Cross-Market Signals

Economic Spillovers Across Markets

a thesis presented
by
Hamzeh Baseem Hamdan
to
The Statistics and Computer Science Departments

in partial fulfillment of the requirements
for the joint degree of
Bachelor of Arts (Honors)
in the subjects of
Statistics and Computer Science

Harvard University
Cambridge, Massachusetts
May 2025

Thesis advisor: Emil Siriwardane Hamzeh Baseem Hamdan

Cross-Market Signals

Abstract

This thesis explores the growing interconnectedness of global markets, with a focus on the explanatory power U.S. and China macroeconomic factors have on each other's stock market returns. Using the JKP Global Factors Dataset with 13 themes and 153 factors, we model this relationship using linear regressions, sparse additive models, and kernel regressions. The analyses were conducted using the Wilshire 5000 and Shanghai Composite returns.

Across all models, U.S. economic data consistently improved the predictions of Chinese market returns, but Chinese economic data rarely added improved the predictions of U.S. market returns. Linear regressions revealed decent R^2 values using domestic data, with $0.70 \leq R^2 \leq 0.80$ for the U.S. market and $0.55 \leq R^2 \leq 0.60$ for the Chinese market. Sparse additive models and kernel regressions achieved higher R^2 values in the data. They were more likely to overfit to the data, with the partial dependence plots sometimes not following economic intuition. Future directions include accounting for interaction effects, attempting rolling-window methods, and exploring sector-level analyses to obtain more granular insights into market spillovers.

Contents

List of Tables

Listing of figures

This thesis is dedicated to my twin Ahmed;
I love you, Hamoode.

Acknowledgments

Thank you to everyone who has supported me during the process of researching and writing my thesis, and to everyone who has played a role in my college journey. I'm blessed to have had such a wonderful four years at Harvard, and feel fulfilled by this experience. I would like to thank Professor Emil Siriwardane for introducing the topic of replicating portfolios to me, and for inspiring and supporting this journey of intellectual curiosity.

Specifically, I would like to thank Samantha Wilhoit, who has been integral to my Harvard experience. Thank you for always pushing me to do better and for helping me do the things I set my mind on—from planning to execution. I couldn't have written this without you. I would like to thank Chinomso Okafor, who has for several years been a great role model and an even better friend, for his tremendous support throughout this process.

I would like to thank my family, who has always pushed me in my intellectual pursuits, for their support and encouragement throughout the duration of this project.

Lastly, thank you to everyone who has supported me and has rooted for me; I'm incredibly lucky to have such an amazing support system, and I appreciate every one of you.

What we have seen in financial markets should bring home to us all that the central organizing principle of this 21st century is interdependence.

Kevin Rudd

1 Introduction

In the era of globalization, financial markets have become increasingly interconnected. Economies are linked by multinational companies and global supply chains. Economic conditions in one region of the world typically also influence others around the world. Financial shocks in one country can quickly impact markets in other countries, increasing the complexity of markets.

As global interconnectedness increases, understanding the mechanisms by which national markets are connected is an important task for investors, policymakers, and financial researchers. Investors wanting to diversity their portfolio may question to what degree markets can be independent of each other. Policymakers must understand the impact of international regulations on the markets. Financial researchers attempting to understand markets must now

consider links with other international markets, increasing the complexity of the research.

1.1 Background Information

The 1980s and 1990s were fueled by market liberalization, in which many countries relaxed capital regulations and allowed money to flow more easily across borders.[2] This caused investors to start investing in foreign stocks and bonds, leading to a period of rapid growth in international equity and debt markets.[2] In the 1970s, the annual foreign direct investment by multinational corporations was about $13 billion; by 2023, it had increased to $1.37 trillion. The result of this era was the normalization of global portfolios and significantly more interconnected financial markets.

In the resulting markets, price movements, trends, and financial crises in one country's market can quickly advance to other markets. This is seen at a large scale in cases as early as the 1997 Asian Financial Crisis, during which a currency crisis in Thailand led to steep market declines in Malaysia, Indonesia, and South Korea and led to spillover effects in Russia, Latin America, and Eastern Europe.[12] The 2008 Financial Crisis began as the collapse of the U.S. housing bubble and evolved into the worst global recession in decades, leading to the major financial institutions across Europe requiring government bailouts within a year.[12] In 2015, a sharp selloff in the Chinese market resulted in the Dow Jones falling by over 1,000 points at the next trade open.[12] Most recently with the COVID-19 Pandemic in early 2020, uncertainty and fear by the pandemic sparked market crashes all around the world.[12]

More importantly, studies have shown that spillovers between the U.S. market and other regions increased after these global crises. The markets are so interconnected that they cannot live in isolation, and

their interconnectedness will only keep growing.

The structure of this interdependence is an important field of study in international finance, typically focusing on links between two specific countries. Much of the literature has focused on measuring the integration of markets through correlations, co-movements, and global factors. We do not fully understand how information from one market can be used to statistically replicate the return structure of another.

These strategies are especially important in the case of the U.S. and Chinese equity markets, which make up over 60% of the global stock market capitalization. Historically, U.S. market signals influenced markets in Europe and Asia, with China remaining isolated and not impacted by foreign changes. This has changed dramatically over the past decade as China has now integrated itself in the global markets. China allowed international investors to trade certain Chinese A-shares for the first time in 2014.[8] More than a decade later, U.S. investors hold increasing stakes in Chinese companies and Chinese firms are listed on American exchanges, demonstrating an interconnectedness that impacts both equity markets.

This thesis addresses the information gap by investigating the extent to which U.S. and Chinese macroeconomic factors and market-based signals can be used to replicate returns in the U.S. and Chinese equity markets. As a measure of economic co-integration, this thesis employs methods to estimate how much additional explanatory power is gained by adding one country's economic factors to a model that estimates the other country's market returns using it's own economic factors. If the additional set of factors add a lot of explanatory power, then there must be some level of market spillovers, with spillovers in both directions implying co-integration. This analysis is conducted using linear models with regularization, generalized additive models, and kernel regression. The investigations in this thesis will also include data from 13 themes and over 150

factors representing the U.S. and Chinese economies, with some measurements dating back to the 1920s. The data is also available at daily and monthly granularity, with 1,164 rows in the monthly dataset and nearly 3 million rows in the daily dataset.

1.2 Literature Review

1.2.1 Foundational Asset Pricing Theories

Introduced by Harry Markowitz in 1952, the Markowitz model was a revolutionary portfolio optimization model that maximized expected return at set levels of risk, typically measured by standard deviation.[11] The model showed that by diversifying a portfolio, investors can take advantage of a fronteir of optimal risk-return trade-offs. This model was so impactful that it started an entire new field: Modern Portfolio Theory.

The Capital Asset Pricing Model was introduced independently by Treynor, Sharpe, Lintner, and Mossin between 1961 and 1966.[14] It extended the pricing model by separating systematic risk from idiosyncratic risk. In 1964 and 1965, Sharpe and Lintner showed that the market portfolio, a portfolio in which each asset is weighted proportionally to its presence in the market, is mean-variance efficient under a few assumptions.[14] This implied that all investors should hold a combination of the market and a risk-free asset.

However, the model did not hold empirically, and in 1976, Ross developed the Arbitrage Pricing Theory, which generalized this from a singular market risk factor to multiple risk factors.[18] The Arbitrage Pricing Theory does not specify the factors a priori and assumes that the relationship between expected returns and factor exposures is linear.

In the 1990s, Fama and French showed that a couple systematic

factors can help explain movements in equity returns better than any single market beta.[3] The factors were initially size and value, though they added market in their three-factor model. Carhart introduced momentum as a fourth factor in 1997, and Fama and French proposed a five-factor model in 2015 with profitability and investment.[1, 4] This has sparked an interest in using factors to understand equity market returns.

1.2.2 Global Market Integration

Whether or not a set of economic factors from one country can replicate the returns depends crucially on the degree to which the markets are globally integrated. If they are highly integrated, they can be impacted by the same global factors. Otherwise, the markets might be more impacted by local idiosyncratic factors.

In 1995, Bekaert and Harvey found time-varying world market integration among the 12 emerging markets they examined, many of which exhibited only partial integration.[5] They noted that integration is expected to increase with financial liberalization. During the late 20th century, developed markets in the West were fairly integrated and emerging markets in Asia and Latin America were more segmented.

Pukthuanthong and Roll showed in 2009 that correlation alone is not a valid measure of integration, arguing that markets that are perfectly integrated can have low correlations if they are impacted differently by global factors.[10] They proposed using measure of the variance explained by models using the factors as predictors to predict the returns of the market—they specifically suggested R^2 in multi-factor models.[10]

It's important to note that China had a largely isolationist policy until 2014, historically operating with tighter capital controls. As

such, China was for a long time partially segmented from the world, and its markets were no different. Analyses done in the 2000s revealed little integration with the global economy.

1.2.3 Cross-Market Return Replication as a Measure of Integration

An emerging approach in the literature is to assess market integration by testing how well one market's return can be replicated using factors from another market. If two markets are integrated, then the economic factors that drive one market should explain some variability in the other.

An attempt to explain Chinese A-share portfolio returns using U.S. risk factors for 1993-2006 data by Brooks et al. (2010) found that the Chinese market was almost completely segmented, with U.S. factors adding no explanatory power over the Chinese factors.[17] As China's markets liberalized, replication studies found stronger links between the U.S. and Chinese markets. Goh et al. (2013) found no link between the countries before China entered the World Trade Organization in 2001 and a strong link after, characterized by a set of U.S. economic factors showing significant predictive power for Chinese stock returns.[7] This trend is increasing, with recent literature showing that the Chinese market becoming so integral to the global economy that its lagged returns can predict returns in numerous other markets worldwide. Chinese market integration is expected to continue to grow, extending trends from the past decade.

1.2.4 Comparison with Traditional Integration Metrics

Classical approaches for market integration metrics have relied on correlational studies bewteen the markets, cointegration testing for long-term relationships, and comparing factor exposures in global

CAPM or mutli-factor models.

Correlation-Based Measures

Although Pukthuanthong and Roll showed that correlation is not a valid measure of integration, Billio et al. found that many advanced integration measures yielded the same patterns as rolling correlations.[13] Regardless of effectiveness, correlation-based measures only analyze co-movement strength and do not dive into the underlying drivers behind the movements of each market. Correlational studies also assume linear and uniform relationships between the markets.

Cointegration and Long-Run Parity Tests

Another common approach is to test whether markets are cointegrated, typically assessed by the Johansen test. For two markets to be co-integrated, their divergences must revert to the mean. Though cointegration studies between China and the U.S. show cointegration only after China's market started liberalizing, cointegration is often a binary indicator with no value representing the strength of the relationship.

Shared Factor Exposure Models

The most recent methodology is to replicate each market using a subset of global factors and compare the two sets of factors that explain the most variability in the returns. If the same global CAPM or multi-factor model can represent both markets without needing country-specific data, integration between the markets is implied. The Fama-French models are a popular attempt to do this with preset factors, often comprising of three or five factors. These are often

limiting as they are a static set of factors that are limited in complexity.

1.2.5 Advantages of the Return Replication Approach

Returns replication has several advantages over these methods. First, it analyzes the underlying effects of the integration and determines the relative strength of different factors. Second, the lack of a standardized set of factors allows for a more robust measurement that is more capable of capturing the true market relationship. The results naturally follow a practical economic meaning. Lastly, this replication approach is capable of taking advantage of modern econometric and machine learning techniques like linear models with regularization, generalized additive models, and kernel regressions.

2

Statistical Framework

2.1 Theoretical Framework and Statistical Formulation

2.1.1 General Model

A primitive approach to the portfolio replication problem is to re-frame it as an optimization problem in which asset weights are chosen to minimize a loss function that quantifies the error of the constructed portfolio. While a more primitive approach, linear models have a deep root in the statistical applications in finance.

Probabilistic Setup

Let $(\Omega, \mathcal{F}, P)$ be the probability space representing the financial markets. Define the following two random variables on this space: R_t^M representing the target market return and R_t^P representing the replicating portfolio return at time t. Assume that both R_t^M and R_t^P are well-defined distributions with finite mean and variance and that they are mean-centered.

Let $\mu_M = \mathbb{E}[R_t^M] = 0$ and $\mu_P = \mathbb{E}[R_t^P] = 0$ for their expected values, $\sigma_M^2 = \text{Var}(R_t^M)$ and $\sigma_P^2 = \text{Var}(R_t^P)$ for their variances, and $\text{Cov}(R_t^M, R_t^P)$ for their covariance.

Portfolio Replication Problem

The portfolio replication problem is one of constructing a *replicating portfolio* with returns that closely match the target market's returns distribution. Suppose we have a vector of N asset returns or set of factors $\vec{R}_t = (R_{1,t}, R_{2,t}, \ldots, R_{N,t})^\top$ that serve as building blocks; note that these are assumed to be linearly independent and mean-centered.

A replicating portfolio is defined by a weight vector $\vec{w} = (w_1, \ldots, w_N)^\top$ such that the returns of the replicating portfolio are $R_t^P = \vec{w}^\top \vec{R}_t = \sum_{i=1}^N w_i R_{i,t}$. There are several constraints that can be placed on the weight vector: (1) for a fully invested portfolio, the weights must sum to 1, (2) if short-selling is prohibited, the weights must all be non-negative.

Optimal Linear Replication

We can think of the portfolio replication problem as a stochastic optimization problem, in which the objective function is the mean squared tracking error. More formally, we must find $\vec{w}^*$ that minimizes $\mathbb{E}\left[\left(R_t^M - \vec{w}^\top \vec{R}_t\right)^2\right]$ subject to any portfolio constraints.

Imposing the expectation matching constraint to ensure that the replicating portfolio is unbiased, we have:

$$\mathbb{E}[R_t^P] = \mathbb{E}[R_t^M]$$

Since $R_t^P = \vec{w}^\top \vec{R}_t$, we have:

$$\vec{w}^\top \mathbb{E}[\vec{R}_t] = \mu_M$$

This is naturally satisfied under the assumption that asset returns are mean-centered.

Thus, we select:

$$\vec{w}^* = \arg\min_{\vec{w}} \mathbb{E}\left[\left(R_t^M - \vec{w}^\top \vec{R}_t\right)^2\right]$$

Expanding, we get:

$$\vec{w}^* = \arg\min_{\vec{w}} \mathbb{E}\left[\left(R_t^M\right)^2 - 2R_t^M \vec{w}^\top \vec{R}_t + \left(\vec{w}^\top \vec{R}_t\right)^2\right]$$

Applying linearity of expectation:

$$\vec{w}^* = \arg\min_{\vec{w}} \left\{\mathbb{E}\left[(R_t^M)^2\right] - 2\vec{w}^\top \mathbb{E}[R_t^M \vec{R}_t] + \vec{w}^\top \mathbb{E}[\vec{R}_t \vec{R}_t^\top]\vec{w}\right\}$$

Because $\mathbb{E}\left[(R_t^M)^2\right]$ does not depend on $\vec{w}$, it can be ignored when finding the minimum.

Let $\vec{\mu} = \mathbb{E}[\vec{R}_t]$ be the mean vector of basis returns; note that because $\vec{R}_t$ is mean-centered, this is a zero vector. We can calculate

the covariance as:

$$\begin{aligned}\Sigma &= \mathrm{Cov}(\vec{R}_t, \vec{R}_t) \\ &= \mathbb{E}[(\vec{R}_t - \vec{\mu})(\vec{R}_t - \vec{\mu})^\top] \\ &= \mathbb{E}[\vec{R}_t \vec{R}_t^\top]\end{aligned}$$

The covariance vector between basis returns and target returns is given by:

$$\begin{aligned}\vec{\sigma}_M &= \mathrm{Cov}(R_t^M, \vec{R}_t) \\ &= \mathbb{E}[(R_t^M - \mu_M)(\vec{R}_t - \vec{\mu})] \\ &= \mathbb{E}[R_t^M \vec{R}_t] - \mu_M \mathbb{E}[\vec{R}_t] - \mathbb{E}[R_t^M]\vec{\mu} + \mu_M \vec{\mu} \\ &= \mathbb{E}[R_t^M \vec{R}_t] \qquad\qquad (\ \vec{\mu} = \mathbb{E}[\vec{R}_t] = \vec{0}\ )\end{aligned}$$

Thus, the minimization can be simplified to:

$$\vec{w}^* = \arg\min_{\vec{w}} \left\{ -2\vec{w}^\top \vec{\sigma}_M + \vec{w}^\top \Sigma \vec{w} \right\}$$

Taking the derivative with respect to $\vec{w}$ and setting the derivative as zero, we get:

$$-2\vec{\sigma}_M + 2\Sigma\vec{w} = 0$$

Implying:

$$\Sigma\vec{w} = \vec{\sigma}_M$$

Assuming Σ is invertible, the optimal vector is given by:

$$\vec{w}^* = \Sigma^{-1}\vec{\sigma}_M$$

We can verify this solution is a minimum and not a maximum by taking the second derivative with respect to $\vec{w}$ we get the Hessian matrix $H = 2\Sigma$, which is independent of $\vec{w}$. If Σ is positive definite,

then H will also be positive definite. The covariance matrix is symmetric and positive semidefinite by definition, and is positive definite if and only if the returns are linearly independent; this follows from our assumption that Σ is invertible. The function is thus strictly convex, and the optimal weights are at a global minimum.

Note that assuming linear independence works in theory, but we might encounter near-collinearity in the data, especially as the number of factors increases. This will be addressed in the computational section of this thesis.

This vector represents the best linear unbiased replication of the target portfolio's return, and it has the highest possible correlation with R_t^M using a linear combination of the factors. We can define the residual error as $\epsilon_t = R_t^M - R_t^P$; it has minimal variance at $\vec{w}^*$. We can evaluate the quality of the replicating portfolio by comparing ϵ_t with the variance of R_t^M.

Distributional Assumptions

This model includes several important distributional assumptions. The return distributions R_t^M and R_t^P are assumed to be well-defined, with finite mean and variance. They are also assumed to be stationary. The target returns R_t^M and the factor returns $\vec{R}_t$ are assumed to be mean-centered; this implies that R_t^P is also mean-centered. The factors are assumed to be linearly independent, so that their covariance matrix is invertible.

Note that joint normality of returns is not required for the derivations above. However, if assumed, the conditional expectation $\mathbb{E}[R_t^M|\vec{R}_t]$ is a linear function of $\vec{R}_t$, and the best linear predictor generated from $\vec{w}^*$ is guaranteed to be the best predictor.

U.S. Macroeconomic Factors and Market Returns

This framework will will be applied to model the relationship between U.S. macroeconomic factors and stock market indexes in both the U.S. and China. Setting the target return R_t^M to the Wilshire 5000, we can construct R_t^P from a set of U.S. macroeconomic factors $\vec{X} \in \mathbb{R}^m$. Our basis vector thus becomes $\vec{R}_t \equiv \vec{X}_t$ and the optimal weight vector $\vec{w}^*$ represents the optimal weighted sum of factors that best approximates the US market returns. Similarly, we can set R_t^M to a Chinese market index that represents Chinese market returns, and we can analyze the optimal weight vector.

The validity of these models should be closely assessed as some of the model assumptions might not hold. First, the macroeconomic variables in $\vec{X}_t$ might be collinear, leading to unstable coefficient estimates. Second, economic factors might not be stationary. Third, the assumption that the covariance between the macroeconomic factors and market returns should remain relatively stable over time might not hold empirically.

It is important to note that in the general model the primary goal is to construct a replicating portfolio that tracks the index by minimizing the MSE. Unlike the future models, it does not decompose the risks associated with the investment.

2.1.2 Generalized Additive Models

So far, we've assumed a linear relationship among the factors in which each factor linearly impacts market returns. Generalized Additive Models (GAMs) allow for nonlinear relationships between each factors and market returns. In the context of the U.S. and Chinese markets, GAMs help in quantifying how U.S. economic conditions nonlinearly influence asset returns.

Model Formulation and Statistical Background

A generalized additive model (GAM) is different from traditional linear regressions as it allows for nonlinear transformations of each predictor. Formally, it allows us to model returns as

$$R_t^M = \alpha + f_1(X_{t,1}) + f_2(X_{t,2}) + \cdots + f_d(X_{t,d}) + \epsilon_t$$

where α is an intercept and f_j are arbitrary smooth functions representing the effect of predictor X_j on R_t^M. As usual, ϵ_t represents the error term and has a zero mean.

Generally, R_t^M can follow any exponential-family distribution with a link function g such that $g(\mathbb{E}[R_t^M|X]) = \alpha + \sum_j f_j(X_j)$. Additionally, the following constraint is imposed for identifiability: each smooth function must have zero mean over the sample, formally shown as $\sum_{t=1}^{d} f_j(X_{t,j}) = 0$ for all j. This nesures that α captures the overall mean of R_t^M, making the additive decomposition unique.

Connection to Linear Models and Basis Expansions

If we restrict each smooth function to be linear, such that $f_j(X_{t,j}) = \beta_j X_{t,j}$, then the GAM reduces to a multiple linear regression, in which $R_t^M = \alpha + \beta_1 X_{t,1} + \beta_2 X_{t,2} + \cdots + \beta_d X_{t,d} + \epsilon_t$. GAMs are extensions of linear models in which the effects are still additive but can be nonlinear.

The smooth functions $f_j(x)$ are unknown and are estimated from the data. The traditional way to do this is to represent each one using a set of basis function, converting the problem into a linear regression in those basis terms. More formally, we can write $f_j(x) = \sum_{k=1}^{K_j} \beta_{j,k} b_{j,k}(x)$, where $\beta_{j,k}$ are estimated from the data. This

gives us the following overall model:

$$R_t^M = \alpha + \sum_{j=1}^{d} \sum_{k=1}^{K_j} \beta_{j,k} b_{j,k}(X_{t,j})$$

This model is linear in the parameters $\beta_{j,k}$, but is not linear in terms of the original variables X_j. The choice of basis complexity is crucial for avoiding overfitting to the noise. Including a smoothness penalty can help ensure smooth trends are captured. A common approach is to use penalized splines: for each f_j, add a penalty term to the loss function that penalizes roughness, which we'll denote as $J(f_j)$. The optimization problem thus becomes

$$\min_{\{f_j\}} \sum_{t=1}^{N} \left(R_t^M - \alpha - \sum_{j=1}^{d} f_j(X_{t,j}) \right)^2 + \sum_{j=1}^{d} \lambda_j J(f_j)$$

where $\lambda_j \geq 0$ are smoothing parameters to control the roughness of f_j. Selecting these values can be done using cross-validation.

Limitations and Practical Considerations

The first limitation of GAMs is that they assume that there are no interactions amongst the factors. A GAM can include interaction terms, but this increases complexity and the required data size. In practice, interaction terms should be manually included in variables that are plausible.

GAMs also require a generous amount of data as they use up degrees of freedom for each smooth term. Having too many predictors or allowing the f_j's to be too flexible can lead to overfitting. Additionally, choosing how smooth each f_j should be is not trivial, adding another layer of complexity in comparison to standard linear regression. Fitting a GAM is much more computationally expensive

than fitting a linear regression.

Finally, in the context of Chinese market linkages with the U.S. economy, it is important to note that linkages are heavily impacted by trade policies. As such, the stationary assumption should be taken into account when choosing how far back to include data from. One way to tackle this is to include indicator variables for different time periods. Additionally, macroeconomic factors and returns might have inherent lags; we can handle this by including lagged predicted as separate inputs.

2.1.3 Nonparametric Kernel Regression

The previous models introduced assume some parametric relationship between market return and the factors used. The linearity assumption may be too restrictive in practice. To capture complexities in the data, we introduce a nonparametric regression model that does not impose a predetermined functional form on the data.

Model Setup and Motivation

Similar to the previous sections, let R_t^M represent the market return at time t and let X_t be a vector of factors at time t. We are interested in the conditional expectation of R_t^M given $X_t = x$, which we'll write as $m(x) = \mathbb{E}[R_t^M | X_t = x]$. Unlike the previous models, there is no assumption that the conditional expectation is linear in x. It is simply assumed that the relationship can be expressed as

$$R_t^M = m(X_t) + \epsilon_t$$

where m is an unknown smooth function and ϵ_t is an error term with mean zero. The goal is to estimate $m(x)$ from the data in a generalized way that can capture interactions among factors. This

generality is the reason we cannot solve for a closed-form solution for m. Instead, we can construct an estimator inspired by the k nearest neighbors algorithm.

Deriving the Nadaraya-Watson Kernel Estimator

Drawing inspiration from the k-NN model, we can estimate $m(x) = \mathbb{E}[R_t^M | X_t = x]$ by averaging the R_t^M values in our data with factors X_t close to x. The idea here is that the true value of $m(x)$ should be close to the average of its surrounding neighbors.

Consider the following estimator, in which a simple rolling sample mean for a small neighborhood around x:

$$\hat{m}_{window}(x) = \frac{\sum_{t=1}^{n} \mathrm{I}\{||X_t - x|| \leq h\} R_t^M}{\sum_{t=1}^{n} \mathrm{I}\{||X_t - x|| \leq h\}}$$

where I is an indicator function and $h > 0$ is the radius of the neighborhood. While this estimator is powerful, it is not typically smooth. We can make it smooth by replacing the indicator function with a kernel weight that decays with distance from x.

Let $K(u)$ be a kernel function, a smooth, non-negative, and symmetric function that gives higher values for u near 0 and satisfies $\int K(u)\, \mathrm{d}u = 1$. Let the kernel weight for observation t at the target point x be $K_h(x - X_t)$, given by

$$K_h(u) = \frac{1}{h^d} K\left(\frac{u}{h}\right)$$

where h is the neighborhood radius (also called the bandwidth) and $d = \dim(X_t)$ is the number of factors. The kernel weight is large when X_t is close to x since $\frac{x - X_t}{h}$ is near 0, implying that $K\left(\frac{x - X_t}{h}\right)$ is large.

Using these weights, we adjust our estimator to be:

$$\hat{m}_h(x) = \frac{\sum_{t=1}^{n} K_h(x - X_t) R_t^M}{\sum_{t=1}^{n} K_h(x - X_t)}$$

This estimator was proposed by Nadaraya and Watson independently in 1964 and is called the Nadaraya-Watson kernel regression estimator.[20, 21]

Note that this does not assume a specific kernel function. Studies have also shown that the choice of the kernel is less critical than the choice of the bandwidth.

Bandwidth Selection and the Bias-Variance Trade-Off

The choice of the bandwidth is critical in kernel regression. A small h only uses observations very close to x, which reduces bias and increases the variance of the estimates. A large h gives a smoother, low-variance estimate but can blend points that may have different true means together.

This is an example of the classic bias-variance trade-off. We want to capture local detail with a smaller h, but we don't want it to be so small that $\hat{m}_h(x)$ becomes volatile. Cross-validation can be helpful in practice for selecting the bandwidth, typically using the MSE as the objective function.

Assumptions and Theoretical Properties

The first assumption is that the data samples are expected to be independent across t. This assumption justifies using the empirical local average to estimate the conditional expectation. Typically, identical distributions are also assumed in theory, but in practice a weaker assumption of stationary distributions has been shown to suffice.

The true conditional mean function $m(x)$ is assumed to be smooth in x. This ensures that observations near x do have similar expected values of R_t^M, making local averaging effective.

The kernel is typically assumed to be symmetric about 0 and has a finite variance. The Gaussian and the Epanechnikov kernels satisfy these conditions and are common choices for a kernel. The bandwidth sequence is typically chosen such that h shrinks as the number of data samples increases.

Under these conditions, the Nadaraya-Watson estimator is consistent for $m(x)$.[20, 21] More formally, if $h_n \to 0$ as $n \to \infty$, $m_{h_n}(x) \to m(x)$ in probability for each fixed point x where $f_X(x) > 0$.

Application to Market Return Prediction with Macroeconomic Factors

We can apply this model to the market by letting R_t^M represent the return of a market index in month t and X_t be a vector of macroeconomic indicators observed at month t. By using a kernel regression, we can estimate the impact of specific factors without assuming linearity in the underlying relationship.

The advantages of flexibility and minimal assumptions come at a cost of interpretability and scalability. Unlike a traditional linear model, which returns a set of easily-interpretable coefficients, kernel regression returns the estimated $\hat{m}(x)$, which is harder to interpret without visualizations or further examination into other factors. Kernel regression also struggles in cases where the dimension of x is large or in cases with limited data. In practice, this limits kernel regression to a small number of key factors or requires some form of dimension reduction, which can lose more interpretability. Kernel regression is also far more computationally expensive than OLS regression, especially for larger values of n.

2.2 Statistical Modeling Frameworks

This section covers the empirical frameworks that this thesis considers employing to analyzing the impact of the U.S. and Chinese economies on their markets. These frameworks build on the statistical models previously discussed.

2.2.1 Linear Regression with Regularization

In practice, linear regression can be improved by adding regularization terms to the loss function. Techniques like Ridge Regression and Lasso Regression can help with multicollinearity and overfitting.

The main idea is to start with the same OLS objective and add a penalty term on the coefficients, making the minimizing function

$$\vec{w}^* = \arg\min_{\vec{w}} \mathbb{E}\left[\left(R_t^M - \vec{w}^\top \vec{R}_t\right)^2\right] + \lambda P(\vec{w})$$

where $P(\vec{w})$ is a penalty function with $\lambda \geq 0$ representing its strength.

In Ridge regression, $P(\vec{w}) = \sum_{j=1}^{d} w_j^2$. The penalty shrinks all coefficients towards zero. In Lasso regression, $P(\vec{w}) = \sum_{j=1}^{d} |w_j|$. The penalty here can set some coefficients to zero and can act as some form of variable selection. Empirically, Lasso can be unstable in highly correlated features.

In Python, Lasso and Ridge regression can be implemented using scikit-learn. After standardizing the factors, we can choose the optimal λ using cross-validation using RidgeCV or LassoCV.

2.2.2 Sparse Additive Models

Standard GAMs are prone to overfitting when many factors are irrelevant, especially as the dimensionality of the data increases. Sparse Additive Models (SpAM) impose a sparsity constraint on

GAMs so that only a subset of the f_j functions have non-zero value, analogous to the effect of Lasso regularization on a linear regression.

It applies a Lasso penalty to each f_j. If we represent each f_j as a basis, we can apply a group lasso penalty to highlight the most important f_j factors. SpAMs have been found to be very effective in this method, correctly identifying the relevant components with high probability given enough data.

While there isn't a specific library for SpAMs in Library, one could use pyGAM to create a basis expansion for each feature and then apply the Lasso penalty across the groups of basis coefficients for each feature.

2.2.3 Kernel Regression

In the theoretical modeling section, we introduced kernel regression and the Nadaraya-Watson estimator as a method for accounting for nonlinearities.

Using the statsmodels Python package, we can use the KernelReg class, which fits the model using the Nadaraya-Watson estimator. We can also use scikit-learn's kernel-based algorithms, which include the kernel ridge regression, Gaussian process regressors, and support vector regressors.

An important choice in kernel regression is the selection of the bandwidth. In order to ensure comparability in distance measurements across factors, it's important to first standardize the data. After that, cross-validation can be used to choose a bandwidth.

2.2.4 Empirical Considerations

There are several challenges to be aware of when moving from theory to empirical implementation. Although they vary in importance, testing for the impact of the challenges and accounting for them is

central towards achieving an accurate model.

Non-Stationarity

The models above all assume stationarity in the underlying distributions. This is not a natural assumption in financial data, which naturally have cycles of expansion and contraction. Beyond that, changes in regimes and global policy can impact the underlying distributions of economies.

Non-stationarity can be handle with data transformations. Incorporating time indicators can also help adjust for shifts in the data, potentially marked by regime changes.

Factor Relevance and Spanning

When analyzing economic conditions, the data is naturally limited as it is typically measured monthly. With many potential factors impacting the market, identifying significant factors is important for applying these models. This is particularly important in models that require numerous data, like kernel regression. Multicollinearity among the factors can also be harmful, leading to unstable and unreliable coefficient estimates.

Economic factors, especially when reduced to smaller set, likely will not fully capture the risks in the market. Thus, factor selection is critical to the success of the models. Selecting too many factors makes some models impossible to estimate accurately, but not selecting enough significant factors can prevent the degree to which the models can accurately represent the market.

The data used in this thesis has 13 themes and over 150 factors. To combat this issue, I will use the themes data in the kernel regression as they typically aggregate several similar factors' information.

3 Data

3.1 Data Description

The data used in this thesis can be split into two portions: the economic data and the market returns.

3.1.1 Data Origin

The economic data is retrieved from the Jensen, Kelly, and Pedersen dataset. The dataset includes a comprehensive collection of equity market factors described in their paper "Is There a Replication Crisis in Finance?" (2023). The data is publicly available with extensive documentation of factor definitions and construction methods. The JKP factors are constructed using data from the Center for Research in Security Prices (CRSP), Compustat, I/B/E/S, and OptionMetrics.

The data is available for both the U.S. economy and the Chinese economy.

The market data is retrieved from Finaeon, a platform that provides financial and economic data, including over 307,000 global data series spanning markets from 1000 AD to the present. They compile their data from a variety of historical and current sources, including financial reports, news periodicals, sectoral journals, and other publications, all of which are meticulously transcribed and verified.

3.1.2 Factor Data

Data is provided on 153 unique factors representing the U.S. and Chinese economies, measured at both a daily and monthly granularity. The factors are value-weighted by a method called *capped market capitalization*, meaning that the impact of a company on the measure is weighted by its size, winsorized at the 80th percentile level. Using market capitalization as a weight is standard in the market as it better reflects investable opportunities and is less volatile than equally-weighted values. The capping methodology is implemented by JKP to cap the impact a singular company can have on the measure to 80%.

The factors can be split into the following categories: enterprise value, earnings momentum, earnings quality, sales growth, volatility measures, liquidity, market size, and technical factors like Amihud illiquidity and maximum returns.

The U.S. monthly data has 1,188 rows and covers January 1926 to December 2024; of those, 638 include measurements of all 153 factors. The U.S. daily data has 26,051 rows and covers January 2, 1926 to December 31, 2024; of those, 13,407 have measures of all 153 factors.

The Chinese monthly data has 348 rows and covers November 1993 to December 2024; of those, 112 include measurements of all 153

factors. The Chinese daily data has 7,109 rows and covers November 1, 1993 to December 31, 2024; of those, 2,266 have measures of all 153 factors.

3.1.3 Themes Data

The U.S. and Chinese factor data are provided as 13 themes, each aggregating multiple factors; as such, the theme data is also value-weighted by market capitalization. The theme data is available at a daily and monthly granularity. The themes are: accruals, debt issuance, investment, low leverage, low risk, momentum, profit growth, profitability, quality, seasonality, short term reversal, size, and value. Descriptions for each of these themes are provided below.

Accruals Measures the difference between a company's reported earnings and its cash flows. Lower values signal a higher quality of sustainable earnings.

Debt Issuance Tracks changes in corporate debt.

Investment Measures corporate spending on assets, capital expenditures, and acquisitions.

Low Leverage Measures absolute debt levels.

Low Risk Captures stocks with lower risk measures, often represented by volatility.

Momentum Measures return persistence and trend-following behavior.

Profit Growth Measures change in profit margins and earnings.

Profitability Measures levels of corporate profitability.

Quality Combines measures of financial stability, earnings consistency, governance, and operational efficiency.

Seasonality Captures recurring calendar-based patterns in market returns.

Short Term Reversal Measures the tendency of stocks to reverse their previous 1-4 week performance.

Size Measures market capitalization, allowing to distinguish between company sizes.

Value Identifies stocks trading at lower prices relative to their values of fundamental measures (book value, earnings, cash flows, or sales).

The U.S. monthly data has 1,188 rows and covers January 1926 to December 2024; of those, 878 include measurements of all 13 themes. The U.S. daily data has 26,051 rows and covers January 2, 1926 to December 31, 2024; of those, 18,443 have measures of all 13 themes.

The Chinese monthly data has 348 rows and covers November 1993 to December 2024; of those, 336 include measurements of all 13 themes. The Chinese daily data has 7,109 rows and covers November 1, 1993 to December 31, 2024; of those, 6,787 have measures of all 13 themes.

3.1.4 Market Returns

The Wilshire 5000 Total Market Index and the Shanghai SE Composite were chosen to represent the equity markets in the U.S. and China, respectively, because of their broad coverage of the markets. The Wilshire 5000 is a market-capitalization-weighted index of the market value of all American stocks actively traded in the

United States. The Shanghai SE Composite is an index of all stocks traded at the Shanghai Stock Exchange.

The data is available at the daily and monthly levels. The Wilshire 5000 data dates back to 1970 and is available until May, 2024, and the Shanghai Composite data dates back to its opening in 1990 and is available until December, 2024.

The close values at the end of the market day are used. The returns are also given as percent changes from the previous period, either daily or monthly. The Shanghai Composite is given in U.S. dollars.

3.2 Exploratory Data Analysis

3.2.1 Data Cleaning

The data was provided in high quality, as JKP cleaned the data before publishing it. The market data provided by Finaeon was also of great quality, which is expected as they are providers of data for financial researchers and professionals. Most of the data was provided in a stacked format. After pivoting the datasets, I analyzed it for missingness and temporal coverage. While the missingness and temporal coverage are explored for the entire dataset, the trends are explored for 2001 to 2024, the data available after China joined the World Trade Organization.

3.2.2 Data Trends

Market Trends

Data trends are explored on a monthly granularity to reduce the impact of noise. Looking at the market returns for the U.S. and China, shown in Figure 3.2.1, we see that the Chinese market is much more volatile than the U.S. market.

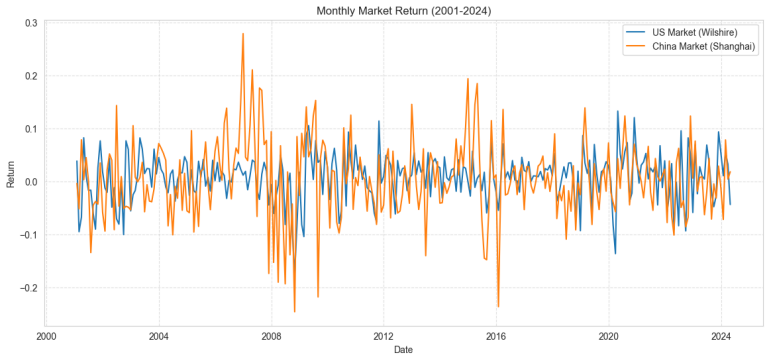

Figure 3.2.1: Monthly Market Returns

Even after calculating the 6-month moving average, displayed in Figure 3.2.2, the volatilities are clearly different. Note that the greatest differences appear around the 2005-2008 period, the 2014-2016 period, and the 2018-2019 period.

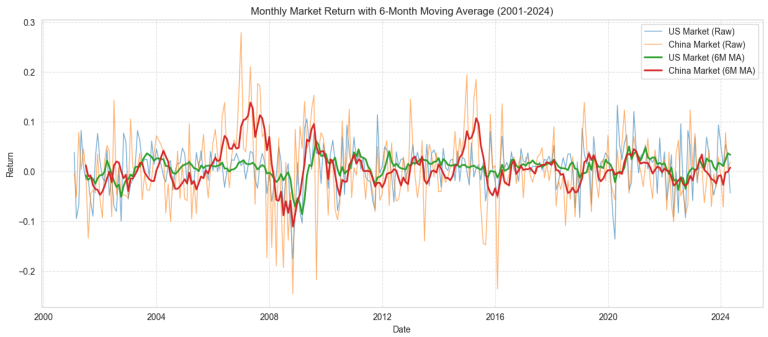

Figure 3.2.2: Monthly Market Returns with 6-Month Moving Averages

During the 2005-2008 period, China's economy experienced rapid growth and market liberalization. This explains the period of higher returns that was not exhibited in the U.S. market. The 2008-2010 period had similar impacts on the U.S. and China as financial crises led to sharp market corrections in both markets. In the U.S., the

markets were crashing because of the bursting of the housing bubble and the failures of the financial sector. In China, investors fear surrounding the slowing exports and accumulated debt led to a market correction.

The 2014-2016 period had large increases and decreases in returns in China, while the U.S. economy remained relatively stable. The trends in China were mainly due to the retail bubble and its burst.

In 2018-2019, the Chinese market corrected after growing too quickly.

Economic Trends

Economic trends are explored at the monthly level once again to avoid volatility and noise. Themes are visualized for greater explainability as there are 13 themes and 153 factors.

The 6-month moving averages of the 13 themes are plotted in Figure 3.2.3 from 2001 to 2024 to reduce noisy peaks and increase visibility in trends. We notice, consistent with the market data, that the Chinese economy is less stable than the U.S. economy, with themes generally being more volatile.

We do also notice general trends between the market returns and the economic themes within countries as periods with high volatility in markets typically align with high volatility in economic themes.

3.2.3 Correlations and Collinearity

Themes correlations are analyzed using a correlation heatmap, shown in Figure 3.2.4 for the U.S. and in Figure 3.2.5 for China. Note that there appear to be several strong correlations in the U.S. and Chinese data, but the correlations are not necessarily the same for both countries. Some of the strongest correlations in the U.S. market are

Figure 3.2.3: Monthly Theme 6-Month Moving Averages

uncorrelated in the Chinese market, and vice versa. This implies that the markets have structural differences, and that the themes important for market signals in the U.S. might not be important signals for the Chinese market.

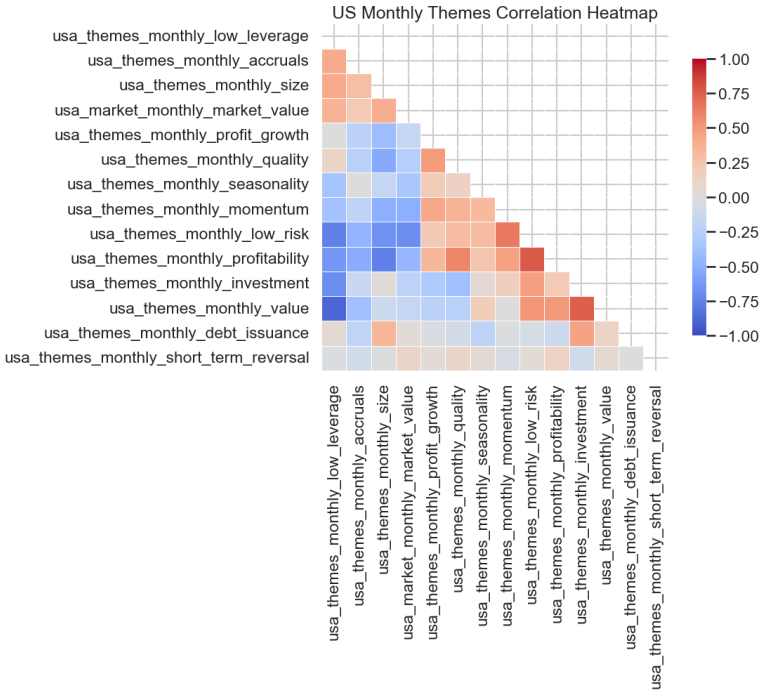

Figure 3.2.4: U.S. Monthly Themes Correlation Heatmap

Correlation networks were also analyzed in the U.S. and China Themes and Factors datasets at a daily granularity. The themes network in the U.S., shown in Figure B.0.5, reveals only correlations ($r \geq 0.70$) among the low leverage, value, and investment themes. The themes network in China, shown in Figure B.0.6, reveals a similar trend in which value and low leverage are correlated ($r \geq 0.70$). It also shows that, in China, size is correlated to profitability, profitability is correlated to quality, and quality is correlated to

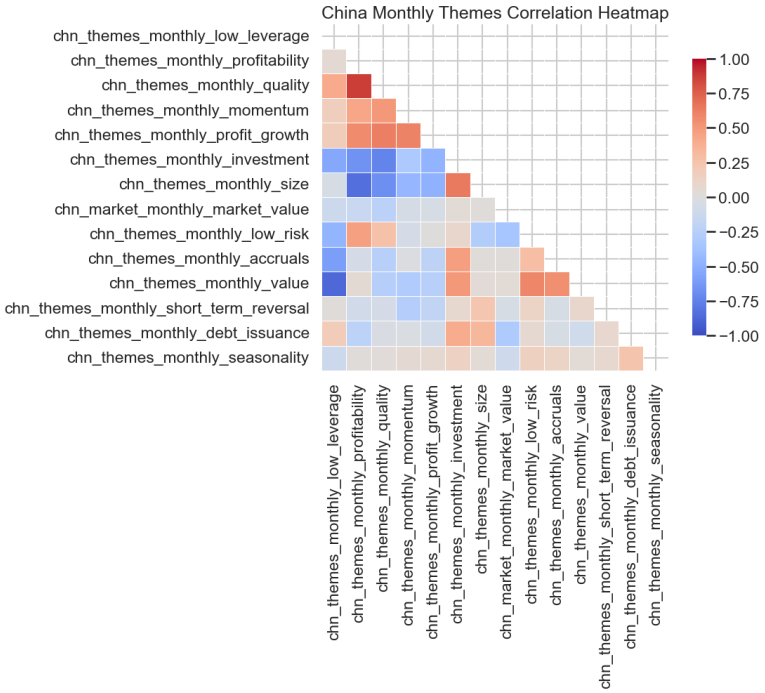

Figure 3.2.5: China Monthly Themes Correlation Heatmap

investment. Figure B.0.7 and Figure B.0.8 show the factor correlation networks for the U.S. and China, respectively.

3.2.4 Missingness

Factor Data

While some U.S. factors have data points as early as 1926, they only comprise 29 of the 153 factors. Data is available for 40 factors starting in 1930, 49 factors starting in 1950, 136 factors starting in 1960, and all 153 factors are available after 1972. These trends for missing values in the daily and monthly factors for U.S. data are shown in the top subplot in Figure 3.2.6.

For the Chinese factors, we see a similar decreasing trend, in which more data is available for more recent years. However, the Chinese data availability is much more volatile. Note that the data is much more stable post 2001, the year in which China joined the WTO. This is typically the earliest that the literature begins considering China a liberalized market, and is the earliest data that will be included in our models. These trends for missing values in the daily and monthly factors for China data are shown in the bottom subplot in Figure 3.2.6.

Themes Data

The U.S. themes data is much more available, with no missingness after 1960 in either the daily data or the monthly data. These are shown in the top subplot in Figure 3.2.7.

The Chinese daily themes are still volatile before 2002 in terms of missingness, but the monthly data is complete from 1999 onward. These trends are shown in the bottom subplot in Figure 3.2.7.

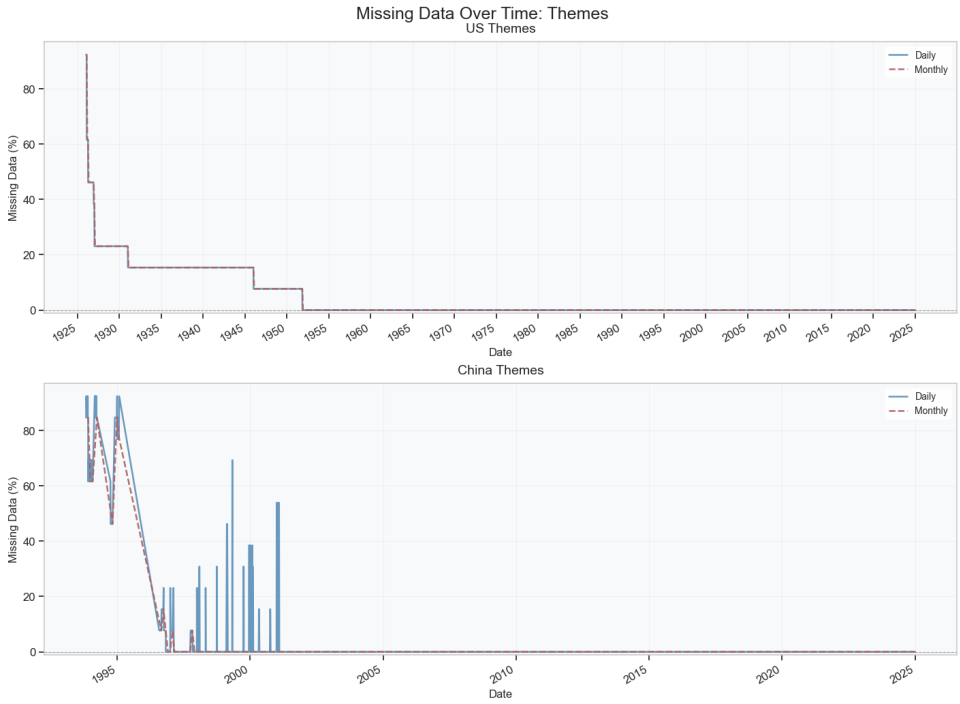

Figure 3.2.6: Factors Missing Values Over Time

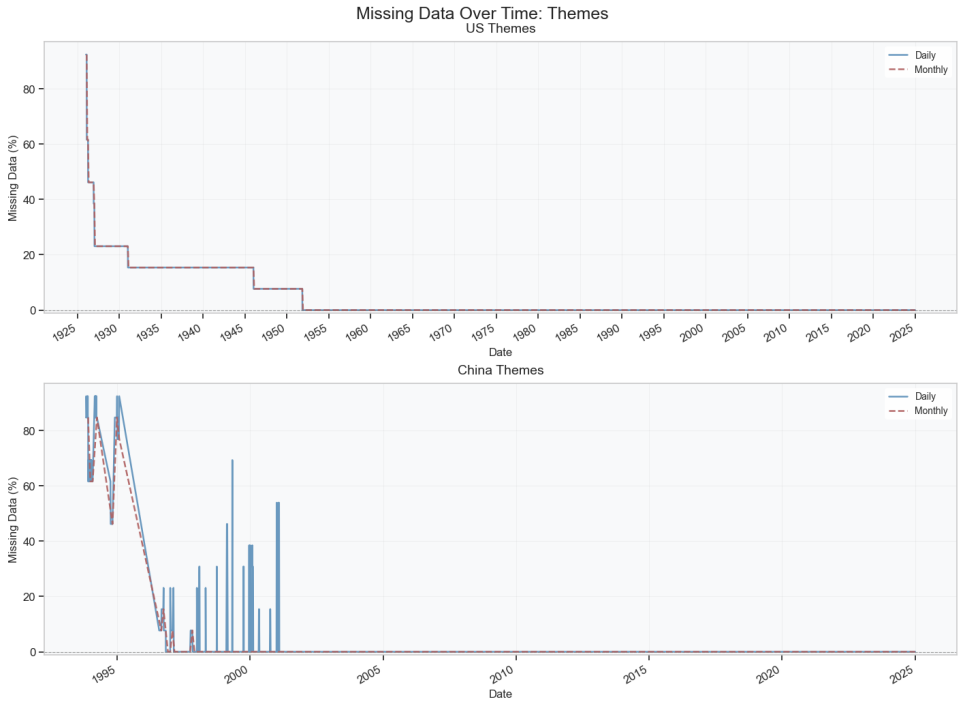

Figure 3.2.7: Themes Missing Values Over Time

A few trends come out in the missingness of the data. First, data missingness is generally decreasing over time. Second, themes data is more complete than the factors, with no missingness after 1960 for the U.S. and 2002 for China. Third, data with monthly granularity tends to have less missingness than daily data with daily granularity.

Handling Missingness

The monthly themes data does not exhibit any missingness in the data. The daily themes, on the other hand, have some sporadic missingness in the data. Figure 3.2.8 shows the missingness observed in the daily themes for the 40 themes with the most missingness. The missingness is sporadic, and will be imputed using linear imputation.

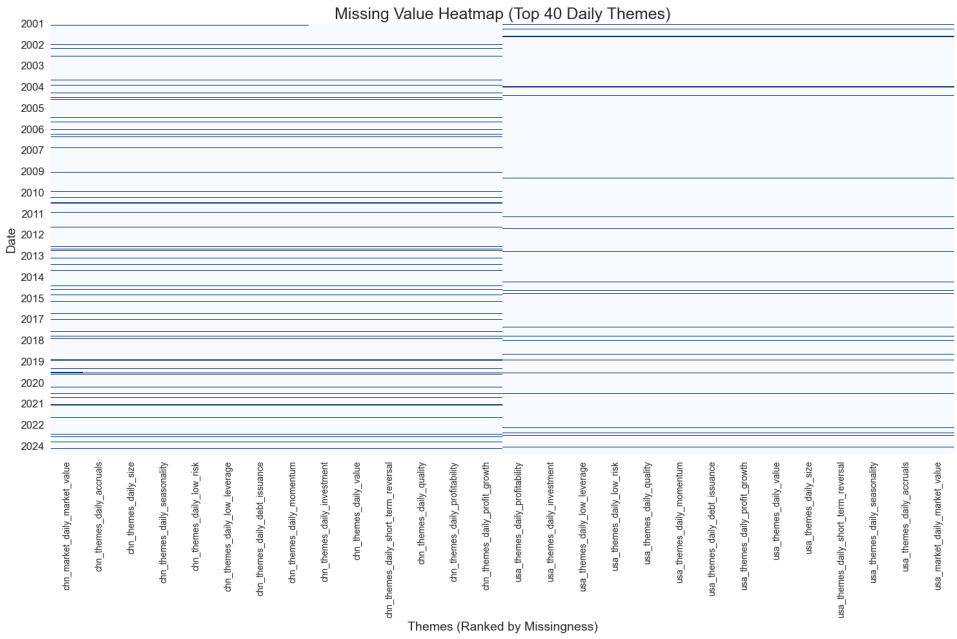

Figure 3.2.8: Missing Themes Heatmap (Top 40 Themes, Daily Data)

The factor data, on the other hand, exhibits much more severe

missingness. Figure 3.2.9 shows the 40 factors with the most missingness on a daily factor. The missingness can be decomposed into two types: (1) sporadic missingness similar to that observed with the daily themes data, which can be linearly imputed, and (2) large chunks of missing data for some of the Chinese factors. These factors, listed below and shown in Figure 3.2.9, have too much missing data to be considered, often not having any data for the first 4-13 years of the data. In order to avoid bias, I've excluded the factors from both countries' data.

The factors I've excluded from the data are: rd5_at, seas_16_20an, seas_16_20na, rd_me, rd_sale, emp_gr1, sale_emp_gr1, seas_11_15an, seas_11_15na, seas_6_10na, seas_6_10an, iskew_hxz4_21d, ivol_hxz4_21d, ocfq_saleq_std, saleq_su, eqnetis_at, saleq_gr1, ni_inc8q, capex_abn, capx_gr3, resff3_6_1, resff3_12_1, niq_su.

Figure B.0.4 shows the missingness on the monthly level, which is similar to the daily plot in Figure 3.2.9 but without the dates where data is missing for all factors. The same factors are excluded from the monthly data.

Market Returns Data

While no missingness was initially exhibited in the data at either the daily or the monthly granularity, further analyses showed some gaps in temporal coverage at the daily granularity. Figure 3.2.10 shows the gaps. Gaps were calculated as periods of more than two business days without data for the daily level and any month without data for the monthly level. Gaps do not account for federal holidays or other yearly missingness in the data. This is not central to the analyses, so it was not explored in more depth. The largest gaps are shown, with 23 business days in the U.S. daily data and 14 business days in the

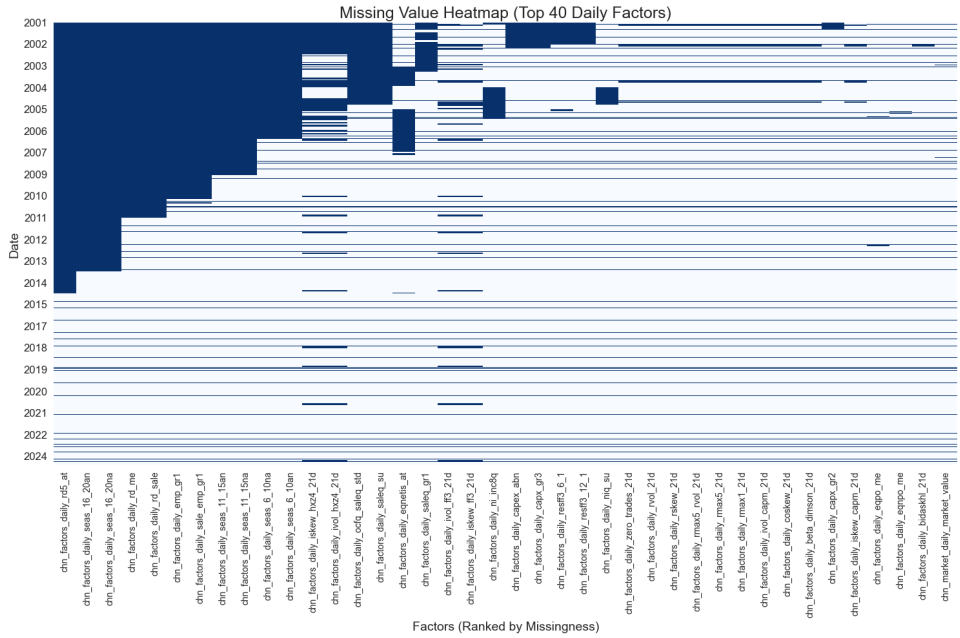

Figure 3.2.9: Missing Factors Heatmap (Top 40 Factors, Daily Data)

Chinese daily data.

Additionally, there were some days where the markets were closed but the factor and theme data were available. I chose to omit these from the data since the observed outcome is not measured.

3.2.5 Consistency and Outlier Analysis

Market Outliers

I first merged the data to create four Pandas dataframes: (1) Factors with Market Returns for the U.S. and China at a daily granularity, (2) Factors with Market Returns for the U.S. and China at a monthly granularity, (3) Themes with Market Returns for the U.S. and China at a daily granularity, and (4) Themes with Market Returns for the U.S. and China at a monthly granularity.

Figure 3.2.10: Market Data Time Coverage Analysis

Using the IQR method, 6 outliers were detected in the U.S. market returns and 11 outliers were detected in the Chinese market returns. Figure 3.2.11 shows the dates that the outliers appeared. Of the 17 total outliers, 10 occurred from 2007 to 2010. Three occurred in the 2014-2016 time range, in which the Chinese market was volatile as a result of the retail bubble. Three occurred in the 2020-2021 range in the U.S. market, displaying volatility around the widespread fear caused by the COVID-19 pandemic.

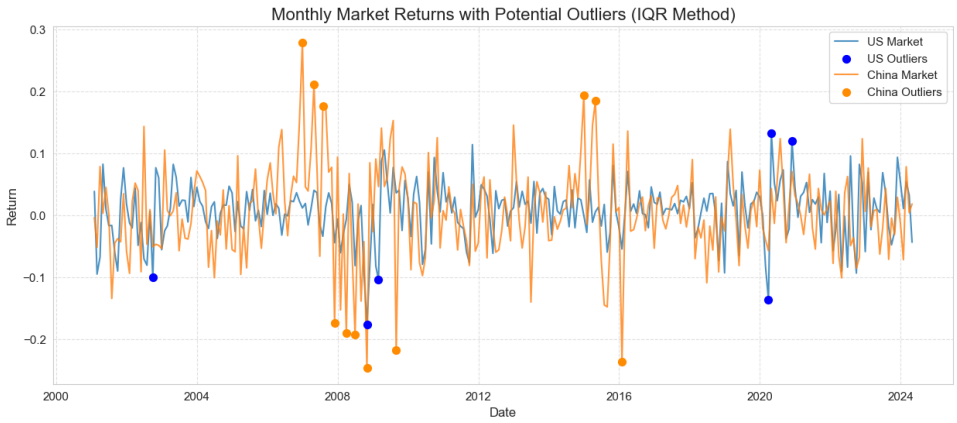

Figure 3.2.11: Monthly Market Returns with Potential Outliers

Economic Outliers

Theme values in the U.S. and China were also analyzed for outliers using the IQR method. Figure 3.2.12 shows box plots of the themes for each country, measured at the monthly granularity. The plot reveals some similarities in theme variance across the countries, but it also reveals great differences in others. Accruals Debt Issuance, Profit Growth, and Seasonality were among the most compact Themes for both countries. Momentum was one of the most varying themes, with severe outliers in both countries. Size was the most varied data in the

Chinese economy, though much more stable in the U.S. economy. Most other factors were generally more varied in the Chinese economy than the U.S. economy.

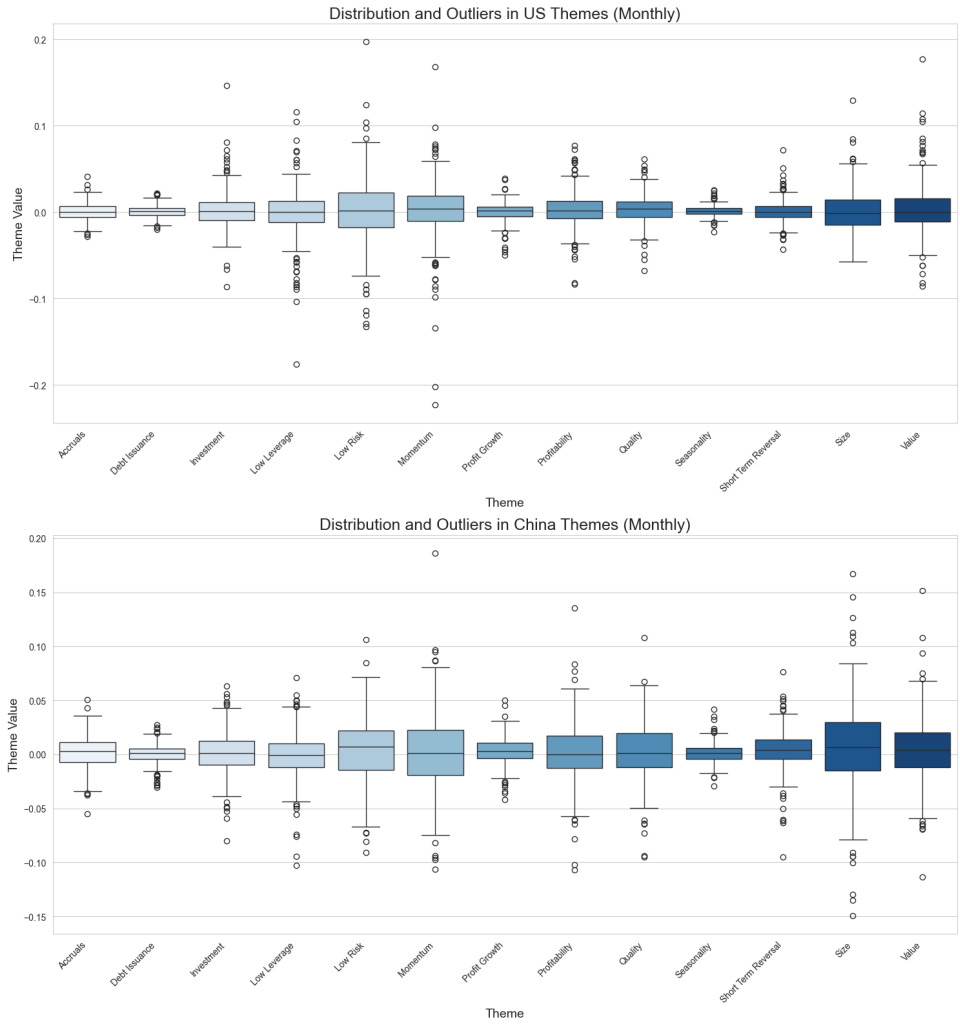

Figure 3.2.12: Box Plots of U.S. and China Themes

Theme and factor volatilities were calculated as 6-month rolling standard deviations for all columns, either all themes or all factors, for each country and then averaged. Examining the results, we see

similar trends over time for themes and factors. Theme volatilities are placed in the appendix (see Figure B.0.1). Figure 3.2.13 shows the factor volatilities over time for the U.S. and Chinese economies. It confirms our suspicions of the Chinese economy generally having a greater volatility than the U.S. economy, and it shows periods of divergence and near-convergence of the volatilities. Most recently, the volatilities have generally been the same since 2020.

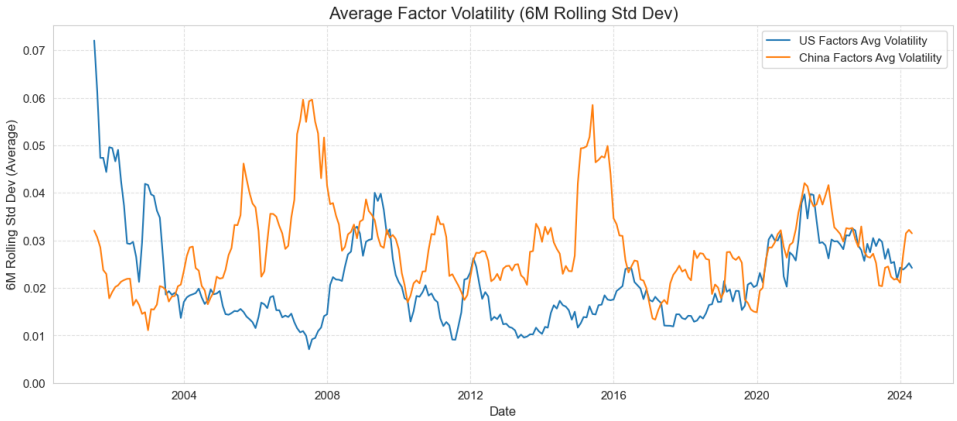

Figure 3.2.13: Factor Volatility (6-Month Rolling Standard Deviation)

I also examined the number of outliers over time in each country for factors and months. The results were consistent with the volatilities above, with areas with greater volatility having more outliers. The plot of the number of factor outliers over time is shown in Figure B.0.2, and the plot of the number of theme outliers over time is shown in Figure B.0.3. We note that it is very rare for factors and themes to have no outliers at a given time point.

When deciding what to do with outliers, we have a few options. We can retain outliers, using the data as it appears. These cases do not appear to be miscalculations or as a result of human error; they align with historically significant scenarios. Keeping the data can impact

the result of the models, especially models that use MSE as an objective function. Significant scenarios in either country might significantly impact the regression. As a result, one potential method would be to fit a model on the entire 2001 to 2024 region, and then again post-2010 for more recent events; still, the models would likely be influenced by the 2015 bubble in the Chinese market.

We can exclude the outliers, following some criteria for when to include or exclude a time period. However, these outliers are not errors and are not irrelevant to the relationship between the markets. Excluding outliers or idiosyncratic scenarios biases the models by omitting important market scenarios.

We can winsorize outliers to reduce their impact. We might set any data above the 99th percentile to be the same value as the 99th percentile (and similarly for the 1st percentile). This complicates the interpretation of the effect size and does not exclude the impact of the outliers in the models.

As a result, I've chosen to winsorize the all columns in the data at the 0.05% and 99.95% levels. In the monthly data, this will impact 1.430% of the cells in the factors dataset, 1.429% of the cells in the themes dataset, and 1.429% of the cells in the returns dataset. For the daily data, it will impact 1.021% of the cells in the factors dataset, 1.021% of the cells in the themes dataset, and 1.21% of the cells in the returns dataset. In total,

This choice was made in order to limit the impact of extreme values in the data. It's important to note that it still preserves the information from the outliers, but it caps the amount of effect a single point can have. The literature also shows that this does not fully eliminate outlier bias, showing that the signals of the winsorized outliers remain.

3.2.6 Standardizing the Data

Applying the Shapiro test for normality at an $\alpha = 0.05$ level shows that none of the columns in the daily data is normal or standardized. Of the monthly datasets, no columns are standardized, but 34 of the 262 factor columns are normal and 4 of the 28 theme columns are normal. I checked for standardization in the data by checking if the mean of a column's data μ satisfies $|\mu| \leq 0.05$ and its standard deviation σ satisfies $0.95 \leq \sigma \leq 1.05$.

The themes and factors are standardized as follows using sklearn's StandardScaler. Suppose that $x_{i,j}$ represents the value of row i in column j. Let μ_j and σ_j be the mean and standard deviation of the data in column j. The standardized value of $x_{i,j}$, denoted as $x_{i,j}^{(std)}$, is given by $x_{i,j}^{(std)} = \frac{x_{i,j} - \mu_j}{\sigma_j}$.

Note that, as outlined in the statistical formulation of the models, the factors and themes are not assumed to be normally distributed. As such, standardizing the data is enough.

3.2.7 Stationarity and Structural Breaks

Testing for Stationarity and Structural Breaks

As described in the statistical formulation section, stationarity is an important assumption of the models Two tests are used to assess stationarity: the Augmented Dickey-Fuller (ADF) Test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test. I used both at the $\alpha = 0.05$ level.

The ADF test has the null hypothesis that the time series is non-stationary, and an alternative hypothesis that it is. It can be implemented using adfuller() from statsmodels in Python. The series is said to be stationary when the null is rejected when the p-value is less than 0.05.

The KPSS test has the null hypothesis that the time series is

stationary, and an alternative hypothesis that it is not. It can be implemented using kpss() from statsmodels in Python. The series is said to be stationary when we fail to reject the null hypothesis, when the p-value is greater than 0.05.

In order to detect structural breaks in the data, the Pruned Exact Linear Time (PELT) algorithm is also used. It detects multiple change points in time series data, and it can be implemented using the ruptures package. It efficiently searches for optimal segmentation using the squared error cost function.

The time series were plotted with vertical lines marking detected break points.

I also used the CUSUM (Cumulative Sum) test for breaks in the US-China relationship, which looked for structural changes in regression relationships between the market returns.

Testing Results

In the monthly factors dataset, the ADF test concluded that 261 of the 262 columns were stationary and the KPSS test concluded that 240 of the 262 columns were stationary. They both agreed on 239 columns being stationary. Using the CUSUM test, a break was detected in the relationship, with $p < 0.0001$.

In the monthly themes dataset, the ADF test concluded that 27 of the 28 columns were stationary and the KPSS test concluded that 26 of the 28 columns were stationary. They both agreed on 25 columns being stationary. Using the CUSUM test, a break was detected in the relationship, with $p = 0.0002$.

In the daily factors dataset, the ADF test concluded that all 262 columns were stationary and the KPSS test concluded that 197 of the 262 columns were stationary. Using the CUSUM test, a break was detected in the relationship, with $p < 0.0001$.

In the daily themes dataset, the ADF test concluded that all 28 columns were stationary and the KPSS test concluded that 22 of the 28 columns were stationary. Using the CUSUM test, a break was detected in the relationship, with $p < 0.0001$.

Structural breaks were analyzed in the factors and themes as well, at both the daily and monthly granularities. Two common trends amongst those with structural breaks were in the middle of 2002 and in 2022.

Making the Data Stationary

While every column in the data was described as stationary by at least one of the tests, I decided to only count data as stationary if both models agree on it. As such, there were 97 columns that needed to be changed in some way.

I wrote a function that attempted a few transformations, checking for stationarity as it applied different ones. First, it attempts differencing, in which observations are subtracted from their previous values. This is generally great for series with upward or downward trends. If that doesn't work, the natural logarithm is used instead; this accounts for exponential growth in the data. The square root transformation is attempted next. Lastly, the Box-Cox and Yeo-Johnson transformations are attempted, which normalize the data. Lastly, the function standardizes any data it is able to make stationary.

In the monthly factors data, 23 columns were identified as non-stationary by either test. Of those, 21 were transformed into stationary columns, 20 by differencing and 1 using the Yeo-Johnson transformation. The two columns that didn't achieve stationarity by any method were usa_factors_monthly_aliq_mat and chn_factors_monthly_age.

In the daily factors data, 65 columns were identified as non-stationary by either test. Of those, 61 were transformed into stationary columns, 59 by differencing and 2 using the Yeo-Johnson transformation. The four columns that didn't achieve stationarity by any method were usa_factors_daily_lnoa_gr1a, chn_factors_daily_ocf_at, usa_factors_daily_fnl_gr1a, and usa_factors_daily_nncoa_gr1a.

In the monthly theme data, 3 columns were identified as non-stationary by either test. All three were transformed into stationary columns by differencing. In the daily theme data, 6 columns were identified as non-stationary by either test. All six were transformed into stationary columns by differencing.

The Cumulative Sum Test on Market Returns

The cumulative sum (CUSUM) test is a technique used to detect structural breaks or regime shifts in financial data. It looks at the cumulative sum of deviations from a target value. When it exceeds the target value, it signals a change in the underlying process.

The U.S. results, shown in Figure 3.2.14, reveal a few potential structural breaks: 2001, 2007-2008, and 2011-2013. The China results, shown in Figure 3.2.15, reveal a few other potential structural breaks: 2001, 2007-2009, and 2015. These results align with the historical contexts presented earlier in the thesis.

3.2.8 Limitations in the Data

There are a few considerations we must take when analyzing the data with our models. First, stationarity was resolved in some factors, but those will likely pose an issue with interpretation. Second, some factors were removed due to data missingness and a lack of

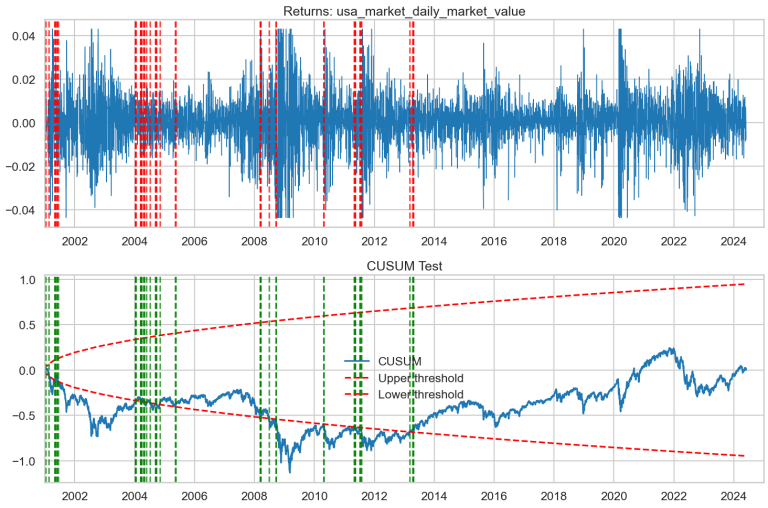

Figure 3.2.14: CUSUM Test on U.S. Market Returns

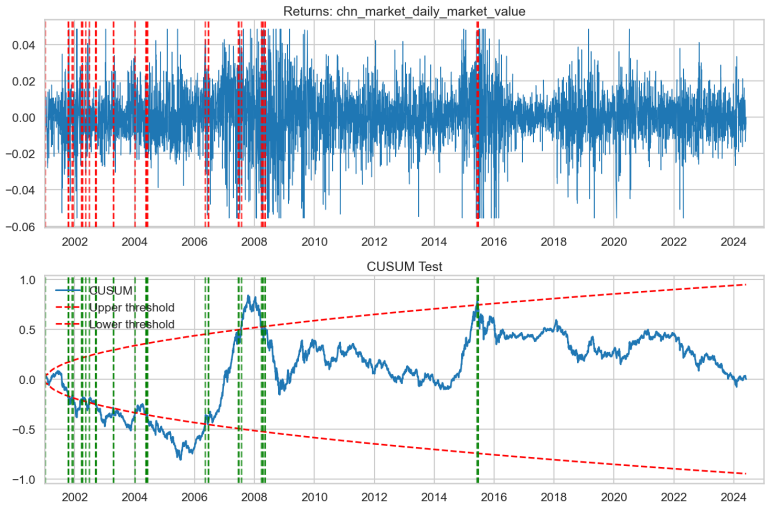

Figure 3.2.15: CUSUM Test on Chinese Market Returns

stationarity, even after attempting transformations. Lastly, there are some issues with structural breaks in the markets, typically representative of shifts in regimes or policy. With the lastest structural break being a decade ago, we can still analyze the integration from 2016-2024.

4

Fitting the Models

4.1 Linear Regression

Linear regressions were fit, with and without regularization. Of the 192 models fit, 48 models did not use regularization and the other 144 were split amongst ridge, lasso, and elastic net. The models were combinations of granularity (daily or monthly), type of predictors (factors or themes), data time frame (2001-2024 or 2016-2024), market returns being predicted (U.S. or China), origin of the predictors used (U.S. data, China data, or both).

As suggested by the literature, the R^2 value is computed as a measure of explainability. I've also computed the Adjusted-R^2 value as an extra measure that considers the number of variables used. Adjusted-R^2 is used to penalize using more predictors, which can help

in finding models that were most efficient.

4.1.1 Assumptions

The main assumptions on the data in a regression are the lack of multicollinearity, homoskedasticity, and normality and independence in the errors.

The data has so far been standardized. Testing for multicollinearity revealed one near-perfect correlation in the daily factors datasets and in the monthly factors dataset, and no near-perfect collinearity in the themes datasets. The correlation pair for both was turnover_126d and zero_trades_126d, both of which were removed from the datasets for the U.S. and China.

For each of the models trained, residual and Q-Q plots are assessed. Trends are described below.

4.1.2 Without Regularization

The linear regression without regularization results are shown in Table A.0.1 and Table A.0.2, separated by the market being replicated. Note that six models returned an R^2 of 1 and an Adjusted-R^2 of 0. Those cases resulted from models in which the number of rows in the data was lower than the number of predictors. This was the case when estimating 2016-2024 monthly data using facts; recall that there are about 140 factors remaining in the data for each country and there are about 114 months in the analyzed time frame. The Adjusted-R^2 of 0 was a safeguard in my code to prevent an erroneous calculation.

Analyzing the models that predicted U.S. returns using the daily data, we note that there was no scenario in which using Chinese economic data alone had an $R^2 > 0.10$. Using economic data from both countries also produced negligible effects on the R^2 value, as

compared to using only U.S. economic data. Analyzing the models that predicted Chinese returns using the daily data, we note that the best performing model using only U.S. economic data had $R^2 = 0.19$. Using U.S. economic data along with Chinese economic data seems to have bigger impacts; in the factors model on 2016-2024 data, it increased the R^2 from 0.5948 to 0.6728 (see Table 4.1.1).

Table 4.1.1: Chinese Returns Regression using U.S. and Chinese Factors

Time Period	Granularity	Predictors	Adjusted R^2	R^2
2016–2024	Daily	usa_factors	0.0815	0.1910
2016–2024	Daily	chn_factors	0.5599	0.5948
2016–2024	Daily	both_factors	0.5918	0.6728

The fitted models typically had four types of Q-Q and residual plots. In the first case, the residual plots (see Figure B.0.9 for an example) seemed normally distributed and the Q-Q plots (see Figure B.0.10) displayed some deviations towards the extremes in the data.

In the second case, the residual plots (see Figure 4.1.1 for an example) seemed Normally distributed, but there were clear linear boundaries in the residuals, and the Q-Q plot (see Figure B.0.11) displayed some greater deviations towards the ends of the data. The residual plots' trends suggest that there might be an underlying constraint or structure in the data that isn't fully captured by the model, potentially as a result of a transformation.

In the third case, the residual plots (see Figure 4.1.2 for an example) showed clear trends in the data and the Q-Q plot (see Figure B.0.12) only showed issues on the most extreme data points. The residual plots' trends suggest that the linear assumption in the data might be very limiting, and that the underlying data might not

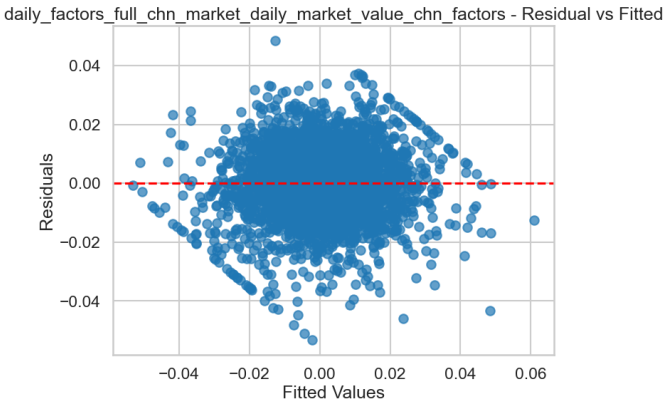

Figure 4.1.1: Case 2: Residual Plot

follow the necessary structure.

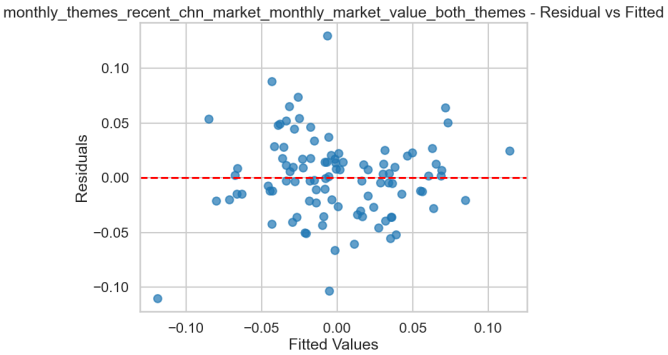

Figure 4.1.2: Case 3: Residual Plot

In the fourth case, the residual plots (see Figure 4.1.3 for an example) showed signs of increasing variance over time and the Q-Q plot (see Figure B.0.13) displayed some deviations towards the ends of the data. The residual plots' trends suggest heteroscedasticity in the data, which can impact the reliability of the model. The Q-Q plot suggests that this is mainly at the extreme ends of the data.

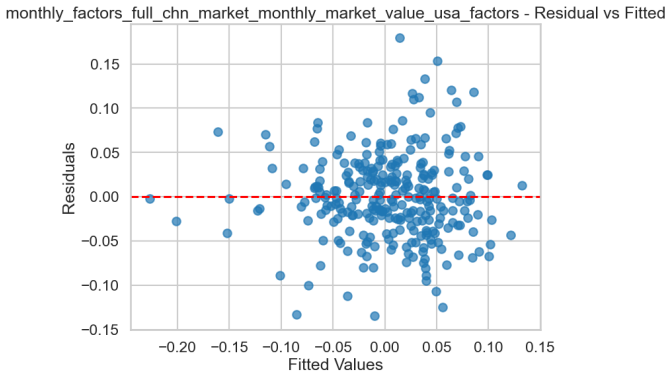

Figure 4.1.3: Case 4: Residual Plot

These plots suggest that the linear regressions are not capturing the full relationship between the data. Before moving onto models that do not assume linearity, let's assess regressions with regularization.

4.1.3 With Regularization

Using sklearn, I fit a total of 144 models using either ridge, lasso, or elastic net regularization. The standardscaler was used once again to ensure that everything is standardized; note that while this process might be redundant, it does not impact the data. Grid search cross-validation is used to tune the hyperparameters for each regression based on minimizing the MSE. The TimeSeriesSplit method was used to account for the temporal structure of the data, ensuring that the training and validation data split is somewhat uniform across time.

The best model of the three regularization methods was used. Of the 48 models, Ridge was the best model by R^2 for 45; Lasso was the best for two and Elastic Net was the best for one. Similar to the previous regression section, the results are split between those predicting U.S. returns (see Table A.0.3) and those predicting Chinese

returns (see Table A.0.4).

We note no significant gain was added to the R^2 value when predicting U.S. returns when Chinese predictors were also included. When Chinese predictors were used to predict U.S. returns, they had at most an R^2 of 0.1341; this was also a scenario in which adding the Chinese predictors to the U.S. predictors actually decreased the R^2 of the general model.

Table 4.1.2: Best Regularized Model for U.S. Returns

Time Period	Granularity	Predictors	Technique	Adjusted R^2	R^2
2001–2024	Daily	both_factors	Ridge	0.7385	0.7510
2001–2024	Daily	chn_factors	Ridge	0.0036	0.0272
2001–2024	Daily	usa_factors	Ridge	0.7378	0.7441

We see similar trends in the Chinese market regressions, in which the highest R^2 using only U.S. predictors was 0.2210. This was also a scenario in which adding the U.S. predictors to the Chinese predictors decreased the R^2 of the model. Generally, we also note lower R^2 values in the models than in those predicting U.S. returns.

Table 4.1.3: Best Regularized Model for Chinese Returns

Time Period	Granularity	Predictors	Technique	Adjusted R^2	R^2
2016–2024	Daily	both_factors	Lasso	0.4712	0.5761
2016–2024	Daily	chn_factors	Ridge	0.5162	0.5545
2016–2024	Daily	usa_factors	Ridge	-0.0105	0.1100

A few trends were exhibited in the regularized models. There were trends similar to Case 1 from above, in which the residual plot seemed scattered with normality holding for most of the data, tapering off towards the ends of the data. There were many cases similar to Case

2, in which the residual plot had some linear boundaries. Interestingly, those were sometimes accompanied with far worse Q-Q plots (see Figure 4.1.4)) in which most of the data was not normal and the outliers were very extreme. This suggests that the residuals are far from a normal distribution, showing violations in the assumption of a normally distributed error term.

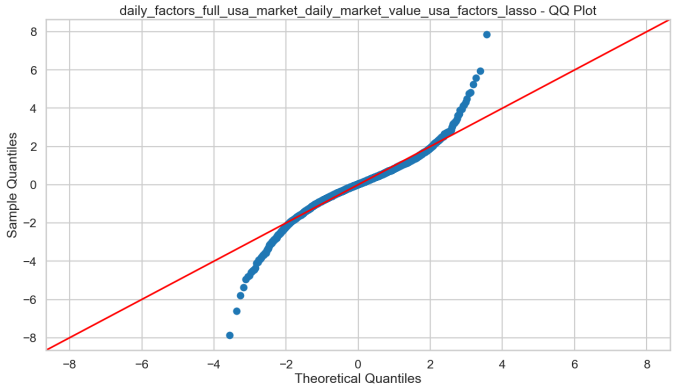

Figure 4.1.4: Curved Q-Q Plot

One model's residual plot exhibited linear constraints from all four edges, forming an outlining square (see Figure 4.1.5)). This was associated with curved Q-Q plots that generally followed the theoretical quantiles (see Figure B.0.16).

These plots motivate the next attempts with Sparse Additive Models (SpAMs), which do not assume a linear relationship between the predictors and the returns. Note that it does assume an additive effect, not accounting for interactions between the themes.

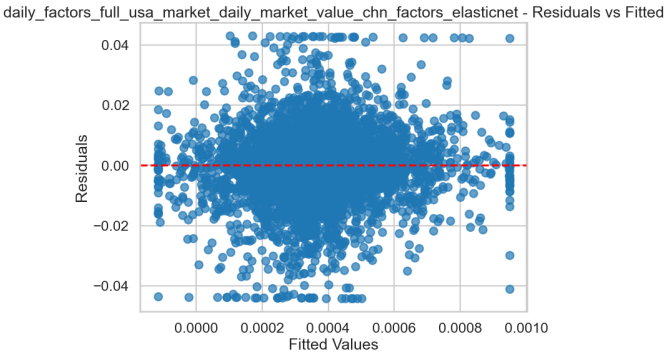

Figure 4.1.5: Square Outline in Residual Plot

4.2 Sparse Additive Models (SpAMs)

Using the pyGAM package, I ran 24 SpAMs on the theme data. Each predictor is assigned a spline term with 10 splines. The additive model is the sum of the spline terms. It is fitted using grid search to find the optimal parameters.

Note that I decided to run the SpAMs on only the theme data as one of the main advantage of a SpAM is to determine the marginal relationship between a predictor and the outcome variables. The factors are too granular to provide a strong economic interpretation unless there is a specific set of factors being investigated.

The R^2 and Adjusted-R^2 measures for the SpAM models for U.S. market returns are shown in Table 4.2.1. We note that the Chinese themes alone were able to explain at most 6.33% of the variation in the U.S. market returns, yet again hinting at a lack of signaling from the Chinese to U.S. markets. We note that the model fit on 2016-2024 data using both themes yielded an R^2 of 0.9999; while SpAM does integrate some form of regularization, this is likely a case of overfitting.

Observing some of the partial effects of the themes in the model

Table 4.2.1: SpAM Regressions Predicting U.S. Market Returns

Time Period	Granularity	Predictors	Adjusted R^2	R^2
2001–2024	Daily	both_themes	0.5949	0.6032
2001–2024	Daily	chn_themes	0.0078	0.0120
2001–2024	Daily	usa_themes	0.5970	0.6022
2016–2024	Daily	both_themes	0.5507	0.5687
2016–2024	Daily	chn_themes	0.0123	0.0222
2016–2024	Daily	usa_themes	0.5579	0.5714
2001–2024	Monthly	both_themes	0.6140	0.6538
2001–2024	Monthly	chn_themes	0.0100	0.0633
2001–2024	Monthly	usa_themes	0.6127	0.6359
2016–2024	Monthly	both_themes	0	0.9999
2016–2024	Monthly	chn_themes	0.0196	0.0459
2016–2024	Monthly	usa_themes	0.4725	0.5723

support our speculations of the model being overfit. This is frequently detected by very wavy curves that do not follow economic intuition. The marginal effect for Chinese theme for debt issuance, seen in Figure 4.2.1, is very wavy in a way that does not match economic intuition. You would expect that any relationship between Chinese debt issuance to not be so wavy; according to this model, changing debt issuance by -1.2, -0.3, 0.3, 0.8, and 1.4 all have the same zero-effect.

The best model for U.S. market returns by R^2 is thus the model trained on 2001 to 2024 using U.S. and Chinese themes, with $R^2 = 0.6538$. Most of the partial dependence plots of the themes follow a linear structure, with the exception of Chinese Value and U.S. Momentum.

Chinese value (see Figure 4.2.2) is shown to have a strongly curved decreasing relationship, in which negative changes in Chinese Value estimates are estimated to be related to raises in the U.S. market.

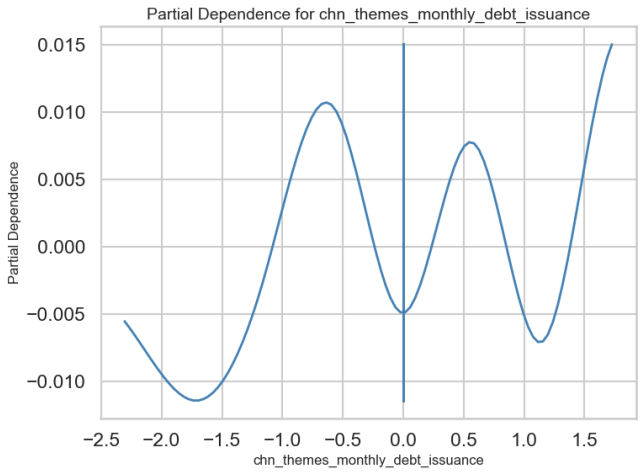

Figure 4.2.1: Partial Dependence of Chinese Debt Issuance on Overfit SpAM Model

Recall that Value measures how many stocks are trading at lower prices relative to their values of fundamental measures. As stocks in China become overvalued, it makes sense that investors might shift their portfolios to U.S. stocks. On the other hand, increases in Chinese Value are related to diminishing effects on the U.S. market; this makes sense when you consider how much easier it has historically been to invest in the U.S. market as a foreigner than in the Chinese market. Still, we see a negative impact when Chinese value is increased, which aligns with economic intuition.

U.S. Momentum is expected to have a positive relationship with U.S. market returns. The model does not show such a relationship (see Figure 4.2.3). Instead, there is a negative estimated relationship that contradicts economic interpretation. This could be a result of an issue with model specification, since SpAMs assume there is no interaction between the themes and can have erroneous estimations of the dependence. Still, there are cases in which a partially negative

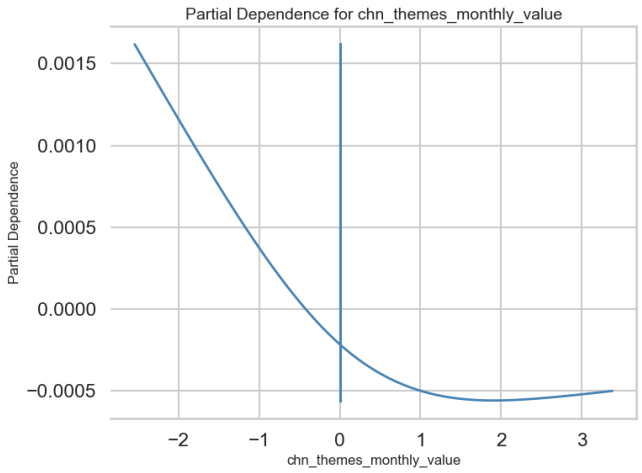

Figure 4.2.2: Partial Dependence of Chinese Value on U.S. SpAM Model

relationship can make sense, such as in cases of high returns reverting to the mean or in cases where the general market trend is negative.

Looking at the SpAMs for Chinese market returns, we see that the same model that overfit the U.S. market has an $R^2 = 0.9999$, which is likely also overfit. Generally, we see less variation in Chinese market returns being explained by the theme data, with $R^2 < 0.50$ across all models. The best fitting model has $R^2 = 0.4968$ and uses both themes and is trained on 2001 to 2024. Interestingly, we've seen this trend both times; it seems that introducing more data in these cases is far more valuable than segmenting the data due to detected structural breaks.

The Chinese model shows many more curves in the factors. In particular, we will examine Chinese Accruals, U.S. Profit Growth, and Chinese Momentum. Recall that the Accruals theme measures the difference between a company's reported earnings and its cash flows; lower values signal healthier earnings.

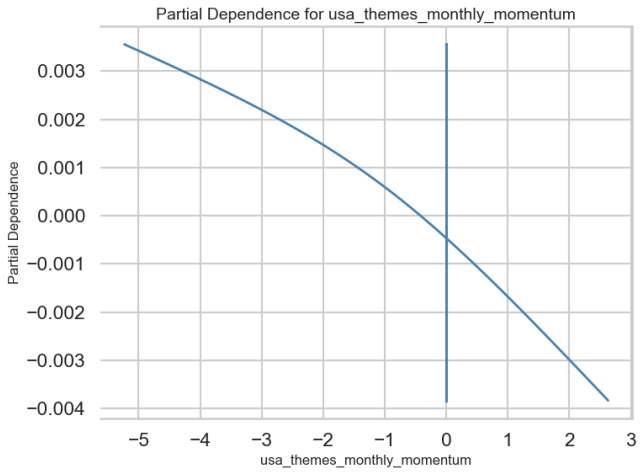

Figure 4.2.3: Partial Dependence of U.S. Momentum on U.S. SpAM Model

Table 4.2.2: SpAM Regressions Predicting Chinese Market Returns

Time Period	Granularity	Predictors	Adjusted R^2	R^2
2001–2024	Daily	both_themes	0.4357	0.4421
2001–2024	Daily	chn_themes	0.4278	0.4324
2001–2024	Daily	usa_themes	0.0173	0.0212
2016–2024	Daily	both_themes	0.3906	0.4038
2016–2024	Daily	chn_themes	0.3678	0.3763
2016–2024	Daily	usa_themes	0.0502	0.0618
2001–2024	Monthly	both_themes	0.4090	0.4968
2001–2024	Monthly	chn_themes	0.3903	0.4564
2001–2024	Monthly	usa_themes	0.0553	0.1055
2016–2024	Monthly	both_themes	0	0.9999
2016–2024	Monthly	chn_themes	0.0036	0.0053
2016–2024	Monthly	usa_themes	0.0752	0.1173

Chinese Accruals (see Figure 4.2.4) have a negative relationship with the Chinese market. This matches our economic intuition. Interestingly, it seems as though the theme has increasing effects as the value increases; this matches our intuition of investors detecting signals and market behavior.

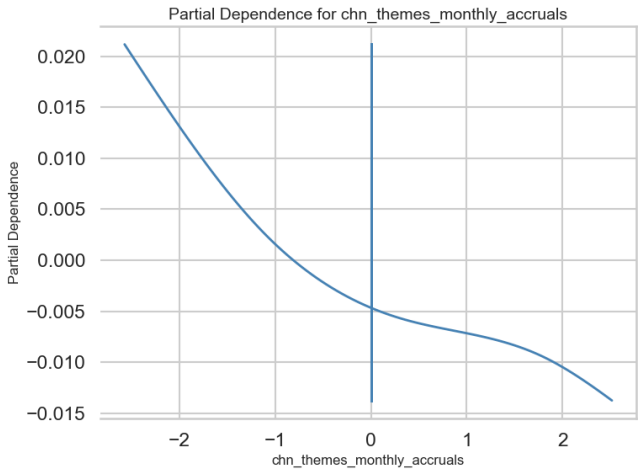

Figure 4.2.4: Partial Dependence of Chinese Accruals on Chinese SpAM Model

U.S. Profit Growth (see Figure 4.2.5) also has a negative relationship with the Chinese market. Larger values of negative profit growth in the U.S. is related to increasing Chinese market returns; this likely is a result of investors diversifying their portfolios into the Chinese market. Even more interestingly, even at 0, the partial dependence of U.S. profit growth has a negative effect on the Chinese returns.

Chinese Momentum (see Figure B.0.17) also has a negative relationship with the Chinese market. This is interesting as the previous SpAM model also detected such a relationship. While this

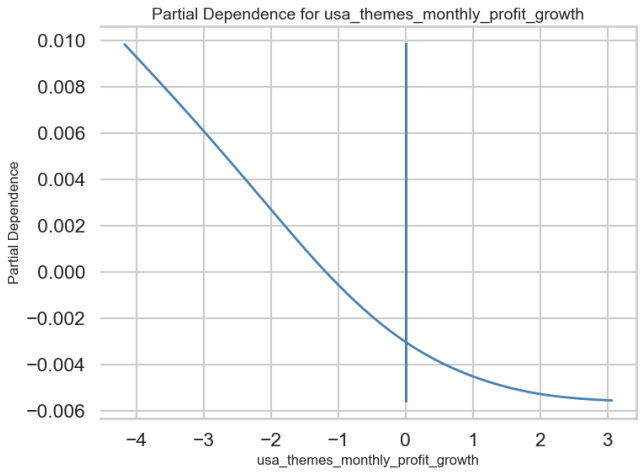

Figure 4.2.5: Partial Dependence of U.S. Profit Growth on Chinese SpAM Model

doesn't match basic economic intuition, it poses an interesting trend in the data.

I also ran 24 other SpAMs on the factor data (see Table A.0.5 for results). The monthly data with the factors all overfit to the data, showing an $R^2 = 1$ and having more factors than observations in the dataset. Similar trends occurred in cases where daily data was used for the 2016-2024 time period. Of the models that used the 2001-2024 data at a daily granularity, the U.S. factors had an $R^2 = 0.7900$ on the U.S. market and the Chinese factors had an $R^2 = 0.6021$ on the Chinese market. The results for the 2001-2024 models using the daily data are also shown in Table 4.2.3.

The SpAM models we've fit have shown nonlinear relationships among the predictors, but they assume an additive relationship and do not account for interactions between the terms. Kernel regressions are explored next as a nonparametric way to explore the dependencies between themes and market returns.

Table 4.2.3: 2021-2024 Factor-Based SpAM Regression Summary

Time Period	Granularity	Target Market	Predictors	Adjusted R^2	R^2
2001–2024	Daily	USA	usa_factors	0.7768	0.7900
2001–2024	Daily	USA	chn_factors	0.0216	0.0565
2001–2024	Daily	USA	both_factors	0.7606	0.7767
2001–2024	Daily	China	usa_factors	0.0564	0.0871
2001–2024	Daily	China	chn_factors	0.5815	0.6021
2001–2024	Daily	China	both_factors	0.5874	0.6152

4.3 Kernel Regression

Kernel regression assumes nothing about the form of the relationship between the predictors and the outcomes, making it the most flexible of all models used so far.

Using statsmodels' KernelReg with a Gaussian kernel and the Nadaraya-Watson estimator, I fit 24 kernel regressions total on the U.S. and Chinese markets using only the themes data. The themes data was chosen to increase interpretability of the model. Of the 24 models, all 12 using data at the monthly granularity overfit (see Table A.0.6). Of the 6 models assigned to each countries' returns, 2 only include the U.S. themes, 2 only include the China themes, and 2 include both. The results of the models are shown in Table 4.3.1.

Recall that the bandwidth, which represents the length of the boundary by which points are designated as neighboring, is a very important hyperparameter in kernel regressions. A very small bandwidth causes the model to overfit and a very large bandwidth causes the model to underfit. To decrease the search space for datasets at the daily granularity, bandwidth selection was done by randomly choosing 25% of the data and using that subsample to determine the optimal bandwidth using cross validation. This is likely

Table 4.3.1: Kernel Regression Summary

Time Period	Market	Data Used	# Predictors	Adjusted R^2	R^2
2001–2024	China	both_themes	26	0.9836	0.9837
2001–2024	China	chn_themes	13	0.9041	0.9043
2001–2024	China	usa_themes	13	0.1101	0.1122
2001–2024	USA	both_themes	26	0.9913	0.9914
2001–2024	USA	chn_themes	13	-0.0024	4.6e-15
2001–2024	USA	usa_themes	13	0.8821	0.8824
2016–2024	China	both_themes	38	0.9992	0.9992
2016–2024	China	chn_themes	15	0.8505	0.8516
2016–2024	China	usa_themes	23	0.9056	0.9067
2016–2024	USA	both_themes	38	0.9998	0.9998
2016–2024	USA	chn_themes	15	0.7900	0.7916
2016–2024	USA	usa_themes	23	0.9523	0.9529

why the fifth model in Table 4.3.1 has such a low R^2. The R^2 would have likely been low as we've seen that Chinese themes very weakly explain the variability in U.S. returns and do so at a weaker trend than the vice versa, which had an $R^2 = 0.1122$.

Analyzing plots that show the predicted and actual results of each model, we draw important insights on how the models performed with the different information they were provided.

Looking at the kernel regressions fit on the Chinese market returns first, we see that the model was able to reproduce some of the general trends in the data using the Chinese themes only (see Figure 4.3.1). The model that used both the U.S. and Chinese themes performed even better, picking up on better trends in the data and making more accurate predictions (see Figure 4.3.2). This shows that there are some signals being picked up by the data, though their estimated functions might be overfit.

Most interesting is the kernel regression on the Chinese market that

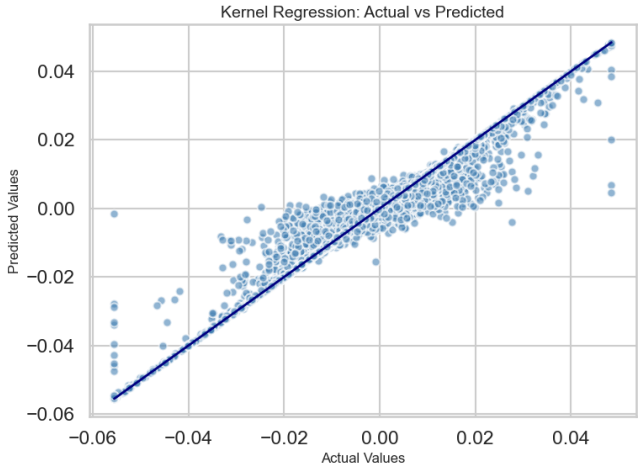

Figure 4.3.1: Chinese Market Kernel Regression (Chinese Themes): Actual vs Predicted

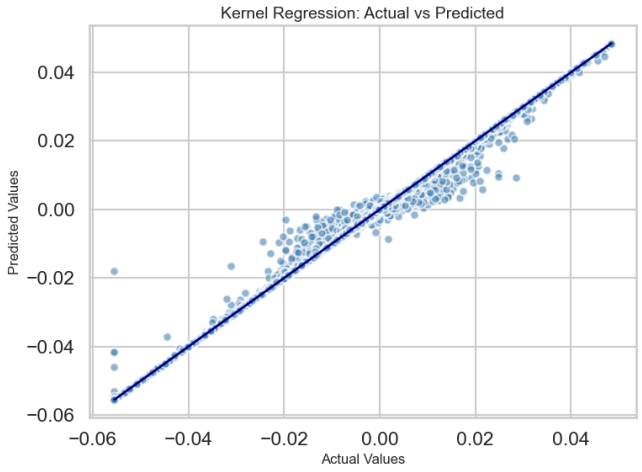

Figure 4.3.2: Chinese Market Kernel Regression (Chinese and U.S. Themes): Actual vs Predicted

uses only U.S. themes, shown in Figure 4.3.3. Even with the economic themes from the U.S., the model frequently predicted values very close to 0, even when the actual values were quite far. We note that some of the trends are picked up by the signals; we see that points tend to be estimated somewhere between 0 and the actual value for estimates within [-0.02, 0.02].

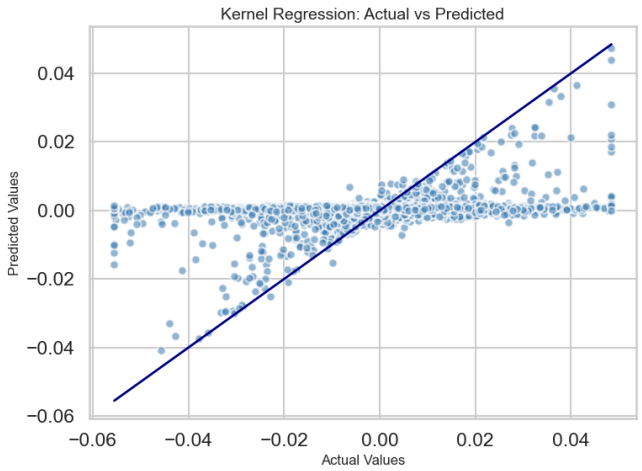

Figure 4.3.3: Chinese Market Kernel Regression (U.S. Themes): Actual vs Predicted

The U.S. market exhibited similar trends in the kernel regressions where it only had U.S. themes and themes from both countries (see Figure B.0.18 and Figure B.0.19). The most interesting portion of this analysis was when looking at the kernel regression on the U.S. market that only had access to the Chinese economic themes. The model consistently predicted 0 for returns (see Figure 4.3.4), indicating no signal being picked up from the Chinese themes.

We know this to be consistent with our other models. Generally, the Chinese market is related to changes in signals from the U.S.

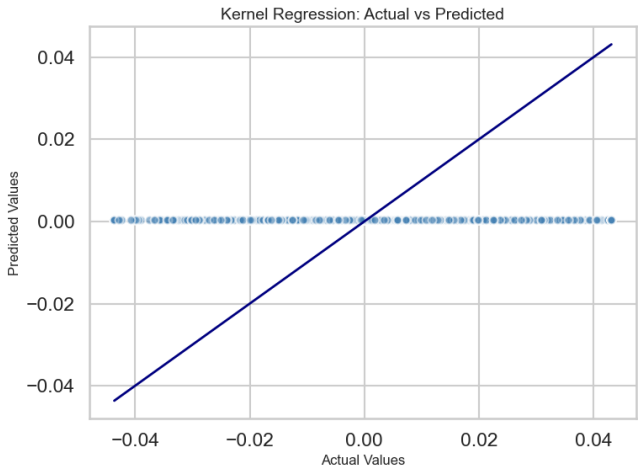

Figure 4.3.4: U.S. Market Kernel Regression (Chinese Themes): Actual vs Predicted

economy at a rate much higher than the U.S. market's equivalent with the Chinese economy.

Analyzing the Chinese kernel model that used both Chinese and U.S. economic themes over the 2001 to 2024 time period, we notice a case of potentially conflicting signals being picked up from the U.S. market from profitability and profit growth, several curves that do not align with economic intuition, and other curves that do align with economic intuition.

Of the curves that do match economic intuition, we notice a few trends that have opposite directions in the data. Chinese value has a positive linear relationship with the Chinese returns, shown in Figure B.0.20. An opposite reaction is shown in the U.S. value, which is negative relationship with Chinese returns, shown in Figure B.0.21. These match our economic intuition. The Chinese market returns should increase when Chinese companies have more value and decrease when U.S. companies have more value.

The partial dependence plots for U.S. profit growth and profitability were also conflicting. Chinese market returns had a smooth negative relationship with U.S. profitability (see Figure 4.3.5), which matches our economic intuition. However, the relationship with U.S. profit growth is seen as a generally positive one, with a negative effect in [-2, 1] (see Figure 4.3.6). Although this seems to be potentially conflicting information, it does not go against our economic intuition. Cases in which U.S. companies are either experiencing tremendous change can be caused by significant global situations that impact both countries.

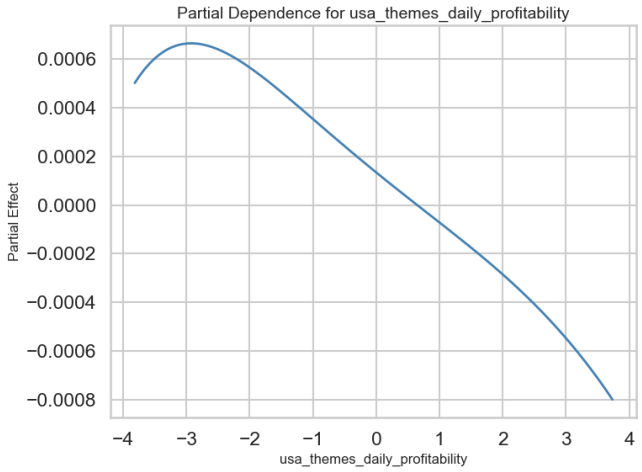

Figure 4.3.5: Partial Dependence for Chinese Kernel Regression: U.S. Profitability

There were also two cases of potential overfitting. The U.S. Accruals and U.S. Low Leverage Growth showed rocky curves with un-intuitive trends towards the extremes of the data, see Figure 4.3.7 and Figure B.0.22.

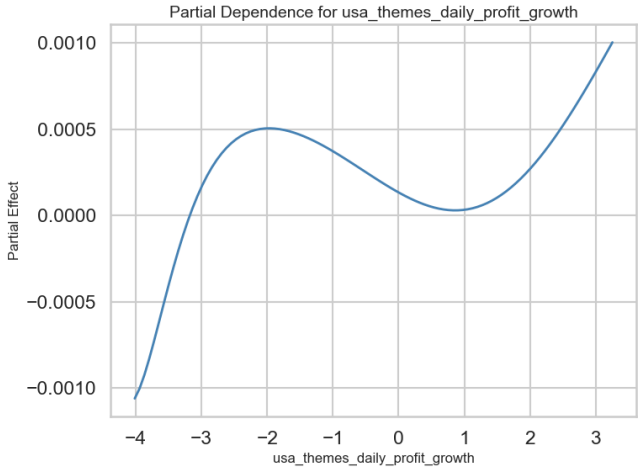

Figure 4.3.6: Partial Dependence for Chinese Kernel Regression: U.S. Profit Growth

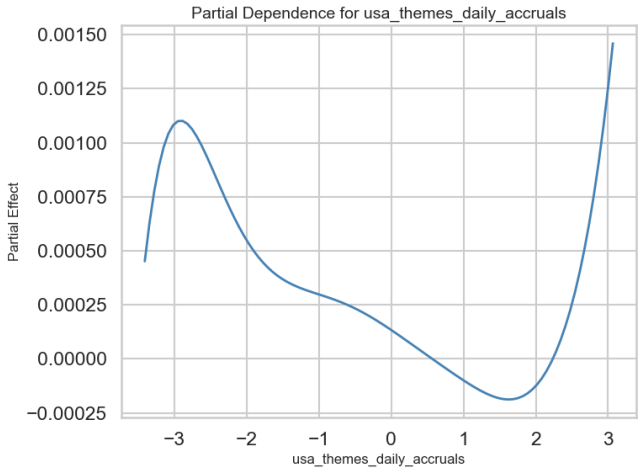

Figure 4.3.7: Partial Dependence for Chinese Kernel Regression: U.S. Accruals

5

Discussion

Reviewing the trends in the models, we recap several important trends about the methodologies used and their results. The linear regressions and regularized regressions achieved moderate R^2 and Adjusted-R^2 values for predicting U.S. returns using only U.S. economic data and for predicting Chinese returns using only Chinese economic data, achieving $0.70 \leq R^2 \leq 0.80$ for the U.S. market and $0.55 \leq R^2 \leq 0.60$ for the Chinese market. While the Chinese economic data gave no additional explanatory power to the U.S. market models, the U.S. economic data more consistently improved the predictions of Chinese returns. Still, there were some issues with the linear models, with residual analyses revealing heteroskedasticity and otherwise suggesting nonlinearity in the factors.

The SpAM models often saw similar R^2 and Adjusted-R^2 values,

showing that the generalization beyond linearity did not hurt the models ability to pick up on trends. In some cases, the R^2 and Adjusted-R^2 values were even higher; however, the model was more prone to overfitting on the data, especially in cases where the time frames were short or when using larger number of predictors. The model suffered from the lack of interpretability as well, where a plot was required to understand the relationship between an economic measure and the returns.

The kernel regression models often had very high R^2 and Adjusted-R^2 scores, indicating that the model could almost replicate the in-sample data. This was, in a lot of cases, the result of severe overfitting. Still, there were some interesting results captures in some of the economic curves, and even more interesting results in cross-country prediction. Though consistent with previous approaches, cross-country prediction trends were very interesting to examine in the nonparametric model. The partial dependence plots revealed some interesting trends in the factors that aligned with economic intuition and were very helpful in assessing overfitting.

As far as general trends within the results, domestic economic data was the strongest predictor for both the U.S. and Chinese market returns. The Chinese market gained more explanatory power when U.S. economic data was introduced, which supports the view that global markets receive stronger signals from the U.S. market than in the opposite direction.

The models with the least assumptions also had a strong tendency to overfit to the data, especially in scenarios where there were limited samples and a larger number of predictors. While logical in principle, this posed a great challenge when analyzing market data, especially when segmenting for structural breaks in the data.

Throughout the models, there is a trade off between flexibility and the ease of interpretation. Linear models worked really well, achieving

high R^2 values while maintaining a good level of easy interpretability. As the models became less structures, and thus more capable of handling more complex data, they also became more difficult to interpret. The partial effect plots in the SpAMs and kernel regressions also suffered from interpretability as it makes it even harder to understand how a factor is impacted with other factor movement.

When using less-structured models, it is more difficult to get enough data without breaking the stationarity assumption in the data. Assessments of structural breaks in the market yielded several scenarios over a few decades where the underlying structural factors had shifted behind the data. This is a significant challenge when attempting to understand the intricacies of the market.

There are several ways to build off this work. The first next step is to introduce models that can naturally capture interactions among the economic factors without requiring predefined functional forms, perhaps with tree-based methods. Alternatively, using the several decades of data available to better estimate the GAMs and SpAMs to allow for pairwise interactions. In order to deal with structural breaks in the data, one can try to implement indicator variables in their analyses. If attempting to work on SpAMs or kernel regressions, more nuanced regularization can help mitigate the large amount of overfitting seen in the models.

Another natural first step would be to attempt the same methods on rolling-window estimates, which might capture the trends of the data better and reduce the noise in the model estimates. Analyses in individual sectors might benefit from stronger and more interpretable insights as sector-specific spillovers are easier to detect, model, and understand.

The economic signals picked up can also be tested for their market power, creating specific portfolio strategies using the indicators from each country. This would be an interesting way to quantity the

strength of certain signals and the benefits of accounting for information from other markets.

The relationship between the U.S. and Chinese markets has been long studied by researchers. The empirical challenges of quantifying such a relationship have changed significantly over the past decade. Economic and financial data are now publicly accessible at a rate greater than ever before, allowing for exploration in the markets.

The research done in this thesis should serve as a stepping stone for testing for market integrations and cross-market signals. Mis-specified models and overfit partial effects provide more information about the models and about their relations to the markets. With global markets trending towards increasing co-integration, they will reveal more and more information about their intricacies as time goes by.

A

Tables

Table A.0.1: Regressions Predicting U.S. Market Returns

Time Period	Granularity	Predictors	Adjusted R^2	R^2
2001–2024	Daily	usa_factors	0.7465	0.7526
2001–2024	Daily	chn_factors	0.0164	0.0398
2001–2024	Daily	both_factors	0.7483	0.7603
2016–2024	Daily	usa_factors	0.6982	0.7342
2016–2024	Daily	chn_factors	0.0202	0.0977
2016–2024	Daily	both_factors	0.6978	0.7577
2001–2024	Daily	usa_themes	0.5784	0.5794
2001–2024	Daily	chn_themes	0.0074	0.0098
2001–2024	Daily	both_themes	0.5795	0.5815
2016–2024	Daily	usa_themes	0.5218	0.5274
2016–2024	Daily	chn_themes	0.0126	0.0201
2016–2024	Daily	both_themes	0.5224	0.5316
2001–2024	Monthly	usa_factors	0.7494	0.8670
2001–2024	Monthly	chn_factors	0.1197	0.4952
2001–2024	Monthly	both_factors	0.7699	0.9761
2016–2024	Monthly	usa_factors	0	1
2016–2024	Monthly	chn_factors	0	1
2016–2024	Monthly	both_factors	0	1
2001–2024	Monthly	usa_themes	0.6043	0.6227
2001–2024	Monthly	chn_themes	0.0104	0.0565
2001–2024	Monthly	both_themes	0.6092	0.6456
2016–2024	Monthly	usa_themes	0.6139	0.6724
2016–2024	Monthly	chn_themes	0.1886	0.3279
2016–2024	Monthly	both_themes	0.5822	0.7172

Table A.0.2: Regressions Predicting Chinese Market Returns

Time Period	Granularity	Predictors	Adjusted R^2	R^2
2001–2024	Daily	usa_factors	0.0490	0.0719
2001–2024	Daily	chn_factors	0.5509	0.5616
2001–2024	Daily	both_factors	0.5656	0.5864
2016–2024	Daily	usa_factors	0.0815	0.1910
2016–2024	Daily	chn_factors	0.5599	0.5948
2016–2024	Daily	both_factors	0.5918	0.6728
2001–2024	Daily	usa_themes	0.0139	0.0163
2001–2024	Daily	chn_themes	0.4168	0.4182
2001–2024	Daily	both_themes	0.4225	0.4252
2016–2024	Daily	usa_themes	0.0467	0.0578
2016–2024	Daily	chn_themes	0.3586	0.3635
2016–2024	Daily	both_themes	0.3825	0.3944
2001–2024	Monthly	usa_factors	0	1
2001–2024	Monthly	chn_factors	0	1
2001–2024	Monthly	both_factors	0	1
2016–2024	Monthly	usa_factors	0.1830	0.3068
2016–2024	Monthly	chn_factors	0.2571	0.3847
2016–2024	Monthly	both_factors	0.3444	0.5563
2001–2024	Monthly	usa_themes	0.0518	0.0960
2001–2024	Monthly	chn_themes	0.2677	0.3018
2001–2024	Monthly	both_themes	0.3147	0.3786
2016–2024	Monthly	usa_themes	0.1830	0.3068
2016–2024	Monthly	chn_themes	0.2571	0.3847
2016–2024	Monthly	both_themes	0.3444	0.5563

Table A.0.3: Best Regressions Predicting U.S. Market Returns by Regularization Type

Time Period	Granularity	Predictors	Technique	Adjusted R^2	R^2
2001–2024	Daily	both_factors	Ridge	0.7385	0.7510
2001–2024	Daily	chn_factors	Ridge	0.0036	0.0272
2001–2024	Daily	usa_factors	Ridge	0.7378	0.7441
2016–2024	Daily	both_factors	Ridge	0.6343	0.7069
2016–2024	Daily	chn_factors	Ridge	-0.0358	0.0462
2016–2024	Daily	usa_factors	Ridge	0.6494	0.6912
2001–2024	Daily	both_themes	Ridge	0.5795	0.5815
2001–2024	Daily	chn_themes	Ridge	0.0057	0.0081
2001–2024	Daily	usa_themes	Ridge	0.5783	0.5793
2016–2024	Daily	both_themes	Ridge	0.5224	0.5316
2016–2024	Daily	chn_themes	Ridge	0.0053	0.0129
2016–2024	Daily	usa_themes	Ridge	0.5218	0.5273
2001–2024	Monthly	both_factors	Ridge	-0.9439	0.7979
2001–2024	Monthly	chn_factors	Ridge	-0.6224	0.0696
2001–2024	Monthly	usa_factors	Ridge	0.6642	0.8219
2016–2024	Monthly	both_factors	Lasso	0	0.8773
2016–2024	Monthly	chn_factors	Ridge	0	0.1341
2016–2024	Monthly	usa_factors	Ridge	0	0.8919
2001–2024	Monthly	both_themes	Ridge	0.6088	0.6452
2001–2024	Monthly	chn_themes	Ridge	-0.0354	0.0128
2001–2024	Monthly	usa_themes	Ridge	0.6043	0.6227
2016–2024	Monthly	both_themes	Ridge	0.5781	0.7145
2016–2024	Monthly	chn_themes	Ridge	-0.1724	0.0289
2016–2024	Monthly	usa_themes	Ridge	0.6138	0.6723

Table A.0.4: Best Regressions Predicting Chinese Market Returns by Regularization Type

Time Period	Granularity	Predictors	Technique	Adjusted R^2	R^2
2001–2024	Daily	both_factors	Ridge	0.5313	0.5537
2001–2024	Daily	chn_factors	Ridge	0.5242	0.5354
2001–2024	Daily	usa_factors	Ridge	0.0308	0.0542
2016–2024	Daily	both_factors	Lasso	0.4712	0.5761
2016–2024	Daily	chn_factors	Ridge	0.5162	0.5545
2016–2024	Daily	usa_factors	Ridge	-0.0105	0.1100
2001–2024	Daily	both_themes	Ridge	0.4225	0.4252
2001–2024	Daily	chn_themes	Ridge	0.4168	0.4182
2001–2024	Daily	usa_themes	Ridge	0.0126	0.0149
2016–2024	Daily	both_themes	Ridge	0.3815	0.3935
2016–2024	Daily	chn_themes	Ridge	0.3580	0.3629
2016–2024	Daily	usa_themes	Ridge	0.0295	0.0409
2001–2024	Monthly	both_factors	Ridge	-4.5676	0.4213
2001–2024	Monthly	chn_factors	Ridge	-0.1556	0.3373
2001–2024	Monthly	usa_factors	Ridge	-0.6283	0.1362
2016–2024	Monthly	both_factors	Ridge	0	0.3703
2016–2024	Monthly	chn_factors	Elastic Net	0	0.3409
2016–2024	Monthly	usa_factors	Ridge	0	0.2210
2001–2024	Monthly	both_themes	Ridge	0.2404	0.3112
2001–2024	Monthly	chn_themes	Ridge	0.2179	0.2543
2001–2024	Monthly	usa_themes	Ridge	0.0335	0.0785
2016–2024	Monthly	both_themes	Ridge	-0.0357	0.2991
2016–2024	Monthly	chn_themes	Ridge	0.1933	0.3318
2016–2024	Monthly	usa_themes	Ridge	0.0271	0.1745

Table A.0.5: Factor-Based SpAM Regression Summary

Time Period	Granularity	Target Market	Predictors	Adjusted R^2	R^2
2001–2024	Daily	USA	usa_factors	0.7768	0.7900
2001–2024	Daily	USA	chn_factors	0.0216	0.0565
2001–2024	Daily	USA	both_factors	0.7606	0.7767
2001–2024	Daily	China	usa_factors	0.0564	0.0871
2001–2024	Daily	China	chn_factors	0.5815	0.6021
2001–2024	Daily	China	both_factors	0.5874	0.6152
2016–2024	Daily	USA	usa_factors	0.7246	0.7592
2016–2024	Daily	USA	chn_factors	0.0398	0.1235
2016–2024	Daily	USA	both_factors	0.8806	0.9931
2016–2024	Daily	China	usa_factors	0.1174	0.1727
2016–2024	Daily	China	chn_factors	0.5781	0.6148
2016–2024	Daily	China	both_factors	0.8638	0.9922
2001–2024	Monthly	USA	usa_factors	0	1
2001–2024	Monthly	USA	chn_factors	0	1
2001–2024	Monthly	USA	both_factors	0	1
2001–2024	Monthly	China	usa_factors	0	1
2001–2024	Monthly	China	chn_factors	0	1
2001–2024	Monthly	China	both_factors	0	1
2016–2024	Monthly	USA	usa_factors	0	1
2016–2024	Monthly	USA	chn_factors	0	1
2016–2024	Monthly	USA	both_factors	0	1
2016–2024	Monthly	China	usa_factors	0	1
2016–2024	Monthly	China	chn_factors	0	1
2016–2024	Monthly	China	both_factors	0	1

Table A.0.6: Month-Granularity Kernel Regression Summary

Time Period	Market	Data Used	# Predictors	Adjusted R^2	R^2
2001–2025	China	both_themes	26	0.9999	0.9999
2001–2025	China	chn_themes	13	0.9400	0.9428
2001–2025	China	usa_themes	13	0.4555	0.4809
2016–2025	China	both_themes	32	1.0000	1.0000
2016–2025	China	chn_themes	17	0.9975	0.9980
2016–2025	China	usa_themes	15	0.9910	0.9923
2001–2025	USA	both_themes	26	0.9999	0.9999
2001–2025	USA	chn_themes	13	0.7315	0.7440
2001–2025	USA	usa_themes	13	0.9678	0.9693
2016–2025	USA	both_themes	32	1.0000	1.0000
2016–2025	USA	chn_themes	17	0.9957	0.9965
2016–2025	USA	usa_themes	15	0.9983	0.9985

B

Figures

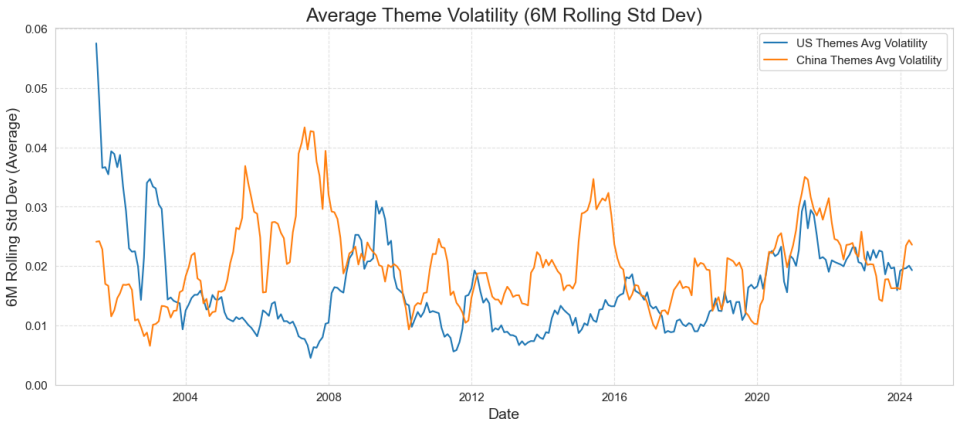

Figure B.0.1: Theme Volatility (6-Month Rolling Standard Deviation)

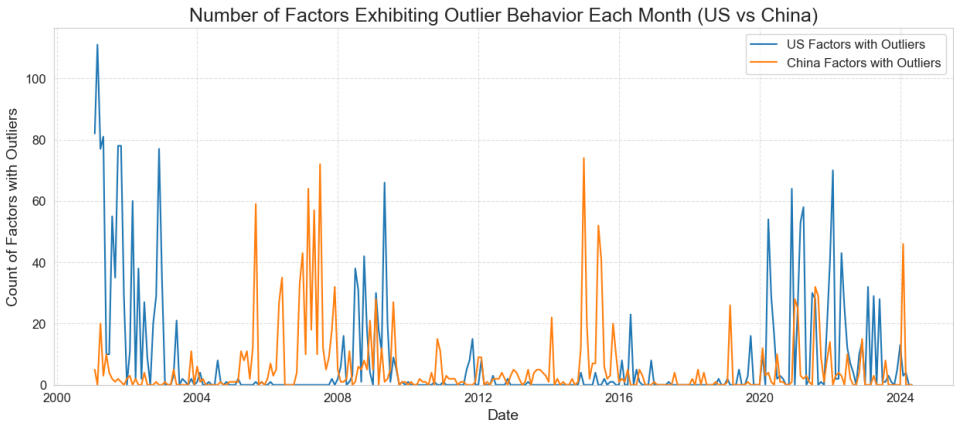

Figure B.0.2: Number of Factor Outliers Over Time

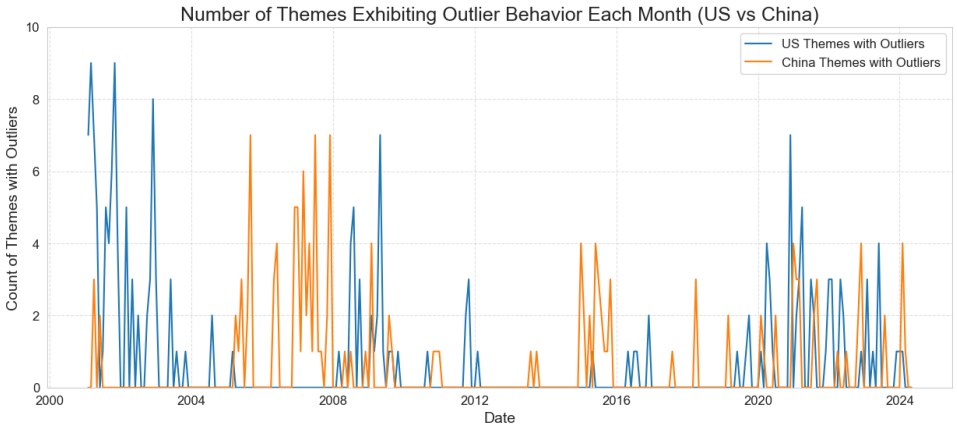

Figure B.0.3: Number of Theme Outliers Over Time

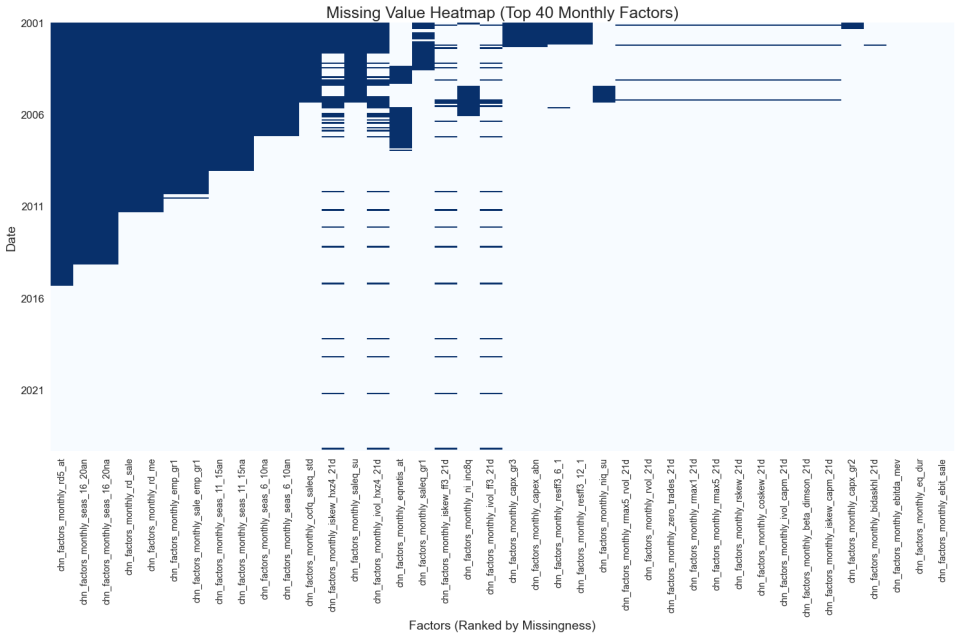

Figure B.0.4: Missing Factors Heatmap (Top 40 Factors, Monthly Data)

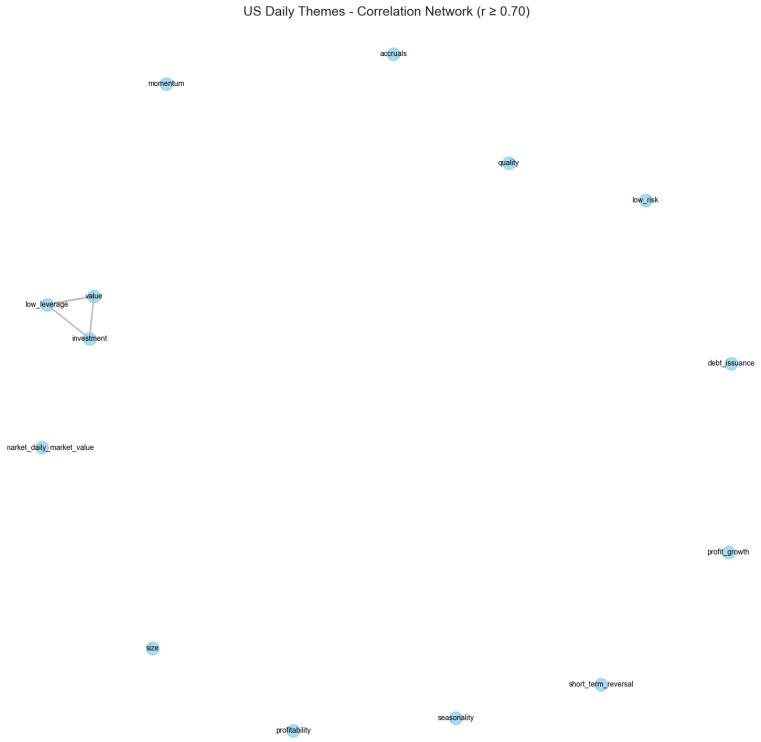

Figure B.0.5: U.S. Daily Themes Correlation Network

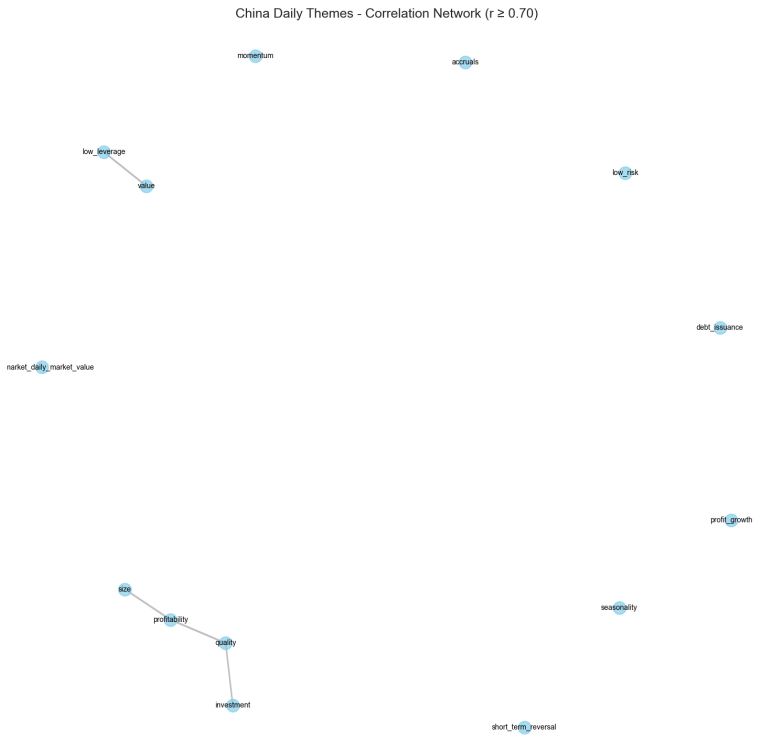

Figure B.0.6: China Daily Themes Correlation Network

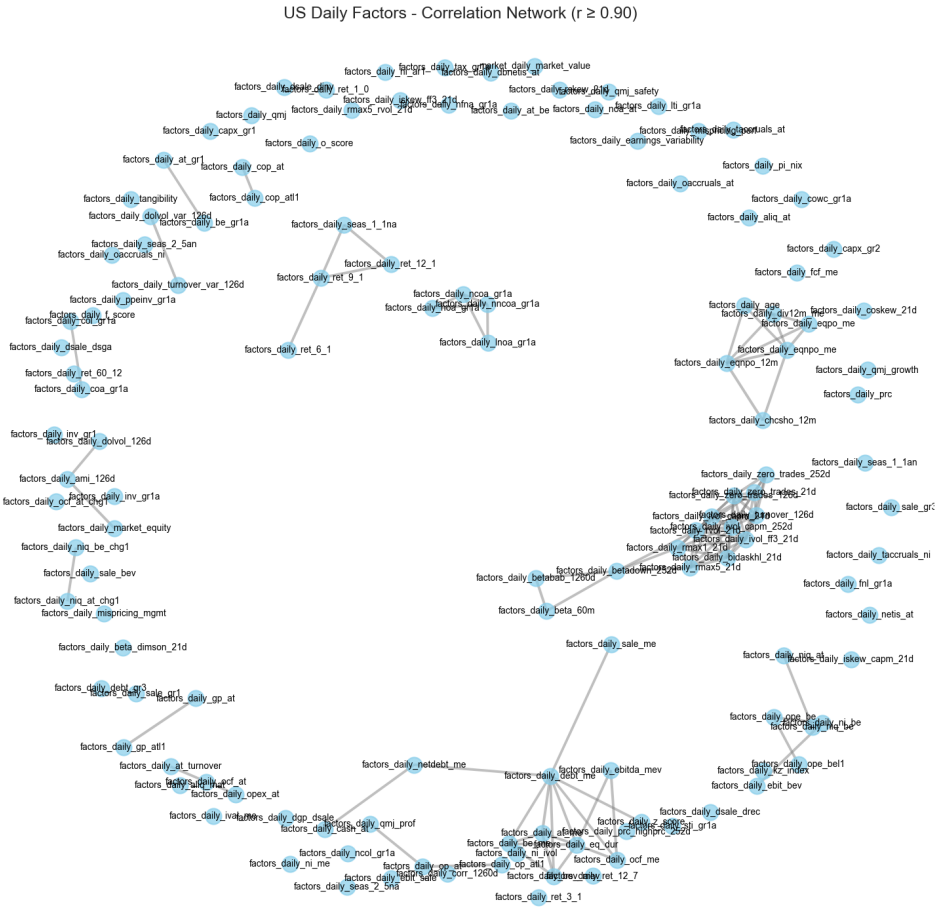

Figure B.0.7: U.S. Daily Factors Correlation Network

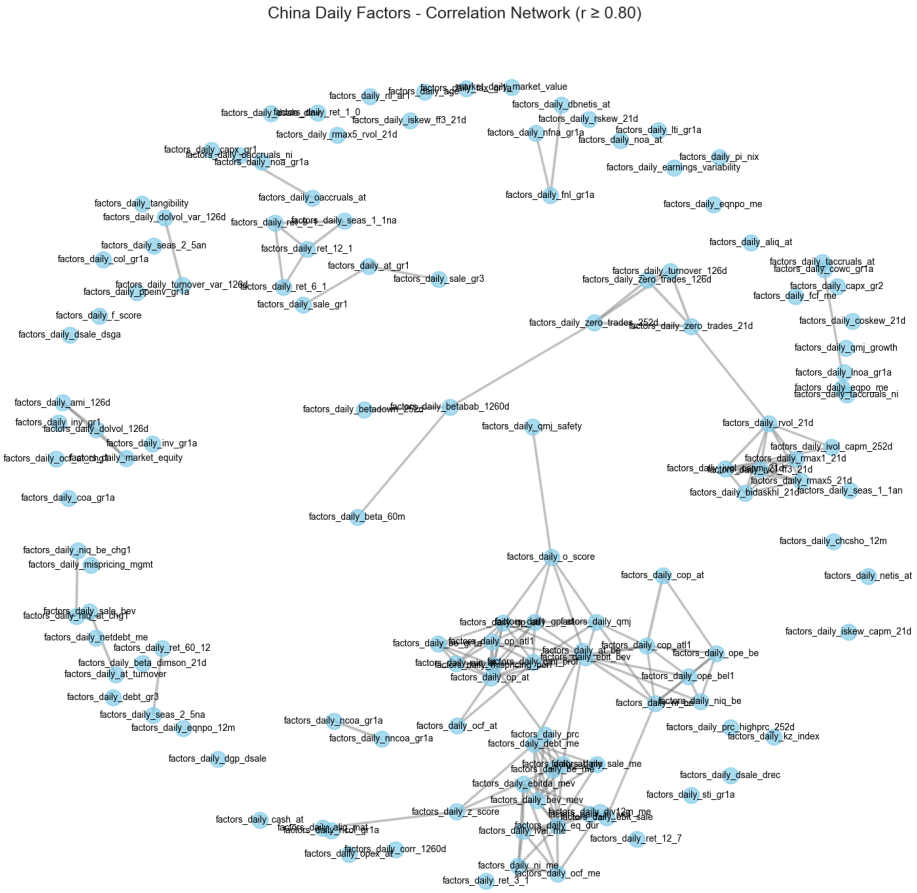

Figure B.0.8: China Daily Factors Correlation Network

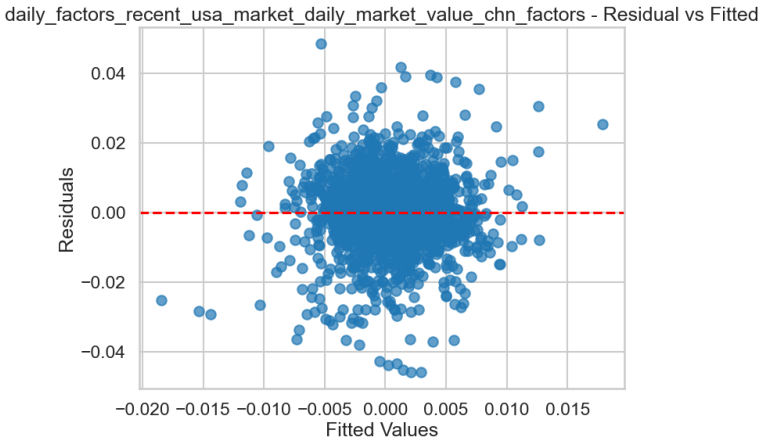

Figure B.0.9: Case 1: Residual Plot

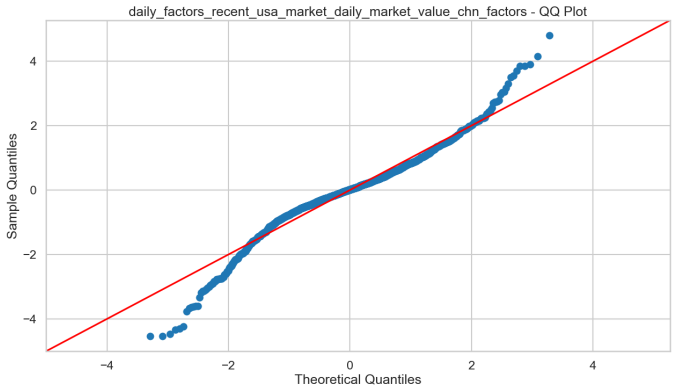

Figure B.0.10: Case 1: Q-Q Plot

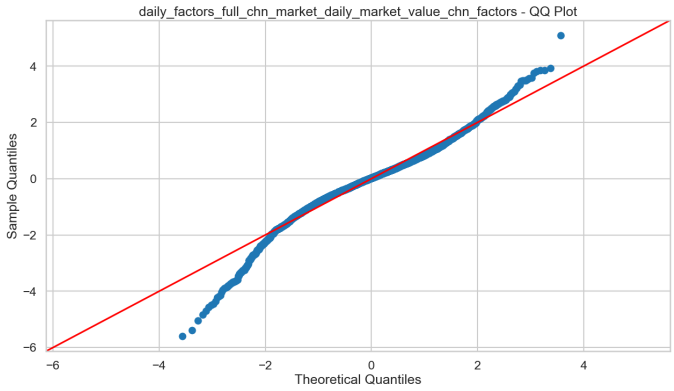

Figure B.0.11: Case 2: Q-Q Plot

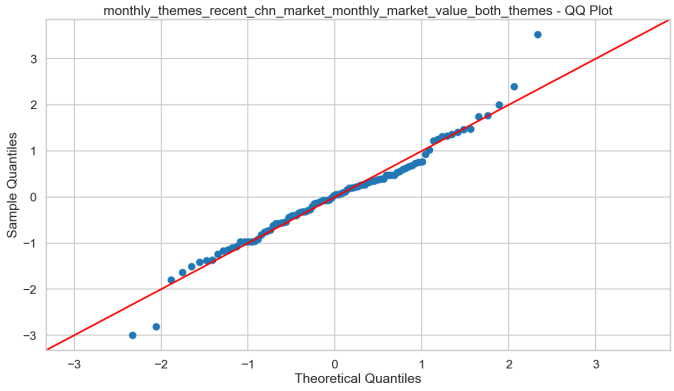

Figure B.0.12: Case 3: Q-Q Plot

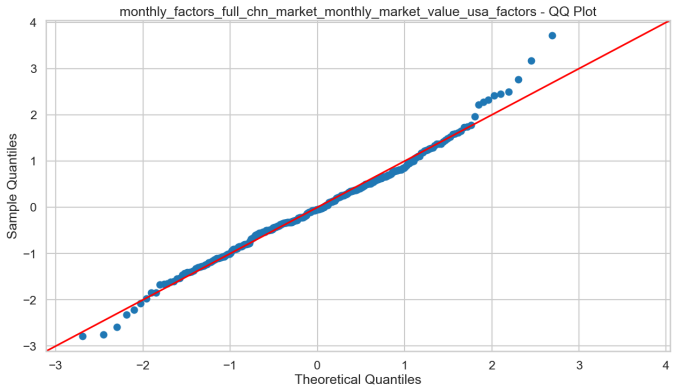

Figure B.0.13: Case 4: Q-Q Plot

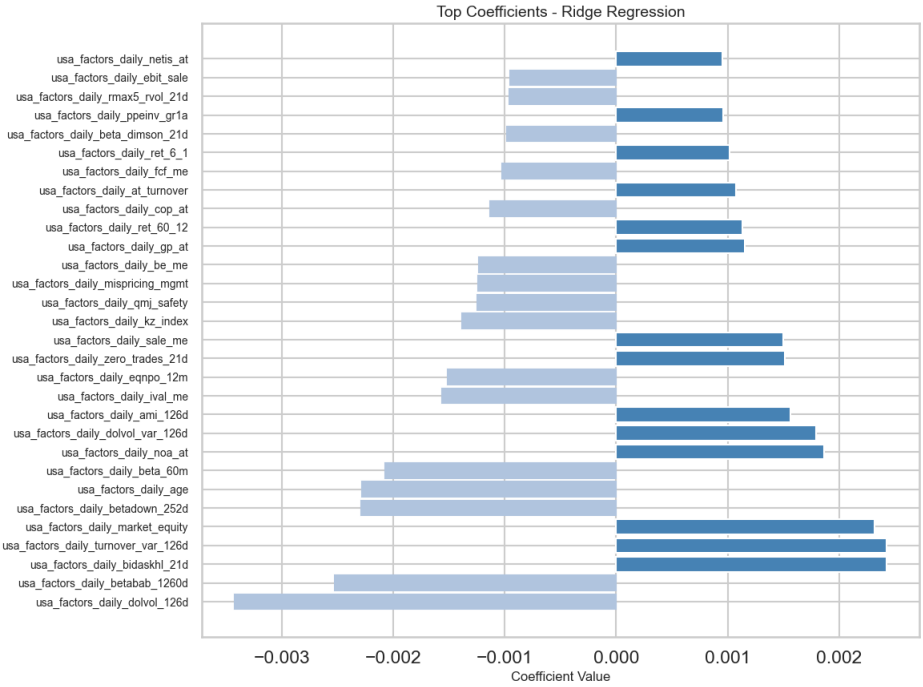

Figure B.0.14: Top Coefficients in Best U.S. Market Regularized Regression

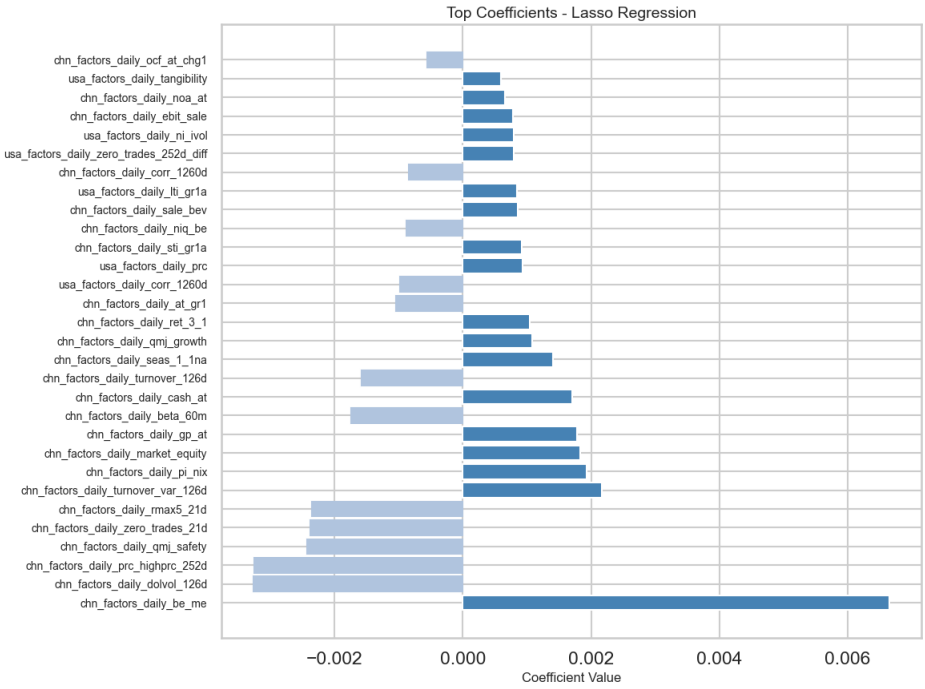

Figure B.0.15: Top Coefficients in Best Chinese Market Regularized Regression

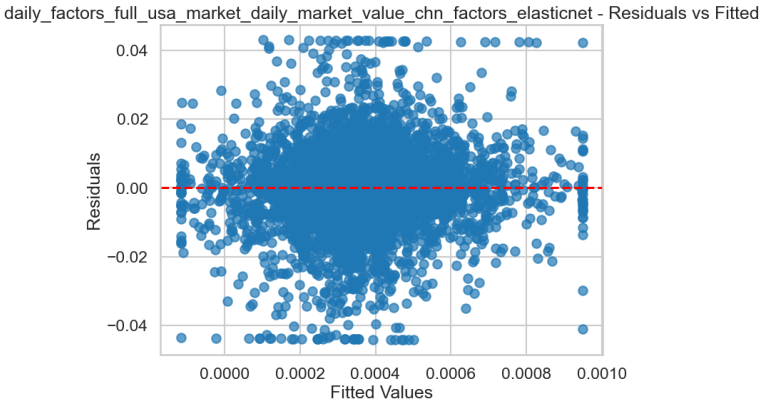

Figure B.0.16: Case 5: Q-Q Plot

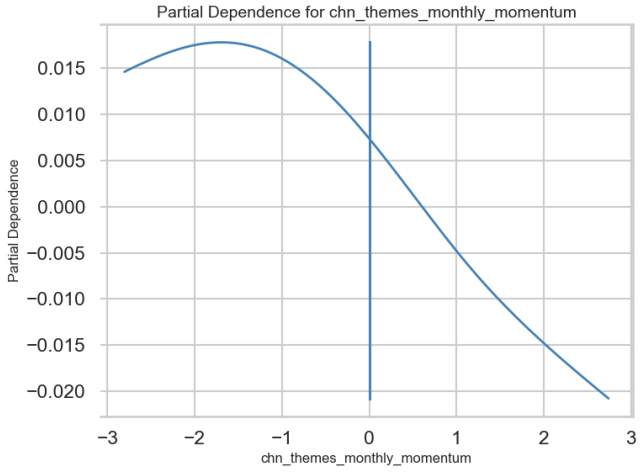

Figure B.0.17: Partial Dependence of Chinese Momentum on Chinese SpAM Model

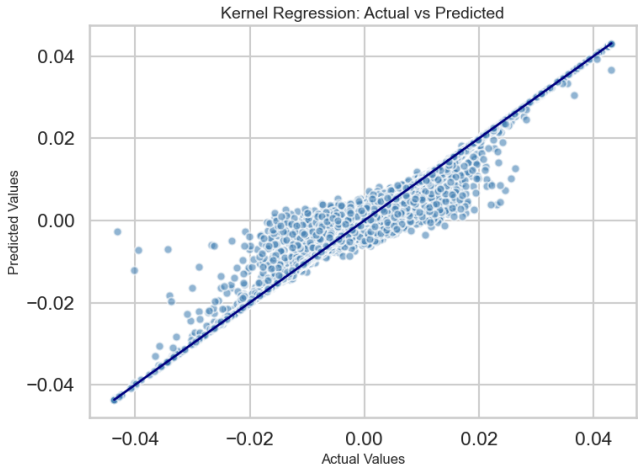

Figure B.0.18: U.S. Market Kernel Regression (U.S. Themes): Actual vs Predicted

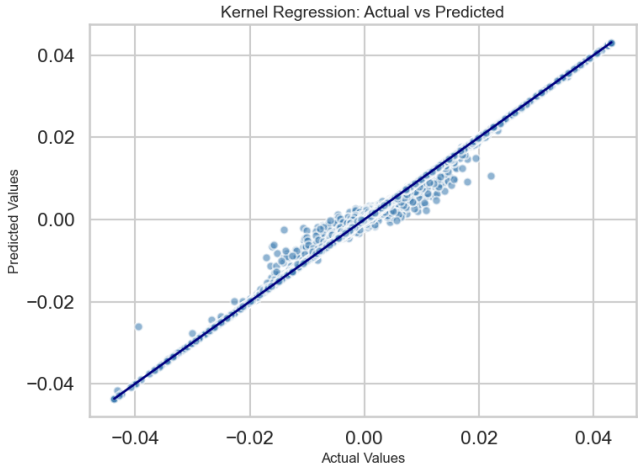

Figure B.0.19: U.S. Market Kernel Regression (U.S. and Chinese Themes): Actual vs Predicted

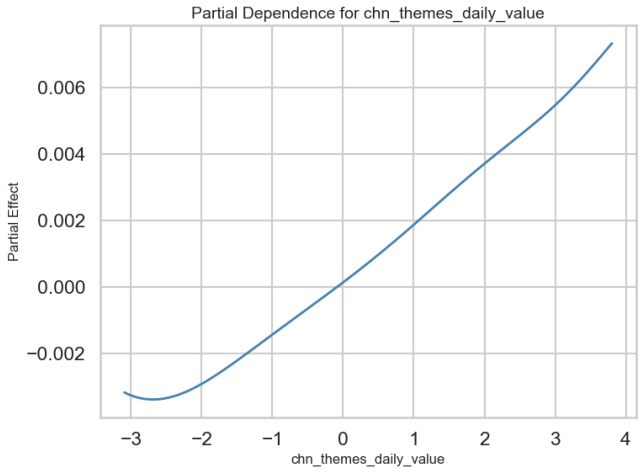

Figure B.0.20: Partial Dependence for Chinese Kernel Regression: Chinese Value

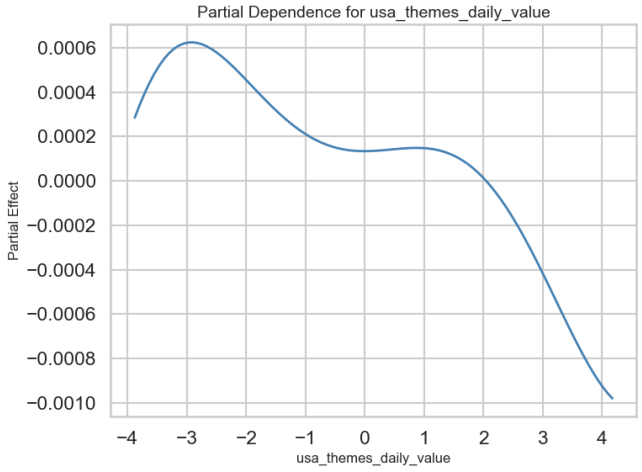

Figure B.0.21: Partial Dependence for Chinese Kernel Regression: U.S. Value

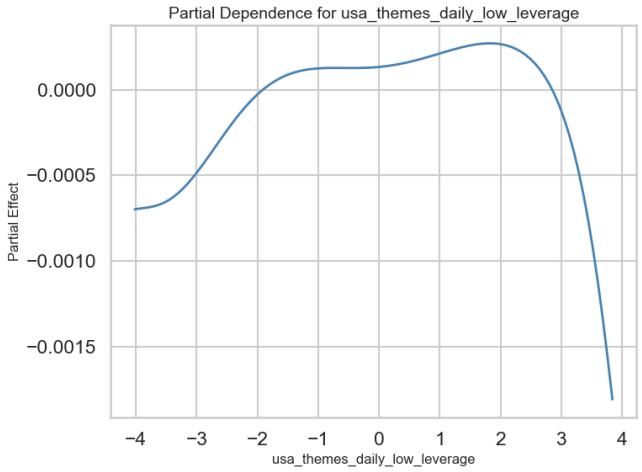

Figure B.0.22: Partial Dependence for Chinese Kernel Regression: U.S. Low Leverage Growth

References

[1] Mark M. Carhart. On persistence in mutual fund performance. *The Journal of Finance*, 2012.

[2] National Research Council. *Following the Money: U.S. Finance in the World Economy*. National Academies Press, 1995.

[3] Kenneth R. French Eugene F. Fama. The cross-section of expected stock returns. *The Journal of Finance*, 1992.

[4] Kenneth R. French Eugene F. Fama. A five-factor asset pricing model. *The Journal of Finance*, 2015.

[5] Campbell P. Harvey Geert Bekaert. Time-varying world market integration. *The Journal of Finance*, 1995.

[6] Dacheng Xiu Guanhao Feng, Stefano Giglio. Taming the factor zoo: A test of new factors. *The Journal of the American Finance Association*, 2020.

[7] Jun Tu Yuchen Wang Jeremy C. Goh, Fuwei Jiang. Can us economic variables predict the chinese stock market? *Pacific-Basin Finance Journal*, 2013.

[8] Sergio L. Schmukler Juan J. Cortina, Maria Soledad Martinez Peria and Jasmine Xiao. The international of china's

equity markets. A 2023 International Monetary Fund Working Paper.

[9] Junbo Wang Tengfei Zhang Kuntara Pukthuanthong, Richard Roll. A tool kit for factor-mimicking portfolios.

[10] Richard Roll Kuntara Pukthuanthong. Global market integration: An alternative measure and its application. *Journal of Financial Economics*, 2009.

[11] Harry Markowitz. Portfolio selection. *The Journal of Finance*, 1952.

[12] Duc Hong Vo Minh Phuoc-Bao Tran. Market return spillover from the us to the asia-pacific countries: The role of geopolitical risk and the information communication technologies. *Public Library of Science One*, 2023.

[13] Andrew W. Lo Loriana Pelizzan Monica Billio, Mila Getmansky. Econometric measures of connectedness and systemic risk in the finance and insruance sectors. *Journal of Financial Econometrics*, 2012.

[14] André F. Perold. Thecapital asset pricing model. *Journal of Economic Perspectives*, 2004.

[15] Pradeep Ravikumar. Sparse additive models. *Journal of the Royal Statistical Society*, 2009.

[16] Akshay Srivastava Richard Roll. Mimicking portfolios. *The Journal of Portfolio Management*, 2018.

[17] Robert Faff Yuenan Wang Robert Brooks, Amalia Di Iorio. Testing the integration of the us and chinese stock markets in a fama-french framework. *Journal of Economic Integration*, 2009.

[18] Stephen A. Ross. The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 1976.

[19] Robert Tibshirani Trevor Hastie. Generalized additive models. *Journal of Statistical Science*, 1986.

[20] Geoffrey Stuart Watson. Smooth regression analysis. *Sankhya: The Indian Journal of Statistics*, 1964.

[21] Èlizbar Nadaraya. On estimating regression. *Theory of Probability Its Applications*, 1964.