times for each set of parameter inputs. Source of graphics: [Mallinger-Dogan and Szigety, 2014]



single parameter being changed. The research runs the following three tests:

1. *Proxy Goodness Test*: determine the impact of a proxy goodness on inference method performance. Proxy goodness $\rho$ ranges from zero (poor proxy) to one (perfect proxy). Specifically, the parameter $\rho$ is varied in the range $\rho \in [0, 1]$ in increments of 0.05.

2. *Market Tail Risk Test*: determine the impact of market tail risk on inference method performance. A spectrum of zero, low, medium, and high tail risk scenarios are simulated by varying $\lambda \in [0, 1]$ in increments of 0.05.

3. *Data Autocorrelation Test*: determine the impact of data autocorrelation on inference method performance. To simulate a spectrum of zero, low, medium, and high autocorrelation scenarios, the parameter $H$ varies in the range $H \in [0.5, 0.95]$ in increments of 0.05[1].

In addition to these three main tests, the research also includes a single test using the default parameter values shown in Figure 5.1. This test, called the "Normal Market Conditions Test", highlights which inference method is preferred under normal market conditions. Table 5.2 contains a qualitative summary of all of the simulation tests in the study.

| Simulation Test | Proxy Goodness $\rho$ | Tail Risk $\lambda$ | Autocorr $H$ |
|---|---|---|---|
| Proxy Goodness Test | Varies | Low | Low |
| Market Tail Risk Test | Moderate | Varies | Low |
| Data Autocorrelation Test | Moderate | Low | Varies |
| Normal Market Conditions | Moderate | Low | Low |

**Table 5.2** – Simulation Tests Summary

A qualitative description of the simulations tests as a function of proxy correlation, jump risk, and data autocorrelation. Normal market conditions assume a proxy is available with 40% correlation, market dislocations have low probability, and there is mild autocorrelation in the illiquid returns.

## 5.3 Portfolio Allocation

Asset allocation is the decision faced by the investor to divide her portfolio across various assets classes, like stocks, bonds, and commodities. For example, a globally-invested pension fund must choose how much capital to allocate to each major country or region. In the research, portfolio weights are determined for seven major asset classes, four liquid (equities,

[1]When $H$ is high ($\approx 1$), fBm simulation is unstable. Thus, the $H$ parameter is capped at 0.95

fixed income, commodities, and hedge funds) and three illiquid (private equity, venture capital, and real estate), as described in Section 4.4.

Modern Portfolio Theory, hereafter MPT (also known as mean-variance optimization or Markowitz optimization), first introduced by Harry Markowitz in a 1952 essay entitled "Portfolio Selection" offers a solution to this problem in which the expected returns and covariance of assets are known[2]. It assembles a portfolio of assets, such that expected return is maximized for a given level of risk. Markowitz's key insight is that an asset's risk and return should not be assessed by itself but rather by how it contributes to a portfolio's overall risk and return [Markowitz, 1952]. The research elected to use MPT rather than more sophisticated portfolio allocation tools for two main reasons. First, MPT remains widely used in industry. The primary intended audience for the research is large-scale asset allocators, such as pensions or endowments, and many of these individuals use MPT to help inform their portfolio management decisions. Second, the expected returns and covariance of assets are known and constant throughout time in the simulation framework. This feature of the simulation framework alleviates many of the criticisms of applying MPT, which are discussed below.

MPT makes three main assumptions. The first and largest assumption (and often the method's largest criticism[3]) is that the expected returns and covariance of assets are known. In reality, the investor must substitute predictions based on historical returns for these measurements, which have a high level of error. In using historical data to estimate the go-forward covariance matrix and returns, one implicitly assumes returns in different periods are independent and drawn from the same statistical distribution, which is not the case in reality [Markowitz, 1952]. Past performance alone is not a good predictor of future performance. The research simulates data with mean returns and covariance that are known and static through time. Therefore, this assumption is not an issue for applying MPT in the framework.

The second assumption is that the risk of a portfolio is defined by the volatility of returns of the portfolio. MPT models risk in terms of likelihood of losses but says nothing about why these losses might occur; it assumes risk measurements are probabilistic rather

[2]Markowitz's contributions to modern investment decision making led to his sharing the 1990 Nobel Prize in Economics, along with William Sharpe and Merton Miller.

[3]In mean-variance optimization, small changes in the inputs can give rise to large changes in the portfolio. Since expected return and covariance are never known in practice, they must be estimated from historical data, often with noise. This noise, coupled with instability in the optimization, often outputs non-intuitive portfolios, which have large transaction costs or shake the confidence of the portfolio manager in the model. For this reason, mean-variance optimization has been dubbed an "error maximization" device [Bernd, 2002].

than structural in nature. In reality, portfolio managers often look at risk in a much more comprehensive way, often through drawdown analysis or value at risk frameworks. However, a portfolio manager's primary goal is often to maximize Sharpe Ratio [Sharpe, 1966], defined as excess portfolio return over the risk-free rate per unit of volatility. MPT, by equating risk with volatility, outputs a portfolio that maximizes Sharpe Ratio.

The third assumption is that the investor is risk averse and always prefers to increase consumption. Risk aversion implies that the investor must be compensated with higher expected return for taking on increased risk. Increasing consumption implies that the investor always seeks a portfolio with higher return for the same level of risk. These two properties imply that the investor's utility function is assumed to be concave and increasing, which allows the optimization problem to be solved quickly with commercial convex optimization solvers.

Applying MPT for asset allocation involves two main steps. First, the allocator must determine the efficient frontier, which is the set of all fully-invested portfolios in risky assets that maximize return for given level of variance. Finding the efficient frontier involves solving an optimization problem overall levels of investor risk aversion. Next, the investor must select the best portfolio on the efficient frontier. To make allocations realistic, the research assumes there exists a risk-free asset in the market. Therefore, the investor has the opportunity to earn interest on unallocated cash positions at the risk-free rate. Also, the research assumes the investor has access to leverage, which gives the ability to borrow at the risk-free rate to increase capital allocated in risky assets.

To determine the efficient frontier, an allocator solves the optimization problem specified as follows [Kempthorne, 2013]. Assume there are $n$ risky assets $i = 1, 2, \ldots, n$ with single-period returns given by the $n$-variate random vector $\boldsymbol{R} = \begin{bmatrix} R_1 & R_2 & \ldots & R_n \end{bmatrix}^T$. The mean and covariance of returns are given by

$$\mathbb{E}\left[\boldsymbol{R}\right] = \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}, \qquad \operatorname{cov}\left[\boldsymbol{R}\right] = \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{1,1} & \cdots & \boldsymbol{\Sigma}_{1,n} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{n,1} & \cdots & \boldsymbol{\Sigma}_{n,n} \end{bmatrix}$$

Define the portfolio as the length-$n$ vector of weights indicating the fraction of portfolio wealth held in each asset

$$\boldsymbol{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} : \ \boldsymbol{w}^T \mathbf{1}_n = 1$$

Portfolio return is thus defined by $\boldsymbol{R}_p = \boldsymbol{w}^T \boldsymbol{R}$, with

$$\begin{aligned} \mu_w &= \mathbb{E}\left[\boldsymbol{R}_p\right] &= \boldsymbol{w}^T \mu \\ \sigma_w^2 &= \operatorname{var}\left[\boldsymbol{R}_p\right] &= \boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w} \end{aligned}$$

One way to determine the portfolio $\boldsymbol{w}$ involves solving a quadratic program to minimize variance while maximizing expected return for each level of investor risk aversion $\gamma$. This parameter gauges the trade-off between risk and return. The long-only constraint ($w_i \geq 0, \forall i$) is also enforced[4]. Mathematically, this is

$$\begin{aligned} &\min_{\boldsymbol{w}} \frac{\gamma}{2} \boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w} - \boldsymbol{w}^T \boldsymbol{\mu} \\ &\text{subject to} \begin{cases} \boldsymbol{G}\boldsymbol{w} \leq \boldsymbol{h} & \text{long-only} \\ \boldsymbol{A}\boldsymbol{w} = b & \text{weights sum to 1} \end{cases} \end{aligned} \tag{5.1}$$

where $\boldsymbol{G} = -I_{n \times n}$, $\boldsymbol{h} = \mathbf{0} \in \mathbb{R}^n$, $A = \mathbf{1}^T \in \mathbb{R}^n$, and $b = 1$. As $\gamma$ increases, the investor becomes more risk-averse, preferring to minimize variance rather than maximize return. Solving this program over multiple levels of $\gamma$ generates the efficient frontier – the set of portfolios with highest expected return per level of risk. The efficient frontier has a convex bullet shape, due to the assumption that the investor is risk averse and always prefers to increase consumption. An investor's optimal portfolio is found at the point of tangency of the efficient frontier with the investor's indifference curve. When a risk-free asset is involved, the investor's indifference curve is a straight line, referred to as the capital allocation line [Sharpe, 1966]. It is tangent to the efficient frontier at the pure risky portfolio with the highest Sharpe ratio. Its vertical intercept represents a portfolio with 100% of holdings in the risk-free asset and its tangency with the efficient frontier represents a portfolio with no risk-free holdings and 100% of assets held in risky assets. Portfolios between these two points contain positive amounts of both the risky tangency portfolio and the risk-free asset,

[4] Without the long-only constraint, mean-variance optimization often finds highly leveraged portfolios, which are infeasible in practice.

and points beyond the tangency point are leveraged portfolios. For leveraged portfolios, the amount invested in the tangency portfolio is equal to more than 100% of the investor's initial capital due to borrowing, causing negative holdings of the risk-free asset.

Deriving the formula for the capital allocation line [Sharpe, 1966] starts with the realization that if investors can purchase a risk-free asset with some return $r_f$, then all correctly priced risky portfolios will have expected return in the form

$$\mathbb{E}\left[\boldsymbol{R}_p\right] = r_f + d\sigma_p \tag{5.2}$$

where $d$ is some incremental return to offset the additional risk (also called a risk premium), and $\sigma_p = \sqrt{\operatorname{var}\left[\boldsymbol{R}_p\right]}$ is the risk itself, expressed as a standard deviation. Consider another portfolio that is a combination of the risk-free asset and the correctly priced portfolio above (which is itself just another risky portfolio). If correctly priced, it will have exactly the same form

$$\mathbb{E}\left[\boldsymbol{R}_C\right] = r_f + d\sigma_C \tag{5.3}$$

Substituting the definition of the risk premium $d$ from (5.2) into (5.3) yields the capital allocation line

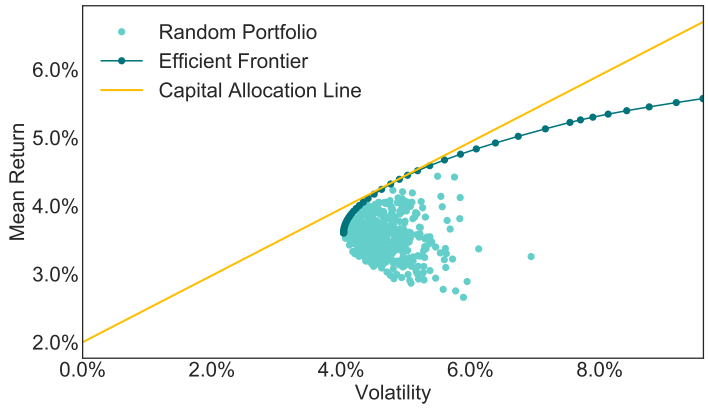
$$\text{CAL}: \; \mathbb{E}\left[\boldsymbol{R}_C\right] = r_f + \sigma_C \frac{\mathbb{E}\left[\boldsymbol{R}_p\right] - r_f}{\sigma_p} \tag{5.4}$$

where $p$ is the risky portfolio, $r_f$ is the risk-free rate, and $C$ is a combination of these portfolios. The research chooses portfolio along the capital allocation line with 8% annualized volatility, as would a typical institutional investor.

Figure 5.2 shows an example visualization of the Markowitz efficient frontier and capital allocation line generated from simulated data. As expected, it has a characteristic convex bullet shape due to the investor risk aversion assumption. The capital allocation line intercepts the y-axis at the risk-free rate, assumed to be 2% annualized, and is tangent to the efficient frontier at the point that maximizes Sharpe ratio. The research chooses portfolio along the capital allocation line that achieves 8% annualized volatility.

**Figure 5.2** – Example Markowitz Optimization

An example visualization of 500 random portfolios and the corresponding Markowitz mean-variance frontier and capital allocation line for a simulated basket of asset returns. The risk-free rate is assumed to be 2% annualized.




It is also possible to directly solve for the final allocation with a risk-free asset without generating the entire efficient frontier if a target mean return or volatility is specified, as is the case in the research [Kempthorne, 2013]. An optimal portfolio with target volatility $\sigma_0$ achieves the maximum expected return with risk no greater than $\sigma_0^2$. Consider the same formulation as previously, except also assume that in addition to risky assets, there exists a single risk-free asset $\boldsymbol{R}_f$ for which $\boldsymbol{R}_f \equiv r_f$, (i.e. $\mathbb{E}[\boldsymbol{R}_f] = r_f$ and $\mathrm{var}(\boldsymbol{R}_f) = 0$). Therefore, the investor has $\boldsymbol{w}^T\mathbf{1}_n = \sum_{i=1}^n w_i$ invested in risky assets and $1 - \boldsymbol{w}^T\mathbf{1}_n$ invested in the risk-free asset. Since the investor has access to leverage, these weights need no longer sum to one (i.e. $\boldsymbol{w}^T\mathbf{1}_n \neq 1$ ). Since borrowing is allowed, the quantity $\left(1 - \boldsymbol{w}^T\mathbf{1}_n\right)$ can be negative. The portfolio return is therefore $\boldsymbol{R}_w = \boldsymbol{w}^T\boldsymbol{R} + \left(1 - \boldsymbol{w}^T\mathbf{1}_n\right)\boldsymbol{R}_0$, with

$$
\begin{aligned}
\mu_w &= \mathbb{E}[\boldsymbol{R}_w] &= \boldsymbol{w}^T\boldsymbol{\mu} + \left(1 - \boldsymbol{w}^T\mathbf{1}_n\right)\boldsymbol{r}_0 \\
\sigma_w^2 &= \mathrm{var}[\boldsymbol{R}_w] &= \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}
\end{aligned}
$$

One way to determine the portfolio $\boldsymbol{w}$ consists of maximizing return for a given choice of target volatility $\sigma_0$. This is specified as

$$\begin{aligned}
&\max_{\boldsymbol{w}} \ \boldsymbol{w}^T\boldsymbol{\mu} + \left(1 - \boldsymbol{w}^T\mathbf{1}_n\right)\boldsymbol{r}_0 \\
&\text{subject to} \begin{cases} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w} = \sigma_0^2 & \text{volatility target is met} \\ \boldsymbol{G}\boldsymbol{w} \leq \boldsymbol{h} & \text{long-only} \end{cases}
\end{aligned} \tag{5.5}$$

where $\boldsymbol{G} = -\boldsymbol{I}_{n\times n}$, $\boldsymbol{h} = \mathbf{0} \in \mathbb{R}^n$. In the research, this parameter $\sigma_0 = 8\%$, as is typical of an institutional investor.

## 5.4 Performance Criteria

To assess the allocated portfolios, the research uses a variety of common performance measures. For the return of the allocated portfolios, mean return is computed. For risk, volatility and maximum drawdown are used. For risk-adjusted performance, the research calculates Sharpe Ratio, and Sortino Ratio. In discussing these measures, let the series of annualized monthly portfolio returns be the vector $\mathbf{R}_p = \begin{bmatrix} p_0 & p_1 & \cdots & p_{n-1} \end{bmatrix}^T$, where $n = 120$, since ten years of data are simulated. The research defines the following portfolio metrics:

- *Mean return*:

$$\mu_p = \mathbb{E}\left[\mathbf{R}_p\right] = \frac{1}{n}\sum_{i=1}^{n} p_i$$

- *Volatility of returns*:

$$\begin{aligned}
\sigma_p &= \sqrt{\mathbb{E}\left[\left(\mathbf{R}_p - \mathbb{E}\left[\mathbf{R}_p\right]\right)^2\right]} \\
&= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(p_i - \mu_p\right)^2}
\end{aligned}$$

- *Sharpe Ratio*:

$$S = \frac{\mu_p - r_f}{\sigma_p}$$

where $r_f$ is the risk-free rate, which is assumed to be 2% annualized.

- *Sortino Ratio*:
$$\mathrm{SOR}(\tau) = \frac{\mu_p - \tau}{\sqrt{\mathrm{LPM}_2(\tau)}}$$
where $\tau$ is a target threshold, which is assumed equal to the risk-free rate (2% annualized) and $\mathrm{LPM}_2$ represents the 2nd lower partial moment, given by $\mathrm{LPM}_j = \frac{1}{n}\sum_{i=1}^{n} \max(\tau - p_i, 0)^j$ where $j = 2$.

- *Maximum drawdown*:
$$\mathrm{MD} = \max_{0 \leq i < n} \left\{ 0, \max_{0 \leq j \leq i} \{S_j - S_{j-i}\} \right\}$$
where $S_j$ represents cumulative return, $S_j = \sum_{0 \leq k \leq j} p_k$. Since the research uses logarithmic returns, returns are additive over time [Hudson and Gergoriou, 2010].

The research looks at mean return because it gives an estimate of overall portfolio performance, and it is important that the inference methods yield high-performing portfolios. Volatility is included because it is commonly used to assess broader portfolio risk. The research calculates Sharpe Ratio, because it gives a measure of risk-adjusted performance that is the industry standard for benchmarking portfolios. Risk management is generally concerned with avoiding large downside risk, and Sharpe Ratio equally penalizes upside and downside risk. Therefore, the Sortino Ratio is included, which measures risk-adjusted return while only penalizing downside risk [Sortino and Price, 1994]. Drawdown, defined as the peak-to-trough decline of a portfolio, estimates the amount of capital that is lost when the portfolio experiences negative returns. The research calculates maximum drawdown, the maximum percentage decline in portfolio value, which is a commonly used measure of downside portfolio risk.

Reported portfolio metrics are the absolute errors between the statistics based on true monthly data and those based on quarterly data inferred to monthly frequency, $|s^A - s^I|$, where $A$ and $I$ stand for actual and inferred, respectively and $s \in \{$mean return, volatility, Sharpe Ratio, Sortino Ratio, maximum drawdown$\}$. The absolute error represents the loss of information due to a specific inference methodology and allows performance to be benchmarked with respect to the baseline portfolio.

In Markowitz optimization, which is used for portfolio allocation, small changes in the inputs can yield large changes in portfolio weights [Ang, 2014]. This effect means that small variations in illiquid return inference can dramatically change the performance of the allocated portfolio. Therefore, the research also looks at the overall similarity of the actual and inferred returns prior to allocation. Define the root mean squared error of two series $\boldsymbol{x} = x_0, \ldots, x_n$ and $\boldsymbol{y} = y_0, \ldots, y_n$ as

$$\text{RMSE}\left(\boldsymbol{x}, \boldsymbol{y}\right) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i\right)^2}$$

The portfolio has three illiquid investments. To determine the overall pre-allocated error of a given inference methodology, the research uses the average RMSE between the inferred and actual monthly return series across illiquid investments, defined as

$$\bar{E} = \frac{1}{m}\sum_{j=1}^{m}\text{RMSE}\left(\boldsymbol{r}_j^A, \boldsymbol{r}_j^I\right)$$

where $\boldsymbol{r}_j^I$ and $\boldsymbol{r}_j^A$ represent the inferred and actual monthly return series of illiquid investment $j$ and $m = 3$.

## 5.5 Research Assumptions

The research made a number of design choices to simply the simulation environment. This section discusses each of the assumptions surrounding design choices in the framework.

- *A single proxy is used for each illiquid asset:* in reality, there could exist several good proxies available for an illiquid asset which an asset allocator would wish to use. To simplify the framework, a single proxy is generated for each illiquid asset, even though each of the research's proxy-based inference methods is capable of incorporating multiple proxies. It is assumed that using a single highly-correlated proxy would yield similar effects as using several proxies each with lower correlation. As a next step in the research, it would make sense to expand the methodology to incorporate multiple proxies with varying levels of goodness.

- *Proxies are exposed to the same tail risks as their related asset class*: an independent jump-diffusion process is generated for each asset, and proxies are generated from the same jump-diffusion processes as their corresponding asset. This assumption allows proxies to supply higher-frequency information about the timing of dislocation events to improve inference. In reality, it would be important to select a proxy that is exposed to the same tail risks as the series being inferred to achieve this effect. The research also assumes that the jump intensity, size, and volatility parameters are the same across assets[5]. In reality, proxies and asset classes might come from different markets and therefore be exposed to different tail risks. However, an asset allocator should do her best to select a proxy that is exposed to similar tail risks.

- *Proxies contain information of true higher-frequency values of illiquid assets:* proxies are generated by perturbing the true higher-frequency returns of the illiquid assets (which are not known in reality) with varying levels of noise. In reality, proxy returns are not driven by the exact same factors as their corresponding asset. This design choice was made to simplify the proxy generation process and make it easier to define proxy "goodness".

- *Proxy returns have low (zero) autocorrelation* : one would not expect high-frequency returns to be smoothed to the same extent as illiquid returns. For instance, reported private equity returns often have lower volatility than respective cash flows, suggesting that firms mark companies to smooth out returns [Jenkinson et al., 2013]. As a result, this research makes the design choice to generate proxies from an independent geometric Brownian motion process, where increments have zero autocorrelation, rather than a fractional Brownian motion process with positive autocorrelation.

- *Expected asset return and covariance are static throughout time:* asset returns are generated according to historically-fit mean returns and covariance that are static throughout time. In reality, the covariance structure among asset classes fluctuates, which makes forecasting expected return and covariance for use in mean-variance optimization so difficult. This design choice ensures that mean-variance optimization, a widely used asset allocation tool, will output reliable portfolio weights that are reflective of true expected return and covariance expectations.

[5]Fitting these parameters on historical data is a problem in of itself and is out of the scope of this research.

# Chapter 6

# Results

The study consisted of three main tests – a proxy goodness test, market tail risk test, and data autocorrelation test, as described in Section 5.2. Each of these tests varies a single parameter, proxy correlation $\rho$, jump risk $\lambda$, or data autocorrelation $H$ respectively, while setting the remaining parameters to their default values. These tests determine how each of these dimensions independently impact the performance of the data inference models. In normal market conditions, it is assumed there exists a reasonably good proxy available, low chance of market dislocations, and a small degree of autocorrelation of returns. This parameterization is based on literature about illiquid returns autocorrelation as well as intuition. In addition to these three main tests, the research includes single test using default parameter values called the "Normal Market Conditions Test".

As described in Section 5.1, ten years of monthly asset returns are generated (and corresponding monthly proxies) according to specified parameters using fractional Brownian motion in conjunction with Merton's jump-diffusion. Returns are generated for four liquid asset classes, observed monthly, and three illiquid asset classes, observed quarterly. A single monthly proxy is generated for each illiquid asset, and inference of the illiquid asset uses the corresponding monthly proxy.

Portfolios are allocated via Markowitz mean-variance optimization using a rolling window of three years. This assumption is in line with a representative investor who re-balances the portfolio every three months and keeps the target allocation constant within the quarter. Allocations are long-only and fully-invested, with an annualized target volatility of 8%. The baseline portfolio is allocated from the original monthly data while the experimental portfolio is allocated using partially-inferred data. Allocations are out-of-sample, using weights fit on the previous three years of data to predict performance over the next quarter.

Results are composed of five portfolio metrics, as described in Section 5.4– mean return, volatility, Sharpe ratio, Sortino ratio, and maximum drawdown. These reported metrics are the absolute errors between the statistics based on true monthly data and those based on quarterly data inferred to monthly frequency, $|s^A - s^I|$, where $A$ and $I$ stand for actual and inferred, respectively. The absolute error represents the loss of information due to a specific inference methodology and allows performance to be benchmarked with respect to the baseline portfolio. Lower errors indicate superior performance. In addition to these portfolio metrics, RMSE between illiquid returns in inferred and actual portfolios is computed to gain an overall sense of fit prior to allocation. RMSE, by definition, represents the error with respect to the baseline portfolio, so it is not normalized like the other five statistics. All reported statistics are the sample means across 1000 trials, which converge to true values by the law of large numbers.

The next sections explore proxy goodness, market tail risk, and data autocorrelation individually and conclude with a holistic assessment of which methods to use and when, based on a single composite score weighting each method's performance according to perceived importance to investors. The results imply that the recommendation of which inference method to use changes with the goodness of the available proxy and data autocorrelation but is not significantly impacted by market tail risk. In general, the relative performance of the proxy-based inference methods improves as proxy goodness increases. For investors lacking a good proxy available, it is recommended they use back fill. If the goodness of available proxy is above 0.3, then they should switch to the CL Method. If data autocorrelation is high, autoregressive Kalman filters, which are capable of explicitly modeling the autocorrelation outperform CL Method but perform worse than simple back fill. In normal market conditions, the CL Method gives the best overall performance, indicated by low RMSEs and reliable forecasts for mean return, volatility, Sharpe Ratio, and drawdown.

## 6.1 Results: Proxy Goodness Test

This test varies the proxy goodness parameter $\rho$ from 0 to 1 to determine the impact of proxy goodness on inference method performance. Market tail risk and data autocorrelation are fixed to their default values. The results imply that the optimal choice of inference method across statistics changes with the goodness of the available proxy. For Sharpe Ratio and RMSE, back fill gives the best forecasts when proxy correlation is very low (below 0.25) while CL Method is superior for all other correlation levels. For all other statistics – mean return, volatility, Sortino Ratio, and maximum drawdown – CL Method is the preferred choice across all levels of correlation.

Figure 6.1 shows Sharpe Ratio ($SR$), mean return ($\mu$), volatility ($\sigma$), Sortino Ratio ($S$), maximum drawdown ($DD$), and pre-allocated RMSE (RMSE), averaged across 1000 simulations, when the illiquid asset is interpolated with a proxy whose correlation is in the range $\rho \in [0, 1]$. For each simulated portfolio, $\rho$, and inference technique, the plot reports the average absolute error between the actual ($A$) and the inferred ($I$) portfolio's statistics, with lower errors indicating superior performance. RMSE does not need to be normalized. Non-proxy-based methodologies are plotted with dashed lines while proxy-based ones are plotted with solid lines. Table A.1, located in Appendix A, reports the same results in raw form.

Panel (f) of Figure 6.1 indicates that the overall fit of all proxy-based methods improves as proxy goodness increases. For each of the methods, the RMSE decreases exponentially and tends to 0, as proxy goodness tends to 1. This result is expected, since increasing proxy goodness should provide improve the quality of proxy-based inference. Also, for a perfect proxy equal to the true monthly returns, each proxy-based method perfectly recovers the true monthly returns, since the errors across statistics are 0. As a sanity check, the performance of non-proxy-based methodologies should not change with proxy goodness, since these methods do not utilize proxies. Indeed, the RMSEs (and errors of other statistics) for all non-proxy-based methods appear consistent over all levels of proxy goodness.

Panel (a) shows that the performance of all four proxy-based methods for predicting Sharpe ratio improves as proxy correlation increases. For proxy goodness below 0.2, back fill gives the best Sharpe, until it is taken over by the CL Method for all higher levels of proxy goodness. For proxy goodness above 0.75, all proxy-based methods are better for predicting Sharpe Ratio than their non-proxy-based counterparts. The overall poor performance of the MIDAS framework can likely be attributed to the fact that it is not designed for infra-period inference. While applications of MIDAS to forecast the next quarterly data point are common [Armesto et al., 2010, Bai et al., 2013, Eric Gysels, 2006], these models are rarely used to infer monthly data points within the quarter.

While CL Method falls short in predicting Sharpe ratio for low levels of correlation, it better predicts mean performance and volatility alone across all levels of correlation, shown in panels (b) and (c) in the same figure. Back fill's superior Sharpe Ratio performance could be attributed to the method's skewing prediction of mean performance and volatility in the same direction, an effect that is washed out by taking their ratio. Cubic spline interpolation happens to be the best non-proxy-based method for predicting Sortino ratio, seen in panel (d). The CL Method is also the superior choice for predicting maximum drawdown across all levels of proxy correlation, seen in panel (e).

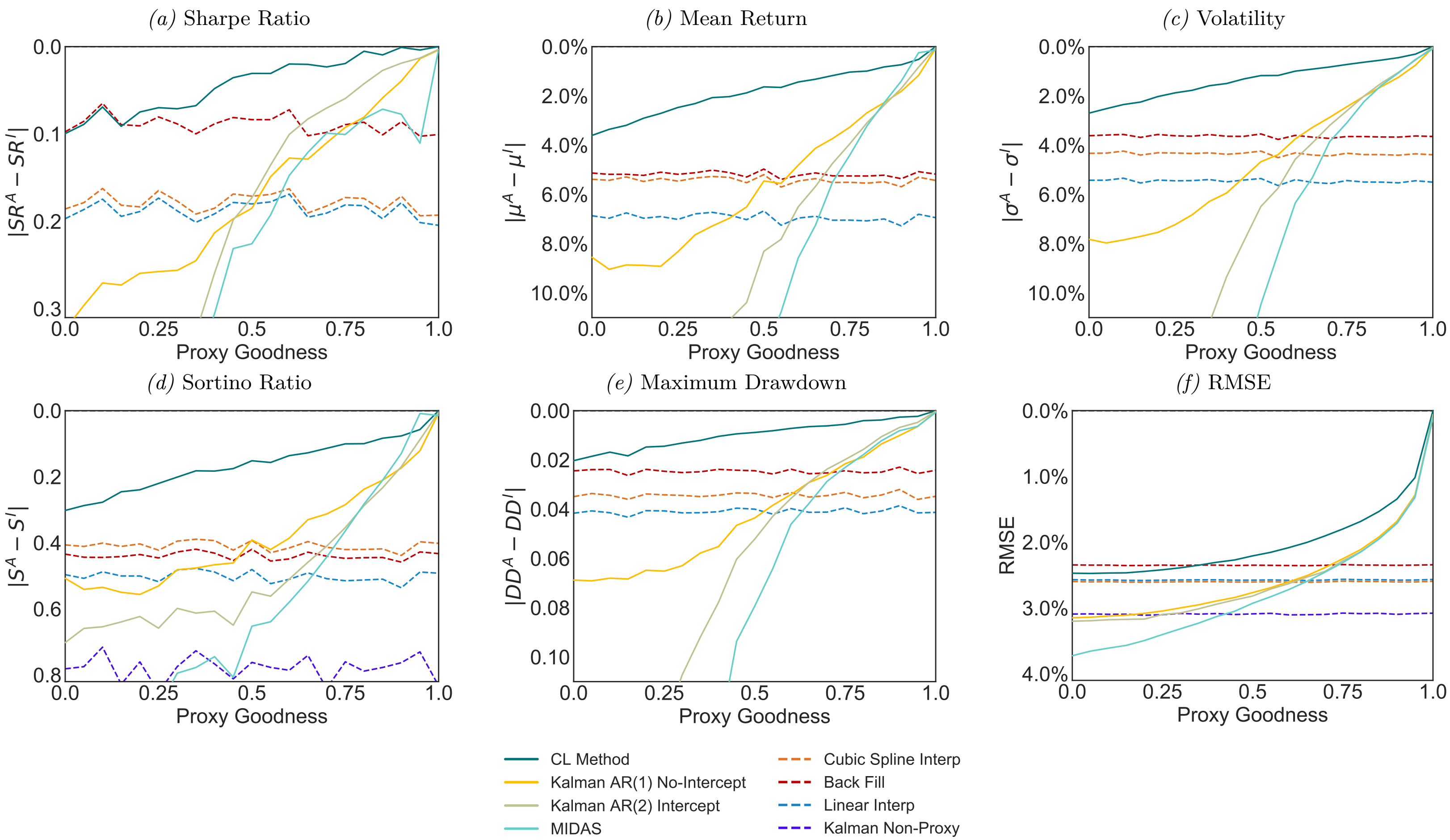
## 6.2 Results: Market Tail Risk Test

This test varies the market tail risk parameter $\lambda$ from 0 to 1 to determine the impact of tail risk on inference method performance. As $\lambda$ increases, market dislocations become more frequent. A value of $\lambda = 0$ corresponds to no risk of market dislocations while a value of $\lambda = 1$ corresponds to one market dislocation in expectation per year. Proxy goodness and data autocorrelation parameters are fixed to their default values. It is important to note that since proxy correlation is fixed at 0.6, CL Method is expected to outperform the other methods, as revealed by the proxy goodness results. These market tail risk results do not control for this factor. Rather than looking for which method is best, this analysis instead focuses on how the relative performance of the methods changes with market tail risk. The results imply that the relative performance of the CL Method in predicting Sharpe Ratio declines as market tail risk increases and does not meaningfully change across the other statistics.

Figure 6.2 shows Sharpe Ratio ($SR$), mean return ($\mu$), volatility ($\sigma$), Sortino Ratio ($S$), maximum drawdown ($DD$), and pre-allocated RMSE (RMSE), averaged across 1000 simulations, when market tail risk varies in the range $\lambda \in [0, 1]$. For each simulated portfolio, $\lambda$, and inference technique, the plot reports the average absolute error between the actual ($A$) and the inferred ($I$) portfolio's statistics, with lower errors indicating superior performance. RMSE does not need to be normalized. Non-proxy-based methodologies are plotted with dashed lines while proxy-based ones are plotted with solid lines. Table A.2, located in Appendix A, reports the same results in raw form.

Panel (f) of Figure 6.2 indicates that the overall fit of all methods declines as market tail risk increases. For each method, RMSE increases as market dislocations become more frequent. This result is expected, since increasing market tail risk increases the variability of

**Figure 6.1** – Portfolio Performance: Proxy Goodness and Loss of Information



the true returns, which should decrease the quality of inference for moderately good proxies. For each method, the degradation in RMSE appears to be linear with approximately the same slope.

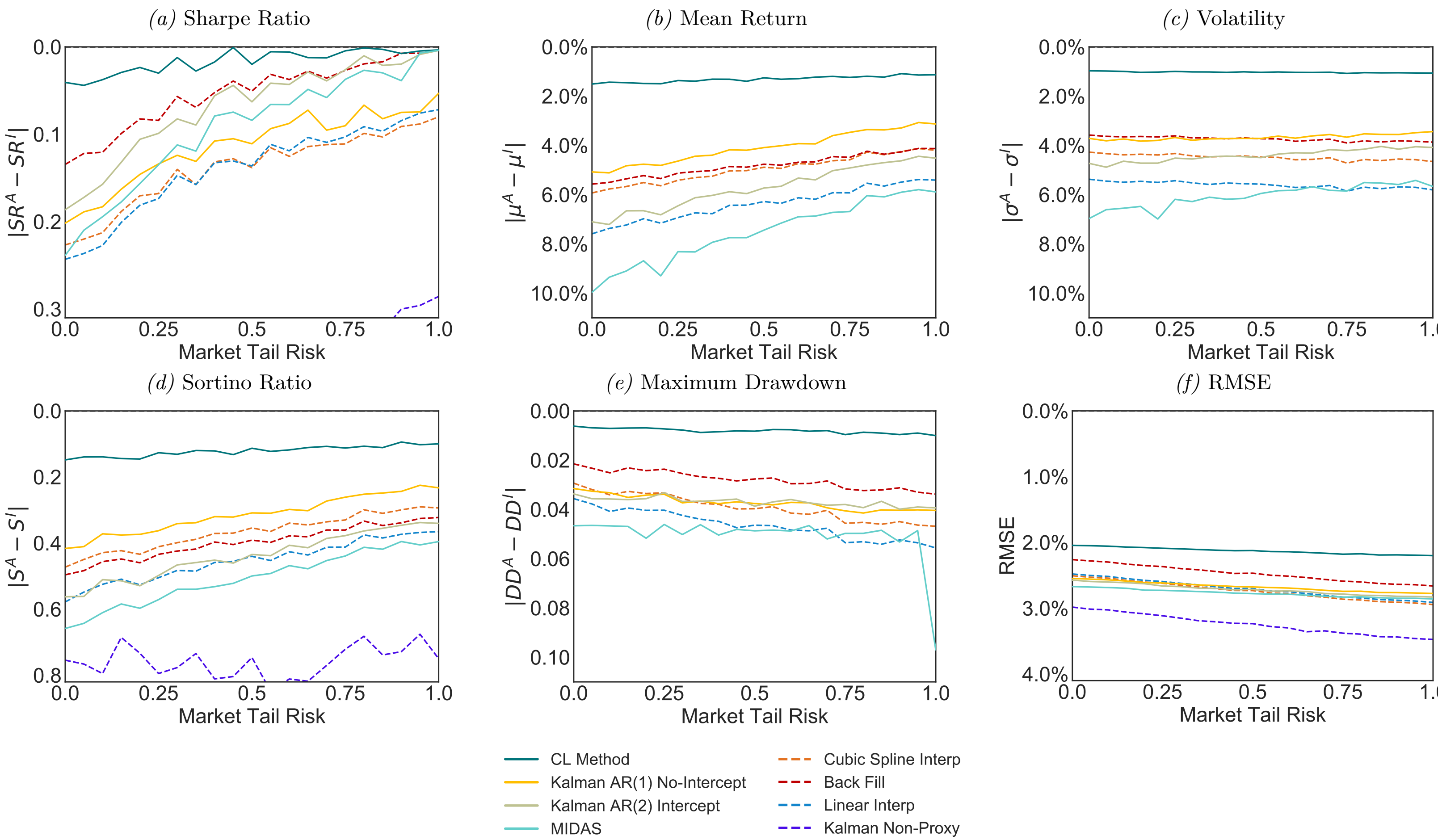
Despite RMSEs decreasing across methods, the ability to forecast mean return improves as tail risk increases, demonstrated by Panel (b) of the same figure. This effect appears to be linear, with the Kalman filters showing the highest slope. One hypothesis is that this effect can be explained by the research assumption that proxies are exposed to the same tail risks as their corresponding illiquid assets. Because of this assumption, proxies are generated using the same jump-diffusion process as their illiquid assets, which allows proxies to inject meaningful high-frequency information about the timing of market dislocations. As discussed in Section 4.3, the correlation of the proxy increases as market tail risk increases. Perhaps this increased correlation manifests in improved abilities to forecast mean return but is still detrimental to overall fit indicated by RMSE.

On a relative basis, CL Method's dominance in predicting mean return stays proportional as tail risk increases. For instance, when market tail risk is 0, CL Method's mean return error is 0.15, which is about 3x as small as the error of its closest competitor, Kalman AR(1) No-Intercept, with an error of 0.051. When $\lambda = 1$, CL Method's error is 0.011, which is again roughly 3x as small as its closest competitor, Kalman AR(1) No-Intercept, with an error of 0.031. Panel (c) shows that volatility forecasting abilities of the methods stays relatively constant across levels of tail risk.

The Sharpe Ratio results, in Panel (a), appear to be driven by the mean return forecasting results. However, on a relative basis, back fill and Kalman AR(2) Intercept vastly outperform the CL Method as market tail risk increases. For instance, as $\lambda$ increases from 0 to 1, the CL Method sees roughly a 10x improvement, with error decreasing from 0.041 to 0.003. However, back fill sees roughly a 30x improvement (error decreasing from 0.134 to 0.004), while Kalman AR(2) Intercept sees roughly a 45x improvement (error decreasing from 0.186 to 0.004).

This effect for Sharpe Ratio does not occur with Sortino Ratio. Panel (d) indicates that the relative performance of the CL Method for forecasting Sortino Ratio stays roughly constant across levels of market tail risk. The effect of market tail risk on the ability to recover maximum drawdown (Panel (e)) appears similar to the effect on RMSE. Maximum drawdown error steadily increases for all methods as tail risk increases.

**Figure 6.2** – Portfolio Performance: Market Tail Risk and Loss of Information



## 6.3 Results: Data Autocorrelation Test

This test varies the data autocorrelation Hurst index $H$ from 0.5 to 0.95 to determine the impact of data autocorrelation on inference method performance. At a value of $H = 0.5$, returns have no autocorrelation. For a value of $H = 0.95$, returns are strongly autocorrelated, as discussed in Section 4.2. Proxy goodness and market tail risk are fixed to their default values. Similar to the market tail risk results, these results do not control for the tendency of the CL Method is to outperform the other methods when proxy goodness is 0.6. Therefore, this analysis focuses on how the relative performance of the methods changes with data autocorrelation. The results imply that the overall fit, indicated by RMSE, of all non-proxy-based methods improves relative to the CL Method as data autocorrelation increases. Additionally, autoregressive Kalman filters, which are capable of explicitly modeling data autocorrelation, outperform the CL Method for high levels of autocorrelation in overall fit and Sharpe Ratio forecasting.

Figure 6.3 shows the performance statistics averaged across 1000 simulations, when the Hurst index varies in the range $H \in [0.5, 0.95]$ in a similar manner as the previous two figures. Table A.3, located in Appendix A, reports the same results in raw form.

Panel (f) of this figure shows the overall fit of non-proxy-based methods increases with data autocorrelation. This result makes intuitive sense, as autocorrelation is expected to be volatility reducing. The non proxy-based methods produce inference based on observed quarterly returns divided by 3, and these returns become closer to true monthly returns as volatility decreases. In fact, back fill's RMSE surpasses CL Method's RMSE when $H = 0.75$, and both linear and cubic spline interpolation surpass CL Method's RMSE when $H = 0.85$.

Since proxies in the framework lack the same autocorrelation structure as their corresponding illiquid returns (as one would expect deliberate return smoothing to take place with illiquid assets but not liquid ones), it makes sense that models like MIDAS[1] and the CL Method that cannot explicitly capture autocorrelation in the illiquid data would underperform. Indeed, as data autocorrelation increases, MIDAS's RMSE increases, while the CL Method's RMSE decreases only marginally. In contrast, the autoregressive Kalman filter implementations, which explicitly model autocorrelation in their state equations, outperform the CL Method as autocorrelation increases. At high levels of $H$, their RMSEs decrease to

[1] One possible next step in the research is to extend the MIDAS framework to incorporate autoregression, a desirable but not straightforward feature. Ghysels, Santa-Clara and Valkanov explore the idea of incorporating AR dynamics as a regressor into MIDAS, producing the MIDAS-AR model [Ghysels et al., 2004].

levels below the CL Method but still higher than non-proxy-based methods like back fill.

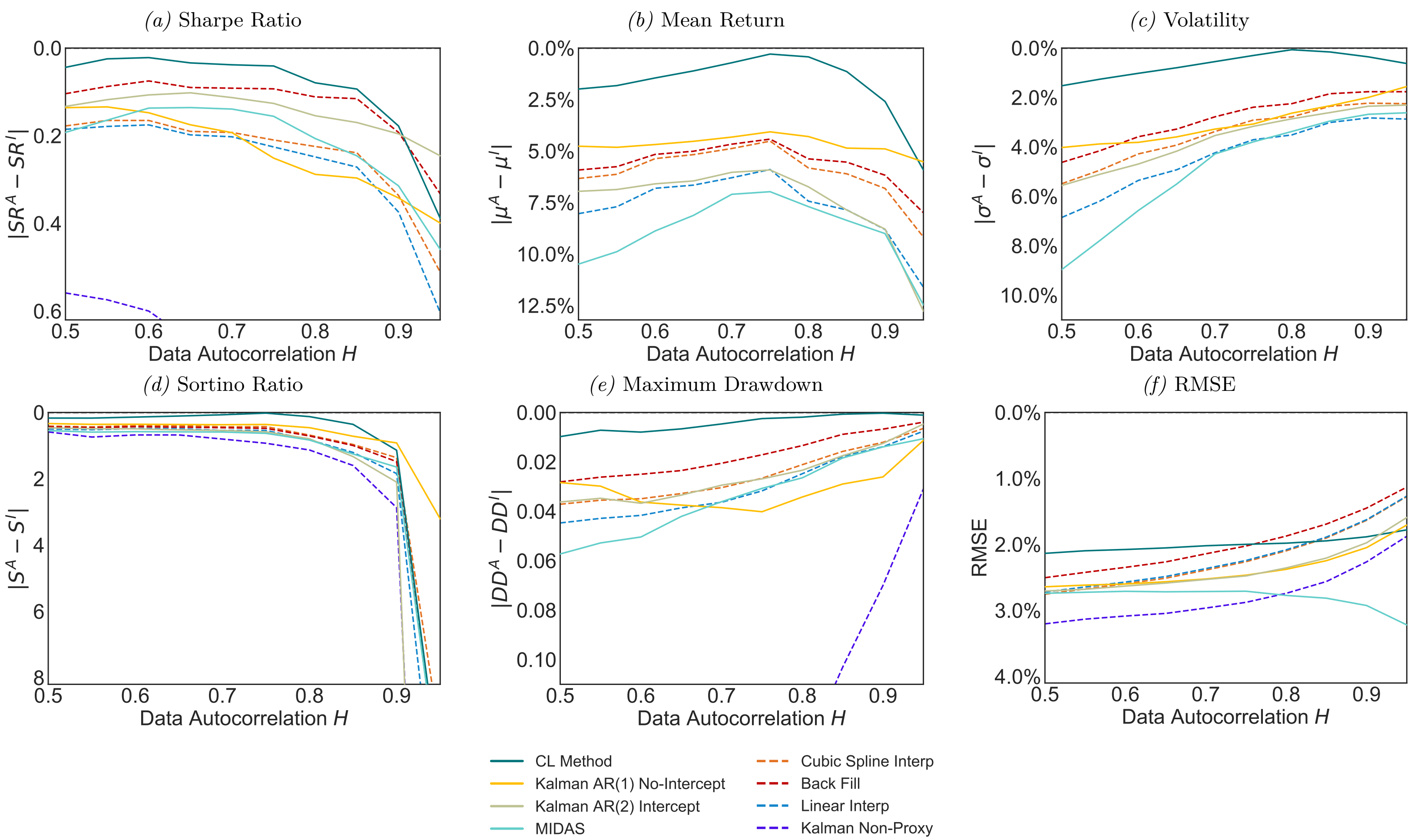
Panel (c) shows that for all methods, the ability to forecast volatility improves with data autocorrelation. Since autocorrelation is volatility reducing, one would expect the error in forecasting volatility with respect to the baseline portfolio to be smaller as $H$ increases. Similarly, since all asset classes in the framework have positive expected return and volatility decreases with $H$, one would expect maximum drawdown to decrease with $H$ as well. Panel (e) confirms this effect, which is persistent across methods.

Autocorrelation has a nonlinear effect on mean return, shown in Panel (b). For all methods, the ability to forecast mean return peaks for moderate Hurst index values and decreases for large Hurst index values. CL Method experiences steepest decline in mean return forecasting performance for high values of $H$. As a result, both back fill and Kalman AR(2) Intercept outperform the CL Method in forecasting Sharpe Ratio when $H = 0.95$. In general, the ability to forecast Sharpe ratio for all methods declines as $H$ increases, shown in Panel (a), primarily driven by the decline in mean return forecasting performance at high autocorrelation. Reduced downside volatility can make Sortino Ratio computations highly unstable. Panel (d) shows that this effect manifests in Sortino Ratio errors that spike for all methods when $H > 0.9$.

## 6.4 Results: Overall Performance

In this section, each inference method is assigned a single composite score, I-Score for "Investor's Score", according to its weighted performance on the metrics according to perceived importance to investors. I-Score collapses the results from a given test – proxy goodness, market tail risk, or data autocorrelation – and helps to quantitatively compare the overall performance of the methodologies. Investors' main goal is achieving superior risk-adjusted return, so Sharpe Ratio, the industry standard, is assigned the highest weight. Investors are likely concerned with the overall goodness of fit of the inference models, so pre-allocated RMSE is included. Investors also seek to minimize downside risk, so maximum drawdown is included in the calculation. Mean return and volatility, two standard performance measures, are also included. I-Score is a normalized weighted sum of scenario Z-scores across Sharpe ratio, pre-allocated RMSE, maximum drawdown, mean return, and volatility in a 3:2:1:1:1 ratio for a given test. These weights are based on intuition. This section first discusses the derivation of I-Score and then applies it to the research's results.

**Figure 6.3** – Portfolio Performance: Data Autocorrelation and Loss of Information



### 6.4.1 Scoring System Derivation

The steps for constructing I-Score are as follows:

1. Find absolute error $d_{s,m,\theta} = \left|s^A - s^I\right|_{m,\theta}$ between statistics ($s \in S =${Sharpe ratio, pre-allocated RMSE, maximum drawdown, mean return, volatility}) based on actual ($A$) and inferred ($I$) data for each method $m$ and parameterization $\theta$ from the family[2] of possible parameterizations $\Theta$.

2. Z-Score these absolute errors so that statistics can be more directly compared:

$$\mu_{m,\theta} = \frac{1}{|S|}\sum_{s=1}^{|S|} d_{s,m,\theta}$$
$$Z_{s,m,\theta} = \frac{d_{s,m,\theta} - \mu_{m,\theta}}{\sqrt{\frac{1}{|S|-1}\sum_{s=1}^{|S|}\left(d_{s,m,\theta} - \mu_{m,\theta}\right)^2}}$$

   where $|S| = 3$ is the number of statistics that are considered.

3. Average Z-Scores across parameterizations $\theta \in \Theta$ into low, mid, and high buckets for a given parameter. This step simplifies the recommendation of which inference method to use into three scenarios. In the case of the proxy goodness, this looks like

$$Z_{s,m,\text{high}} = \frac{1}{7}\sum_{\tilde{\rho}\in\{1,0.95,0.9,0.85,0.8,0.75,0.7\}} Z_{s,m,\rho=\tilde{\rho}}$$
$$Z_{s,m,\text{mid}} = \frac{1}{7}\sum_{\tilde{\rho}\in\{0.65,0.6,0.55,0.5,0.45,0.4,0.35\}} Z_{s,m,\rho=\tilde{\rho}}$$
$$Z_{s,m,\text{low}} = \frac{1}{7}\sum_{\tilde{\rho}\in\{0.3,0.25,0.2,0.15,0.1,0.05,0\}} Z_{s,m,\rho=\tilde{\rho}}$$

   where $Z_{s,m,\rho=j}$ indicates that the parameter $\rho$ has been set to $j$ and the remaining parameters are set to default values. Similar definitions follow for the market tail risk and data autocorrelation tests.

4. Calculate the score $S_{m,\text{bucket}}$ as a weighted sum of these across statistics $s \in S$, weighting Sharpe ratio, pre-allocated RMSE, maximum drawdown, mean return, and

[2] In the case of the proxy correlation test, $\Theta = \{\theta_{\rho=0}, \theta_{\rho=0.05}, \ldots, \theta_{\rho=1}\}$ where $\theta_{\rho=j}$ indicates the parameter $\rho$ is set to $j$ while the remaining parameters are set to default values. The families for the market tail risk test and data autocorrelation test are defined similarly.

volatility in a 3:2:1:1:1 ratio. More formally, this step looks like

$$
\begin{aligned}
\boldsymbol{w} &= \begin{bmatrix} 3 & 2 & 1 & 1 & 1 \end{bmatrix}^T \\
\boldsymbol{s}_{m,\text{low}} &= \begin{bmatrix} Z_{\text{SR},m,\text{low}} & Z_{\text{RMSE},m,\text{low}} & Z_{\text{MD},m,\text{low}} & Z_{\text{mean},m,\text{low}} & Z_{\text{vol},m,\text{low}} \end{bmatrix}^T \\
S_{m,\text{low}} &= \frac{\boldsymbol{w}^T \boldsymbol{s}_{m,\text{low}}}{\sum_i \boldsymbol{w}_i}
\end{aligned}
$$

5. Normalize these scores across methods between $[0, 1]$, where a larger number means better performance, using the formula

$$
\begin{aligned}
S_{\text{bucket}} &= \{S_{\text{CL,bucket}}, S_{\text{ForwardFill,bucket}}, \ldots, S_{\text{MIDAS,bucket}}\} \\
\text{I-Score}_{m,\text{bucket}} &= 1 - \frac{S_{m,\theta} - \min\left(S_{\text{bucket}}\right)}{\max\left(S_{\text{bucket}}\right) - \min\left(S_{\text{bucket}}\right)}
\end{aligned}
$$

### 6.4.2 Relative Performance Results

The following discussion applies I-Score to the research's main results. Figure 6.4 visualizes the I-Score of each inference method for low, medium, and high parameter buckets across research tests: proxy goodness, market tail risk, and data autocorrelation. I-Score is computed by weighting each method's performance statistics according to perceived importance to investors and are normalized between 0 and 1. A higher I-Score score indicates superior relative performance.

Panel (a) shows that for low values of proxy goodness, back fill has the highest I-Score, closely followed by CL Method. For moderate and high levels of proxy goodness, CL Method performs best overall according to this metric. In general, as proxy goodness increases, the I-Scores of the proxy-based methods increase, as expected. Since I-Score computes relative performance, the performance of non proxy-based methods correspondingly decreases as proxy goodness increases. For example, for low proxy goodness, three of the four top performing methods are non-proxy-based. For high proxy goodness, all four top-performing methods are proxy-based. At high levels of data autocorrelation, Kalman AR(2) Intercept is the second-best performing method, behind CL Method.

Note that the market tail risk and data autocorrelation results do not control for the CL Method's expected outperformance at default parameter values. Therefore, in analyzing Panels (b) and (c), simply looking at which methods have the highest I-Scores is unhelpful.

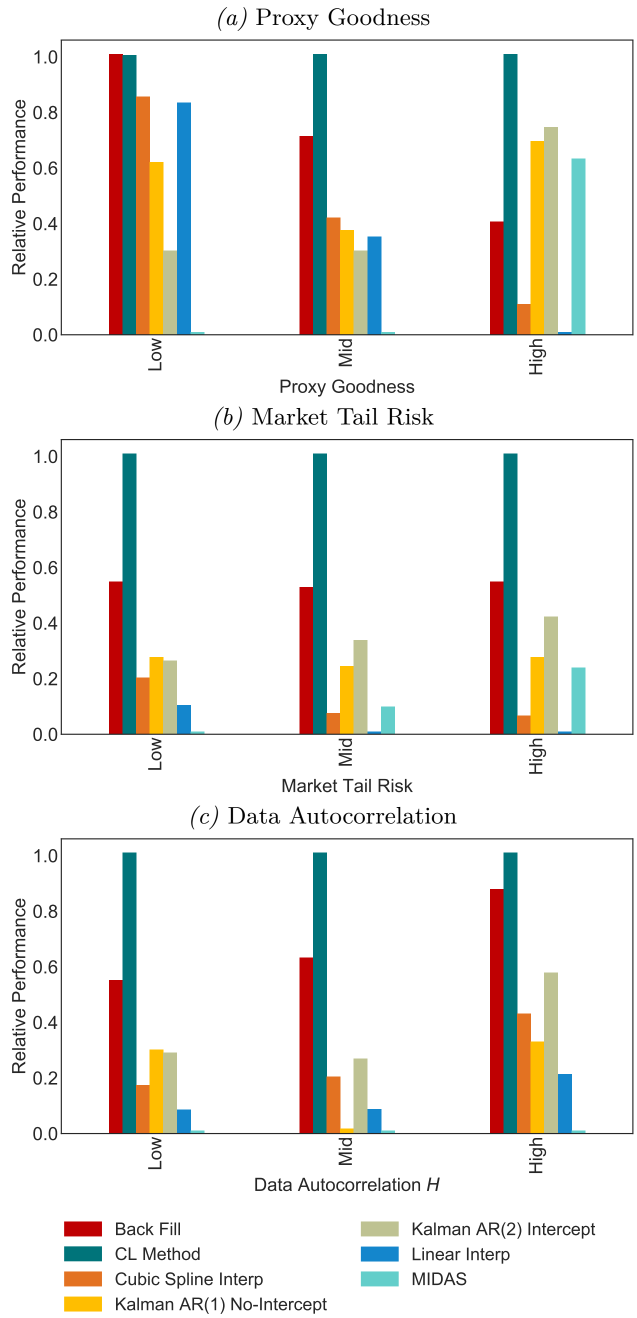
One should instead look at how the scores for a given method change across parameter buckets.

Panel (b) shows that market tail risk does not meaningfully impact the I-Scores. For instance, in all three states of the world – low, mid, and high market tail risk – the CL Method is the clear favorite over back fill, by about a factor of 2. The I-Score of Kalman AR(2) Intercept increases slightly as market tail risk increases, though likely not enough to be meaningful.

The data autocorrelation results, shown in Panel (c), are more interesting. Though CL Method has the highest I-Score across parameter buckets, the I-Score of back fill grows as autocorrelation increases. For high autocorrelation, back fill's I-Score is about 0.9, compared to CL Method's 1.0, which is quite close. Also, the I-Score of Kalman AR(2) Intercept increases with data autocorrelation, driven by the filter's explicit modeling of the autocorrelation in the state transition function.

**Figure 6.4** – Overall Relative Performance

The I-Score of each inference method for low, medium, and high parameter buckets across scenarios: (a) proxy goodness, (b) market tail risk, and (c) data autocorrelation. I-Score is computed by weighting each method's performance statistics according to perceived importance to investors and are normalized between 0 and 1. A higher I-Score score indicates superior relative performance.



## 6.5 Results: Normal Market Conditions Test

This section shows results from a single simulation where the market parameters are set to their default values. This test is meant to highlight performance of the methods under *typical* market conditions. Parameters are fixed according to assumptions that in normal market conditions there exists a moderately correlated proxy available, low-mid chance of market dislocations, and a small degree of autocorrelation of returns [Getmansky et al., 2004, Khandani and Lo, 2011]. This parameterization is based on literature about illiquid returns autocorrelation as well as intuition.

Table 6.1 shows the raw performance across metrics in this scenario, measured as the average absolute error between the actual ($A$) and the inferred ($I$) portfolio's statistics, with lower errors indicating superior performance. Figure 6.5 shows the relative performance of each method according to an unweighted I-Score computation[3].

Consistent with previous results, the results show that the CL Method performs best overall in a typical state of the world. In fact, the CL Method has the highest unweighted I-Score across all portfolio performance measures. The second-best overall method is back fill, which reliably recovers Sharpe Ratio and maximum drawdown. On a relative basis, back fill provides a much lower RMSE than its counterparts except for the CL Method.

It is clear that the performance of the CL Method in this test and across the previous simulations makes it an all-weather method and the best choice for use when at least a moderately correlated proxy is available. When a good proxy is unavailable, back fill is a convenient and high-performing choice. In conditions of extremely high data autocorrelation, it makes sense to switch to models like the Kalman filter that can more appropriately model the autocorrelation. However, the CL Method is a safe bet for the allocator under normal market conditions.

[3] The steps for this computation are the same as described in Section 6.4.1, but leave out step (4) which performs the weighting across statistics.

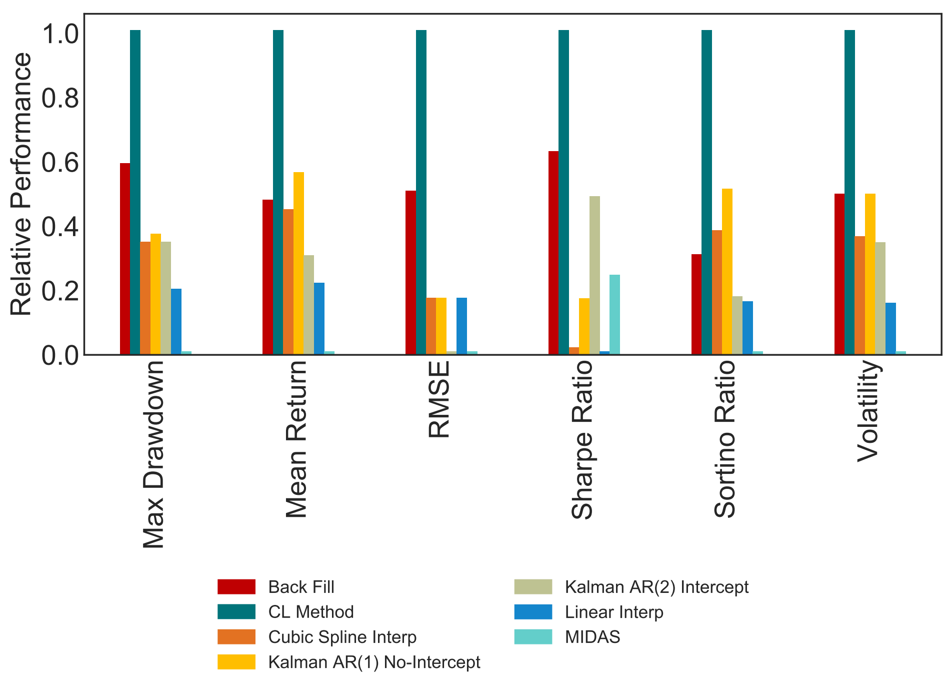
| Method | Statistic | | | | | |
|---|---|---|---|---|---|---|
| | Max DD | Mean | RMSE | Sharpe | Sortino | Vol |
| Back Fill | 0.025 | 0.051 | 0.024 | 0.096 | 0.432 | 0.037 |
| CL Method | 0.008 | 0.014 | 0.021 | 0.039 | 0.124 | 0.010 |
| Cubic Spline Interp | 0.035 | 0.053 | 0.026 | 0.188 | 0.399 | 0.044 |
| Kalman AR(1) No-Intercept | 0.034 | 0.045 | 0.026 | 0.165 | 0.342 | 0.037 |
| Kalman AR(2) Intercept | 0.035 | 0.063 | 0.027 | 0.117 | 0.490 | 0.045 |
| Kalman Non-AR Intercept | 0.032 | 0.065 | 0.025 | 0.079 | 0.523 | 0.044 |
| Kalman Non-Proxy | 0.504 | 0.799 | 0.031 | 0.606 | 0.766 | 0.904 |
| Linear Interp | 0.041 | 0.069 | 0.026 | 0.190 | 0.497 | 0.055 |
| MIDAS | 0.049 | 0.084 | 0.027 | 0.154 | 0.566 | 0.063 |

**Table 6.1** – Realistic State of the World Performance

The performance of each inference method in a realistic state of the world, in which there is a good proxy available, low-medium market dislocation risk, and slight data autocorrelation. Scores are normalized error of performance from actual $(A)$ and inferred $(I)$ data for each methodology.

**Figure 6.5** – Realistic State of the World Relative Performance

The relative performance of each inference method in a realistic state of the world, in which there is a good proxy available, low-medium market dislocation risk, and slight data autocorrelation. Scores are normalized distance of performance from actual $(A)$ and inferred $(I)$ data for each methodology. A higher score indicates superior relative performance.



# Chapter 7

# Empirical Analysis

The following is an empirical analysis showcasing the research's results on real data, tying together the loose ends of Chapter 2. Consider the same hypothetical investor who holds liquid index positions in stocks (MSCI World Investable Markets), bonds (Barclays U.S. Aggregate Bond Total Return), hedge funds (HFRI Fund Weighted Composite), and commodities (S&P Goldman Sachs Commodities Total Return), observed monthly. Her historical data for each index extends from June 1994 to March 2017. The investor is considering adding an illiquid investment in the form of private equity (Thomson Reuters PE Buyout Index), observable quarterly, to the portfolio and wants to assess (1) whether to add the investment at all and (2) which data inference model to use to infer quarterly returns to monthly frequency.

To address the first point, the investor might consider historical mean returns and the correlation matrix involving private equity and the other assets in the portfolio, shown in Table 7.1. Since private equity is observable quarterly, she must aggregate the other data series to quarterly frequency in order to generate this correlation matrix. The investor first observes that the mean historical performance of private equity is strongly positive, at 6.6% annualized return, which is a good start. She then sees that since the private equity is negatively correlated with bonds (-7%) and weakly correlated with commodities (18%) in her existing portfolio, then it would likely provide meaningful diversification benefit. She decides to go ahead and consider portfolio allocations involving private equity.

| | Stocks | Bonds | Hedge Fnd | Comdity | PE |
|---|---|---|---|---|---|
| Stocks | 1.00 | -0.23 | 0.83 | 0.45 | 0.78 |
| Bonds | -0.23 | 1.00 | -0.21 | -0.23 | -0.07 |
| Hedge Fnd | 0.83 | -0.21 | 1.00 | 0.61 | 0.72 |
| Comdity | 0.45 | -0.23 | 0.61 | 1.00 | 0.18 |
| PE | 0.78 | -0.07 | 0.72 | 0.18 | 1.00 |
| Mean Return | 0.084 | 0.046 | 0.060 | 0.065 | 0.066 |

**Table 7.1** – Empirical Analysis Historical Returns and Correlation Matrix

The historical mean returns and quarterly correlation matrix for a portfolio of liquid assets such as stocks (MSCI ACWI IMI), bonds (Barclays U.S. Agg TR), hedge funds (HFRI Fund Weighted), commodities (GSCI TR), observed monthly, as well as private equity (Thomson Reuters PE Buyout), observed quarterly.

Applying the research, the investor knows that if she can find a proxy with correlation at least 0.25, it is valuable to use the CL Method over back fill in normal market conditions. She finds that the Russell 2000 Index (RPY) is 0.71 correlated with her private equity index at quarterly frequency. It makes sense that small-cap equity would be a good proxy for private equity, since many private equity investments are in markets represented in the Russell 2000, so she is happy with this choice in proxy. She goes forward and infers the quarterly private equity returns to monthly frequency via the CL Method, using the Russell 2000 as a proxy, and re-balances her new portfolio using mean-variance optimization with a rolling window of three years, as was done in the research.

It just so happens that our investor's private equity index, the Thomson Reuters PE Buyout Index, intends to track the returns of the private equity universe via a replicating portfolio of liquid exchange-traded instruments and is therefore available at monthly frequency. As a result, a "full-data" portfolio can be constructed, as done in the research, to benchmark the performance of this data inference choice. This analysis allocates several portfolios. The first portfolio is composed the basket of stocks, bonds, hedge funds, commodities, and private equity observed monthly and serves as the benchmark for the CL Method and other inference methods. The second portfolio is constructed from the same liquid assets, without private equity. This portfolio serves as a second benchmark and helps answer the question of whether it was wise for the investor to include private equity in her asset allocation in the first place. The remaining portfolios are constructed from the original basket of assets, with the private equity index aggregated to quarterly and then inferred back to monthly frequency. The CL Method uses the Russell 2000 index (RTY) as the proxy while back fill does not utilize proxies. The Sharpe and Sortino Ratio computations assume a 3% risk-free rate.

Similar to the research framework, portfolios allocated from inferred data are expected to underperform the portfolio allocated from full data. The portfolio allocated via the CL Method should outperform the investor's original portfolio without illiquid assets, or else the investor made a bad choice in including the illiquid asset. Table 7.2 summarizes these results. As expected, the full data portfolio outperforms all the other portfolios, achieving a Sharpe ratio of 0.630. The portfolio applying the CL Method, with a Sharpe ratio of 0.603, is superior to the portfolio using forward fill, achieving a Sharpe Ratio of 0.554. This result falls in-line with the proxy goodness results discussed in Section 6.1. As a confirmation it was wise adding the illiquid investment in the first place, the portfolio involving the illiquid investment inferred via CL Method outperforms the portfolio without illiquids, which achieves the lowest Sharpe Ratio of all allocated portfolios. As a sanity check, it is confirmed that the portfolio allocated via the CL Method yields similar mean returns, volatility, and maximum drawdown to the full data portfolio. These results increase confidence that the hypothetical investor made the right choice in including illiquid investments in her portfolio with data inferred via the CL Method.

| | Full Data | No Illiquids | CL Method | Forward Fill |
|---|---|---|---|---|
| Mean | 0.051 | 0.046 | 0.052 | 0.053 |
| Vol | 0.034 | 0.033 | 0.036 | 0.041 |
| Min | -0.385 | -0.385 | -0.385 | -0.438 |
| Max | 0.448 | 0.448 | 0.448 | 0.580 |
| Max DD | -0.058 | -0.058 | -0.058 | -0.058 |
| Sharpe | 0.630 | 0.502 | 0.603 | 0.554 |
| Sortino | 0.223 | 0.170 | 0.226 | 0.234 |
| #obs | 125 | 125 | 125 | 125 |
| Uses Proxy | N/A | N/A | Yes | No |

**Table 7.2** – Empirical Analysis Performance Results

The performance of a mean-variance portfolio with and without including an illiquid asset. The portfolio labeled "Full Data" is composed of liquid assets such as stocks (MSCI ACWI IMI), bonds (Barclays U.S. Agg TR), hedge funds (HFRI Fund Weighted), commodities (GSCI TR), and private equity (Thomson Reuters PE Buyout), observed monthly. The portfolio labeled "No Illiquids" is constructed from the same assets, without private equity. The remaining portfolios are constructed from the original basket of assets, with the private equity index aggregated to quarterly and then inferred back to monthly frequency. The CL Method uses the Russell 2000 index (RTY), observed monthly, as the proxy while forward fill does not utilize proxies. The Sharpe and Sortino ratio results assume a 3% risk-free rate.

# Chapter 8

# Conclusion and Future Work

Investing in illiquid assets poses a challenge to investors, as the lack of higher-frequency information makes it difficult to quantify the risks across portfolios and make asset allocation decisions. This work reviewed several commonly used methodologies to infer missing data and tested their implications for asset allocation. It compared these methods by applying them to hypothetical portfolios in a realistic simulation environment, helping allocators decide which option to use and when. The simulation environment generated data that varied along three dimensions – the risk of market dislocations, autocorrelation, and availability of a good proxy – and evaluated portfolios along industry standard performance metrics like Sharpe Ratio and drawdown.

The results imply that the CL Method is an all-weather method and likely the allocator's best choice for typical states of the world with at least a moderately good proxy available. If only a poor proxy is available, back fill is a reasonable choice. These recommendations are consistent across levels of market tail risk. If data autocorrelation is high, models like autoregressive Kalman filters, which are capable of explicitly modeling the autocorrelation, are preferred. To demonstrate these results in practice, an empirical analysis is included, showcasing the research design on historical data.

There are several potential areas for future research. The first is researching other robust proxy-based methods like the Mariano and Marasawa method, explored by Dogan and Szigety [Mallinger-Dogan and Szigety, 2014], as well as more sophisticated filtering and filter parameter estimation techniques. Also, the MIDAS framework could be extended with an autoregression term to better capture data autocorrelation [Ghysels et al., 2004].

The research made the simplifying assumption to utilize a single proxy for inference, and it is reasonable to hypothesize that adding additional proxies could change the results. However, adding too many proxies to the filtering methods will likely decrease accuracy due to overfitting and parameter estimation issues [Mallinger-Dogan and Szigety, 2014]. The research could have also implemented more complex autocorrelation structures observed in illiquid returns rather than using a simple fractional Brownian motion process. Last, the research is limited to inferring quarterly returns to monthly frequency. A high-frequency trader or another investor might find herself in a situation in which it would be useful to infer daily returns to monthly or even more-extreme conversions.

Though the infrequent data common with illiquid investments can challenge investors, there are approaches and tools to help. We hope our work provides useful results for institutions holding or wishing to hold illiquid assets.

# Bibliography


[Almon, 1965] Almon, S. (1965). The distributed lag between capital appropriations and expenditures. 33(1):178–196.

[Ang, 2014] Ang, A. (2014). *Asset Management: A Systematic Approach to Factor Investing*. Oxford University Press.

[Armesto et al., 2010] Armesto, M. T., Engemann, K., and Owyang, M. (2010). Forecasting with mixed frequencies. (Nov):521–536.

[Asmussen et al., 1998] Asmussen, S., universitet. Center for matematisk fysik og stokastik, A., for Mathematical Physics, Å. U. C., and Stochastics (1998). *Stochastic Simulation with a View Towards Stochastic Processes*. Lecture notes. University of Aarhus, Centre for Mathematical Physics and Stochastics.

[Bai et al., 2013] Bai, J., Ghysels, E., and Wright, J. H. (2013). State space models and midas regressions. 32(7):779–813.

[Bernd, 2002] Bernd, S. (2002). Portfolio resampling: Review and critique. 58(6):98–109.

[Carr and Wu, 2007] Carr, P. and Wu, L. (2007). Theory and evidence on the dynamic interactions between sovereign credit default swaps and currency options. 31(8):2383–2403.

[Chow and Lin, 1971] Chow, G. C. and Lin, A.-L. (1971). Best linear unbiased interpolation, distribution, and extrapolation of time series by related series. 53(4):372.

[Cuche and Hess, 2000] Cuche, N. A. and Hess, M. K. (2000). Estimating monthly gdp in a general kalman filter framework: evidence from switzerland. 7(4):153–194.

[Dieker, 2004] Dieker, T. (2004). *Simulation of fractional Brownian motion*. PhD thesis.

[Eric Gysels, 2006] Eric Gysels, Arthur Sinko, R. V. (2006). Midas regressions: Further results and new directions.

[Getmansky et al., 2004] Getmansky, M., Lo, A., and Makarov, I. (2004). An econometric model of serial correlation and illiquidity in hedge fund returns. 74(3):529–609.

[Ghysels et al., 2004] Ghysels, E., Santa-Clara, P., and Valkanov, R. (2004). The midas touch: Mixed data sampling regression models. University of california at los angeles, anderson graduate school of management, Anderson Graduate School of Management, UCLA.

[Hudson and Gergoriou, 2010] Hudson, R. and Gergoriou, A. (2010). Calculating and comparing security returns is harder than you think: A comparison between logarithmic and simple returns.

[Jenkinson et al., 2013] Jenkinson, T., Sousa, M., and Stucke, R. (2013). How fair are the valuations of private equity funds?

[Kalman, 1960] Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. 82(1):35.

[Kelly et al., 2016] Kelly, B. T., Manzo, G., and Palhares, D. (2016). Credit-implied volatility.

[Kempthorne, 2013] Kempthorne, P. (2013). Lecture 14: Portfolio theory - 18.s096. Technical report, Massachusetts Institute of Technology. `ocw.mit.edu/courses/mathematics/`.

[Khandani and Lo, 2011] Khandani, A. E. and Lo, A. W. (2011). Illiquidity premia in asset returns: An empirical analysis of hedge funds, mutual funds, and us equity portfolios. 01(02):205–264.

[Lyuu, 2015] Lyuu, Y.-D. (2015). Merton's jump diffusion. Technical report.

[Mallinger-Dogan and Szigety, 2014] Mallinger-Dogan, M. and Szigety, M. C. (2014). Higher-frequency analysis of low-frequency data. 41(1):121.

[Mandelbrot and Ness, 1968] Mandelbrot, B. B. and Ness, J. W. V. (1968). Fractional brownian motions, fractional noises and applications. 10(4):422–437.

[Mariano and Murasawa, 2010] Mariano, R. S. and Murasawa, Y. (2010). A coincident index, common factors, and monthly real gdp. 72(1):27–46.

[Markowitz, 1952] Markowitz, H. (1952). Portfolio selection. 7(1):77–91.

[Merton, 1976] Merton, R. (1976). Option pricing when underlying stock returns are discontinuous. 3(1-2):125–144.

[Pandas, 2017] Pandas (2017). Working with missing data. Technical report. `pandas.pydata.org/pandas-docs/`.

[Reid and Term, 2001] Reid, I. and Term, H. (2001). Estimation ii. Technical report, Oxford University. `robots.ox.ac.uk/~ian/Teaching/`.

[Sharpe, 1966] Sharpe, W. F. (1966). Mutual fund performance. 39(1):119–138.

[Sortino and Price, 1994] Sortino, F. and Price, L. (1994). Performance measurement in a downside risk framework. 3(3):59–64.

[Tarnopolski, 2006] Tarnopolski, M. (2006). On the relationship between the hurst exponent, the ratio of the mean square successive difference to the variance, and the number of turning points. *Astronomical Observatory*. `https://arxiv.org/pdf/1512.02928.pdf`.

[Thacker and Lacey, 2006] Thacker, N. and Lacey, T. (2006). Tutorial: The likelihood interpretation of the kalman filter. Technical report, University of Manchester. `tina-vision.net/docs/memos/`.

[Treuille, 2010] Treuille, A. (2010). Curves and splines. Technical report. `cs.cmu.edu/afs/cs/`.

[Welch and Bishop, 2006] Welch, G. and Bishop, G. (2006). An introduction to the kalman filter. Technical report, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3175. `cs.unc.edu/~welch/`.

[Young and Mohlenkamp, 2009] Young, T. and Mohlenkamp, M. J. (2009). Polynomial and spline interpolation. Technical report. `math.ohiou.edu/courses/`.

# Appendix A

# Appendix: Raw Results

The following tables report the results displayed in Figures 6.1, 6.2, and 6.3 in raw form.

Results are composed of five portfolio metrics, as described in Section 5.4– mean return, volatility, Sharpe ratio, Sortino ratio, and maximum drawdown. These reported metrics are the absolute errors between the statistics based on true monthly data and those based on quarterly data inferred to monthly frequency, $|s^A - s^I|$, where $A$ and $I$ stand for actual and inferred, respectively. The absolute error represents the loss of information due to a specific inference methodology and allows performance to be benchmarked with respect to the baseline portfolio. Lower errors indicate superior performance. In addition to these portfolio metrics, RMSE between illiquid returns in inferred and actual portfolios is computed to gain an overall sense of fit prior to allocation. RMSE, by definition, represents the error with respect to the baseline portfolio, so it is not normalized like the other five statistics. All reported statistics are the sample means across 1000 trials, which converge to true values by the law of large numbers.

Each table in this Appendix gives Sharpe Ratio, mean return, volatility, Sortino Ratio, maximum drawdown, and pre-allocated RMSE (RMSE), averaged across 1000 simulations, for a single research test specified in 5.2. For each simulated portfolio, parameter values, and inference technique, the tables report the average absolute error between the actual ($A$) and the inferred ($I$) portfolio's statistics. The results reported in each table are as follows:

- Table A.1 – *Proxy Goodness Test:* determine the impact of a proxy goodness on inference method performance. Proxy goodness $\rho$ ranges from zero (poor proxy) to one (perfect proxy). Specifically, the parameter $\rho$ is varied in the range $\rho \in [0, 1]$ in increments of 0.05. All other parameters are set to default values.

- Table A.2– *Market Tail Risk Test*: determine the impact of market tail risk on inference method performance. A spectrum of zero, low, medium, and high tail risk scenarios are simulated by varying $\lambda \in [0, 1]$ in increments of 0.05. All other parameters are set to default values.

- Table A.3 – *Data Autocorrelation Test*: determine the impact of data autocorrelation on inference method performance. To simulate a spectrum of zero, low, medium, and high autocorrelation scenarios, the parameter $H$ varies in the range $H \in [0.5, 0.95]$ in increments of 0.05. All other parameters are set to default values.

**Table A.1** – Proxy Goodness Test Raw Results

| Statistic | Proxy Goodness $\rho$ | Back Fill | CL Method | Cubic Spline Interp | Kalman AR(1) No-Intcpt. | Kalman AR(2) Intercept | Kalman Non-AR Intercept | Kalman Non-Proxy | Linear Interp | MIDAS |
|---|---|---|---|---|---|---|---|---|---|---|
| Max Drawdown | 0.0 | 0.024 | 0.020 | 0.035 | 0.069 | 0.206 | 0.498 | 0.464 | 0.042 | 0.418 |
| | 0.05 | 0.024 | 0.019 | 0.034 | 0.069 | 0.195 | 0.396 | 0.459 | 0.041 | 0.332 |
| | 0.1 | 0.024 | 0.017 | 0.034 | 0.068 | 0.195 | 0.367 | 0.472 | 0.041 | 0.952 |
| | 0.15 | 0.026 | 0.018 | 0.036 | 0.068 | 0.170 | 0.993 | 0.468 | 0.043 | 0.316 |
| | 0.2 | 0.024 | 0.015 | 0.034 | 0.065 | 0.141 | 0.256 | 0.450 | 0.041 | 0.248 |
| | 0.25 | 0.025 | 0.014 | 0.034 | 0.065 | 0.128 | 0.185 | 0.453 | 0.041 | 0.230 |
| | 0.3 | 0.025 | 0.013 | 0.034 | 0.063 | 0.107 | 0.148 | 0.420 | 0.041 | 0.197 |
| | 0.35 | 0.025 | 0.012 | 0.035 | 0.058 | 0.092 | 0.111 | 0.384 | 0.041 | 0.156 |
| | 0.4 | 0.024 | 0.010 | 0.034 | 0.055 | 0.078 | 0.083 | 0.491 | 0.041 | 0.134 |
| | 0.45 | 0.024 | 0.009 | 0.033 | 0.047 | 0.060 | 0.063 | 0.449 | 0.040 | 0.094 |
| | 0.5 | 0.024 | 0.009 | 0.034 | 0.044 | 0.052 | 0.052 | 0.579 | 0.040 | 0.079 |
| | 0.55 | 0.026 | 0.008 | 0.035 | 0.038 | 0.043 | 0.042 | 0.452 | 0.042 | 0.064 |
| | 0.6 | 0.024 | 0.007 | 0.033 | 0.034 | 0.036 | 0.033 | 0.545 | 0.040 | 0.046 |
| | 0.65 | 0.026 | 0.006 | 0.035 | 0.029 | 0.029 | 0.026 | 1.139 | 0.041 | 0.038 |
| | 0.7 | 0.025 | 0.006 | 0.034 | 0.026 | 0.024 | 0.020 | 0.397 | 0.041 | 0.029 |
| | 0.75 | 0.024 | 0.006 | 0.033 | 0.022 | 0.020 | 0.017 | 0.431 | 0.039 | 0.023 |
| | 0.8 | 0.025 | 0.004 | 0.035 | 0.019 | 0.016 | 0.014 | 0.457 | 0.042 | 0.018 |
| | 0.85 | 0.025 | 0.004 | 0.034 | 0.014 | 0.011 | 0.009 | 1.853 | 0.041 | 0.012 |
| | 0.9 | 0.023 | 0.003 | 0.032 | 0.010 | 0.007 | 0.006 | 0.494 | 0.039 | 0.008 |
| | 0.95 | 0.026 | 0.002 | 0.036 | 0.006 | 0.005 | 0.004 | 0.586 | 0.041 | 0.006 |
| | 1.0 | 0.024 | 0.000 | 0.035 | 0.000 | 0.000 | 0.000 | 0.435 | 0.041 | 0.001 |
| Mean Return | 0.0 | 0.051 | 0.036 | 0.054 | 0.085 | 0.285 | 0.835 | 0.708 | 0.069 | 0.639 |
| | 0.05 | 0.052 | 0.034 | 0.054 | 0.090 | 0.261 | 0.778 | 0.709 | 0.070 | 0.576 |
| | 0.1 | 0.052 | 0.032 | 0.053 | 0.089 | 0.241 | 0.674 | 0.684 | 0.068 | 0.607 |
| | 0.15 | 0.052 | 0.029 | 0.055 | 0.089 | 0.222 | 0.646 | 0.758 | 0.070 | 0.524 |
| | 0.2 | 0.051 | 0.027 | 0.053 | 0.089 | 0.187 | 0.405 | 0.709 | 0.069 | 0.459 |
| | 0.25 | 0.052 | 0.025 | 0.055 | 0.083 | 0.171 | 0.305 | 0.763 | 0.070 | 0.387 |
| | 0.3 | 0.051 | 0.023 | 0.053 | 0.076 | 0.142 | 0.216 | 0.702 | 0.068 | 0.324 |
| | 0.35 | 0.050 | 0.021 | 0.053 | 0.073 | 0.127 | 0.166 | 0.686 | 0.067 | 0.281 |
| | 0.4 | 0.051 | 0.020 | 0.053 | 0.070 | 0.111 | 0.128 | 0.671 | 0.068 | 0.185 |
| | 0.45 | 0.053 | 0.019 | 0.055 | 0.065 | 0.104 | 0.117 | 0.673 | 0.070 | 0.169 |
| | 0.5 | 0.050 | 0.016 | 0.052 | 0.055 | 0.083 | 0.090 | 0.614 | 0.067 | 0.125 |
| | 0.55 | 0.054 | 0.017 | 0.057 | 0.056 | 0.078 | 0.082 | 0.750 | 0.073 | 0.108 |
| | 0.6 | 0.052 | 0.014 | 0.055 | 0.048 | 0.065 | 0.067 | 0.689 | 0.070 | 0.086 |
| | 0.65 | 0.051 | 0.013 | 0.054 | 0.041 | 0.057 | 0.057 | 0.627 | 0.069 | 0.073 |
| | 0.7 | 0.052 | 0.012 | 0.055 | 0.037 | 0.047 | 0.047 | 0.729 | 0.070 | 0.055 |
| | 0.75 | 0.053 | 0.010 | 0.055 | 0.033 | 0.039 | 0.038 | 0.633 | 0.070 | 0.044 |
| | 0.8 | 0.053 | 0.010 | 0.055 | 0.027 | 0.031 | 0.030 | 0.947 | 0.071 | 0.032 |
| | 0.85 | 0.052 | 0.008 | 0.055 | 0.023 | 0.024 | 0.023 | 0.630 | 0.070 | 0.023 |
| | 0.9 | 0.054 | 0.007 | 0.057 | 0.018 | 0.017 | 0.017 | 0.668 | 0.073 | 0.014 |
| | 0.95 | 0.051 | 0.005 | 0.053 | 0.012 | 0.008 | 0.008 | 0.693 | 0.068 | 0.002 |
| | 1.0 | 0.052 | 0.000 | 0.054 | 0.001 | 0.000 | 0.001 | 0.673 | 0.069 | 0.001 |

Proxy Goodness Test Raw Results (Continued)

| Statistic | Proxy Goodness $\rho$ | Back Fill | CL Method | Cubic Spline Interp | Kalman AR(1) No-Intcpt. | Kalman AR(2) Intercept | Kalman Non-AR Intercept | Kalman Non-Proxy | Linear Interp | MIDAS |
|---|---|---|---|---|---|---|---|---|---|---|
| RMSE | 0.0 | 0.023 | 0.025 | 0.026 | 0.031 | 0.032 | 0.031 | 0.031 | 0.026 | 0.037 |
| | 0.05 | 0.023 | 0.025 | 0.026 | 0.031 | 0.032 | 0.030 | 0.031 | 0.026 | 0.037 |
| | 0.1 | 0.023 | 0.025 | 0.026 | 0.031 | 0.032 | 0.030 | 0.031 | 0.026 | 0.036 |
| | 0.15 | 0.024 | 0.025 | 0.026 | 0.031 | 0.032 | 0.030 | 0.031 | 0.026 | 0.036 |
| | 0.2 | 0.024 | 0.024 | 0.026 | 0.031 | 0.032 | 0.030 | 0.031 | 0.026 | 0.035 |
| | 0.25 | 0.023 | 0.024 | 0.026 | 0.030 | 0.031 | 0.030 | 0.031 | 0.026 | 0.034 |
| | 0.3 | 0.024 | 0.024 | 0.026 | 0.030 | 0.031 | 0.029 | 0.031 | 0.026 | 0.033 |
| | 0.35 | 0.023 | 0.023 | 0.026 | 0.030 | 0.030 | 0.029 | 0.031 | 0.026 | 0.032 |
| | 0.4 | 0.023 | 0.023 | 0.026 | 0.029 | 0.029 | 0.028 | 0.031 | 0.026 | 0.031 |
| | 0.45 | 0.024 | 0.023 | 0.026 | 0.028 | 0.029 | 0.028 | 0.031 | 0.026 | 0.030 |
| | 0.5 | 0.023 | 0.022 | 0.026 | 0.028 | 0.028 | 0.027 | 0.031 | 0.026 | 0.029 |
| | 0.55 | 0.024 | 0.022 | 0.026 | 0.027 | 0.027 | 0.026 | 0.031 | 0.026 | 0.028 |
| | 0.6 | 0.024 | 0.021 | 0.026 | 0.026 | 0.026 | 0.025 | 0.031 | 0.026 | 0.027 |
| | 0.65 | 0.024 | 0.020 | 0.026 | 0.025 | 0.025 | 0.025 | 0.031 | 0.026 | 0.026 |
| | 0.7 | 0.024 | 0.019 | 0.026 | 0.024 | 0.024 | 0.023 | 0.031 | 0.026 | 0.025 |
| | 0.75 | 0.023 | 0.018 | 0.026 | 0.023 | 0.023 | 0.022 | 0.031 | 0.026 | 0.023 |
| | 0.8 | 0.023 | 0.017 | 0.026 | 0.021 | 0.021 | 0.021 | 0.031 | 0.026 | 0.021 |
| | 0.85 | 0.023 | 0.015 | 0.026 | 0.019 | 0.020 | 0.019 | 0.031 | 0.026 | 0.020 |
| | 0.9 | 0.024 | 0.013 | 0.026 | 0.017 | 0.017 | 0.016 | 0.031 | 0.026 | 0.017 |
| | 0.95 | 0.023 | 0.010 | 0.026 | 0.013 | 0.013 | 0.012 | 0.031 | 0.026 | 0.013 |
| | 1.0 | 0.023 | 0.000 | 0.026 | 0.001 | 0.001 | 0.001 | 0.031 | 0.026 | 0.001 |
| Sharpe Ratio | 0.0 | 0.097 | 0.099 | 0.186 | 0.323 | 0.531 | 0.591 | 0.613 | 0.197 | 0.475 |
| | 0.05 | 0.085 | 0.089 | 0.178 | 0.296 | 0.524 | 0.572 | 0.612 | 0.186 | 0.472 |
| | 0.1 | 0.065 | 0.069 | 0.162 | 0.270 | 0.495 | 0.540 | 0.615 | 0.175 | 0.445 |
| | 0.15 | 0.089 | 0.091 | 0.181 | 0.273 | 0.487 | 0.545 | 0.609 | 0.194 | 0.447 |
| | 0.2 | 0.091 | 0.075 | 0.183 | 0.259 | 0.457 | 0.495 | 0.634 | 0.189 | 0.425 |
| | 0.25 | 0.080 | 0.070 | 0.165 | 0.257 | 0.396 | 0.439 | 0.604 | 0.173 | 0.401 |
| | 0.3 | 0.088 | 0.071 | 0.177 | 0.256 | 0.364 | 0.393 | 0.600 | 0.188 | 0.391 |
| | 0.35 | 0.100 | 0.068 | 0.192 | 0.245 | 0.327 | 0.339 | 0.618 | 0.201 | 0.367 |
| | 0.4 | 0.088 | 0.048 | 0.185 | 0.213 | 0.259 | 0.243 | 0.613 | 0.191 | 0.305 |
| | 0.45 | 0.081 | 0.036 | 0.169 | 0.197 | 0.198 | 0.176 | 0.594 | 0.178 | 0.231 |
| | 0.5 | 0.084 | 0.031 | 0.171 | 0.185 | 0.173 | 0.149 | 0.603 | 0.180 | 0.226 |
| | 0.55 | 0.084 | 0.031 | 0.169 | 0.149 | 0.135 | 0.109 | 0.601 | 0.178 | 0.193 |
| | 0.6 | 0.072 | 0.020 | 0.163 | 0.128 | 0.100 | 0.071 | 0.585 | 0.169 | 0.147 |
| | 0.65 | 0.102 | 0.020 | 0.190 | 0.129 | 0.083 | 0.054 | 0.622 | 0.195 | 0.121 |
| | 0.7 | 0.098 | 0.023 | 0.182 | 0.110 | 0.070 | 0.028 | 0.622 | 0.190 | 0.099 |
| | 0.75 | 0.089 | 0.019 | 0.173 | 0.092 | 0.059 | 0.028 | 0.626 | 0.181 | 0.100 |
| | 0.8 | 0.086 | 0.005 | 0.174 | 0.081 | 0.043 | 0.019 | 0.630 | 0.182 | 0.083 |
| | 0.85 | 0.101 | 0.009 | 0.187 | 0.059 | 0.027 | 0.012 | 0.629 | 0.197 | 0.072 |
| | 0.9 | 0.086 | 0.001 | 0.171 | 0.039 | 0.019 | 0.002 | 0.630 | 0.178 | 0.077 |
| | 0.95 | 0.103 | 0.004 | 0.194 | 0.014 | 0.013 | 0.007 | 0.639 | 0.201 | 0.110 |
| | 1.0 | 0.101 | 0.000 | 0.193 | 0.004 | 0.003 | 0.001 | 0.646 | 0.204 | 0.003 |

Proxy Goodness Test Raw Results (Continued)

| Statistic | Proxy Goodness $\rho$ | Back Fill | CL Method | Cubic Spline Interp | Kalman AR(1) No-Intcpt. | Kalman AR(2) Intercept | Kalman Non-AR Intercept | Kalman Non-Proxy | Linear Interp | MIDAS |
|---|---|---|---|---|---|---|---|---|---|---|
| Sortino Ratio | 0.0 | 0.434 | 0.302 | 0.406 | 0.507 | 0.702 | 0.827 | 0.781 | 0.496 | 0.952 |
| | 0.05 | 0.443 | 0.287 | 0.411 | 0.541 | 0.659 | 0.849 | 0.775 | 0.507 | 0.935 |
| | 0.1 | 0.444 | 0.277 | 0.401 | 0.535 | 0.654 | 0.812 | 0.716 | 0.488 | 0.923 |
| | 0.15 | 0.442 | 0.245 | 0.411 | 0.550 | 0.639 | 0.763 | 0.828 | 0.500 | 0.932 |
| | 0.2 | 0.435 | 0.239 | 0.403 | 0.556 | 0.623 | 0.759 | 0.760 | 0.500 | 0.926 |
| | 0.25 | 0.446 | 0.220 | 0.422 | 0.530 | 0.658 | 0.738 | 0.852 | 0.517 | 0.896 |
| | 0.3 | 0.428 | 0.201 | 0.395 | 0.481 | 0.598 | 0.675 | 0.775 | 0.482 | 0.794 |
| | 0.35 | 0.419 | 0.182 | 0.389 | 0.476 | 0.613 | 0.689 | 0.727 | 0.477 | 0.777 |
| | 0.4 | 0.431 | 0.183 | 0.393 | 0.466 | 0.607 | 0.682 | 0.766 | 0.488 | 0.744 |
| | 0.45 | 0.454 | 0.175 | 0.422 | 0.461 | 0.649 | 0.707 | 0.812 | 0.515 | 0.807 |
| | 0.5 | 0.419 | 0.152 | 0.393 | 0.391 | 0.549 | 0.605 | 0.762 | 0.480 | 0.652 |
| | 0.55 | 0.455 | 0.157 | 0.430 | 0.420 | 0.561 | 0.597 | 0.777 | 0.523 | 0.639 |
| | 0.6 | 0.449 | 0.136 | 0.415 | 0.386 | 0.510 | 0.537 | 0.786 | 0.510 | 0.579 |
| | 0.65 | 0.428 | 0.128 | 0.396 | 0.330 | 0.459 | 0.480 | 0.741 | 0.491 | 0.516 |
| | 0.7 | 0.440 | 0.114 | 0.413 | 0.313 | 0.410 | 0.428 | 0.841 | 0.508 | 0.443 |
| | 0.75 | 0.447 | 0.100 | 0.420 | 0.285 | 0.354 | 0.360 | 0.760 | 0.515 | 0.364 |
| | 0.8 | 0.445 | 0.100 | 0.420 | 0.238 | 0.288 | 0.291 | 0.788 | 0.512 | 0.283 |
| | 0.85 | 0.444 | 0.083 | 0.418 | 0.210 | 0.233 | 0.230 | 0.777 | 0.510 | 0.210 |
| | 0.9 | 0.458 | 0.077 | 0.438 | 0.173 | 0.169 | 0.174 | 0.763 | 0.536 | 0.130 |
| | 0.95 | 0.427 | 0.057 | 0.397 | 0.121 | 0.088 | 0.088 | 0.730 | 0.488 | 0.009 |
| | 1.0 | 0.433 | 0.000 | 0.401 | 0.007 | 0.004 | 0.007 | 0.833 | 0.492 | 0.015 |
| Volatility | 0.0 | 0.036 | 0.027 | 0.043 | 0.078 | 0.286 | 0.900 | 0.770 | 0.054 | 0.619 |
| | 0.05 | 0.036 | 0.025 | 0.043 | 0.080 | 0.266 | 0.779 | 0.851 | 0.054 | 0.545 |
| | 0.1 | 0.036 | 0.024 | 0.042 | 0.078 | 0.250 | 0.687 | 0.777 | 0.053 | 0.585 |
| | 0.15 | 0.037 | 0.023 | 0.044 | 0.077 | 0.223 | 0.627 | 0.905 | 0.055 | 0.493 |
| | 0.2 | 0.036 | 0.020 | 0.043 | 0.075 | 0.180 | 0.402 | 0.844 | 0.054 | 0.409 |
| | 0.25 | 0.036 | 0.019 | 0.043 | 0.072 | 0.161 | 0.300 | 0.843 | 0.054 | 0.365 |
| | 0.3 | 0.036 | 0.018 | 0.043 | 0.068 | 0.132 | 0.210 | 0.763 | 0.054 | 0.293 |
| | 0.35 | 0.036 | 0.016 | 0.043 | 0.063 | 0.112 | 0.149 | 0.719 | 0.054 | 0.232 |
| | 0.4 | 0.036 | 0.015 | 0.044 | 0.059 | 0.093 | 0.107 | 0.795 | 0.055 | 0.171 |
| | 0.45 | 0.036 | 0.013 | 0.043 | 0.053 | 0.079 | 0.089 | 0.766 | 0.054 | 0.130 |
| | 0.5 | 0.035 | 0.012 | 0.042 | 0.047 | 0.065 | 0.068 | 0.747 | 0.054 | 0.105 |
| | 0.55 | 0.038 | 0.012 | 0.045 | 0.044 | 0.057 | 0.057 | 0.826 | 0.056 | 0.084 |
| | 0.6 | 0.036 | 0.010 | 0.043 | 0.037 | 0.046 | 0.045 | 0.771 | 0.054 | 0.063 |
| | 0.65 | 0.037 | 0.009 | 0.044 | 0.033 | 0.039 | 0.038 | 0.702 | 0.055 | 0.053 |
| | 0.7 | 0.037 | 0.008 | 0.044 | 0.029 | 0.032 | 0.030 | 0.842 | 0.056 | 0.039 |
| | 0.75 | 0.036 | 0.007 | 0.043 | 0.024 | 0.026 | 0.024 | 0.695 | 0.054 | 0.031 |
| | 0.8 | 0.037 | 0.006 | 0.044 | 0.020 | 0.020 | 0.018 | 0.974 | 0.055 | 0.022 |
| | 0.85 | 0.037 | 0.005 | 0.044 | 0.016 | 0.015 | 0.014 | 0.732 | 0.055 | 0.016 |
| | 0.9 | 0.037 | 0.004 | 0.044 | 0.012 | 0.010 | 0.010 | 0.746 | 0.055 | 0.011 |
| | 0.95 | 0.036 | 0.003 | 0.044 | 0.007 | 0.005 | 0.005 | 0.787 | 0.054 | 0.005 |
| | 1.0 | 0.036 | 0.000 | 0.044 | 0.000 | 0.000 | 0.000 | 0.820 | 0.055 | 0.001 |

**Table A.2** – Market Tail Risk Test Raw Results

| Statistic | Market Tail Risk $\lambda$ | Back Fill | CL Method | Cubic Spline Interp | Kalman AR(1) No-Intcpt. | Kalman AR(2) Intercept | Kalman Non-AR Intercept | Kalman Non-Proxy | Linear Interp | MIDAS |
|---|---|---|---|---|---|---|---|---|---|---|
| Max Drawdown | 0.0 | 0.024 | 0.020 | 0.035 | 0.069 | 0.206 | 0.498 | 0.464 | 0.042 | 0.418 |
| | 0.05 | 0.024 | 0.019 | 0.034 | 0.069 | 0.195 | 0.396 | 0.459 | 0.041 | 0.332 |
| | 0.1 | 0.024 | 0.017 | 0.034 | 0.068 | 0.195 | 0.367 | 0.472 | 0.041 | 0.952 |
| | 0.15 | 0.026 | 0.018 | 0.036 | 0.068 | 0.170 | 0.993 | 0.468 | 0.043 | 0.316 |
| | 0.2 | 0.024 | 0.015 | 0.034 | 0.065 | 0.141 | 0.256 | 0.450 | 0.041 | 0.248 |
| | 0.25 | 0.025 | 0.014 | 0.034 | 0.065 | 0.128 | 0.185 | 0.453 | 0.041 | 0.230 |
| | 0.3 | 0.025 | 0.013 | 0.034 | 0.063 | 0.107 | 0.148 | 0.420 | 0.041 | 0.197 |
| | 0.35 | 0.025 | 0.012 | 0.035 | 0.058 | 0.092 | 0.111 | 0.384 | 0.041 | 0.156 |
| | 0.4 | 0.024 | 0.010 | 0.034 | 0.055 | 0.078 | 0.083 | 0.491 | 0.041 | 0.134 |
| | 0.45 | 0.024 | 0.009 | 0.033 | 0.047 | 0.060 | 0.063 | 0.449 | 0.040 | 0.094 |
| | 0.5 | 0.024 | 0.009 | 0.034 | 0.044 | 0.052 | 0.052 | 0.579 | 0.040 | 0.079 |
| | 0.55 | 0.026 | 0.008 | 0.035 | 0.038 | 0.043 | 0.042 | 0.452 | 0.042 | 0.064 |
| | 0.6 | 0.024 | 0.007 | 0.033 | 0.034 | 0.036 | 0.033 | 0.545 | 0.040 | 0.046 |
| | 0.65 | 0.026 | 0.006 | 0.035 | 0.029 | 0.029 | 0.026 | 1.139 | 0.041 | 0.038 |
| | 0.7 | 0.025 | 0.006 | 0.034 | 0.026 | 0.024 | 0.020 | 0.397 | 0.041 | 0.029 |
| | 0.75 | 0.024 | 0.006 | 0.033 | 0.022 | 0.020 | 0.017 | 0.431 | 0.039 | 0.023 |
| | 0.8 | 0.025 | 0.004 | 0.035 | 0.019 | 0.016 | 0.014 | 0.457 | 0.042 | 0.018 |
| | 0.85 | 0.025 | 0.004 | 0.034 | 0.014 | 0.011 | 0.009 | 1.853 | 0.041 | 0.012 |
| | 0.9 | 0.023 | 0.003 | 0.032 | 0.010 | 0.007 | 0.006 | 0.494 | 0.039 | 0.008 |
| | 0.95 | 0.026 | 0.002 | 0.036 | 0.006 | 0.005 | 0.004 | 0.586 | 0.041 | 0.006 |
| | 1.0 | 0.024 | 0.000 | 0.035 | 0.000 | 0.000 | 0.000 | 0.435 | 0.041 | 0.001 |
| Mean Return | 0.0 | 0.051 | 0.036 | 0.054 | 0.085 | 0.285 | 0.835 | 0.708 | 0.069 | 0.639 |
| | 0.05 | 0.052 | 0.034 | 0.054 | 0.090 | 0.261 | 0.778 | 0.709 | 0.070 | 0.576 |
| | 0.1 | 0.052 | 0.032 | 0.053 | 0.089 | 0.241 | 0.674 | 0.684 | 0.068 | 0.607 |
| | 0.15 | 0.052 | 0.029 | 0.055 | 0.089 | 0.222 | 0.646 | 0.758 | 0.070 | 0.524 |
| | 0.2 | 0.051 | 0.027 | 0.053 | 0.089 | 0.187 | 0.405 | 0.709 | 0.069 | 0.459 |
| | 0.25 | 0.052 | 0.025 | 0.055 | 0.083 | 0.171 | 0.305 | 0.763 | 0.070 | 0.387 |
| | 0.3 | 0.051 | 0.023 | 0.053 | 0.076 | 0.142 | 0.216 | 0.702 | 0.068 | 0.324 |
| | 0.35 | 0.050 | 0.021 | 0.053 | 0.073 | 0.127 | 0.166 | 0.686 | 0.067 | 0.281 |
| | 0.4 | 0.051 | 0.020 | 0.053 | 0.070 | 0.111 | 0.128 | 0.671 | 0.068 | 0.185 |
| | 0.45 | 0.053 | 0.019 | 0.055 | 0.065 | 0.104 | 0.117 | 0.673 | 0.070 | 0.169 |
| | 0.5 | 0.050 | 0.016 | 0.052 | 0.055 | 0.083 | 0.090 | 0.614 | 0.067 | 0.125 |
| | 0.55 | 0.054 | 0.017 | 0.057 | 0.056 | 0.078 | 0.082 | 0.750 | 0.073 | 0.108 |
| | 0.6 | 0.052 | 0.014 | 0.055 | 0.048 | 0.065 | 0.067 | 0.689 | 0.070 | 0.086 |
| | 0.65 | 0.051 | 0.013 | 0.054 | 0.041 | 0.057 | 0.057 | 0.627 | 0.069 | 0.073 |
| | 0.7 | 0.052 | 0.012 | 0.055 | 0.037 | 0.047 | 0.047 | 0.729 | 0.070 | 0.055 |
| | 0.75 | 0.053 | 0.010 | 0.055 | 0.033 | 0.039 | 0.038 | 0.633 | 0.070 | 0.044 |
| | 0.8 | 0.053 | 0.010 | 0.055 | 0.027 | 0.031 | 0.030 | 0.947 | 0.071 | 0.032 |
| | 0.85 | 0.052 | 0.008 | 0.055 | 0.023 | 0.024 | 0.023 | 0.630 | 0.070 | 0.023 |
| | 0.9 | 0.054 | 0.007 | 0.057 | 0.018 | 0.017 | 0.017 | 0.668 | 0.073 | 0.014 |
| | 0.95 | 0.051 | 0.005 | 0.053 | 0.012 | 0.008 | 0.008 | 0.693 | 0.068 | 0.002 |
| | 1.0 | 0.052 | 0.000 | 0.054 | 0.001 | 0.000 | 0.001 | 0.673 | 0.069 | 0.001 |

Market Tail Risk Test Raw Results (Continued)

| Statistic | Market Tail Risk λ | Back Fill | CL Method | Cubic Spline Interp | Kalman AR(1) No-Intcpt. | Kalman AR(2) Intercept | Kalman Non-AR Intercept | Kalman Non-Proxy | Linear Interp | MIDAS |
|---|---|---|---|---|---|---|---|---|---|---|
| RMSE | 0.0 | 0.023 | 0.025 | 0.026 | 0.031 | 0.032 | 0.031 | 0.031 | 0.026 | 0.037 |
| | 0.05 | 0.023 | 0.025 | 0.026 | 0.031 | 0.032 | 0.030 | 0.031 | 0.026 | 0.037 |
| | 0.1 | 0.023 | 0.025 | 0.026 | 0.031 | 0.032 | 0.030 | 0.031 | 0.026 | 0.036 |
| | 0.15 | 0.024 | 0.025 | 0.026 | 0.031 | 0.032 | 0.030 | 0.031 | 0.026 | 0.036 |
| | 0.2 | 0.024 | 0.024 | 0.026 | 0.031 | 0.032 | 0.030 | 0.031 | 0.026 | 0.035 |
| | 0.25 | 0.023 | 0.024 | 0.026 | 0.030 | 0.031 | 0.030 | 0.031 | 0.026 | 0.034 |
| | 0.3 | 0.024 | 0.024 | 0.026 | 0.030 | 0.031 | 0.029 | 0.031 | 0.026 | 0.033 |
| | 0.35 | 0.023 | 0.023 | 0.026 | 0.030 | 0.030 | 0.029 | 0.031 | 0.026 | 0.032 |
| | 0.4 | 0.023 | 0.023 | 0.026 | 0.029 | 0.029 | 0.028 | 0.031 | 0.026 | 0.031 |
| | 0.45 | 0.024 | 0.023 | 0.026 | 0.028 | 0.029 | 0.028 | 0.031 | 0.026 | 0.030 |
| | 0.5 | 0.023 | 0.022 | 0.026 | 0.028 | 0.028 | 0.027 | 0.031 | 0.026 | 0.029 |
| | 0.55 | 0.024 | 0.022 | 0.026 | 0.027 | 0.027 | 0.026 | 0.031 | 0.026 | 0.028 |
| | 0.6 | 0.024 | 0.021 | 0.026 | 0.026 | 0.026 | 0.025 | 0.031 | 0.026 | 0.027 |
| | 0.65 | 0.024 | 0.020 | 0.026 | 0.025 | 0.025 | 0.025 | 0.031 | 0.026 | 0.026 |
| | 0.7 | 0.024 | 0.019 | 0.026 | 0.024 | 0.024 | 0.023 | 0.031 | 0.026 | 0.025 |
| | 0.75 | 0.023 | 0.018 | 0.026 | 0.023 | 0.023 | 0.022 | 0.031 | 0.026 | 0.023 |
| | 0.8 | 0.023 | 0.017 | 0.026 | 0.021 | 0.021 | 0.021 | 0.031 | 0.026 | 0.021 |
| | 0.85 | 0.023 | 0.015 | 0.026 | 0.019 | 0.020 | 0.019 | 0.031 | 0.026 | 0.020 |
| | 0.9 | 0.024 | 0.013 | 0.026 | 0.017 | 0.017 | 0.016 | 0.031 | 0.026 | 0.017 |
| | 0.95 | 0.023 | 0.010 | 0.026 | 0.013 | 0.013 | 0.012 | 0.031 | 0.026 | 0.013 |
| | 1.0 | 0.023 | 0.000 | 0.026 | 0.001 | 0.001 | 0.001 | 0.031 | 0.026 | 0.001 |
| Sharpe Ratio | 0.0 | 0.097 | 0.099 | 0.186 | 0.323 | 0.531 | 0.591 | 0.613 | 0.197 | 0.475 |
| | 0.05 | 0.085 | 0.089 | 0.178 | 0.296 | 0.524 | 0.572 | 0.612 | 0.186 | 0.472 |
| | 0.1 | 0.065 | 0.069 | 0.162 | 0.270 | 0.495 | 0.540 | 0.615 | 0.175 | 0.445 |
| | 0.15 | 0.089 | 0.091 | 0.181 | 0.273 | 0.487 | 0.545 | 0.609 | 0.194 | 0.447 |
| | 0.2 | 0.091 | 0.075 | 0.183 | 0.259 | 0.457 | 0.495 | 0.634 | 0.189 | 0.425 |
| | 0.25 | 0.080 | 0.070 | 0.165 | 0.257 | 0.396 | 0.439 | 0.604 | 0.173 | 0.401 |
| | 0.3 | 0.088 | 0.071 | 0.177 | 0.256 | 0.364 | 0.393 | 0.600 | 0.188 | 0.391 |
| | 0.35 | 0.100 | 0.068 | 0.192 | 0.245 | 0.327 | 0.339 | 0.618 | 0.201 | 0.367 |
| | 0.4 | 0.088 | 0.048 | 0.185 | 0.213 | 0.259 | 0.243 | 0.613 | 0.191 | 0.305 |
| | 0.45 | 0.081 | 0.036 | 0.169 | 0.197 | 0.198 | 0.176 | 0.594 | 0.178 | 0.231 |
| | 0.5 | 0.084 | 0.031 | 0.171 | 0.185 | 0.173 | 0.149 | 0.603 | 0.180 | 0.226 |
| | 0.55 | 0.084 | 0.031 | 0.169 | 0.149 | 0.135 | 0.109 | 0.601 | 0.178 | 0.193 |
| | 0.6 | 0.072 | 0.020 | 0.163 | 0.128 | 0.100 | 0.071 | 0.585 | 0.169 | 0.147 |
| | 0.65 | 0.102 | 0.020 | 0.190 | 0.129 | 0.083 | 0.054 | 0.622 | 0.195 | 0.121 |
| | 0.7 | 0.098 | 0.023 | 0.182 | 0.110 | 0.070 | 0.028 | 0.622 | 0.190 | 0.099 |
| | 0.75 | 0.089 | 0.019 | 0.173 | 0.092 | 0.059 | 0.028 | 0.626 | 0.181 | 0.100 |
| | 0.8 | 0.086 | 0.005 | 0.174 | 0.081 | 0.043 | 0.019 | 0.630 | 0.182 | 0.083 |
| | 0.85 | 0.101 | 0.009 | 0.187 | 0.059 | 0.027 | 0.012 | 0.629 | 0.197 | 0.072 |
| | 0.9 | 0.086 | 0.001 | 0.171 | 0.039 | 0.019 | 0.002 | 0.630 | 0.178 | 0.077 |
| | 0.95 | 0.103 | 0.004 | 0.194 | 0.014 | 0.013 | 0.007 | 0.639 | 0.201 | 0.110 |
| | 1.0 | 0.101 | 0.000 | 0.193 | 0.004 | 0.003 | 0.001 | 0.646 | 0.204 | 0.003 |

Market Tail Risk Test Raw Results (Continued)

| Statistic | Market Tail Risk $\lambda$ | Back Fill | CL Method | Cubic Spline Interp | Kalman AR(1) No-Intcpt. | Kalman AR(2) Intercept | Kalman Non-AR Intercept | Kalman Non-Proxy | Linear Interp | MIDAS |
|---|---|---|---|---|---|---|---|---|---|---|
| Sortino Ratio | 0.0 | 0.434 | 0.302 | 0.406 | 0.507 | 0.702 | 0.827 | 0.781 | 0.496 | 0.952 |
| | 0.05 | 0.443 | 0.287 | 0.411 | 0.541 | 0.659 | 0.849 | 0.775 | 0.507 | 0.935 |
| | 0.1 | 0.444 | 0.277 | 0.401 | 0.535 | 0.654 | 0.812 | 0.716 | 0.488 | 0.923 |
| | 0.15 | 0.442 | 0.245 | 0.411 | 0.550 | 0.639 | 0.763 | 0.828 | 0.500 | 0.932 |
| | 0.2 | 0.435 | 0.239 | 0.403 | 0.556 | 0.623 | 0.759 | 0.760 | 0.500 | 0.926 |
| | 0.25 | 0.446 | 0.220 | 0.422 | 0.530 | 0.658 | 0.738 | 0.852 | 0.517 | 0.896 |
| | 0.3 | 0.428 | 0.201 | 0.395 | 0.481 | 0.598 | 0.675 | 0.775 | 0.482 | 0.794 |
| | 0.35 | 0.419 | 0.182 | 0.389 | 0.476 | 0.613 | 0.689 | 0.727 | 0.477 | 0.777 |
| | 0.4 | 0.431 | 0.183 | 0.393 | 0.466 | 0.607 | 0.682 | 0.766 | 0.488 | 0.744 |
| | 0.45 | 0.454 | 0.175 | 0.422 | 0.461 | 0.649 | 0.707 | 0.812 | 0.515 | 0.807 |
| | 0.5 | 0.419 | 0.152 | 0.393 | 0.391 | 0.549 | 0.605 | 0.762 | 0.480 | 0.652 |
| | 0.55 | 0.455 | 0.157 | 0.430 | 0.420 | 0.561 | 0.597 | 0.777 | 0.523 | 0.639 |
| | 0.6 | 0.449 | 0.136 | 0.415 | 0.386 | 0.510 | 0.537 | 0.786 | 0.510 | 0.579 |
| | 0.65 | 0.428 | 0.128 | 0.396 | 0.330 | 0.459 | 0.480 | 0.741 | 0.491 | 0.516 |
| | 0.7 | 0.440 | 0.114 | 0.413 | 0.313 | 0.410 | 0.428 | 0.841 | 0.508 | 0.443 |
| | 0.75 | 0.447 | 0.100 | 0.420 | 0.285 | 0.354 | 0.360 | 0.760 | 0.515 | 0.364 |
| | 0.8 | 0.445 | 0.100 | 0.420 | 0.238 | 0.288 | 0.291 | 0.788 | 0.512 | 0.283 |
| | 0.85 | 0.444 | 0.083 | 0.418 | 0.210 | 0.233 | 0.230 | 0.777 | 0.510 | 0.210 |
| | 0.9 | 0.458 | 0.077 | 0.438 | 0.173 | 0.169 | 0.174 | 0.763 | 0.536 | 0.130 |
| | 0.95 | 0.427 | 0.057 | 0.397 | 0.121 | 0.088 | 0.088 | 0.730 | 0.488 | 0.009 |
| | 1.0 | 0.433 | 0.000 | 0.401 | 0.007 | 0.004 | 0.007 | 0.833 | 0.492 | 0.015 |
| Volatility | 0.0 | 0.036 | 0.027 | 0.043 | 0.078 | 0.286 | 0.900 | 0.770 | 0.054 | 0.619 |
| | 0.05 | 0.036 | 0.025 | 0.043 | 0.080 | 0.266 | 0.779 | 0.851 | 0.054 | 0.545 |
| | 0.1 | 0.036 | 0.024 | 0.042 | 0.078 | 0.250 | 0.687 | 0.777 | 0.053 | 0.585 |
| | 0.15 | 0.037 | 0.023 | 0.044 | 0.077 | 0.223 | 0.627 | 0.905 | 0.055 | 0.493 |
| | 0.2 | 0.036 | 0.020 | 0.043 | 0.075 | 0.180 | 0.402 | 0.844 | 0.054 | 0.409 |
| | 0.25 | 0.036 | 0.019 | 0.043 | 0.072 | 0.161 | 0.300 | 0.843 | 0.054 | 0.365 |
| | 0.3 | 0.036 | 0.018 | 0.043 | 0.068 | 0.132 | 0.210 | 0.763 | 0.054 | 0.293 |
| | 0.35 | 0.036 | 0.016 | 0.043 | 0.063 | 0.112 | 0.149 | 0.719 | 0.054 | 0.232 |
| | 0.4 | 0.036 | 0.015 | 0.044 | 0.059 | 0.093 | 0.107 | 0.795 | 0.055 | 0.171 |
| | 0.45 | 0.036 | 0.013 | 0.043 | 0.053 | 0.079 | 0.089 | 0.766 | 0.054 | 0.130 |
| | 0.5 | 0.035 | 0.012 | 0.042 | 0.047 | 0.065 | 0.068 | 0.747 | 0.054 | 0.105 |
| | 0.55 | 0.038 | 0.012 | 0.045 | 0.044 | 0.057 | 0.057 | 0.826 | 0.056 | 0.084 |
| | 0.6 | 0.036 | 0.010 | 0.043 | 0.037 | 0.046 | 0.045 | 0.771 | 0.054 | 0.063 |
| | 0.65 | 0.037 | 0.009 | 0.044 | 0.033 | 0.039 | 0.038 | 0.702 | 0.055 | 0.053 |
| | 0.7 | 0.037 | 0.008 | 0.044 | 0.029 | 0.032 | 0.030 | 0.842 | 0.056 | 0.039 |
| | 0.75 | 0.036 | 0.007 | 0.043 | 0.024 | 0.026 | 0.024 | 0.695 | 0.054 | 0.031 |
| | 0.8 | 0.037 | 0.006 | 0.044 | 0.020 | 0.020 | 0.018 | 0.974 | 0.055 | 0.022 |
| | 0.85 | 0.037 | 0.005 | 0.044 | 0.016 | 0.015 | 0.014 | 0.732 | 0.055 | 0.016 |
| | 0.9 | 0.037 | 0.004 | 0.044 | 0.012 | 0.010 | 0.010 | 0.746 | 0.055 | 0.011 |
| | 0.95 | 0.036 | 0.003 | 0.044 | 0.007 | 0.005 | 0.005 | 0.787 | 0.054 | 0.005 |
| | 1.0 | 0.036 | 0.000 | 0.044 | 0.000 | 0.000 | 0.000 | 0.820 | 0.055 | 0.001 |

**Table A.3** – Data Autocorrelation Test Raw Results

| Statistic | Hurst Index $H$ | Back Fill | CL Method | Cubic Spline Interp | Kalman AR(1) No-Intcpt. | Kalman AR(2) Intercept | Kalman Non-AR Intercept | Kalman Non-Proxy | Linear Interp | MIDAS |
|---|---|---|---|---|---|---|---|---|---|---|
| Max Drawdown | 0.5 | 0.028 | 0.010 | 0.037 | 0.028 | 0.036 | 0.038 | 0.719 | 0.045 | 0.057 |
| | 0.55 | 0.026 | 0.007 | 0.035 | 0.030 | 0.035 | 0.035 | 5.498 | 0.043 | 0.053 |
| | 0.6 | 0.025 | 0.008 | 0.035 | 0.036 | 0.037 | 0.034 | 0.473 | 0.042 | 0.050 |
| | 0.65 | 0.023 | 0.007 | 0.033 | 0.037 | 0.033 | 0.029 | 0.327 | 0.039 | 0.042 |
| | 0.7 | 0.021 | 0.005 | 0.030 | 0.039 | 0.029 | 0.024 | 0.302 | 0.036 | 0.036 |
| | 0.75 | 0.017 | 0.002 | 0.027 | 0.040 | 0.027 | 0.020 | 0.220 | 0.032 | 0.031 |
| | 0.8 | 0.013 | 0.002 | 0.021 | 0.034 | 0.023 | 0.015 | 0.140 | 0.025 | 0.026 |
| | 0.85 | 0.009 | 0.001 | 0.016 | 0.029 | 0.017 | 0.010 | 0.103 | 0.018 | 0.018 |
| | 0.9 | 0.007 | 0.000 | 0.012 | 0.026 | 0.013 | 0.007 | 0.070 | 0.014 | 0.014 |
| | 0.95 | 0.004 | 0.001 | 0.006 | 0.011 | 0.005 | 0.002 | 0.031 | 0.008 | 0.011 |
| Mean Return | 0.5 | 0.059 | 0.020 | 0.063 | 0.048 | 0.070 | 0.080 | 0.813 | 0.080 | 0.105 |
| | 0.55 | 0.058 | 0.018 | 0.061 | 0.048 | 0.069 | 0.076 | 0.879 | 0.077 | 0.099 |
| | 0.6 | 0.052 | 0.015 | 0.054 | 0.047 | 0.066 | 0.066 | 0.624 | 0.068 | 0.089 |
| | 0.65 | 0.050 | 0.011 | 0.052 | 0.045 | 0.065 | 0.062 | 0.521 | 0.067 | 0.081 |
| | 0.7 | 0.047 | 0.007 | 0.049 | 0.043 | 0.060 | 0.055 | 0.435 | 0.063 | 0.071 |
| | 0.75 | 0.044 | 0.003 | 0.045 | 0.041 | 0.059 | 0.051 | 0.317 | 0.059 | 0.070 |
| | 0.8 | 0.054 | 0.004 | 0.058 | 0.043 | 0.067 | 0.054 | 0.294 | 0.074 | 0.077 |
| | 0.85 | 0.055 | 0.011 | 0.061 | 0.049 | 0.079 | 0.058 | 0.284 | 0.079 | 0.084 |
| | 0.9 | 0.062 | 0.026 | 0.068 | 0.049 | 0.088 | 0.058 | 0.286 | 0.088 | 0.090 |
| | 0.95 | 0.080 | 0.059 | 0.092 | 0.055 | 0.128 | 0.078 | 0.414 | 0.116 | 0.125 |
| RMSE | 0.5 | 0.025 | 0.021 | 0.028 | 0.026 | 0.027 | 0.026 | 0.032 | 0.027 | 0.027 |
| | 0.55 | 0.024 | 0.021 | 0.027 | 0.026 | 0.027 | 0.025 | 0.031 | 0.026 | 0.027 |
| | 0.6 | 0.023 | 0.021 | 0.026 | 0.026 | 0.026 | 0.025 | 0.031 | 0.026 | 0.027 |
| | 0.65 | 0.023 | 0.021 | 0.025 | 0.026 | 0.026 | 0.025 | 0.030 | 0.025 | 0.027 |
| | 0.7 | 0.021 | 0.020 | 0.024 | 0.025 | 0.025 | 0.025 | 0.030 | 0.024 | 0.027 |
| | 0.75 | 0.020 | 0.020 | 0.023 | 0.025 | 0.025 | 0.025 | 0.029 | 0.022 | 0.027 |
| | 0.8 | 0.019 | 0.020 | 0.021 | 0.024 | 0.024 | 0.025 | 0.027 | 0.021 | 0.028 |
| | 0.85 | 0.017 | 0.019 | 0.019 | 0.022 | 0.022 | 0.024 | 0.026 | 0.019 | 0.028 |
| | 0.9 | 0.014 | 0.019 | 0.016 | 0.020 | 0.020 | 0.022 | 0.023 | 0.016 | 0.029 |
| | 0.95 | 0.011 | 0.018 | 0.013 | 0.017 | 0.016 | 0.018 | 0.019 | 0.013 | 0.032 |
| Sharpe Ratio | 0.5 | 0.104 | 0.044 | 0.178 | 0.136 | 0.133 | 0.111 | 0.559 | 0.185 | 0.192 |
| | 0.55 | 0.088 | 0.025 | 0.165 | 0.134 | 0.118 | 0.095 | 0.574 | 0.179 | 0.164 |
| | 0.6 | 0.075 | 0.022 | 0.165 | 0.148 | 0.107 | 0.083 | 0.600 | 0.175 | 0.137 |
| | 0.65 | 0.090 | 0.034 | 0.190 | 0.175 | 0.102 | 0.064 | 0.666 | 0.198 | 0.136 |
| | 0.7 | 0.092 | 0.038 | 0.192 | 0.193 | 0.113 | 0.059 | 0.676 | 0.202 | 0.139 |
| | 0.75 | 0.093 | 0.041 | 0.210 | 0.251 | 0.126 | 0.063 | 0.662 | 0.226 | 0.156 |
| | 0.8 | 0.111 | 0.079 | 0.224 | 0.288 | 0.154 | 0.081 | 0.791 | 0.248 | 0.206 |
| | 0.85 | 0.115 | 0.093 | 0.239 | 0.296 | 0.170 | 0.068 | 0.824 | 0.271 | 0.246 |
| | 0.9 | 0.192 | 0.178 | 0.336 | 0.342 | 0.196 | 0.085 | 0.893 | 0.374 | 0.314 |
| | 0.95 | 0.332 | 0.388 | 0.511 | 0.399 | 0.246 | 0.017 | 1.218 | 0.603 | 0.459 |

Data Autocorrelation Test Raw Results (Continued)

| Statistic | Hurst Index $H$ | Back Fill | CL Method | Cubic Spline Interp | Kalman AR(1) No-Intcpt. | Kalman AR(2) Intercept | Kalman Non-AR Intercept | Kalman Non-Proxy | Linear Interp | MIDAS |
|---|---|---|---|---|---|---|---|---|---|---|
| Sortino Ratio | 0.5 | 0.441 | 0.172 | 0.421 | 0.342 | 0.469 | 0.545 | 0.623 | 0.507 | 0.559 |
| | 0.55 | 0.464 | 0.169 | 0.445 | 0.367 | 0.498 | 0.558 | 0.838 | 0.529 | 0.598 |
| | 0.6 | 0.438 | 0.139 | 0.403 | 0.362 | 0.505 | 0.522 | 0.705 | 0.488 | 0.587 |
| | 0.65 | 0.460 | 0.107 | 0.421 | 0.372 | 0.545 | 0.546 | 0.719 | 0.524 | 0.604 |
| | 0.7 | 0.472 | 0.064 | 0.444 | 0.391 | 0.575 | 0.562 | 0.858 | 0.554 | 0.622 |
| | 0.75 | 0.503 | 0.012 | 0.451 | 0.377 | 0.610 | 0.565 | 0.998 | 0.571 | 0.656 |
| | 0.8 | 0.774 | 0.142 | 0.742 | 0.498 | 0.904 | 0.785 | 1.297 | 0.909 | 0.943 |
| | 0.85 | 1.136 | 0.412 | 1.121 | 0.793 | 1.536 | 1.221 | $\infty$ | 1.383 | 1.393 |
| | 0.9 | 3.130 | 1.780 | 2.237 | 1.111 | 5.685 | 2.328 | $\infty$ | 3.151 | 2.304 |
| | 0.95 | $\infty$ | $\infty$ | $\infty$ | $\infty$ | $\infty$ | $\infty$ | $\infty$ | $\infty$ | $\infty$ |
| Volatility | 0.5 | 0.046 | 0.015 | 0.055 | 0.040 | 0.056 | 0.062 | 0.968 | 0.069 | 0.090 |
| | 0.55 | 0.042 | 0.013 | 0.049 | 0.039 | 0.051 | 0.054 | 0.960 | 0.062 | 0.078 |
| | 0.6 | 0.036 | 0.010 | 0.043 | 0.038 | 0.047 | 0.046 | 0.818 | 0.054 | 0.066 |
| | 0.65 | 0.033 | 0.008 | 0.039 | 0.036 | 0.042 | 0.038 | 0.548 | 0.049 | 0.055 |
| | 0.7 | 0.028 | 0.005 | 0.034 | 0.033 | 0.035 | 0.030 | 0.414 | 0.042 | 0.043 |
| | 0.75 | 0.024 | 0.003 | 0.029 | 0.031 | 0.032 | 0.025 | 0.278 | 0.037 | 0.038 |
| | 0.8 | 0.023 | 0.001 | 0.028 | 0.026 | 0.029 | 0.021 | 0.197 | 0.035 | 0.034 |
| | 0.85 | 0.018 | 0.002 | 0.024 | 0.023 | 0.026 | 0.017 | 0.152 | 0.030 | 0.030 |
| | 0.9 | 0.018 | 0.004 | 0.022 | 0.020 | 0.024 | 0.014 | 0.121 | 0.028 | 0.027 |
| | 0.95 | 0.018 | 0.006 | 0.023 | 0.016 | 0.023 | 0.011 | 0.110 | 0.029 | 0.026 |