THE UNIVERSITY OF CHICAGO

ESSAYS IN FINANCIAL ECONOMICS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE UNIVERSITY OF CHICAGO
BOOTH SCHOOL OF BUSINESS
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
RUI DA

CHICAGO, ILLINOIS
JUNE 2024

To my parents, Guoqiang and Renhong.

You never fail until you stop trying.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to extend my deepest appreciation to my advisors, Dacheng Xiu and Zhiguo He. Dacheng introduced me to the field of econometrics during my time as a physics student at Princeton. His impeccable work ethic, dedication to research, and exceptional research taste continue to inspire and drive me. Zhiguo is instrumental in me understanding what great finance research is. His guidance backed by his profound and precise comprehension of economics allows me to grow faster than ever.

I am also grateful to my committee members, Stefan Nagel, Ralph Koijen, and Pietro Veronesi, for their invaluable help on my research. Despite their hectic schedules, they have always made time to provide guidance and feedback, which have been crucial in shaping the direction of my work. Their support has been a constant source of motivation, particularly during challenging times.

My PhD journey would not have been possible without the generous financial support from the Booth School of Business. I am thankful for the exceptional support and care provided by our PhD office. Sharing this journey with my fellow doctoral students, especially, Chaoxing Dai, Shirley Zhang, Zhonglin Li, Philip Xu, and James Traina, has been a privilege.

Special thanks go to my endlessly inspiring friends, particularly, Ruohui, Chengcen, Zhen, Stan, Yuexi, Wanrong, Qixuan, Sixue, and Jiageng, for their perspectives, kindness, and passion.

I am especially indebted to Xiaolu Qi, whose unwavering support and trust have been indispensable during the most critical moments of my PhD journey.

Finally, I reserve my deepest gratitude for my parents. They have given me the utmost freedom and support to forge my own path in life. Their unconditional love is the foundation of everything I am today.

Abstract

This dissertation comprises four essays in financial economics. In the first essay, we conduct inference on volatility with noisy high-frequency data. The second essay proposes a semi-parametric approach to disentangling the autocovariance of equity returns at high frequency. In the third essay, we consider the limit of arbitrage arising from learning difficulty when investors are facing a large number of investment opportunities. Finally, we investigate market efficiency in the presence of many investors.

Chapter 1

Introduction

Interplay of economic theory, rich data, and creative methodology can bring new insights about the economy. This dissertation applies this idea to conduct investigation of different aspects of the financial market.

The availability of intraday transaction-level price data provides chances to understand asset volatility and microstructure noise with much more precision. In Chapter 2 and Chapter 3, we push forward new methodologies along this direction. In Chapter 2, we conduct inference on volatility using transaction-level data. We assume the observed transaction price follows a continuous-time Itô-semimartingale, contaminated by a discrete-time moving-average noise process associated with the arrival of trades. We estimate volatility, defined as the quadratic variation of the semimartingale, by maximizing the likelihood of a misspecified moving-average model, with its order selected based on an information criterion. Our inference is uniformly valid over a large class of noise processes whose magnitude and dependence structure vary with sample size. We show that the convergence rate of our estimator dominates $n^{1/4}$ as noise vanishes, and is determined by the selected order of noise dependence when noise is sufficiently small. Our implementation guarantees positive estimates in finite samples. In Chapter 3, we further demonstrate the same modeling techniques also deepens our understanding of microstructure noise. We establish the model-selection consistency, provide a central limit theory on autocovariance parameters, and show their consistency

uniformly over a large class of models that allow for an arbitrary noise magnitude and a flexible dependence structure. We also provide a quadratic representation of the likelihood estimator, which sheds light on its connection with nonparametric kernel estimators. Our simulation evidence suggests that our estimator outperforms the nonparametric alternatives particularly when noise magnitude is small. We apply this estimator to S&P 1500 index constituents, and find that in recent years the microstructure friction has become smaller but existed in 5-minute returns, particularly in small caps, and that the average duration of autocorrelations for large caps has shrunk considerably to merely 10 seconds.

Richer data allows us to better learn the structure of the market, but limit of our statistical learning occurs as often. In Chapter 4, we investigate how statistical learning difficulties can affect investor decisions. When alphas are weak and rare, arbitrageurs learning from historical data encounter a notable gap between the achievable Sharpe ratio and the infeasible maximum with perfect knowledge of return-generating processes. This reflects a statistical arbitrage limit, expanding equilibrium bounds of alphas beyond what is suggested by arbitrage pricing theory. We derive the optimal feasible Sharpe ratio, demonstrating its variation with the strength and sparsity of alphas. Additionally, we design an "all-weather" strategy that achieves it under any alpha condition. Our empirical investigation using machine learning on equity returns highlights the practical impact of this statistical limit to arbitrage.

The previous chapters have largely focused on utilizing price data to study the financial market. In Chapter 5, we extend our arsenal by incorporating investor holding data. Typically, modern financial markets contain many investors. In this context, we study the role of information in investor decision-making, and the informational efficiency and liquidity of the market. An equilibrium is characterized in closed-form for a continuous-time economy with many market participants and imperfect competition, in which investors receive private information with varying quality, and are heterogeneous in their misperception of the

information quality. In equilibrium, investor heterogeneity in their misperception generates return predictability by investors' trading, and trading of different investors follows a simple factor structure with weak factors. To conduct empirical analysis that builds on these equilibrium implications, we develop a new big-data econometric method that utilizes the factor structure to accommodate the high-dimensionality of these implications. The framework can be applied to price and institution holding data to estimate return predictability of individual institution's trades, dynamic price impact, misperception of institutions on their information quality, and institutions' contributions to the informational efficiency of the market.

Chapter 2

When Moving-Average Models Meet High-Frequency Data: Uniform Inference on Volatility[1]

2.1 Introduction

In this paper, we develop a simple estimator of volatility using high-frequency data in the presence of serially correlated, heteroscedastic, and endogenous microstructure noise. More importantly, we propose uniformly valid inference over a large class of noise processes that allows for simultaneously infinite-order autocorrelation and arbitrarily shrinking magnitude.

Hansen and Lunde (2006) provide empirical evidence that microstructure noise is quite small in Dow Jones Industrial Average stocks. To improve efficiency, one can consider a test of whether noise is present (or rely on an informal volatility signature plot), then decide whether to use a noise-robust estimator or the more efficient realized volatility estimator (Andersen, Bollerslev, Diebold, and Labys (2003)), which assumes noise is absent. Standard (pointwise) inference for this pre-testing approach, however, yields a misleading picture of the actual finite-sample behavior. Moreover, assuming the noise exists, a follow-up issue is to determine its dependence structure. An estimator robust to noise with long-range temporal dependence could be inefficient if the actual noise is simply i.i.d.. To strike a

1. This Chapter is a joint work with Dacheng Xiu.

more desirable trade-off between efficiency and robustness, one can consider modeling the noise as a moving-average process and adopting information criteria to determine its order of dependence. Nevertheless, model-selection mistakes are inevitable in finite samples, so that pointwise inference is again unreliable. The lack of uniformity for pre-testing or post-selection estimators has been widely noted in the classic time-series setting by Shibata (1986); Pötscher (1991); Kabaila (1995); and Leeb and Pötscher (2005).

To remedy this issue, we develop uniformly valid inference in the spirit of Mikusheva (2012); Andrews and Cheng (2012); Andrews, Cheng, and Guggenberger (2020); and Belloni, Chernozhukov, and Hansen (2014) in different contexts, on volatility over a large class of MA(∞) models that allow for an asymptotically vanishing noise with a flexible dependence structure. Our inference is thereby more reliable than that of the realized volatility , which simply ignores the impact of small noise when it is difficult to detect. Our inference also allows for model-selection mistakes, which surely occur in the case of an MA(∞) data-generating process, and is therefore robust to the dependence structure of noise.

The crux of our uniformity results is that the convergence rate of our estimator depends on various sequences of noise DGPs. Similar to our estimator but in the case of white noise, Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008) show that as the variance of the noise vanishes, the convergence rate of the realized kernel estimator improves from $n^{1/4}$ to $n^{1/2}$ with the optimal choice of bandwidth. Their inference, however, is not uniformly valid because a gap remains between the small-noise regime they consider and the no-noise regime. More specifically, they require the noise variance to be greater than n^{-1}. This seemingly small gap is not innocuous, because in a finite sample, the noise magnitude may fall into this gap, resulting in a distortion in the prescribed asymptotic distribution. In contrast, Jacod and Protter (2011) (Theorem 16.5.7) study the pre-averaging estimator as the noise variance vanishes at the rate of $n^{-\eta}$ for $\eta \geq 0$. They do not, however, provide a uniformly valid inference procedure; Also, their inference requires knowledge of η, and the convergence

rate of their estimator cannot exceed $n^{1/3}$.[2]

In the case of serially correlated noise, the unknown dependence structure may further plague the efficiency. The pre-averaging estimator by Jacod, Li, and Zheng (2019) and the flat-top realized kernel estimator by Varneskov (2016) converge at the rate of $n^{1/4}$. They do not consider alternative sequences of noise DGPs (particularly those in which noise magnitude and dependence structure interact) that may influence the convergence rate and validity of their inference. We investigate various noise DGPs with simultaneous infinite-order autocorrelation and arbitrarily shrinking magnitude. We show that the convergence rate of our estimator dominates $n^{1/4}$ as noise vanishes, and is determined by the order of noise dependence when noise is sufficiently small. While it is appealing to consider data-driven order selection, such as information criteria, for efficiency gains, in light of the critique by Leeb and Pötscher (2005), we adopt a slightly more conservative order selection procedure than the Akaike information criterion (AIC). As such, our inference remains uniformly valid with a slight efficiency loss only in the small-noise regime.

The literature on the estimation of quadratic variation using noisy high-frequency data is enormous. Earlier works mainly tackle a white microstructure noise.[3] In this paper, we target serially correlated noise using a likelihood-based approach. Hansen, Large, and Lunde (2008) first shed light on the asymptotic equivalence between the maximum likelihood estimator and MA filters. They implement the MA(q) estimator and demonstrate its desirable performance in extensive simulations with various noise models. Related work that discusses serially

2. Jacod and Protter (2011) impose an "essentially white" noise assumption, which requires that conditionally on the latent process, the noise is centered and independent, although not necessarily identically distributed. For the sake of presentation, we do not distinguish this type of noise from white noise, since it is also serially uncorrelated, unlike the "colored" noise setting we consider.

3. Prominent estimators include, but are not limited to, two-scale or multi-scale estimators by Zhang, Mykland, and Aït-Sahalia (2005) and Zhang (2006); the realized kernel estimator and its extensions by Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008) and Barndorff-Nielsen, Hansen, Lunde, and Shephard (2011); the pre-averaging estimator by Jacod, Li, Mykland, Podolskij, and Vetter (2009) and Jacod, Podolskij, and Vetter (2010); the local method of moments estimator by Reiß (2011); and likelihood-based estimators by Aït-Sahalia, Mykland, and Zhang (2005) and Xiu (2010).

dependent noise also include Kalnina and Linton (2008); Bandi and Russell (2008); Aït-Sahalia, Mykland, and Zhang (2011); Hautsch and Podolskij (2013); Bibinger, Hautsch, Malec, and Reiß (2019); and Li, Laeven, and Vellekoop (2020). Their assumptions on noise, however, are more restrictive than in our setting.

Our paper is organized as follows. Section 2.2 sets up the model, Section 2.3 presents the main results, Section 2.4 provides simulation evidence, and Section 2.5 concludes. The appendix contains the proof of the main theorem, and the online supplemental appendix provides the proofs of the corollary, proposition, and technical lemmas.

2.2 Model Setup

We start with notation. For any matrix A, $A^\intercal$ denotes its transpose. We denote by $\delta_{i,j}$ the Kronecker delta. The imaginary unit and the indicator function are written as $\mathbb{i}$ and $\mathbb{1}_{\{\cdot\}}$, respectively. All vectors are column vectors. We write (a, b, c) in place of $(a^\intercal, b^\intercal, c^\intercal)^\intercal$ for simplicity. d-dimensional vectors of 0s and 1s are written as 0_d and 1_d. We use $\|\cdot\|$ to denote the $\mathbb{L}^2$ norm. We use B to denote the backward (lag) operator associated with discrete-time time series. We use K as a generic positive constant that may vary from line to line but not depend on n. All limits are taken as $n \to \infty$. We use $\xrightarrow{\mathcal{L}}$ to denote convergence in law. We write $a_n \lesssim b_n$ if $a_n \leq K b_n$ for all n. We write $a_n \sim b_n$ if $a_n \lesssim b_n \lesssim a_n$. We use $a \vee b$ and $a \wedge b$ to denote $\max\{a, b\}$ and $\min\{a, b\}$, respectively. We use a superscript (n) to facilitate discussion of uniformity over different sequences of data-generating processes (DGPs) indexed by n.

At each stage $n \geq 1$, transaction prices $\widetilde{X}$ are observed at time points $0 = t_0 < t_1 < \ldots < t_{n_T} \leq T$, where T is fixed. Throughout, we assume n_T, the number of observations within $[0, T]$, is an observed random variable, whereas n is a non-observable mathematical abstraction. We let $\Delta_n = T/n_T$. We assume $\widetilde{X}_{t_i}$ comprises two components:

$$\widetilde{X}_{t_i} = X_{t_i} + U_i, \quad 0 \leq i \leq n_T,$$

where X_{t_i} is (the logarithm of) the efficient equilibrium price and U_i is the microstructure noise associated with the ith observation.

Specifically, with respect to the efficient price, we assume the following:

Assumption 1. *The logarithm of the efficient price process X_t is an Itô-semimartingale defined on some filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$ and satisfies*

$$X_t = X_0 + \int_0^t \mu_s ds + \int_0^t \sigma_s dW_s + (\delta \mathbb{1}_{\{|\delta| \leq 1\}}) \star (\underline{\mu} - \underline{\nu})_t + (\delta \mathbb{1}_{\{|\delta| > 1\}}) \star \underline{\mu}_t,$$

where μ_t and σ_t are adapted and locally bounded, W is a standard Brownian motion, and $\underline{\mu}$ is a Poisson random measure on $\mathbb{R}_+ \times E$, where E is a Polish space. The compensator $\underline{\nu}$ satisfies $\underline{\nu}(dt, du) = dt \otimes \lambda(du)$ for some σ-finite measure λ on E. Moreover, $|\delta(\omega, t, u)| \wedge 1 \leq \Gamma_m(u)$ for all (ω, t, u) with $t \leq \tau_m(\omega)$, where $\{\tau_m\}$ is a localizing sequence of stopping times and $\{\Gamma_m\}$ a sequence of deterministic functions satisfying $\int \Gamma_m^r(u)\lambda(du) < \infty$ for some $r \in [0, 1)$.

In addition, the process $Z_t = (\mu_t, \sigma_t^2)$ is also an Itô-semimartingale on the space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$ with the form

$$Z_t = Z_0 + \int_0^t \widetilde{\mu}_s ds + \int_0^t \widetilde{\sigma}_s d\widetilde{W}_s + (\widetilde{\delta} \mathbb{1}_{\{|\widetilde{\delta}| \leq 1\}}) \star (\underline{\mu} - \underline{\nu})_t + (\widetilde{\delta} \mathbb{1}_{\{|\widetilde{\delta}| > 1\}}) \star \underline{\mu}_t, \tag{2.2.1}$$

where $\widetilde{\mu}_t$ and $\widetilde{\sigma}_t$ are locally bounded adapted processes, $\widetilde{W}$ is a multivariate Brownian motion, potentially correlated with W, and $\widetilde{\delta}$ is a predictable function such that for some deterministic function $\widetilde{\Gamma}_m(u)$, $\|\widetilde{\delta}(\omega, t, u)\| \wedge 1 \leq \widetilde{\Gamma}_m(u)$ for all $\omega \in \Omega$, $t \leq \tau_m(\omega)$, and $\int \widetilde{\Gamma}_m^2(u)\lambda(du) < \infty$.

Assumption 6 allows for the leverage effect and jumps in both the efficient price and its volatility. It accommodates most models in asset pricing and is commonly used to derive in-fill asymptotic results for high-frequency data—for example, Jacod and Protter (2011) and Aït-Sahalia and Jacod (2014), with notable exclusions of long-memory volatility models driven by fractional Brownian motions (Comte and Renault (1996, 1998)).

The parameter of interest is the quadratic variation of X (scaled by T^{-1}), which comprises

both continuous and discontinuous components:

$$C_T = \frac{1}{T}\int_0^T \sigma_t^2 dt + \frac{1}{T}\sum_{0\le t\le T}(\Delta X_t)^2,$$

where $\Delta X_t = X_t - X_{t-}$. Although estimating the integrated volatility or the jump component of the quadratic variation is of tremendous interest, we do not pursue this agenda in this paper, in which we aim for a practical volatility estimate that depends on as few tuning parameters as possible.[4]

Next, we make an assumption on the arrival of trades:

Assumption 2. *For each $n \geq 1$, the sequence of observation times $\{t_i : i \geq 0\}$ satisfies $t_0 = 0$ and $t_i = t_{i-1} + \frac{T}{n}\xi_{t_{i-1}}\chi_i$, where the sequence $\{\chi_i : i \geq 1\}$ is i.i.d., $(0,\infty)$-valued, defined on $(\Omega, \mathcal{F}, \mathbb{P})$, and independent of the σ-field $\mathcal{F}_\infty = \bigvee_{t>0}\mathcal{F}_t$, with $m_j = \mathbb{E}((\chi_i)^j) < \infty$ and $m_1 = 1$, for all $j > 0$. In addition, the process $\xi = (\xi_t)_{t\geq 0}$ is a nonnegative Itô-semimartingale defined on $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$ in the form of (3.2.2), such that neither ξ_t nor ξ_{t-} vanishes.*

Assumption 7 allows the arrival rate of transactions to depend on their prices through ξ_t. It also accommodates regular sampling, time-changed regular sampling, Poisson sampling, modulated Poisson sampling, and predictably modulated random-walk sampling schemes, as discussed in detail by Jacod, Li, and Zheng (2017).[5] We introduce here and below several stochastic processes, e.g., ξ_t and η_t, for which their driving Brownian motions (implicitly

4. To achieve robustness to serially correlated noise, prominent nonparametric estimators require three tuning parameters (see two such estimators in our simulation study). Exploring the finer structure of the quadratic variations would require at least one more—rendering these measures impractical to estimate, in particular for illiquid stocks.

5. Our sampling scheme imposes that conditional on $\mathcal{F}_{t_{i-1}}$, t_i is independent of X_t for $t \geq t_{i-1}$, which we need to derive a desirable central limit theory. This assumption conforms with Assumption O(ii) of Jacod, Li, and Zheng (2017), but is more restrictive than those adopted by Li, Mykland, Renault, Zhang, and Zheng (2014) and Fukasawa (2010), both of which find an asymptotic bias in the CLT of the realized volatility estimator associated with their general sampling scheme (in the absence of microstructure noise). On a related note, Renault and Werker (2011) find that the instantaneous causality relations between price volatility and durations of trades could lead to severely biased volatility estimates.

defined) are different from W in Assumption 6, but possibly correlated. Note that finding a single Poisson measure that drives the jumps of all processes involved is always possible.

Finally, we assume the noise is endogenous, heteroscedastic, and serially correlated.

Assumption 3. *For each $n \geq 1$, the noise sequence $\{U_i : i \geq 0\}$ consists of random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which have the following $MA(\infty)$ representation:*

$$U_i = \eta_{t_i} \iota^{(n)} \theta^{(n)}(\mathrm{B}) \varepsilon_i, \quad \text{with} \quad \theta^{(n)}(x) = 1 + \sum_{j=1}^{\infty} \theta_j^{(n)} x^j,$$

where $\varepsilon_i \overset{i.i.d.}{\sim} (0,1)$, defined on $(\Omega, \mathcal{F}, \mathbb{P})$, is independent of $\mathcal{F}_\infty$ and $\{\chi_i : i \geq 1\}$, and has finite moments of all orders; $(\eta_t)_{t \geq 0}$ is an $(\mathcal{F}_t)$-adapted nonnegative Itô-semimartingale that satisfies the same form of (3.2.2); and $\iota^{(n)}$ is a deterministic nonnegative number that characterizes the noise magnitude and satisfies $\iota^{(n)} \leq K$.[6]

Assumption 8 accommodates several empirical features of the microstructure noise. The noise process depends on price X through η, since η_t is $\mathcal{F}_t$-adapted, which may be driven by a Brownian motion and a Poisson random measure that are correlated with X. Such dependence is potentially driven by comovement between the price and bid-ask spread or the discreteness of the observed price.[7] That said, this assumption implies zero correlation between any function of the path of X and U_i for each i—the key identifying assumption that separates efficient returns from noise. A fully specified structural microstructure model would be necessary, along with additional observables (e.g., bid-ask prices), if some non-vanishing correlation between X and U were allowed for. In this paper, we avoid imposing

6. The probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$ can be constructed more explicitly. Specifically, we define X, Z, ξ, and η (which satisfy the relevant assumptions) on a space $(\Omega_{(0)}, \mathcal{F}_\infty, (\mathcal{F}_t), \mathbb{P}_{(0)})$, and define $\{\chi_i\}$ and $\{\varepsilon_i\}$ on a different space $(\Omega_{(1)}, \mathcal{F}_{(1)}, \mathbb{P}_{(1)})$. We then set $\Omega = \Omega_{(0)} \times \Omega_{(1)}$, $\mathcal{F} = \mathcal{F}_\infty \otimes \mathcal{F}_{(1)}$, and $\mathbb{P}(d\omega_{(0)}, d\omega_{(1)}) = \mathbb{P}_{(0)}(d\omega_{(0)}) \mathbb{P}_{(1)}(d\omega_{(1)})$.

7. No one has yet built a perfect model that is meant to be a full describer of reality. Like Jacod, Li, and Zheng (2017) and Jacod, Li, and Zheng (2019), the type of rounding errors our setting allows for is still somewhat restrictive. In the simulation analysis, we investigate a particular form of rounding that appears more realistic but violates Assumption 8, and we show that the rounding effect is negligible.

additional structural assumptions, and instead focus on the reduced-form model of $\widetilde{X}$, while being agnostic about the economic implications of reduced-form parameters; e.g., θ and ι^2. Many structural models yield specific reduced-form ARMA models of returns; for example, Hasbrouck (2007), with differences only in how the reduced-form parameters relate to structural parameters. Estimating and interpreting structural parameters in a microstructure model is interesting, but we leave this for future work.

The noise process features flexible serial correlations through its $\theta^{(n)}(\mathrm{B})\varepsilon$ component, as specified by an MA(∞) model. The next assumption spells out restrictions on its spectral density function, $g(\lambda;\theta^{(n)}) = |\theta^{(n)}(e^{\mathbb{i}\lambda})|^2$, such that the sequence of MA processes is uniformly invertible and their long-range serial dependence cannot be arbitrarily strong.

Assumption 4. *For each $n \geq 1$, the spectral density function of $\theta^{(n)}(\mathrm{B})\varepsilon$ satisfies, for some fixed $\alpha > 3$,*

$$\inf_{\lambda} g(\lambda;\theta^{(n)}) \geq \frac{1}{K} \quad \textit{and} \quad \left| \int_{-\pi}^{\pi} g(\lambda;\theta^{(n)}) e^{\mathbb{i}\lambda j} d\lambda \right| \leq K j^{-\alpha}, \quad \forall j \geq 0.$$

We next introduce our likelihood-based estimator.

2.3 Main Results

2.3.1 Likelihood-based Estimation

In contrast to existing nonparametric estimators, we construct a quasi-maximum likelihood estimator (QMLE) in the spirit of White (1982) by imposing a misspecified parametric model, for which the likelihood function is available:

$$dX_t = \sigma dW_t; \quad U_i = \iota\theta(\mathrm{B})\varepsilon_i, \quad \text{with} \quad \theta(x) = 1 + \sum_{j=1}^{q} \theta_j x^j, \quad \text{and} \quad \varepsilon_i \sim \mathcal{N}(0,1).$$

In other words, we pretend the efficient price (in logarithm) is a Brownian motion with constant volatility but no drift, and that the noise follows a Gaussian MA(q) model with

the order q to be determined. Under this model, the observed log-return vector $Y_n = (Y_{n,1}, Y_{n,2}, \ldots, Y_{n,n_T})^\intercal$, which is defined as

$$Y_{n,i} = X_{t_i} - X_{t_{i-1}} + U_i - U_{i-1}, \quad 1 \leq i \leq n_T, \tag{2.3.2}$$

follows a reduced-form Gaussian MA$(q+1)$ model. Its $n_T \times n_T$ covariance matrix Σ_n is given by

$$\Sigma_n(\sigma^2, \iota^2, \theta) = \sigma^2 \Delta_n \mathbb{I}_n + \sum_{h=0}^{n_T-1} (2\gamma_h - \gamma_{h+1} - \gamma_{h-1}) \mathbb{G}_n^h,$$

where $(\mathbb{I}_n)_{ij} = \delta_{i,j}$, $(\mathbb{G}_n^h)_{ij} = \delta_{h,|i-j|}$, and γ_h is the h-th order autocovariance of U:

$$\gamma_h = \frac{\iota^2}{2\pi} \int_{-\pi}^{\pi} g(\lambda; \theta) e^{\mathbb{i}\lambda h} d\lambda, \quad \text{where} \quad g(\lambda; \theta) = \left|\theta(e^{\mathbb{i}\lambda})\right|^2.$$

Because θ is a nuisance parameter for volatility estimation and is unidentified if $\iota = 0$, we reparameterize the likelihood function in terms of strongly identified parameters (σ^2, γ):

$$L_n(\sigma^2, \gamma) = -\frac{1}{2} \log \det(\Sigma_n(\sigma^2, \gamma)) - \frac{1}{2} \mathrm{tr}(\Sigma_n(\sigma^2, \gamma)^{-1} Y_n Y_n^\intercal),$$

where $\Sigma_n(\sigma^2, \gamma) := \Sigma_n(\sigma^2, \iota^2, \theta)$ and γ is the $(q+1)$-dimensional vector of noise autocovariances.[8] We define $(\widehat{\sigma}_n^2(q), \widehat{\gamma}_n(q))$ as the maximizer of $L_n(\sigma^2, \gamma)$:

$$(\widehat{\sigma}_n^2(q), \widehat{\gamma}_n(q)) = \arg \max_{(\sigma^2, \gamma) \in \Pi_n(q)} L_n(\sigma^2, \gamma).$$

The parameter space of (σ^2, γ), denoted by $\Pi_n(q)$, can be derived from the usual condition that $(\sigma^2, \iota^2, \theta)$ satisfy, i.e., $\inf_\lambda f(\lambda; \gamma) \geq 0$, where $f(\lambda; \gamma) = \iota^2 g(\lambda; \theta)$. However, it is not ideal for reasons we now explain.

8. Note that $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_q)$, which is different from how vectors are typically indexed. For convenience, we often treat γ as an infinite dimensional vector, with 0s filled beyond the $(q+1)$th entry of γ when no ambiguity exists.

Aït-Sahalia and Xiu (2019) show that in the white-noise case, if the noise magnitude is small, the noise variance estimator $\widehat{\iota}_n^2$ hits the boundary zero, so that the asymptotic distribution of the volatility estimator, $\widehat{\sigma}_n^2$, becomes nonstandard. A similar yet more severe issue occurs here: The estimate $\widehat{\gamma}$ hits the boundary $\inf_\lambda f(\lambda;\widehat{\gamma}) = 0$ with nontrivial probability.

An easy solution in the white-noise case is to enlarge the parameter space of the nuisance parameter, allowing for negative values of $\widehat{\iota}_n^2$, so that the asymptotic distribution of $\widehat{\sigma}_n^2$ is not affected by confinement of the parameter space for ι^2. We adopt a similar strategy to enlarge the parameter space of (σ^2,γ) to $\{(\sigma^2,\gamma)\in\mathbb{R}^{q+2} : \inf_\lambda f(\lambda;\sigma^2,\gamma,\Delta_n)\geq 0\}$, where $f(\lambda;\sigma^2,\gamma,\Delta_n) = \sigma^2\Delta_n + |1-e^{\mathbb{i}\lambda}|^2 f(\lambda;\gamma)$ is the spectral density of Y_n under the quasi model. In other words, the parameter space is enlarged such that only a reduced-form MA$(q+1)$ model of observed returns is required to be well defined (see Section 3.3.2), and that a well-defined decomposition of observed returns, as in (3.3.6), may not exist. On the other hand, the parameter space must be sufficiently "local" to the true value to avoid spurious estimates due to potential use of an overly flexible quasi-model (e.g., q is too large). For this purpose, we define a set $\Pi_n(q)$ that imposes constraints on the lower bound of the spectral density function and the decay of autocovariances:[9]

$$\left\{(\sigma^2,\gamma)\in\mathbb{R}^{q+2} : \inf_\lambda f(\lambda;\sigma^2,\gamma,\Delta_n)\geq\frac{\Delta_n}{K},\quad \sigma^2+|\gamma_0|+\frac{\sum_{j=1}^\infty j^2|\gamma_j|}{\inf_\lambda|\sigma^2\Delta_n+f(\lambda;\gamma)|}\leq K\right\}. \tag{2.3.3}$$

This parameter space depends on the order of the MA model, q, which we next discuss how to select.

2.3.2 Model Selection

To determine an appropriate order q, we use AIC, which in our setting can be written as

$$\mathrm{AIC}_n(q) = 2q - 2\max_{(\sigma^2,\gamma)\in\Pi_n(q)} L_n(\sigma^2,\gamma).$$

9. The constraint (3.3.11) is essential for proofs. We do not find it critical to impose in our implementation.

Our choice of order q will be based on (but not necessarily identical to)

$$\widehat{q}_{n,\mathrm{AIC}} = \arg \min_{q \leq n_T^{1/3}} \mathrm{AIC}_n(q). \tag{2.3.4}$$

More generally, in Theorem 1 below, we spell out the conditions a desirable order $\widehat{q}_n$ must satisfy in order to accommodate uniformly valid inference on volatility for a large class of DGPs.

Similar to the case of AR(∞) in Shibata (1980), the upper bound on q precludes MA models with too many parameters from estimation. Asymptotically, this upper bound is not binding, because for all sequences of noise DGPs we consider, $\widehat{q}_{n,\mathrm{AIC}} = o_{\mathrm{P}}(n^{1/6})$—a claim we prove in the online supplemental appendix.

In a companion paper, Da and Xiu (2021a) prove model selection consistency (based on BIC) and provide pointwise asymptotic inference on noise autocovariance parameters when the noise follows an MA($q^\star$) model with a finite $q^\star$. The pointwise asymptotic theory relies on this fixed DGP, as well as this unrealistic result of perfect model selection; hence, it provides a misleading picture of the actual finite-sample behavior of the inference. As shown in the classic time-series setting of Leeb and Pötscher (2005), conducting uniformly valid post-selection inference on parameters over a nontrivially large class of DGPs is generally impossible. For volatility estimation in our setting, however, uniformly valid inference is possible for a wide class of DGPs, which we turn to next.

2.3.3 Uniform Inference on Volatility

Obviously, the class of DGPs cannot be arbitrarily large, so we need restrictions on how the magnitude of the noise and its autocorrelation structure vary with sample size. We denote by $\kappa^{(n)}$ the ∞-dimensional vector of autocovariances of $\theta^{(n)}(\mathrm{B})\varepsilon$, whose components are given

by

$$\kappa_j^{(n)} = \frac{1}{2\pi}\int_{-\pi}^{\pi} g(\lambda;\theta^{(n)})e^{\mathbb{i}\lambda j}d\lambda, \quad j \geq 0.$$

The class of noise models we consider satisfies:

Assumption 5. *For any $0 < k < K$ and any sequence $\alpha_n \to \infty$, we have*

$$q_n^\star(k) = o(n^{1/3}(\iota^{(n)} \vee n^{-1/2})^{4/9}), \quad \psi_n^2 \sum_{j=q_n^\star(k)}^{\infty} |\kappa_j^{(n)}| = o\left(\frac{q_n^\star(k)^{1/2} + \alpha_n}{n^{1/2}} + \frac{\sqrt{\iota^{(n)}}}{n^{1/4}}\right),$$

where

$$q_n^\star(k) := \min q, \quad \textit{subject to} \quad n\psi_n^4 \sum_{j=q+1-\alpha_n}^{2q} |\widetilde{\kappa}_j^{(n)}|^2 \leq kq,$$

$\psi_n := (1 + n^{-1/2}/\iota^{(n)})^{-1}$, *and* $\widetilde{\kappa}_j^{(n)} := \sum_{i=0}^{\infty}(i+1)\psi_n^i(2\kappa_{j+1+i}^{(n)} - \kappa_{j+i+2}^{(n)} - \kappa_{j+i}^{(n)})$.

Intuitively, $q_n^\star(k)$ mimics the "oracle" order that AIC selects. Assumption 10 effectively requires that this order cannot be too large (the first equation above) and that the approximation error induced by selection (the left-hand side of the second equation above) is asymptotically dominated by the estimation error (the right-hand side). Next, we provide two examples to demonstrate that the conditions in Assumption 10 are not restrictive from a practical point of view.

Example 1: Suppose $n^{1/2}\iota^{(n)} \to \infty$ and $\theta^{(n)}(\mathrm{B})\varepsilon$ follows an MA(∞) model with $|\kappa_j^{(n)}| \sim j^{-\alpha}$ for some $\alpha > 3 \vee \frac{2}{1+2\log\iota^{(n)}/\log n}$. It is easy to show that Assumption 10 holds, because

$$q_n^\star(k) \sim n^{1/(2\alpha)} \quad \text{and} \quad \psi_n^2 \sum_{j=q_n^\star(k)}^{\infty} |\kappa_j^{(n)}| \sim n^{-1/2+1/(2\alpha)} = o\big(n^{-1/4}(\iota^{(n)})^{1/2}\big).$$

Jacod, Li, and Zheng (2017) assume $|\kappa_j^{(n)}| \sim j^{-\alpha}$ with $\alpha > 3$ and a fixed $\iota^{(n)}$. Our condition

further sheds light on a trade-off between $\iota^{(n)}$ and α: As $\iota^{(n)}$ shrinks, $|\kappa_j^{(n)}|$ must decay faster.

EXAMPLE 2: Suppose $\theta^{(n)}(\mathrm{B})\varepsilon$ follows an arbitrary ARMA(p,q) process with finite p and q. Assumption 10 holds because in this case, as long as $\iota^{(n)} \lesssim 1$ (it can shrink arbitrarily fast),

$$q_n^\star(k) \lesssim \log n \quad \text{and} \quad \psi_n^2 \sum_{j=q_n^\star(k)}^{\infty} |\kappa_j^{(n)}| \lesssim n^{-1/2}\big(1 + o(q_n^\star(k)^{1/2})\big).$$

Now we are ready to present the main theoretical result, based on which we build uniformly valid inference on volatility:

Theorem 1. *Suppose we select an order $\widehat{q}_n = \widehat{q}_{n,\mathrm{AIC}} \vee \alpha_n$ with $\alpha_n = O(\log n)$. Then, for all sequences of DGPs satisfying Assumptions 6 - 10, and as $\widehat{q}_n \vee (n^{1/2}\iota^{(n)}) \to \infty$, we have*[10]

$$\frac{\widehat{\sigma}^2(\widehat{q}_n) - C_T}{\sqrt{\mathrm{AVAR}(\widehat{q}_n, n)_T}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

where $\mathrm{AVAR}(q,n)_T$ *is given by*

$$\mathrm{AVAR}(q,n)_T = \frac{1}{n}\left[(4q+6)\,E(4,\xi)_T + \Delta_n^{-1/2}\zeta^{(n)}\left(5E(4,\xi)_T C_T^{-1/2} + C_T^{3/2}B(\xi)_T\right)\right], \tag{2.3.5}$$

$(\zeta^{(n)})^2$ *is the "long-run variance" of the general noise process, given by*

$$(\zeta^{(n)})^2 = (\iota^{(n)})^2 g(0,\theta^{(n)}) \frac{\int_0^T \eta_s^2 \xi_s^{-1} ds}{\int_0^T \xi_s^{-1} ds},$$

10. To ensure $\widehat{q}_n \vee (n^{1/2}\iota^{(n)}) \to \infty$ holds without worrying about $n^{1/2}\iota^{(n)}$ and $\widehat{q}_{n,\mathrm{AIC}}$, we can select α_n such that $\alpha_n \to \infty$, say, $\alpha_n \sim \log n$ as in our implementation, although the statement of Theorem 1 does not require this.

$E(4,\xi)_T$ is a general "quarticity" in the presence of random sampling and jumps, given by

$$E(4,\xi)_T = \frac{1}{T}\int_0^T \xi_s \sigma_s^4 ds + \frac{1}{T}\sum_{s\leq T}(\Delta X_s)^2(\xi_s\sigma_s^2 + \xi_{s-}\sigma_{s-}^2), \tag{2.3.6}$$

and

$$B(\xi)_T = \frac{2\int_0^T \eta_s^2\sigma_s^2 ds + \sum_{s\leq T}(\Delta X_s)^2(\eta_s^2+\eta_{s-}^2)}{C_T \times \int_0^T \eta_s^2\xi_s^{-1}ds} + \frac{T\int_0^T \eta_s^4\xi_s^{-1}ds}{\left(\int_0^T \eta_s^2\xi_s^{-1}ds\right)^2}. \tag{2.3.7}$$

Combining with asymptotic variance estimators in Section 2.3.4, we immediately obtain:

Corollary 1. *Suppose the same assumptions as those in Theorem 1 hold. Let $c_{1-\alpha} = F^{-1}(1-\alpha/2)$, where $F(\cdot)$ is the standard Gaussian cumulative distribution function. We have*

$$\lim_{n\to\infty}\mathbb{P}\left(C_T \in CI_n(\alpha)\right) = 1-\alpha,$$

where, using $\widehat{\zeta}_n^2 = \sum_{j=-\widehat{q}_n}^{\widehat{q}_n}\widehat{\gamma}_n(\widehat{q}_n)_{|j|}$, the uniformly valid confidence interval $CI_n(\alpha)$ is constructed as

$$\left[\widehat{\sigma}_n^2(\widehat{q}_n) \pm \frac{c_{1-\alpha}}{n_T^{1/2}}\sqrt{(4\widehat{q}_n+6)\,\widehat{E}_n(4)_T + \frac{\widehat{\zeta}_n^2}{\Delta_n^{1/2}}\left(5\widehat{E}_n(4)_T\widehat{\sigma}_n^2(\widehat{q}_n)^{-1/2} + \widehat{\sigma}_n^2(\widehat{q}_n)^{3/2}\widehat{B}_n(\widehat{q}_n)_T\right)}\right].$$

To shed light on the asymptotic behavior of our estimator, we examine two special DGP sequences:

i. Under $n^{1/2}\iota^{(n)}/(4\widehat{q}_n+6)\to\infty$,

$$\mathrm{AVAR}(\widehat{q}_n,n)_T = n^{-1/2}T^{-1/2}\zeta^{(n)}\left(5E(4,\xi)_T C_T^{-1/2} + C_T^{3/2}B(\xi)_T\right) + o_{\mathrm{P}}(n^{-1/2}\iota^{(n)}). \tag{2.3.8}$$

ii. Under $n^{1/2}\iota^{(n)}/(4\widehat{q}_n+6)\to 0$,[11]

$$\mathrm{AVAR}(\widehat{q}_n,n)_T=\frac{1}{n}(4\widehat{q}_n+6)E(4,\xi)_T+o_{\mathrm{P}}\Big(\frac{\widehat{q}_n+1}{n}\Big). \tag{2.3.9}$$

Case i describes the behavior of our estimator in the presence of "large" noise. The convergence rate is $(\iota^{(n)})^{-1/2}n^{1/4}$, which varies within $[n^{1/4},n^{1/2}\widehat{q}_n^{-1/2}]$. This result echoes and extends that of Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008) for the realized kernel estimator whose rate varies within $[n^{1/4},n^{1/2}]$ in the case of i.i.d. noise. As to the colored noise, our rate dominates $n^{1/4}$—the convergence rate of flat-top realized kernel and pre-averaging estimators by Varneskov (2016) and Jacod, Li, and Zheng (2019).

In the case of small noise (Case ii), the convergence rate is prescribed by $n^{1/2}\widehat{q}_n^{-1/2}$. When noise is absent, Case ii also shows that the efficiency loss compared with the realized volatility estimator is given by a factor $2\widehat{q}_n+3$, because realized volatility has knowledge of the absence of noise. Moreover, the bias of the realized volatility estimator is of order $(\iota^{(n)})^2n$, which may not vanish in Case ii, because noise is not entirely negligible in this regime.

We now explain our choice of $\widehat{q}_n$. Recall that the noise-dependence structure follows MA(∞). Intuitively, a smaller choice of $\widehat{q}_n$ leads to a more efficient estimator at the risk of a larger bias due to model misspecification ($\widehat{q}_n<\infty$). In contrast to the somewhat ad hoc tuning parameters other approaches rely on, our estimate $\widehat{q}_{n,\mathrm{AIC}}$ is informative about the minimal order using which the model misspecification bias is negligible. The importance of this guidance on q is manifested in Case ii, in which the convergence rate clearly improves as $\widehat{q}_n$ decreases.

Nonetheless, instead of fully relying on $\widehat{q}_{n,\mathrm{AIC}}$, Theorem 1 requires the use of a certain $\widehat{q}_n=\widehat{q}_{n,\mathrm{AIC}}\vee\alpha_n$ that also approaches ∞ slowly if $n^{1/2}\iota^{(n)}$ is bounded, even when the true

11. In this case, Theorem 1 requires that $\widehat{q}_n$ approaches ∞, so that $4\widehat{q}_n+6$ and $\widehat{q}_n$ are in fact of the same order. That said, we prefer this small-sample adjustment that can be established in the case of a finite $\widehat{q}_n$.

model may be of a finite order (and hence $\widehat{q}_{n,\mathrm{AIC}}$ is small). Indeed, if the true model is a finite-order MA(q), we can show that QMLE based on $\widehat{q}_{n,\mathrm{AIC}}$ can achieve a convergence rate as fast as $n^{1/2}$ in Case ii. However, the asymptotic distribution is highly nonstandard, because the model-selection bias is of an order comparable to the estimation error. For this reason, we intentionally inflate the order of the employed model, requiring $\widehat{q}_n \to \infty$, so that a standard asymptotic normal distribution is available in Case ii. α_n is the single tuning parameter required by our procedure. One possible choice of α_n is $\log n_T$, which (potentially) inflates $\widehat{q}_{n,\mathrm{AIC}}$ by $\log n_T$. The choice of $\widehat{q}_n$ does not affect the asymptotic variance in Case i, since (2.3.8) in fact does not rely on $\widehat{q}_n$, but it may hurt the efficiency of our estimator in Case ii. As a result, our rate in Case ii is strictly smaller than $n^{1/2}$ under the conditions in Theorem 1. This efficiency cost is in fact unavoidable for the sake of uniformity, because of the following "impossibility" result in the spirit of Leeb and Pötscher (2008).

To demonstrate this result, we consider a simple setting in which the noise process has no autocorrelation beyond the first lag (so that we use AIC to select q from $\{0, 1\}$), and the noise magnitude $(\iota^{(n)})^2$ is dominated by n^{-1} (so that the optimal rate of the volatility estimator is $n^{1/2}$). The next proposition shows that even with constant volatility, no uniformly consistent estimator exists for the cumulative distribution function $G_n(x)$, where

$$G_n(x) = \mathbb{P}\left(n^{1/2}(\widehat{\sigma}_n^2(\widehat{q}_{n,\mathrm{AIC}} \wedge 1) - C_T) \leq x\right).$$

Proposition 1. *For each $x \in \mathbb{R}$, there exists a DGP sequence satisfying Assumptions 6 - 10 with $\sigma_t^2 = C_T$ for some C_T fixed and all $t \in [0, T]$, $n(\iota^{(n)})^2 \leq K$, and a single parameter $\theta^{(n)}$, such that*

$$\liminf_{n\to\infty} \inf_{\widehat{G}_n(x)} \mathbb{P}\left(|\widehat{G}_n(x) - G_n(x)| > \frac{1}{K}\right) > 0,$$

where the infimum extends over all estimators $\widehat{G}_n(x)$ of $G_n(x)$.

On a different note, Theorem 1 establishes that our asymptotic distribution is condition-

ally Gaussian, which is typically the case for estimating quadratic variation in the presence of noise. Nonetheless, in the absence of noise, the limiting distribution for the realized volatility estimator is a mixture of Gaussian random variables with square root of uniform random variables around the jump times instead of Gaussian, unless the volatility and price processes do not jump together; see e.g., Theorem 5.4.2 of Jacod and Protter (2011). Because of this, their inference procedure is sufficiently complicated that simulations are entailed in order to achieve a sharp confidence interval; see page 349 of Aït-Sahalia and Jacod (2014). When $n^{1/2}\iota^{(n)} \leq K$, our asymptotic distribution would run into the same issue if $\widehat{q}_n$ is finite. Interestingly, while the requirement on $\widehat{q}_n \to \infty$ is motivated from uniformity considerations, we show that this condition also leads to a conditional Gaussian limiting distribution, which facilitates our inference procedure.

Another related point is that the asymptotic distribution of our volatility estimator does not depend on the true distribution of ε. This is not surprising in the case of large noise; see, e.g., Xiu (2010). However, when noise is small, it is possible that certain moments of ε might affect the asymptotic variance of volatility as the convergence rate improves. In this regime, the reason our volatility estimator has the same asymptotic variance regardless of the true distribution of ε is again due to $\widehat{q}_n \to \infty$.

2.3.4 Asymptotic Variance Estimators

In this section, we develop pre-averaging-based estimators of asymptotic variances. We need two sequences of integers k_n and k_n', satisfying $k_n \sim n^{2/3}, k_n' \sim n^{7/8}$, and a nonzero real-valued function $g : \mathbb{R} \to \mathbb{R}$, supported on $[0, 1]$, which is continuous and piecewise C^1 with a piecewise Lipschitz derivative g' and $g(0) = g(1) = 0$. We also adopt a truncation strategy (Mancini (2001)) to handle jump-related quantities, for which we define:

$$v_n = \alpha(k_n\Delta_n)^{\varpi}, \quad \text{for some } \alpha > 0, \varpi \in (0, 1/2)\,.$$

We construct the estimator of $E(4,\xi)_T$ in (2.3.6) as $\widehat{E}_n(4)_T = \widehat{C}_n(4)_T + \widehat{D}_n(4)_T$ using the pre-averaging approach:

$$\begin{aligned}
\widehat{C}_n(4)_T &= \frac{1}{Tk_n^2\Delta_n\left(\int_0^1 g(s)^2ds\right)^2}\sum_{m=1}^{n_T-2k_n}(\bar{Y}(g)^n_m\bar{Y}(g)^n_{m+k_n})^2\mathbb{1}_{\{|\bar{Y}(g)^n_m|\leq v_n,|\bar{Y}(g)^n_{m+k_n}|\leq v_n\}},\\
\widehat{D}_n(4)_T &= \frac{1}{Tk_n\int_0^1 g(s)^2ds}\sum_{m=k_n'+1}^{n_T-k_n-k_n'}(\bar{Y}(g)^n_m)^2\mathbb{1}_{\{|\bar{Y}(g)^n_m|>v_n\}}\left(\widetilde{c}(g)^n_m+\widetilde{c}(g)^n_{m-k_n'}\right),
\end{aligned}$$

where pre-averaged returns and spot volatilities are given by, respectively,

$$\bar{Y}(g)^n_i = \sum_{j=1}^{k_n-1} g\left(\frac{j}{k_n}\right)Y_{n,i+j},\quad \widetilde{c}(g)^n_i = \frac{1}{k_n'k_n\Delta_n\int_0^1 g(s)^2ds}\sum_{m=1}^{k_n'}(\bar{Y}(g)^n_{i+m})^2\mathbb{1}_{\{|\bar{Y}(g)^n_{i+m}|\leq v_n\}}. \tag{2.3.10}$$

These estimators are the same as those constructed by Aït-Sahalia and Xiu (2016) for i.i.d. noise. Despite their low convergence rate, these estimators are also consistent in this more general setting, because of the choice of a large local window size k_n which averages out the impact of the dependent noise. Because of the jump truncation, Assumption 6 imposes that $r < 1$, which is necessary for consistency.

Finally, we provide the estimator of $B(\xi)_T$ in (2.3.7) using $\widehat{B}_n(\widehat{q}_n)_T$ defined as

$$\left|\frac{1}{\widehat{\sigma}^2_n(\widehat{q}_n)}\left(\frac{1}{\widehat{\gamma}_n(\widehat{q}_n)_0-\widehat{\gamma}_n(\widehat{q}_n)_1}\right)(\widehat{B}'_n(1)+\widehat{B}'_n(2))+\left(\frac{1}{\widehat{\gamma}_n(\widehat{q}_n)_0-\widehat{\gamma}_n(\widehat{q}_n)_1}\right)^2\widehat{B}'_n(3)\right|\wedge\log n_T,$$

where, with $\bar{Y}(g)_m^n$ and $\widetilde{c}(g)_m^n$ defined in (2.3.10),

$$\begin{aligned}
\widehat{B}'_n(1) &= \frac{1}{n_T}\sum_{m=1}^{n_T-k_n-k'_n}(Y_{n,m})^2\widetilde{c}(g)_m^n, \\
\widehat{B}'_n(2) &= \frac{1}{2Tk_n\int_0^1 g(s)^2ds}\sum_{m=1}^{n_T-k'_n}((Y_{n,m})^2+(Y_{n,m+k'_n})^2)(\bar{Y}(g)_m^n)^2\mathbb{1}_{\{|\bar{Y}(g)_m^n|>v_n\}}, \\
\widehat{B}'_n(3) &= \frac{1}{4n_T}\sum_{m=1}^{n_T-k'_n}(Y_{n,m})^2(Y_{n,m+k'_n})^2.
\end{aligned}$$

2.3.5 Implementation

We discuss the implementation of QMLE in this section. Apparently, directly calculating the inverse of $\Sigma_n(\sigma^2,\gamma)$ would be computationally expensive when evaluating the likelihood function at each stage of an optimization routine. To avoid this problem, the classic time-series literature adopts an approximation approach of Whittle (1951). Unfortunately, we can show the Whittle estimator is inconsistent in our in-fill asymptotic setting, even if the noise is i.i.d. Gaussian and the efficient price is a Brownian motion with constant volatility (hence, our QMLE is in fact the MLE).

We instead implement exact likelihood through the state-space representation of an MA model. To avoid the issue of weakly identified parameters, our implementation leverages an auxiliary reduced-form MA$(q+1)$ model of the observed noisy returns:

$$Y_{n,i}=\phi(\mathrm{B})\epsilon_i,\quad \text{with}\quad \phi(x)=1+\sum_{j=1}^{q+1}\phi_j x^j,\quad 1\le i\le n,\quad \epsilon\sim\mathcal{N}(0,\chi^2). \tag{2.3.11}$$

Algorithm 1. *Our algorithm starts as follows:*

1. *Select the optimal order, $\widehat{q}_{n,\mathrm{AIC}}$, of the MA process (3.3.13) for Y_n using AIC, defined by (2.3.4) but rewritten in terms of χ^2 and ϕ.*

2. *Obtain exact quasi-likelihood estimates of $\widehat{\chi}^2$ and $\widehat{\phi}_j$ for $1\le j\le\widehat{q}_n+1$, using the state-*

space representation of (3.3.13) and Kalman filtering (see, e.g., Gardner, Harvey, and Phillips (1980)), where $\widehat{q}_n = \widehat{q}_{n,\mathrm{AIC}} \vee \log n_T$.[12]

3. *Construct volatility and noise autocovariance estimators using the above estimates:*

$$\widehat{\sigma}_n^2(\widehat{q}_n) = \Delta_n^{-1}\widehat{\chi}^2 \left(1 + \sum_{j=1}^{\widehat{q}_n+1} \widehat{\phi}_j\right)^2,$$

$$\widehat{\gamma}_n(\widehat{q}_n)_j = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\widehat{\chi}^2 e^{\mathbb{i}j\lambda}}{|1-e^{\mathbb{i}\lambda}|^2} \left(\left| 1 + \sum_{l=1}^{\widehat{q}_n+1} \widehat{\phi}_l e^{\mathbb{i}l\lambda} \right|^2 - \left(1 + \sum_{l=1}^{\widehat{q}_n+1} \widehat{\phi}_l\right)^2 \right) d\lambda, \quad 0 \le j \le \widehat{q}_n,$$

which are obtained by comparing different parameterizations of the return autocovariances.

Algorithm 2 is sufficient for estimating volatility and noise autocovariance. If we are further interested in (ι^2, θ), a unique solution $(\widehat{\iota}_n^2(q), \widehat{\theta}_n(q))$ exists with probability approaching one when noise is sufficiently large relative to sample size.[13] When noise is small, however, these parameters are weakly identified, and there may not be any solution such that $\widehat{\iota}_n^2(q)$ is positive and $\widehat{\theta}_n(q)$ is real.

2.4 Monte Carlo Simulations

In this section, we examine the finite-sample performance of our volatility estimator and compare it with alternative nonparametric estimators in the literature. Throughout, we fix $T = 1$ day and the average sampling frequency every 5 seconds. We conduct $1,000$ Monte Carlo trials in total.

12. Packages of standard programming software (e.g., R and Matlab) are available that implement a likelihood estimator for MA models, despite the fact that some packages rely on Whittle approximations.

13. Da and Xiu (2021a) suggest a Newton-Raphson algorithm based on Wilson (1969) to solve for $\widehat{q}_n + 1$ model parameters (ι^2, θ) of the MA($\widehat{q}_n$) noise process from up-to-$\widehat{q}_n$th-order autocovariances $\widehat{\gamma}_n(\widehat{q}_n)_j$, $0 \le j \le \widehat{q}_n$.

We simulate X_t and σ_t^2 according to the same log-volatility model as in Li and Xiu (2016):

$$\begin{cases} dX_t &= (0.05 + 0.5\sigma_t^2)dt + \sigma_t dW_t + J^X dN_t, \\ \sigma_t^2 &= D_t \exp\left(-2.8 + 6F_t\right), \quad dF_t = -4F_t dt + 0.8d\widetilde{W}_t + J^F dN_t - 0.02\lambda_N dt, \end{cases}$$

where $\mathbb{E}[dW_t d\widetilde{W}_t] = -0.8dt$, $J^X \sim \mathcal{N}(0, 0.02^2)$, $J^F \sim \mathcal{N}(0.02, 0.02^2)$, N_t is a Poisson process with intensity $\lambda_N = 25$, and D_t captures the diurnal effect:

$$D_t = 0.75\exp(-10t/T) + 0.25\exp(-10(1 - t/T)) + 0.8.$$

The arrival of trades follows an inhomogeneous Poisson process with rate $nT^{-1}\xi_t^{-1} = nT^{-1}(1 + \cos(2\pi t/T)/2)$, so that fewer trades arrive in the middle of the day.

With respect to noise, we simulate an MA(∞) model with heteroscedastic variance:

$$U_i = \iota\eta_{t_i}(1 - 0.4B)^{-1}(1 + 0.2B)\varepsilon_i.$$

We vary the magnitude of the noise, ι, which takes values from: 10^{-4} (small noise), 5×10^{-4} (median noise), and 2.5×10^{-3} (large noise). These noise levels are common choices in the literature and also relevant for empirical data. η_t captures the heteroscedasticity of the noise, which follows

$$d\eta_t = 10 \times \left(\left(1 + 10^{-1}\cos(2\pi t/T)\right) - \eta_t\right)dt + 0.1dW_t,$$

where W_t is the same Brownian motion that drives X. We round the observed prices to the nearest cent: $\tilde{X}_{t_i} = \log\left(\left[100 \times \exp(X_{t_i} + U_i)\right]\right) - \log 100$, where $[\cdot]$ means rounding to the nearest integer.[14]

14. Although our theory does not allow for this type of rounding errors, we simulate this model to demonstrate that the effect of rounding appears negligible in a finite sample.

We first compare the performance of the central limit theory using (i) (2.3.8), (ii) (2.3.9), and (iii) (2.3.5), given by Theorem 1. Recall that (i) works when noise is large and (ii) works when noise is small, whereas only (iii) works uniformly. Figure 2.1 compares the histograms of the standardized estimates of AIC*-QMLE that employ $\widehat{q}_{n,\mathrm{AIC}^\star} := \widehat{q}_{n,\mathrm{AIC}} \vee \log n_T$, using the corresponding asymptotic variances of these different scenarios. The histograms that correspond to (i) (resp. (ii)) on the top (resp. middle) panels do not match the standard normal density when noise is small (resp. large). By contrast, histograms on the bottom panel match the normal density uniformly well.

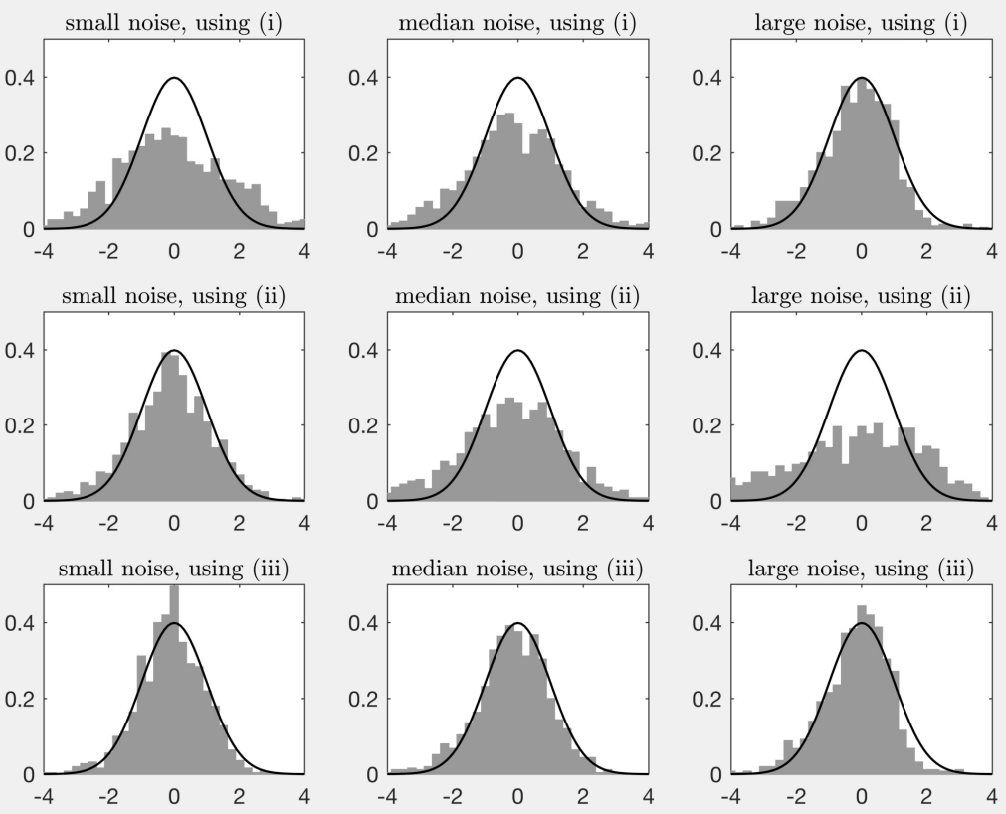


Figure 2.1: Histograms of Standardized Volatility Estimates

Note: This figure plots the histograms of standardized estimates of AIC*-QMLE using the central limit results given by (i) (2.3.8, top), (ii) (2.3.9, middle), and (iii) (2.3.5, bottom) of Theorem 1. Solid lines plot the density of the standard normal distribution. The noise magnitude parameter ι takes three values: 10^{-4} (small), 5×10^{-4} (median), and 2.5×10^{-3} (large).

We then compare a variety of volatility estimators, including the usual realized volatility

estimators using all returns and 5-minute returns; the MA(1)-based QMLE of Xiu (2010); the recent pre-averaging estimator (PVG) proposed by Jacod, Li, and Zheng (2019); and the flat-top realized kernel estimator (FRK) proposed by Varneskov (2016). The last two estimators are consistent, even in the presence of colored noise. Other estimators that are robust to i.i.d. noise, such as two-scales realized variance and realized kernels, have very similar performance compared with the MA(1)-based QMLE, and hence are omitted.

Table 2.1: Simulation Results for Volatility Estimation

	Small Noise		Median Noise		Large Noise	
$\times 10^{-2}$	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
AIC*-QMLE	-0.02	1.09	-0.03	1.28	0.03	2.35
AIC-QMLE	0.06	0.75	0.11	1.40	0.07	2.34
PVG	-0.70	2.37	-0.24	2.34	2.08	3.50
FRK	-0.01	1.02	0.45	1.18	0.09	2.30
RV (5 min)	-0.02	1.68	1.62	2.58	42.24	43.44
MA(1) - QMLE	1.13	1.21	22.51	22.70	477.31	480.96
RV	3.11	3.12	47.99	48.02	1168.96	1169.42

Note: This table compares various volatility estimators in different simulation scenarios. "AIC*-QMLE" and "AIC-QMLE" are MA(q)-likelihood estimators using $\widehat{q}_{n,\text{AIC}^\star} = \widehat{q}_{n,\text{AIC}} \vee \log n_T$ and $\widehat{q}_{n,\text{AIC}}$ as the selected orders, respectively. "PVG" refers to the pre-averaging estimator of Jacod, Li, and Zheng (2019). We report only PVG estimates with the smallest RMSE of all 27 versions, in which case $h_n = [n^{1/2}]$, $h_n' = 2$, and $h_n'' = 3$ when noise is small; $h_n = [n^{1/2}]$, $h_n' = 2$, and $h_n'' = 9$ in the case of a median noise level; and $h_n = [n^{1/2}]$, $h_n' = 2$, and $h_n'' = 3$ when noise is large. "FRK" refers to the flat-top realized kernel estimator of Varneskov (2016). We only report FRK estimates with the smallest RMSE of all 27 versions, in which case $H_n = 25$, $\bar{h}_n = [0.5 \times H_n^{-0.6}]$, and $\bar{h}_n' = 0$ when noise is small; $H_n = 25$, $\bar{h}_n = [0.5 \times H_n^{-0.6}]$, and $\bar{h}_n' = 10$ in the case of a median noise level; and $H_n = 75$, $\bar{h}_n = [H_n^{-0.6}]$, and $\bar{h}_n' = 10$ when noise is large. "5-min RV" is the popular realized volatility estimator based on the 5-minute subsample. "MA(1)-QMLE" uses the MA(1) likelihood. "RV" is the realized volatility based on the full sample.

To construct the PVG, Jacod, Li, and Zheng (2019) propose the following:

$$\widehat{\sigma}_n^{2,\text{PVG}} = \left(T \sum_{i=0}^{h_n} (g_i^n)^2 \right)^{-1} \left(\sum_{i=0}^{n_T - h_n} (\widetilde{Y}_i^n)^2 - n_T \sum_{j=-h_n''}^{h_n''} \widehat{\gamma}_{|j|}^{\text{JLZ}} \sum_{i=0}^{h_n} \bar{g}_i^n \bar{g}_{i-j}^n \right),$$

where g is defined in Section 2.3.4, $g_i^n = g\left(\frac{i}{h_n}\right), \widetilde{Y}_i^n = \sum_{j=1}^{h_n-1} g_j^n Y_{n,i+j}, \bar{g}_i^n = g_{i+1}^n - g_i^n$, and $\widehat{\gamma}_j^{\text{JLZ}}$ is the estimator of the jth autocovariance proposed by Jacod, Li, and Zheng (2017):

$$\widehat{\gamma}_j^{\text{JLZ}} = \frac{1}{n_T} \sum_{i=0}^{n_T+1-j-4h_n'} \left(\widetilde{X}_{t_i} - \frac{1}{h_n'} \sum_{l=0}^{h_n'-1} \widetilde{X}_{t_{i+j+l+h_n'}}\right)\left(\widetilde{X}_{t_{i+j}} - \frac{1}{h_n'} \sum_{l=0}^{h_n'-1} \widetilde{X}_{t_{i+j+l+3h_n'}}\right).$$

The PVG estimator depends on three tuning parameters:

$$h_n' \sim \frac{1}{\Delta_n^{\eta}} \quad \text{with} \quad \frac{1}{2v+1} < \eta < \frac{1}{2}, \quad h_n \sim \frac{1}{\Delta_n^{1/2}}, \quad h_n'' \sim \frac{1}{\Delta_n^{1/8}},$$

where v is the ρ-mixing exponent of ε. h_n' determines the local window size used to estimate realization of the noise, h_n is the usual local window size for averaging returns, and h_n'' determines the maximum number of lags of nonzero noise autocovariances. Jacod, Li, and Zheng (2017) suggest $h_n' = 6$ in simulations with 1-second data. According to their criterion, when data are sampled at 5-second frequency, h_n' must be an even smaller integer in a finite sample, so we choose h_n' from $\{2, 4, 6\}$. Jacod, Li, and Zheng (2019) suggest $h_n'' = 3$ and $h_n = [0.8 \times n^{1/2}]$ based on their simulation setting, so we choose h_n'' from $\{3, 6, 9\}$ and h_n from $\{[0.6 \times n^{1/2}], [0.8 \times n^{1/2}], [1.0 \times n^{1/2}]\}$ for robustness. In total, we consider $3 \times 3 \times 3 = 27$ combinations of tuning parameters. To save space, we report only the best pre-averaging volatility estimate in terms of root-mean-square error (RMSE), despite the fact that this is not feasible beyond simulations and that the choice of tuning parameters matters quite a bit.

Regarding the FRK, Varneskov (2016) proposes $\widehat{\sigma}_n^{2,\text{FRK}}$ defined as:

$$\frac{1}{T}\gamma_0(Y_n^*) + \frac{1}{T}\sum_{h=1}^{\bar{h}_n}(\gamma_h(Y_n^*) + \gamma_{-h}(Y_n^*)) + \frac{1}{T}\sum_{h=\bar{h}_n+1}^{n_T-2\bar{h}_n'} k\left(\frac{h-\bar{h}_n}{H_n}\right)(\gamma_h(Y_n^*) + \gamma_{-h}(Y_n^*)),$$

where $k(\cdot)$ is the Parzen kernel, $\gamma_h(\cdot)$ is the hth-lag sample autocovariance function,

$$\gamma_h(Y_n^*) = \sum_{j=\bar{h}'_n+(0\vee h)}^{n_T-\bar{h}'_n+(0\wedge h)} Y_{n,j}^* Y_{n,j-h}^*,$$

and Y_n^* is a vector of returns after jittering,

$$Y_{n,j}^* = \begin{cases} \widetilde{X}_{t_{\bar{h}'_n}} - \dfrac{1}{\bar{h}'_n} \displaystyle\sum_{k=0}^{\bar{h}'_n-1} \widetilde{X}_{t_k}, & j = \bar{h}'_n; \\ -\widetilde{X}_{t_{n_T-\bar{h}'_n-1}} + \dfrac{1}{\bar{h}'_n} \displaystyle\sum_{k=n_T-\bar{h}'_n}^{n_T} \widetilde{X}_{t_k}, & j = n - \bar{h}'_n; \\ Y_{n,j}, & \text{Otherwise.} \end{cases}$$

There are also three tuning parameters:

$$H_n \sim \frac{1}{\Delta_n^{1/2}}, \quad \bar{h}_n \sim \Delta_n^{(\eta-1)/2} \text{ with } 0 < \eta < \frac{1+2\bar{r}}{2+2\bar{r}}, \quad \bar{h}'_n = \Delta_n^{-\varpi} \text{ with } \frac{1}{4} < \varpi < \frac{1}{2},$$

where $\bar{r}$ is the α-mixing exponent of ε. H_n is the usual bandwidth for kernel estimators, $\bar{h}_n$ controls the flatness of the kernel, and $\bar{h}'_n$ controls jittering. We choose H_n from $\{25, 50, 75\}$, $\bar{h}_n$ from $\{[0.5 \times H_n^{-0.6}], [1 \times H_n^{-0.6}], [1.5 \times H_n^{-0.6}]\}$, and $\bar{h}'_n$ from $\{0, 5, 10\}$. We report only the best estimate (in terms of the RMSE) of all 27 combinations.

Table 2.1 presents the comparison results.[15] When noise is small, all estimators work well, except that RV is a bit worse. This finding is not surprising, because all other estimators are somewhat robust to noise, and noise is present despite its small magnitude. When noise becomes larger and its dependence is thus more evident, the MA(1)-QMLE deteriorates, because it is only robust, essentially, to the white noise. The 5-min RV is better, but still has a substantial bias and a large RMSE, to the extent that using it is not reasonable in practice. AIC*- and AIC-QMLEs, FRK, and PVG are all well behaved, of which the PVG

15. We thank Yingying Li, Xinghua Zheng, and Rasmus Varneskov for sharing their codes with us.

is the worst. That said, of all the estimators we compare, only AIC*-QMLE is uniformly valid. Our experiment also shows that FRK yields negative volatility estimates less often than the PVG estimator, whereas all other estimators guarantee positivity.

2.5 Conclusion

We propose a simple volatility estimator based on the likelihood of an MA model, whose order is selected based on AIC. We establish uniformly valid inference on volatility over a large and flexible class of noise DGPs, featuring autocorrelations of an infinite order and an arbitrarily vanishing noise magnitude. The convergence rate of our estimator is greater than or equal to $n^{1/4}$, which depends on the noise magnitude and its dependence structure. Our estimator requires a single tuning parameter in order selection, and it always guarantees the positivity of volatility estimates. For these reasons, it delivers more desirable finite-sample performance than alternative nonparametric estimators.

Chapter 3

Disentangling Autocorrelated Intraday Returns[1]

3.1 Introduction

Autocorrelations in stock returns are ubiquitous. The earlier literature regards such autocorrelations as evidence against market efficiency. Nonetheless, as market efficiency has improved over past decades, autocorrelations have remained a salient feature of intraday stock returns sampled at sufficiently high frequencies. The modern view of such autocorrelations is that they arise from market microstructure frictions, such as bid-ask bounces, nonsynchronous trading, price discreteness, etc, which coalesce into efficient equilibrium prices and lead to the convoluted dynamics of returns.

To disentangle the observed autocorrelations in intraday returns, we model the transaction price as a discretized continuous-time semimartingale process plus a discrete-time moving-average process. The former represents the efficient price process that features return heteroscedasticity in the form of stochastic volatility and jumps, but does not contribute to any autocovariance; the latter serves as a reduced-form description of the microstructure friction that is the main driver behind the observed autocovariances.

To conduct inference on various model components and parameters, we construct a tractable quasi-maximum likelihood estimator (QMLE), pretending that the transaction

1. This Chapter is a joint work with Dacheng Xiu.

price arrives regularly and comprises a Brownian motion with constant volatility and an MA(q) noise. We select q based on the Akaike/Bayesian information criteria (AIC/BIC). While our estimator shares the same likelihood with that from an MA($q+1$) model, our asymptotic design is in-fill, i.e., the number of observations increases within a fixed window–say, a trading day–which renders our analysis rather different from the usual long-span asymptotics in the classic time-series setting.

In a related paper, Da and Xiu (2021b) show how to conduct uniformly valid inference on volatility over a large class of MA(∞) models that allow for an asymptotically vanishing noise with a flexible dependence structure. In this paper, our main objective is to develop asymptotic properties of the estimator for noise parameters. When the noise data-generating process (DGP) follows a finite-order moving-average model, we show that our quasi-likelihood estimator, combined with BIC, recovers the true model asymptotically, is consistent with respect to the noise parameters, and achieves a pointwise central limit theory at the usual rate of $n^{1/2}$. Moreover, we develop uniform consistency results when noise follows an MA(∞) process. As alternatives to our semiparametric approach, Jacod, Li, and Zheng (2017) and Li and Linton (2021) provide nonparametric estimators of the serial correlations of the microstructure noise based on local averaging and differencing strategies, respectively. They focus on the case in which noise is large, whereas we also allow for vanishing noise. More importantly, our likelihood-based approach provides a benchmark on the efficiency of noise parameters.

We apply our estimator to analyze all intraday returns of S&P 1,500 index constituents from 1996 to 2016. Several interesting findings emerge. The microstructure noise is present in 5-minute returns, at least for small and mid caps, though it is an order of magnitude smaller in recent years than at the beginning of the sample, thanks to the improvement in market efficiency. For a sizable portion of stock-day pairs, it appears that the noise is either absent or approximately follows an i.i.d. assumption. For the remaining stocks with

autocorrelated noise, the duration of autocorrelations has been on the decline, from several minutes in 1996 to merely 10 seconds on average for large caps and 100 seconds for small caps in 2016.

Empirical evidence of autocorrelations in the returns of transaction prices goes back to as early as Niederhoffer and Osborne (1966), Simmons (1971), and Garbade and Lieber (1977). Among others, Hasbrouck and Ho (1987) document positive autocorrelations in intraday stock returns, in returns of quote midpoints, and in the arrival of buy and sell orders. They thus propose a model of the return-generating process, which is observationally equivalent to an ARMA$(2, 2)$ model. While classical time-series models such as ARMA are convenient for dependent data, they are not appropriate for intraday returns because of the heteroscedasticity in returns.

Why do higher-order autocorrelations of returns exist? There are many economic hypotheses, such as strategic order splitting (Garbade and Lieber (1977)); optimal control of execution cost (Bertsimas and Lo (1998)); price impact and inventory control (Kyle (1985a), Amihud and Mendelson (1980)); the crowd effect or herding (Tóth, Palit, Lillo, and Farmer (2015)); and high-frequency trading Brogaard, Hendershott, and Riordan (2014). Our objective here is modest. We aim to estimate parameters in a general class of reduced-form models, since many structural economic models yield specific reduced-form models–see, for example, Hasbrouck (2007)–with differences only in how the reduced-form parameters relate to structural parameters.

There is an enormous literature on the estimation of quadratic variation or its components using noisy high-frequency data; e.g., the two-scale or multi-scale estimators by Zhang, Mykland, and Aït-Sahalia (2005) and Zhang (2006); the realized kernel estimator and its extensions by Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008) and Barndorff-Nielsen, Hansen, Lunde, and Shephard (2011); the pre-averaging estimator by Jacod, Li, Mykland, Podolskij, and Vetter (2009) and Jacod, Podolskij, and Vetter (2010); the quasi-maximum

likelihood estimator (QMLE) by Xiu (2010); and the local method of moments estimator by Reiß (2011). An "essentially" white noise assumption is most common in this strand of the literature, with the exception of Jacod, Li, and Zheng (2019), Varneskov (2016), and Da and Xiu (2021b), who tackle general colored-noise processes for the purpose of volatility estimation. Related work also include Aït-Sahalia, Mykland, and Zhang (2005), Aït-Sahalia, Mykland, and Zhang (2011), Kalnina and Linton (2008), and Bibinger, Hautsch, Malec, and Reiß (2019). Unlike the above papers, which treat noise as nuisance parameters in the estimation of quadratic variation, our target here is mainly the temporal dependence of intraday returns beyond the first-order autocorrelations. Chang, Delaigle, Hall, and Tang (2018) also focus on analyzing the statistical properties of the noise process and propose an estimator of noise density and noise moments in an i.i.d. noise setting.

The paper is organized as follows. Section 3.2 sets up the model. Section 3.3 introduces the estimator and provides the main asymptotic results. Section 3.4 reports Monte Carlo simulations. We analyze volatilities and noise for S&P Composite 1,500 index constituents in Section 3.5, and Section 3.6 concludes. The online supplemental appendix contains all mathematical proofs.

3.2 Model Assumptions

We assume that transaction prices $\widetilde{X}$ are observed at t_i, for $i = 1, 2, \ldots, n_T$, within a fixed window $[0, T]$. They comprise two components: $\widetilde{X}_{t_i} = X_{t_i} + U_i$, where X_{t_i} is (the logarithm of) the efficient equilibrium price and U_i is the microstructure noise associated with the ith observation. Furthermore, the efficient price satisfies:

Assumption 6. *The logarithm of the efficient price process X_t is an Itô-semimartingale defined on some filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$ and satisfies*

$$X_t = X_0 + \int_0^t \mu_s ds + \int_0^t \sigma_s dW_s + (\delta \mathbb{1}_{\{|\delta| \leq 1\}}) \star (\underline{\mu} - \underline{\nu})_t + (\delta \mathbb{1}_{\{|\delta| > 1\}}) \star \underline{\mu}_t, \tag{3.2.1}$$

where μ_t and σ_t are adapted and locally bounded, W is a standard Brownian motion, and $\underline{\mu}$ is a Poisson random measure on $\mathbb{R}_+ \times E$, where E is a Polish space. The compensator $\underline{\nu}$ satisfies $\underline{\nu}(dt, du) = dt \otimes \lambda(du)$ for some σ-finite measure λ on E. Moreover, $|\delta(\omega, t, u)| \wedge 1 \leq \Gamma_m(u)$ for all (ω, t, u) with $t \leq \tau_m(\omega)$, where $\{\tau_m\}$ is a localizing sequence of stopping times and $\{\Gamma_m\}$ a sequence of deterministic functions satisfying $\int \Gamma_m^2(u)\lambda(du) < \infty$.

In addition, the process $Z_t = (\mu_t, \sigma_t^2)$ is also an Itô semimartingale on the space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$ with the form

$$Z_t = Z_0 + \int_0^t \widetilde{\mu}_s ds + \int_0^t \widetilde{\sigma}_s d\widetilde{W}_s + (\widetilde{\delta}\mathbb{1}_{\{|\widetilde{\delta}|\leq 1\}}) \star (\underline{\mu} - \underline{\nu})_t + (\widetilde{\delta}\mathbb{1}_{\{|\widetilde{\delta}|>1\}}) \star \underline{\mu}_t, \tag{3.2.2}$$

where $\widetilde{\mu}_t$ and $\widetilde{\sigma}_t$ are locally bounded adapted processes, $\widetilde{W}$ is a multivariate Brownian motion, potentially correlated with W, and $\widetilde{\delta}$ is a predictable function such that for some deterministic function $\widetilde{\Gamma}_m(u)$, $\|\widetilde{\delta}(\omega, t, u)\| \wedge 1 \leq \widetilde{\Gamma}_m(u)$ for all $\omega \in \Omega$, $t \leq \tau_m(\omega)$, and $\int \widetilde{\Gamma}_m^2(u)\lambda(du) < \infty$.

While the efficient prices are defined in continuous time, we only observe their noisy versions at discrete time points. We now describe the assumption of the arrival times of transactions:

Assumption 7. *The sequence of observation times $\{t_i : i \geq 0\}$ satisfies $t_0 = 0$ and $t_i = t_{i-1} + \frac{T}{n}\xi_{t_{i-1}}\chi_i$, where the sequence $\{\chi_i : i \geq 1\}$ is i.i.d., $(0, \infty)$-valued, defined on $(\Omega, \mathcal{F}, \mathbb{P})$, and independent of the σ-field $\mathcal{F}_\infty = \bigvee_{t>0} \mathcal{F}_t$, with $m_j = \mathbb{E}((\chi_i)^j) < \infty$ and $m_1 = 1$, for all $j > 0$. In addition, the process $\xi = (\xi_t)_{t\geq 0}$ is a nonnegative Itô-semimartingale defined on $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$ in the form of (3.2.2), such that neither ξ_t nor ξ_{t-} vanishes.*

The intervals between adjacent transactions are determined by a continuous-time process, ξ_t, and a discrete-time process, χ_i, jointly. This assumption allows for dependence between trading times and the underlying driving forces of efficient prices, and thereby accommodates a large class of sampling schemes; see Jacod, Li, and Zheng (2017) for detailed discussions.

Next, we impose a discrete-time moving-average process for the microstructure noise to capture the potential temporal dependence in the transaction prices:[2]

Assumption 8. *The noise sequence* $\{U_i : i \geq 0\}$ *consists of random variables defined on the probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *such that* $\{U_i : i \geq 0\}$ *has an MA*(∞) *representation with mean* 0*:*

$$U_i = \eta_{t_i} \iota^{(n)} \theta^{(n)}(\mathrm{B}) \varepsilon_i, \quad \textit{with} \quad \theta^{(n)}(x) = 1 + \sum_{j=1}^{\infty} \theta_j^{(n)} x^j, \tag{3.2.3}$$

where B *is the lag operator;* $\varepsilon_i \overset{i.i.d.}{\sim} (0,1)$*, defined on* $(\Omega, \mathcal{F}, \mathbb{P})$*, is independent of* $\mathcal{F}_\infty$ *and* $\{\chi_i : i \geq 1\}$*, and has finite moments of all orders;* $(\eta_t)_{t \geq 0}$ *is an* $(\mathcal{F}_t)$*-adapted nonnegative Itô-semimartingale that satisfies the same form of* (3.2.2)*; and* $\iota^{(n)}$ *is a deterministic nonnegative number that characterizes the noise magnitude and satisfies* $\iota^{(n)} \leq K$*.*

The noise again depends on a continuous-time process η_t and a discrete-time moving-average process U. The former introduces dependence between noise and the underlying efficient price, whereas the latter dictates the temporal dependence of the noise. Combining the two allows for heteroscedastic, temporally dependent, and endogenous noise.

The parameters of interest in our study are autocovariances $\{\gamma_j^{(n)} : j \geq 0\}$ and autocorrelations $\{\rho_j^{(n)} : j \geq 1\}$ of the noise process, defined as

$$\gamma_j^{(n)} = (\iota^{(n)})^2 \frac{\int_0^T \eta_s^2 \xi_s^{-1} ds}{\int_0^T \xi_s^{-1} ds} \times \kappa_j^{(n)}, \quad j \geq 0, \quad \text{and} \quad \rho_j^{(n)} = \kappa_j^{(n)} / \kappa_0^{(n)}, \quad j \geq 1, \tag{3.2.4}$$

where $\kappa_j^{(n)}$ is given by

$$\kappa_j^{(n)} = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(\lambda; \theta^{(n)}) e^{\mathbb{i}\lambda j} d\lambda, \quad j \geq 0, \tag{3.2.5}$$

2. We use a superscript (n) on noise parameters to facilitate discussion of uniformity over different sequences of data-generating processes (DGPs) of noise indexed by n. n is a nonobservable mathematical abstraction. All limits are taken as $n \to \infty$. K is a generic n-independent positive constant that may vary from line to line.

and $g(\lambda;\theta^{(n)}) = |\theta^{(n)}(e^{\mathbb{i}\lambda})|^2$ is the spectral density of $\theta^{(n)}(\mathrm{B})\varepsilon$. While the autocovariances depend on the underlying processes η_t and ξ_t that drive the sampling times and noise magnitudes, respectively, the autocorrelations are entirely determined by the set of parameters $\{\theta_j^{(n)} : j = 1, 2, \ldots, \infty\}$ in the MA process.

Finally, we need some regularity assumption on the behavior of the spectral density of the noise process so that it is uniformly invertible and its long range dependence cannot be overly strong.

Assumption 9. *For each $n \geq 1$, the spectral density function of $\theta^{(n)}(\mathrm{B})\varepsilon$ satisfies for some fixed $\alpha > 3$,*

$$\inf_{\lambda} g(\lambda;\theta^{(n)}) \geq \frac{1}{K} \quad \textit{and} \quad \left| \int_{-\pi}^{\pi} g(\lambda;\theta^{(n)}) e^{\mathbb{i}\lambda j} d\lambda \right| \leq K j^{-\alpha}, \quad \forall j \geq 0.$$

3.3 Main Results

In what follows, we will discuss the constructed estimators and their asymptotic properties.

3.3.1 Quasi-likelihood Estimation

To estimate volatility, Da and Xiu (2021b) propose a quasi-likelihood approach based on a misspecified model for observed returns. We adopt the same estimator here, but focus on the noise parameters. Specifically, we pretend that the efficient price X (in logarithm) is a Brownian motion with constant volatility but no drift, and that the noise U follows a Gaussian MA(q) model with the order q to be determined:

$$dX_t = \sigma dW_t; \quad U_i = \iota\theta(\mathrm{B})\varepsilon_i, \quad \text{with} \quad \theta(x) = 1 + \sum_{j=1}^{q} \theta_j x^j, \quad \text{and} \quad \varepsilon_i \sim \mathcal{N}(0,1).$$

Under this model, the observed log-return vector $Y_n = (Y_{n,1}, Y_{n,2}, \ldots, Y_{n,n_T})^\intercal$,

$$Y_{n,i} = X_{t_i} - X_{t_{i-1}} + U_i - U_{i-1}, \quad 1 \leq i \leq n_T. \tag{3.3.6}$$

follows a reduced-form Gaussian MA$(q+1)$ model, whose $n_T \times n_T$ covariance matrix Σ_n is given by

$$\Sigma_n(\sigma^2, \iota^2, \theta) = \sigma^2 \Delta_n \mathbb{I}_n + \sum_{h=0}^{n_T-1} (2\gamma_h - \gamma_{h+1} - \gamma_{h-1}) \mathbb{G}_n^h, \tag{3.3.7}$$

where $(\mathbb{I}_n)_{ij} = \delta_{i,j}$, $(\mathbb{G}_n^h)_{ij} = \delta_{h,|i-j|}$, and γ_h is the h-th order autocovariance of U:

$$\gamma_h = \frac{\iota^2}{2\pi} \int_{-\pi}^{\pi} g(\lambda; \theta) e^{\mathbb{i}\lambda h} d\lambda, \quad \text{where} \quad g(\lambda; \theta) = \left|\theta(e^{\mathbb{i}\lambda})\right|^2. \tag{3.3.8}$$

Since we are interested in the noise autocovariances, we reparameterize the likelihood function in terms of (σ^2, γ):

$$L_n(\sigma^2, \gamma) = -\frac{1}{2} \log\det(\Sigma_n(\sigma^2, \gamma)) - \frac{1}{2} \mathrm{tr}(\Sigma_n(\sigma^2, \gamma)^{-1} Y_n Y_n^\intercal), \tag{3.3.9}$$

where $\Sigma_n(\sigma^2, \gamma) := \Sigma_n(\sigma^2, \iota^2, \theta)$ and γ is the $(q+1)$-dimensional vector of the noise autocovariances.

We define $(\widehat{\sigma}_n^2(q), \widehat{\gamma}_n(q))$ as the maximizer of $L_n(\sigma^2, \gamma)$:

$$(\widehat{\sigma}_n^2(q), \widehat{\gamma}_n(q)) = \arg \max_{(\sigma^2, \gamma) \in \Pi_n(q)} L_n(\sigma^2, \gamma), \tag{3.3.10}$$

where, following Da and Xiu (2021b), the parameter space $\Pi_n(q)$ is defined as

$$\left\{(\sigma^2, \gamma) \in \mathbb{R}^{q+2} : \inf_\lambda f(\lambda; \sigma^2, \gamma, \Delta_n) \geq \frac{\Delta_n}{K}, \quad \sigma^2 + |\gamma_0| + \frac{\sum_{j=1}^\infty j^2 |\gamma_j|}{\inf_\lambda |\sigma^2 \Delta_n + f(\lambda; \gamma)|} \leq K\right\}. \tag{3.3.11}$$

Here $f(\lambda; \sigma^2, \gamma, \Delta_n)$ stands for the spectral density of Y_n under the quasi-model: $f(\lambda; \sigma^2, \gamma, \Delta_n) = \sigma^2 \Delta_n + (2 - 2\cos\lambda) f(\lambda; \gamma)$, with $f(\lambda; \gamma) = \sum_{j=-\infty}^\infty \gamma_{|j|} e^{\mathbb{i}j\lambda}$.

To determine an appropriate order q, we use information criteria, such as BIC, which in

our setting can be written as

$$\mathrm{BIC}_n(q) = q \log n_T - 2 \max_{(\sigma^2,\gamma)\in\Pi_n(q)} L_n(\sigma^2, \gamma).$$

Our choice of order q will be based on

$$\widehat{q}_n = \arg \min_{q \le n_T^{1/3}} \mathrm{BIC}_n(q). \tag{3.3.12}$$

We can define a similar criterion based on AIC, by replacing $q \log n_T$ above by $2q$. Hannan (1980) shows that using BIC results in consistent order selection for ARMA models. We demonstrate that a similar result with BIC also holds in our setting. We will therefore focus on BIC in the following discussion.

3.3.2 Implementation

We implement the exact likelihood via an auxiliary reduced-form MA$(q+1)$ model of the observed noisy returns:

$$Y_{n,i} = \phi(B)\epsilon_i, \quad \text{with} \quad \phi(x) = 1 + \sum_{j=1}^{q+1} \phi_j x^j, \quad 1 \le i \le n, \quad \epsilon \sim \mathcal{N}(0, \chi^2). \tag{3.3.13}$$

Algorithm 2. *Our algorithm starts as follows:*

1. *Select the optimal order, $\widehat{q}_n$, of the MA process (3.3.13) for Y_n using BIC, defined by (3.3.12) but rewritten equivalently in terms of χ^2 and ϕ.*

2. *Obtain exact quasi-likelihood estimates of $\widehat{\chi}^2$ and $\widehat{\phi}_j$ for $1 \le j \le \widehat{q}_n + 1$, using the state-space representation of (3.3.13) and Kalman filtering,*

3. *Construct volatility and noise autocovariance estimators using the above estimates:*

$$\widehat{\gamma}_n(\widehat{q}_n)_j = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\widehat{\chi}^2 e^{\mathbb{i}j\lambda}}{|1-e^{\mathbb{i}\lambda}|^2}\left(\left|1+\sum_{l=1}^{\widehat{q}_n+1}\widehat{\phi}_l e^{\mathbb{i}l\lambda}\right|^2 - \left(1+\sum_{l=1}^{\widehat{q}_n+1}\widehat{\phi}_l\right)^2\right)d\lambda, \quad 0\le j\le \widehat{q}_n,$$

$$\widehat{\sigma}_n^2(\widehat{q}_n) = \Delta_n^{-1}\widehat{\chi}^2\left(1+\sum_{j=1}^{\widehat{q}_n+1}\widehat{\phi}_j\right)^2,$$

which are obtained by comparing different parameterizations of the return autocovariances.

4. *Solve* $\widehat{q}_n+1$ *nonlinear equations for* $\widehat{q}_n+1$ *model parameters* $(\widehat{\iota}_n^2(\widehat{q}_n), \widehat{\theta}_n(\widehat{q}_n))$ *from* $\widehat{\gamma}_n(\widehat{q}_n)$ *obtained in Step 3:*

$$\widehat{\gamma}_n(\widehat{q}_n)_j = \widehat{\iota}_n^2(\widehat{q}_n)\sum_{l=0}^{\widehat{q}_n-j}\widehat{\theta}_n(\widehat{q}_n)_l\widehat{\theta}_n(\widehat{q}_n)_{l+j}, \quad 0\le j\le \widehat{q}_n. \tag{3.3.14}$$

A Newton-Raphson algorithm that converges quadratically is available from Wilson (1969).

Effectively, Step 4 is to find $\widehat{q}_n+1$ model parameters of the MA($\widehat{q}_n$) noise process from up-to-$\widehat{q}_n$th-order autocovariances $\widehat{\gamma}_n(\widehat{q}_n)_j$, $0\le j\le \widehat{q}_n$. This practice is common in the classic time-series analysis. For instance, Box, Jenkins, and Reinsel (2007) recommend using this algorithm to find initial values based on autocovariances for the maximum likelihood estimation of an MA model.

Step 3 is sufficient for volatility and noise autocovariance estimation, and it is rather simple to implement. If one is further interested in (ι^2, θ), a unique solution $(\widehat{\iota}_n^2(q), \widehat{\theta}_n(q))$ exists from Step 4, with probability approaching 1 when noise is sufficiently large relative to the sample size. When noise is small, however, these parameters are weakly identified, and (3.3.14) may have no solution such that $\widehat{\iota}_n^2(q)$ is positive and $\widehat{\theta}_n(q)$ is real.

3.3.3 Model Selection Consistency

We now discuss the asymptotic properties of the proposed estimators. The asymptotic analysis here is more involved than the classic time-series analysis, because the DGP of observed returns is misspecified. Moreover, the asymptotic design is in-fill, so that not only the dimensions, but also the entries of the covariance matrix Σ_n in the quasi-likelihood, depend on the sample size n_T; see (3.3.7). Consequently, prior results from classic time-series studies are not applicable. Even worse, the quasi-likelihood estimator does not have an explicit form.

We start with a model selection consistency result based on BIC, which allows us to conduct pointwise inference on autocovariance parameters. We thereby impose a finite-order moving-average model for the DGP of noise. In an in-fill asymptotic experiment, imposing a finite-order MA model for noise independent of the sampling frequency might appear ambiguous, in that observations are filled in between adjacent ones and the dependence structure changes as the sampling frequency approaches 0. However, as Jacod, Li, and Zheng (2017) argue, the frequency of observations in practice is fixed by the available data and does not really go to 0. Therefore, the interpretation of the asymptotic design is that the frequency of our observations is "high enough" to consider that we are "almost" in the asymptotic regime.

Theorem 2. *Suppose Assumptions 6 - 9 hold. We further assume a non-vanishing noise process with an exact MA($q^\star$) structure, i.e., $\iota^{(n)} \geq K^{-1}$ and $\theta^{(n)} \in \mathbb{R}^{q^\star}$ for all $n \geq 1$ and $\sqrt{n}(\log n)^{-1}|\theta^{(n)}_{q^\star}| \to \infty$, for some fixed $q^\star \geq 0$. Then it holds that*

$$\lim_{n\to\infty} \mathbb{P}(\widehat{q}_n = q^\star) = 1.$$

As the sample size increases, the likelihood is asymptotically dominated by that of the noise component. Therefore, the same intuition from the classic time-series result applies

here. The likelihood estimator effectively minimizes the Kullback-Leibler divergence, but only when the selected order is no smaller than the truth. Moreover, the BIC imposes a penalty just large enough to rule out orders that are greater than the truth asymptotically. The combination of these two results leads to the desired consistency in model selection.

3.3.4 Inference on Noise Autocovariances and Autocorrelations

Recall that in (3.3.10) and Step 3 of Algorithm 2, we defined and implemented estimators of noise autocovariances. We now propose estimators of autocorrelations, denoted by $\widehat{\rho}_n(\widehat{q}_n)$, which are defined as follows.

If (3.3.14) has a solution such that $\widehat{\iota}_n^2(\widehat{q}_n)$ is positive and $\widehat{\theta}_n(\widehat{q}_n)$ is real, we set[3]

$$\widehat{\rho}_n(\widehat{q}_n)_j = \frac{\widehat{\gamma}_n(\widehat{q}_n)_j}{\widehat{\gamma}_n(\widehat{q}_n)_0}, \quad j \geq 1.$$

Otherwise, we set

$$\widehat{\rho}_n(\widehat{q}_n) = 0.$$

In light of their definitions, we can regard these estimators as "hard-thresholding" estimators, in that higher-order autocovariance and autocorrelation estimates are truncated to zero beyond the selected order $\widehat{q}_n$.

Next, we prove the pointwise central limit theorem for estimators of noise autocovariances in the finite-order moving average model. The corresponding result for autocorrelations follows straightforwardly.

Theorem 3. *Suppose Assumptions 6 - 9 hold. We further assume* $\iota^{(n)} \geq K^{-1}$ *and* $\theta^{(n)} \in \mathbb{R}^{q^\star}$ *for all* $n \geq 1$ *and some fixed* $q^\star \geq 0$. *Let* $\gamma^{(n)}$ *be the* $(q^\star + 1)$*-dimensional vector of up-to-*$q^\star$*th-order autocovariances of* U*, whose components are defined in equation (3.2.4).*[4]

3. Estimates of autocovariances and autocorrelations are, of course, zero beyond the $\widehat{q}_n$-th lag.

4. Recall that the vectors $\gamma^{(n)}$ and $\gamma^\star$ are indexed from 0. We refer to $\gamma^{(n)}$ here as a $(q^\star+1)$-dimensional vector simply because $\gamma_j^{(n)} = 0$ for all $j > q^\star$, since $\theta^{(n)} \in \mathbb{R}^{q^\star}$. For this reason, in most of our discussions, we do not distinguish it from an ∞-dimensional vector. The same applies to other ∞-dimensional vectors.

Assume there exists a $(q^\star+1)$*-dimensional vector* $\gamma^\star$ *such that* $\gamma^{(n)}-\gamma^\star=o_{\mathrm{P}}(1)$*. Then it holds that*[5]

$$n^{1/2}\big(\widehat{\gamma}_n(q^\star)-\gamma^{(n)}\big)\overset{\mathcal{L}_s-\mathcal{F}_\infty}{\longrightarrow}\mathcal{MN}\left(0_{q^\star+1},\mathrm{AVAR}_1\right),$$

where

$$\mathrm{AVAR}_1=\Big(2W(\gamma^\star)^{-1}+\gamma^\star\gamma^{\star\intercal}cum_4(\varepsilon)\Big)\frac{T\int_0^T\eta_s^4\xi_s^{-1}ds}{\big(\int_0^T\eta_s^2\xi_s^{-1}ds\big)^2},$$

$cum_4(\varepsilon)$ *denotes the fourth cumulant of* ε*,*

$$W(\gamma)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\frac{\partial\log f(\lambda;\gamma)}{\partial\gamma}\Big)^\intercal\frac{\partial\log f(\lambda;\gamma)}{\partial\gamma}d\lambda.$$

This result shows that our estimator achieves the best convergence rate possible—$n^{1/2}$. In addition, the nonparametric estimation of volatility, which serves as a nuisance parameter here, does not influence the asymptotic variance of noise parameters. In fact, the asymptotic variance has the same form as in the classic time-series analysis–e.g., Brockwell and Davis (1991)–barring η and ξ terms, which are irrelevant in discrete time settings, as if the observed prices were purely made of noise. This further suggests that when ε indeed follows a Gaussian distribution, our estimator achieves the optimal efficiency.

The next corollary presents the central limit result for autocorrelations:

Corollary 2. *Suppose the same assumptions as those in Theorem 3 hold. Let* $\rho^{(n)}$ *be the* $q^\star$ *vector of up-to-*$q^\star$*th-order autocorrelations of* U *whose components are defined in equation (3.2.4). Then it holds that*

$$n^{1/2}\big(\widehat{\rho}_n(q^\star)-\rho^{(n)}\big)\overset{\mathcal{L}_s-\mathcal{F}_\infty}{\longrightarrow}\mathcal{MN}\left(0_{q^\star},\mathrm{AVAR}_2\right),$$

5. Here and throughout the appendix, $\overset{\mathcal{L}_s-\mathcal{F}_\infty}{\longrightarrow}$ stands for stable convergence in law with respect to $\mathcal{F}_\infty$.

where the ijth entry of the $q^\star \times q^\star$ matrix $AVAR_2$ is given by

$$\frac{\gamma_i^\star \gamma_j^\star}{\gamma_0^{4\star}} \left(\text{AVAR}_1\right)_{11} + \frac{1}{\gamma_0^{2\star}} \left(\text{AVAR}_1\right)_{i+1,j+1} - \frac{\gamma_i^\star}{\gamma_0^{3\star}} \left(\text{AVAR}_1\right)_{1,j+1} - \frac{\gamma_j^\star}{\gamma_0^{3\star}} \left(\text{AVAR}_1\right)_{1,i+1}.$$

Next, we construct an estimator of the asymptotic variance, AVAR_1, in Theorem 3, which naturally leads to an estimator for AVAR_2 in Corollary 2.

Proposition 2. *Suppose the same assumptions as those in Theorem 3 hold. Define*

$$\widehat{\text{AVAR}_1} = \left(2W(\widehat{\gamma}_n(\widehat{q}_n))^{-1} + \widehat{\gamma}_n(\widehat{q}_n)\widehat{\gamma}_n(\widehat{q}_n)^\intercal \widehat{cum}_4(\varepsilon)\right)\left(\widehat{\gamma}_n(\widehat{q}_n)_0 - \widehat{\gamma}_n(\widehat{q}_n)_1\right)^{-2}\widehat{B}_n,$$

where, with $k_n \sim \log n$,

$$\widehat{cum}_4(\varepsilon) = k_n \widehat{B}_n^{-1}\widehat{B}_n' - 2k_n - \left(\widehat{\gamma}_n(\widehat{q}_n)_0 - \widehat{\gamma}_n(\widehat{q}_n)_1\right)^{-2}\frac{1}{\pi}\int_{-\pi}^{\pi} f(\lambda;\widehat{\gamma}_n(\widehat{q}_n))^2(1-\cos\lambda)^2 d\lambda,$$

$$\widehat{B}_n = \frac{1}{4n_T k_n}\sum_{i=1}^{n_T-2k_n} Y_{n,i}^2 \sum_{j=k_n+1}^{2k_n} Y_{n,i+j}^2, \quad \text{and} \quad \widehat{B}_n' = \frac{1}{4n_T k_n}\sum_{i=1+k_n}^{n_T-k_n} Y_{n,i}^2 \sum_{j=-k_n}^{k_n} Y_{n,i+j}^2.$$

Then, we have

$$\left\|\widehat{\text{AVAR}_1} - n_T n^{-1}\text{AVAR}_1\right\| = o_{\text{P}}(1).$$

With this proposition in place, we can build confidence intervals for noise autocovariances and autocorrelations using $n_T^{-1}\widehat{\text{AVAR}_1}$, which does not involve the unobservable scalar n in the CLT.

3.3.5 Uniform Consistency of Noise Autocovariances and Autocorrelations

The asymptotic inference established here is pointwise, in the sense that it does not allow for model-selection mistakes. As pointed out by Leeb and Pötscher (2005), model selection errors matter in finite samples, to the extent that the prescribed asymptotic distribution could be seriously distorted. Moreover, uniformly valid inference is generally not available.

That said, we establish a uniform consistency result for $\widehat{\gamma}_n(\widehat{q}_n)$ and $\widehat{\rho}_n(\widehat{q}_n)$ with respect to $\gamma^{(n)}$ and $\rho^{(n)}$ under $\mathbb{L}^2$-norm, where all vectors are regarded as ∞-dimensional. This result sheds light on the asymptotic behavior of these estimators when noise DGPs are allowed to vary within a larger class beyond MA(q), allowing for a vanishing magnitude and a more flexible dependence structure. We characterize the class of noise DGPs we consider in the next assumption.

Assumption 10. *Define* $q_n^\star(k) := \min q$, *subject to* $n\psi_n^4 \sum_{j=q}^{2q} |\widetilde{\kappa}_j^{(n)}|^2 \leq kq \log n$, *where* $\psi_n := (1+n^{-1/2}/\iota^{(n)})^{-1}$ *and* $\widetilde{\kappa}_j^{(n)} := \sum_{i=0}^{\infty}(i+1)\psi_n^i(2\kappa_{j+1+i}^{(n)} - \kappa_{j+i+2}^{(n)} - \kappa_{j+i}^{(n)})$. *We assume for any* $0 < k < K$,

$$q_n^\star(k) = o(n^{1/3}(\iota^{(n)} \vee n^{-1/2})^{4/9}), \quad \textit{and} \quad n\psi_n^4 \sum_{j=q_n^\star(k)}^{\infty} |\kappa_j^{(n)}|^2 = O\big(q_n^\star(k)\log n\big).$$

Intuitively, $q_n^\star(k)$ mimics the "oracle" order that BIC selects. Effectively, Assumption 10 requires that this order cannot be too large and imposes an upper bound on the approximation error induced by a selected MA model. Nevertheless, these conditions in Assumption 10 are not restrictive. They accommodate common processes such as MA(∞), with $|\kappa_j^{(n)}| \sim j^{-\alpha}$ for some $\alpha > 3 \vee \frac{3}{2+4\log \iota^{(n)}/\log n}$, as well as any finite order ARMA(p,q) with an arbitrarily shrinking noise magnitude $\iota^{(n)} \lesssim 1$.

We are now ready to present the uniform consistency result for autocovariances and autocorrelations:

Theorem 4. *For any sequence of DGPs that satisfies Assumptions 6 - 10, we have*

$$\big\|\widehat{\gamma}^{(n)}(\widehat{q}_n) - \gamma^{(n)}\big\|^2 = O_{\mathrm{P}}\Big(n^{-1}(\iota^{(n)})^4(\widehat{q}_n+1)^2\log n + n^{-3}(n^{1/2}\iota^{(n)}+1)(\widehat{q}_n+1)^4\log n\Big).$$

If, in addition, we assume $\iota^{(n)} \geq Kn^{-2/3}(\log n)^{1/4}$, *it holds that*

$$\big\|\widehat{\rho}_n(\widehat{q}_n) - \rho^{(n)}\big\|^2 = O_{\mathrm{P}}\Big((\iota^{(n)})^{-4}\big\|\widehat{\gamma}^{(n)}(\widehat{q}_n) - \gamma^{(n)}\big\|^2\Big).$$

In general, the autocorrelation $\rho^{(n)}$ is weakly identified in the presence of small noise. The last part of Theorem 4 rules out this scenario, restricting the class of DGPs such that the noise variance cannot be too small.

Whereas consistent estimation of autocorrelations requires a more restrictive class of DGPs, Theorem 4 allows for arbitrarily small and vanishing noise for autocovariances. The case of small noise is highly relevant in practice, as shown from our empirical study below. Our result is complementary to the asymptotic theory developed by Jacod, Li, and Zheng (2017) and Li and Linton (2021), who focus on the case of non-vanishing noise.

3.3.6 Quadratic Representation

The QMLE estimator appears to have a rather different structure compared with alternative nonparametric estimators in the literature, e.g., realized kernels, which can be regarded as quadratic estimators. In this section, we propose an alternative but equivalent quadratic form of the QMLE, which sheds light on its connection with and distinction from these quadratic estimators. We do so for both volatility and noise autocovariance estimators.

Theorem 5. *Suppose the same assumptions as those in Theorem 3 hold and that* $\gamma^{(n)} = \gamma^\star$. *The QMLE* $(\widehat{\sigma}^2_n(q^\star), \widehat{\gamma}_n(q^\star))$ *satisfies that for* $0 \leq j \leq q^\star$,

$$\widehat{\sigma}^2_n(q^\star) = Y_n^\intercal \mathcal{W}_n(\widehat{\sigma}^2_n(q^\star), \widehat{\gamma}_n(q^\star); 1) Y_n, \quad \widehat{\gamma}_n(q^\star)_j = Y_n^\intercal \mathcal{W}_n(\widehat{\sigma}^2_n(q^\star), \widehat{\gamma}_n(q^\star); j+2) Y_n, \tag{3.3.15}$$

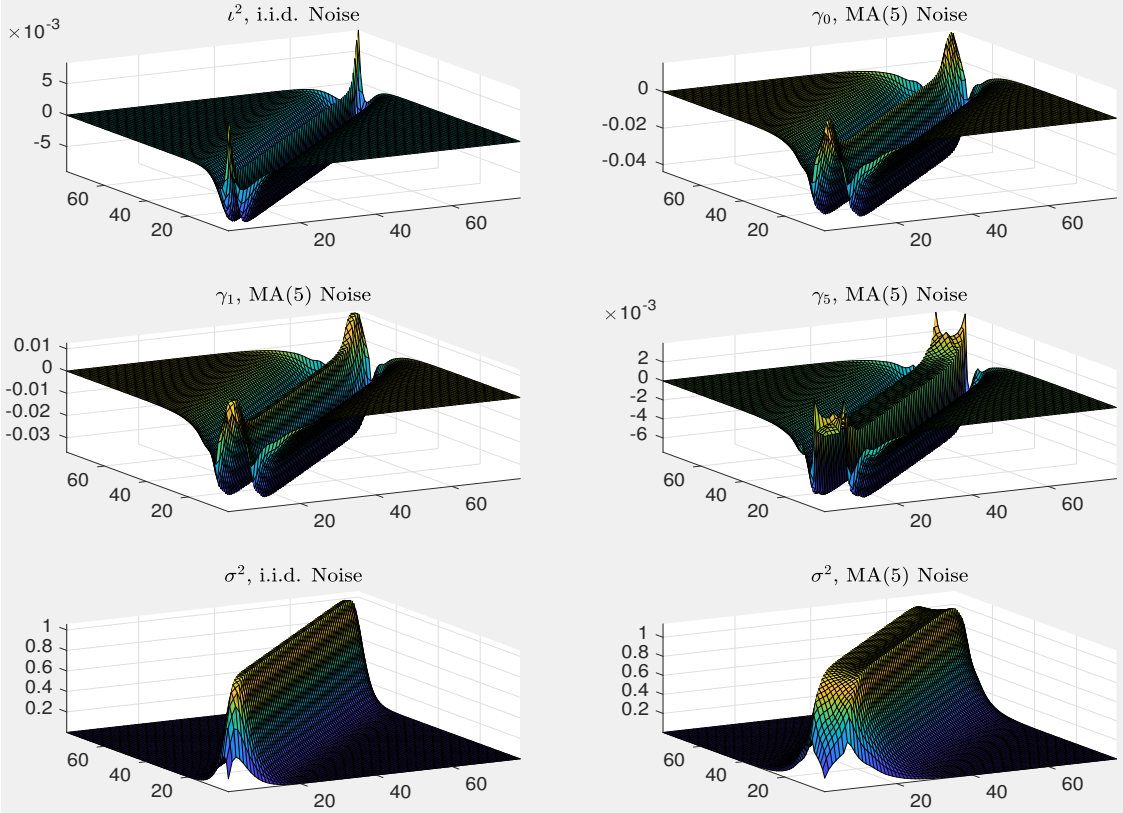
where the set of $n_T \times n_T$ *weighting matrices* $\mathcal{W}_n(\sigma^2, \gamma; l)$, $l = 1, 2, \ldots, q^\star + 2$, *is defined by*[6]

$$vec(\mathcal{W}_n(\sigma^2, \gamma; l)) = \Sigma_n^{-1}(\sigma^2, \gamma) \frac{\partial \Sigma_n(\sigma^2, \gamma)}{\partial (\sigma^2, \gamma)} \Sigma_n^{-1}(\sigma^2, \gamma) \widetilde{W}_n^{-1}(\sigma^2, \gamma) (0_{l-1}, 1, 0_{q^\star + 2 - l}),$$

6. 0_d is the d-dimensional vector of 0s. All vectors are column vectors. We write (a, b, c) in place of $(a^\intercal, b^\intercal, c^\intercal)$ for simplicity.

with $\Sigma_n(\sigma^2, \gamma)$ *given by (3.3.7), and the* $(q^\star + 2) \times (q^\star + 2)$ *matrix* $\widetilde{W}_n(\sigma^2, \gamma)$ *given by*

$$\widetilde{W}_n(\sigma^2, \gamma)_{i,j} = tr\left(\Sigma_n^{-1}(\sigma^2, \gamma) \frac{\partial \Sigma_n(\sigma^2, \gamma)}{\partial (\sigma^2, \gamma)_i} \Sigma_n^{-1}(\sigma^2, \gamma) \frac{\partial \Sigma_n(\sigma^2, \gamma)}{\partial (\sigma^2, \gamma)_j} \right).$$



Note: This figure compares weighting matrices $\mathcal{W}$s in the quadratic representations of the QMLE for σ^2 and ι^2 in the case of i.i.d. noise, as well as those matrices for σ^2, γ_0, γ_1, and γ_5 in the case of MA(5) noise. We scale the volatility weighting matrices by T. In both cases, we fix $\sigma^\star = 0.3$, $\iota^\star = 0.005$, $\Delta = 5$ minutes, $T = 1$ day. The moving-average parameters of the MA(5) process are given by $\theta^\star = (0.25, 0.2, 0.15, 0.1, 0.05)$.

Figure 3.1: Quadratic Representations of the Estimators

Theorem 5 shows that the QMLE can be written as an iterative quadratic estimator. It also suggests an alternative algorithm for estimation. With some initial values given, we can iteratively update parameters via equations given by (3.3.15) until convergence. Figure 3.1 plots these weighting matrices for both volatility and noise parameters, and compares

them in the case of i.i.d. and MA(5) noises. The noise weighting matrices feature a "W" shape along the diagonal, and the magnitude of weighting matrices for autocovariance decays as their order increases. With respect to the volatility estimator, the bottom panel shows notable "flatness" at the top of the volatility weighting matrix for the MA(5) model, which helps cancel out the impact of dependent noise. This patten motivates us to investigate the connection between the QMLE and the flat-top realized kernel introduced by Varneskov (2016) to the high-frequency environment in the context of volatility estimation. We also provide an equivalent kernel for autocovariances.

Theorem 6. *Suppose the same assumptions as those in Theorem 3 hold. In addition, suppose* $q \geq 0$ *is fixed and* $(\sigma^2, \gamma) \in \Pi_n(q)$ *such that* $K^{-1} \leq \inf_\lambda f(\lambda;\gamma) \leq \sup_\lambda f(\lambda;\gamma) \leq K$. *Then for all* $n^{1/2+\alpha} \leq i, j \leq n - n^{1/2+\alpha}$ *with* $0 < a < \frac{1}{2}$, *the weighting matrix* $\mathcal{W}_n(\sigma^2, \gamma; l)$ *satisfies for* $l \geq 1$,

$$
\begin{aligned}
&(i) \quad \mathcal{W}_n(\sigma^2,\gamma;1)_{i,j} = \frac{k}{T}\big(\frac{|i-j|}{H_n}\big)(1+o(1)), \quad \mathcal{W}_n(\sigma^2,\gamma;l)_{i,j} = \lambda_l \widetilde{k}\big(\frac{|i-j|}{H_n}\big) + O(1);\\
&(ii) \quad \sup_{|i-j|\leq q+1} \Big|\mathcal{W}_n(\sigma^2,\gamma;1)_{i,j} - \mathcal{W}_n(\sigma^2,\gamma;1)_{i,i}\Big| = O(\Delta_n^{3/2});\\
&(iii) \quad \sup_{|i-j|\leq q+1} \Big|\mathcal{W}_n(\sigma^2,\gamma;l)_{i,j} + \mathbb{1}_{\{l\leq |i-j|+1\}}\frac{|i-j|+2-l}{2n_T}\Big| = O(\Delta_n^{3/2}),
\end{aligned}
$$

where the implied equivalent kernels are $k(x) = (1+x)e^{-x}$ *and* $\widetilde{k}(x) = xe^{-x}$, *the implied bandwidth is* $H_n = \zeta\sigma^{-1}\Delta_n^{-1/2} + O(1)$ *with* $\zeta^2 = \sum_{|j|\leq q}\gamma_{|j|}$, *and* $\lambda_l = (2\sigma\zeta^3\Delta_n^{1/2}n_T)^{-1}\sum_{r=1}^{q+1}(2-\delta_{r,1})W(\gamma)^{-1}_{r,l-1}$, *with* $W(\gamma)$ *defined in Theorem 3.*

Theorem 6 suggests that the bulk of the QMLE weighting matrices can be approximately written as that of a nonparametric kernel estimator with an implicit bandwidth. Despite this equivalence, it is more convenient to implement the QMLE using Algorithm 2 in Section 3.3.2, which does not require tuning parameters barring order selection, or any special adjustment to the border effect. Also note that this equivalence result is only established under the

assumption that the spectral density of the noise (and hence its magnitude) is bounded from below, which rules out the case of small noise.

3.4 Monte Carlo Simulations

We examine the finite-sample performance of the estimators in a variety of simulation settings. Throughout we fix $T = 1$ day and the average sampling frequency every 5 seconds. We have $1,000$ Monte Carlo trials in total.

3.4.1 Verification of the Asymptotic Results

We simulate X_t and σ_t^2 according to the same log-volatility model as in Li and Xiu (2016):

$$\begin{cases} dX_t &= (0.05 + 0.5\sigma_t^2)dt + \sigma_t dW_t + J^X dN_t, \\ \sigma_t^2 &= D_t \exp\left(-2.8 + 6F_t\right), \quad dF_t = -4F_t dt + 0.8d\widetilde{W}_t + J^F dN_t - 0.02\lambda_N dt, \end{cases} \tag{3.4.16}$$

where $\mathbb{E}[dW_t d\widetilde{W}_t] = -0.8dt$, $J^X \sim \mathcal{N}(0, 0.02^2)$, $J^F \sim \mathcal{N}(0.02, 0.02^2)$, N_t is a Poisson process with intensity $\lambda_N = 25$, and D_t captures the diurnal effect:

$$D_t = 0.75\exp(-10t/T) + 0.25\exp(-10(1 - t/T)) + 0.8.$$

The arrival of trades follows an inhomogeneous Poisson process with rate $nT^{-1}\xi_t^{-1} = nT^{-1}(1 + \cos(2\pi t/T)/2)$, so that fewer trades arrive in the middle of the day.

With respect to the noise, we start with an MA(5) model of U with $\theta^\star = (0.25, 0.2, 0.15, 0.1, 0.05)$, innovation ε_i being Student's t-distribution with 7 degrees of freedom, $\iota = 2.5 \times 10^{-3}$, and η_t following

$$d\eta_t = 10 \times \left(\left(1 + 10^{-1}\cos(2\pi t/T)\right) - \eta_t\right)dt + 0.1dW_t,$$

where W_t is the same Brownian motion that drives X. We also round the observed prices

to the nearest cent: $\tilde{X}_t = \log([100 \times \exp(X_t)]) - \log 100$, where $[\cdot]$ means rounding to the nearest integer.[7]

We first assume that the correct order, namely 5, is known, so that we can verify the CLTs for noise autocovariances given in Section 3.3.4 without worrying about model selection mistakes. Figure 3.2 provides the histograms of the standardized estimates for $\widehat{\gamma}_k(q)$, $k = 0, 2, \ldots, 5$, using estimated asymptotic variances. All histograms match the standard normal density.

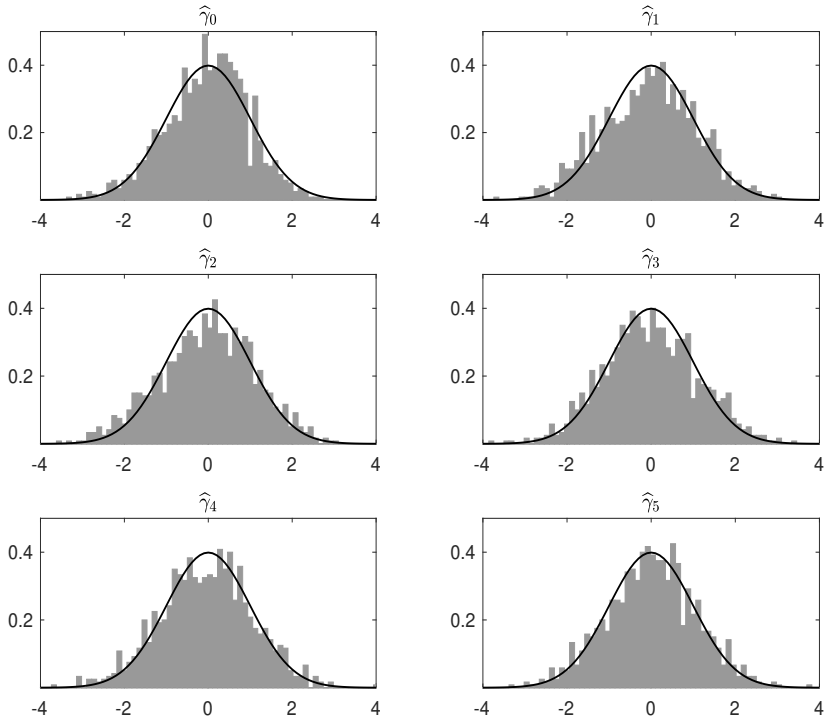


Figure 3.2: Histograms of the Standardized Parameter Estimates

Note: This figure plots the histograms of the standardized estimates for $\widehat{\gamma}_k(q)$, $k = 0, 1, \ldots, 5$, along with the density of the standard normal distribution. The noise is simulated from an MA(5) model with $\theta^\star = (0.25, 0.2, 0.15, 0.1, 0.05)$ and $\iota^\star = 2.5 \times 10^{-3}$. The order of the MA model is known prior to estimation.

7. Our theory does not allow for this type of rounding errors. We simulate this model to demonstrate that the rounding effect appears negligible.

3.4.2 Comparison with Alternative Estimators

We then compare our estimators of noise autocorrelations against alternative nonparametric estimators by Jacod, Li, and Zheng (2017) (JLZ) and Li and Linton (2021) (ReMeDI) in a more challenging MA(∞) setting in which $\theta(B) = (1 - 0.4B)^{-1}(1 + 0.2B)$. To demonstrate the effect of small noise, we consider three different scenarios for the magnitude of the noise, ι, which takes values from 10^{-4} (small noise) to 5×10^{-4} (median noise) and 2.5×10^{-3} (large noise). Our estimator uses either AIC or BIC for model selection, whereas nonparametric estimators involve a tuning parameter.

Jacod, Li, and Zheng (2017) propose to estimate autocovariances, γ, by approximating efficient prices using their local averages:

$$\widehat{\gamma}_j^{\text{JLZ}} = \frac{1}{n_T} \sum_{i=0}^{n_T+1-j-4h_n} \left(\widetilde{X}_{t_i} - \frac{1}{h_n} \sum_{l=0}^{h_n-1} \widetilde{X}_{t_{i+j+l+h_n}} \right) \left(\widetilde{X}_{t_{i+j}} - \frac{1}{h_n} \sum_{l=0}^{h_n-1} \widetilde{X}_{t_{i+j+l+3h_n}} \right).$$

Here h_n is a sequence of integers satisfying $h_n \sim n^{-\eta}$ with $\frac{1}{2v+1} < \eta < \frac{1}{2}$, where v is the ρ-mixing exponent of ε. It determines the local window size used to estimate realization of the noise. Their paper selects $h_n = 6$ in simulations with 1-second data. According to their criterion, when data are sampled at 5-second frequency, h_n must be an even smaller integer in a finite sample, so we report the autocorrelation estimates for $h_n = 2, 4$, and 6.

Li and Linton (2021) suggest an alternative construction that takes the differences of log prices over longer horizons to dampen the impact of efficient prices:

$$\widehat{\gamma}_j^{\text{ReMeDI}} = -\frac{1}{n_T} \sum_{i=1}^{n_T-2k_n-j} (\widetilde{X}_{i+k_n} - \widetilde{X}_i)(\widetilde{X}_{i+j+2k_n} - \widetilde{X}_{i+j+k_n}),$$

where k_n is a tuning parameter that satisfies: $k_n \to \infty$, $k_n n^{-\eta} \to 0$, for $\frac{1}{2v} < \eta < \frac{1}{3}$. We select $k_n = k_n' \log n$, where $k_n' = 0.5, 1$, and 2 in simulations.

With autocovariances given, the autocorrelations can thereby be estimated accordingly:

$\widehat{\rho}_j^{\text{JLZ}} = \widehat{\gamma}_j^{\text{JLZ}}/\widehat{\gamma}_0^{\text{JLZ}}$ and $\widehat{\rho}_j^{\text{ReMeDI}} = \widehat{\gamma}_j^{\text{ReMeDI}}/\widehat{\gamma}_0^{\text{ReMeDI}}$. We prefer autocorrelations (to autocovariances) because their scale is interpretable. However, we find it necessary to winsorize the estimated autocorrelations for AIC-based QMLE and both nonparametric estimators, when the noise magnitude is small, to ensure that their estimates are within the natural bound $[-1, 1]$.[8]

Table 3.1 provides comparison results for autocorrelations among QMLE, JLZ, and ReMeDI estimators across various noise magnitudes. Several points are worth making. For large noise, all estimators work reasonably well, but QMLEs generally outperform nonparametric estimators in terms of RMSE because they are more efficient. AIC slightly outperforms BIC, and ReMeDI appears to outperform JLZ. The latter suffers from a large finite sample bias. In the small noise regime, nonetheless, the biases and RMSEs for both nonparametric estimators deteriorate substantially. For estimation of noise autocovariances, "signal' is the microstructure friction, whereas "noise" is the efficient price. When the signal-to-noise ratio is too low, the error due to estimation is too large to justify doing so. In contrast, the QMLEs either conclude that noise is absent (i.e., θ and ι^2 are not available), in which case all autocorrelations are zeros, or select an MA model with a certain $\widehat{q}_n$, so that any autocorrelation beyond the $\widehat{q}_n$-th order is zero. Because of the rapid decay in autocorrelations and small noise magnitude, 0 is often a better estimate in terms of RMSE than nonparametric estimates, and in particular for larger lags. Comparing AIC with BIC, the latter is more conservative, as it essentially yields 0 autocorrelation estimates for almost all Monte Carlo replications, whereas the former produces many nontrivial estimates. However, doing so seems to increase AIC's RMSE, and AIC does require winsorization for about 5.3% of sample paths, compared with 20.9% for ReMeDi and 4.0% for JLZ. BIC needs no adjustment.

8. If a correlation estimate exceeds 1 (resp. -1), we reset it to be 1 (resp. -1).

Table 3.1: Simulation Results for Noise Autocorrelation Estimation

		QMLE	QMLE	JLZ	JLZ	JLZ	ReMeDI	ReMeDI	ReMeDI
		BIC	AIC	$h_n = 2$	$h_n = 4$	$h_n = 6$	$k_n' = 0.5$	$k_n' = 1$	$k_n' = 2$
		Panel A: Small Noise							
ρ_1	BIAS	-0.309	-0.195	0.586	0.639	0.657	-0.082	0.224	0.387
	RMSE	0.310	0.432	0.587	0.639	0.657	0.577	0.603	0.617
ρ_3	BIAS	-0.163	-0.101	0.716	0.775	0.795	0.008	0.315	0.477
	RMSE	0.163	0.327	0.718	0.775	0.795	0.600	0.691	0.723
ρ_5	BIAS	-0.044	-0.027	0.821	0.885	0.906	0.048	0.369	0.536
	RMSE	0.044	0.242	0.824	0.886	0.906	0.633	0.757	0.801
		Panel B: Median Noise							
ρ_1	BIAS	-0.093	-0.017	0.153	0.300	0.379	-0.006	-0.016	-0.035
	RMSE	0.150	0.075	0.185	0.312	0.386	0.098	0.150	0.251
ρ_3	BIAS	-0.063	-0.010	0.188	0.364	0.459	0.000	-0.012	-0.034
	RMSE	0.094	0.054	0.227	0.379	0.468	0.111	0.170	0.280
ρ_5	BIAS	-0.029	-0.004	0.217	0.416	0.524	0.003	-0.002	-0.033
	RMSE	0.039	0.036	0.262	0.434	0.535	0.115	0.186	0.300
		Panel C: Large Noise							
ρ_1	BIAS	-0.009	-0.001	-0.055	0.010	0.045	0.000	-0.003	-0.001
	RMSE	0.040	0.020	0.073	0.063	0.083	0.036	0.039	0.043
ρ_3	BIAS	-0.008	0.000	-0.066	0.012	0.055	0.001	-0.002	-0.001
	RMSE	0.029	0.019	0.088	0.075	0.100	0.042	0.043	0.048
ρ_5	BIAS	-0.011	0.000	-0.075	0.014	0.062	0.001	0.002	0.000
	RMSE	0.028	0.018	0.101	0.086	0.114	0.044	0.046	0.051

Note: This table compares estimators of 1st-, 3rd-, and 5th-order autocorrelations (ρ_1, ρ_3, ρ_5) in three scenarios of noise magnitude. "QMLE" is an MA($\widehat{q}_n$)-likelihood estimators using either BIC or AIC for order selection. "JLZ" refers to the nonparametric estimator of Jacod, Li, and Zheng (2017). "ReMeDI" refers to the nonparametric estimator of Li and Linton (2021). We report three choices of h_n and k_n' for comparison. The AIC-QMLE, JLZ, and ReMeDI estimates of autocorrelations are winsorized so that their magnitude stays within $[-1, 1]$. The true 1st, 3rd-, and 5th-order autocorrelations are 0.308, 0.163, and 0.04, respectively.

3.5 Empirical Analysis of U.S. Equity

To demonstrate the empirical relevance of the proposed approach, we conduct a large-scale study of noise autocovariances for S&P 1500 index constituents from January 1, 1996, to December 31, 2016. There are approximately 1,500 tickers every day, and about 3,500 tickers

in total due to changes in index constituents. To illustrate, we summarize cross-sectional findings here though all estimates are available upon request. We use BIC-QMLE for noise-related parameters because of the model selection consistency result discussed earlier. We also report volatility estimation results, but with AIC*-QMLE, as suggested by Da and Xiu (2021b).

We download the trades and quotes of all equities at their highest frequency available (up to a millisecond after January 1, 2007, and a microsecond from July 27, 2015) from the TAQ database.[9] Next, we remove trades and quotes with special condition codes or suffix codes, as well as those that occur outside regular trading hours.[10] We then construct national best bid and offer (NBBO) data using quotes from all exchanges at a 1-second frequency.[11] We then match trades with NBBOs by their recorded time points and remove those trades that are outside the range of the corresponding NBBOs.[12] Our approach is less aggressive than that of Barndorff-Nielsen, Hansen, Lunde, and Shephard (2009), in that we maintain trades and quotes from all exchanges, whereas they retain only entries originating from a single exchange. Next, we remove redundant trades, retaining only nonzero returns.[13] This step helps alleviate model misspecification due, for example, to the effect of rounding, latency or delay across exchanges, and so on. Finally, we remove any stock days that have fewer than

9. Because companies change their tickers from time to time for mergers, acquisitions, or other reasons, the same ticker in the TAQ database may correspond to different stocks. We therefore keep track of these changes and use CRSP PERMINOs to index all stocks that do not change over time.

10. We remove trades and quotes with condition codes Z, B, U, T, L, G, W, K, J and corresponding odd-lot trades, which have an additional letter I, as well as those with non-empty suffix codes (preferred shares). We identify opening trades as those with condition codes O, Q, OI, or QI; closing trades with 6, M, 6I, or MI; and remove all trades beyond the window of opening and closing time points. We only keep trades with correction indicator 00 or 01.

11. We construct NBBOs from the millisecond dataset by adapting the SAS codes from https://wrds-web.wharton.upenn.edu/wrds/research/applications/microstructure/NBBO%20derivation/. Although this database has more precise timestamps, we do not construct NBBOs at any frequency higher than every second.

12. For trades that are observed at millisecond or microsecond intervals, we match them with the NBBOs of the previous second. Our SAS codes for cleaning the data are available upon request.

13. This step is called "tick-time sampling" by Griffin and Oomen (2008).

12 observations after cleaning.

We start by examining the time-series behavior of volatility and microstructure noise. The upper panel of Figure 3.3 presents the time series of volatility estimates for constituents of each of the three indices, respectively. The lower panel provides the time series of noise-variance estimates among those constituents whose estimates are available. We use lines to represent the median and shaded areas to represent the lower and upper quartiles in the cross-section. We also smooth these time series using equal weights over a monthly moving window. Although considerable cross-sectional variation is present, the median volatility estimates among constituents of all three indices share a pattern similar to what we usually find from the volatility of the S&P 500 index. That said, the small caps are on average more volatile than the large caps, with the mid caps in between. As to the noise, there is a clear declining pattern in its order of magnitude over time across the entire universe, which is likely because of the improvement in market efficiency. Not surprisingly, the small caps have the largest noise, followed by the mid cap and then the large cap.

Next, we focus on the dependence structure of the noise. As the left panels of Figure 3.4 show, around 30%-60% of stocks have noise that is too small to be estimated. This percentage is higher for large caps than for small caps. For a large percentage of stock-day pairs, the selected orders based on the BIC are 0, so that i.i.d. noise assumption is reasonable for them. That said, about 10%-30% of stock-day pairs remain for which BIC prefers a few more lags. For BIC to select more than 6 lags is rare. We also find more stock-days in 2016 with selected orders greater than or equal to 1, compared with earlier years, particularly for large caps. This finding is due to the availability of data sampled at a frequency even higher than every second, for which we expect to see more autocorrelated lags.

To shed further light on this point, we provide in the right panels of Figure 3.4 histograms of the durations of autocorrelations for those tickers with selected lags greater than or equal to 1. Duration is defined in terms of seconds as the product of the selected order and the

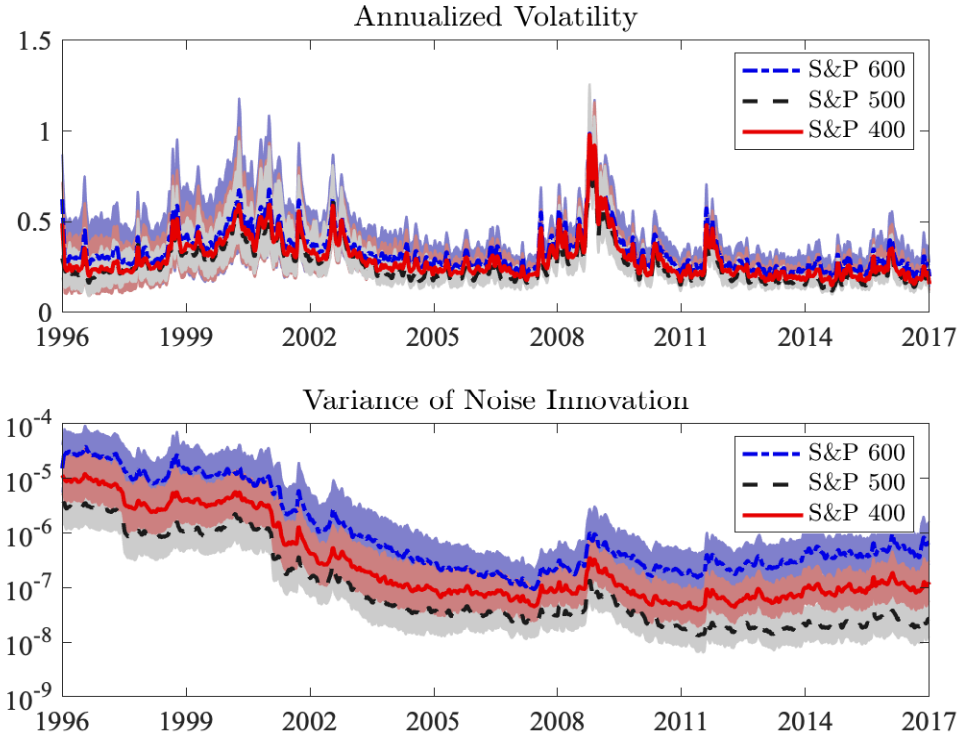


Figure 3.3: Time Series of Volatility and Noise-innovation Variance

Note: The upper panel compares the cross-sectional median (lines), lower, and upper quartiles (shaded areas) of the annualized volatility estimates for S&P Composite 1500 Index constituents (using Algorithm 2.3), and the lower panel presents the variance estimates of noise innovation (using Algorithm 2.4) for those constituents that have large-enough noise. The time series are smoothed with equal weights over a moving window of 21 days. The y-axis of the lower panel is transformed to the logarithm scale for the sake of presentation.

average trading frequency for each stock-day pair. We find that estimated durations are much shorter for large-cap stocks than for smaller caps. Moreover, the average duration of autocorrelations has been decreasing in the past two decades. For instance, the average duration of large caps has decreased from $10^2 \sim 10^3$ to merely 10 seconds.

Finally, we discuss the importance of modeling the microstructure noise through the lens of volatility inference. While there exist informal volatility signature plot or more formal tests of microstructure noise (Aït-Sahalia and Xiu (2019)), such pre-testing-based approaches do not deliver correct volatility inference due to uniformity concerns when noise exists but

is too small to be detected. We compare the biases and RMSEs of the popular realized volatility estimator and the QMLE, to indirectly shed light on the influence of noise. The former estimator, based on data sampled at a prespecified frequency—say, every 5 or 15 minutes—is most commonly adopted in practice.

The left panels of Figure 3.5 compare the cross-sectional medians of realized volatility estimates based on 5-minute and 15-minute subsamples, respectively, with the corresponding medians of the QMLEs. Remarkably, on average, a large upward bias associated with the former estimates is present, potentially due to the presence of noise at the 5-minute frequency. The biases are substantial–over 160% for small caps–compared with noise-robust QMLEs in earlier years. The biases have been decreasing over the past two decades, with a slight increase post-2008. Biases of the small caps are more evident than those of the large caps. On average, the large caps are traded more frequently than every 5 minutes, so their biases in the cross-sectional medians are almost indistinguishable from zero post-2002. This finding does not imply that every 5 minutes is a safe frequency for each individual constituent of the S&P 500 index. At a 15-minute frequency, the biases are clearly smaller–though they have not completely vanished, even in 2016–for these median estimates. The right panels of Figure 3.5 compare the ratios of standard errors between the 5-minute (resp. 15-minute) realized volatility estimator and the QMLE using the entire sample. The larger the ratio, the greater the efficiency loss for the realized volatility. We only report results for 2016, because the quality of the realized volatility estimator is best. We find that when the sampling frequency reaches every 15 minutes, most of the ratios are greater than 1, with some being as large as 10–in particular, for S&P 500 constituents–which suggests substantial efficiency losses.

To sum up, without accounting for noise, the realized volatility estimator faces a bias and variance dilemma. Estimates using 5-minute data are subject to severe biases, whereas 15-minute estimates suffer from considerable efficiency losses. Additionally, the standard errors could still be understated because the noise might not be sufficiently small to the

extent that it can be safely ignored.

3.6 Conclusion

We propose a semiparametric approach to disentangling autocovariances and autocorrelations due to the microstructure frictions associated with observed prices. Our approach resembles a threshold estimator, which gives zero autocovariance estimates beyond the lag selected by the information criteria. This feature delivers superior performance in the finite sample, particularly when noise is relatively small, compared with alternative nonparametric estimators. Our empirical study of S&P 1500 stocks finds that the microstructure noise has shrunk by several orders of magnitude and that its autocovariances have faded more rapidly in recent years than earlier. These findings indicate that market efficiency has improved substantially, potentially due to the popularity of electronic and algorithmic trading. In a cross-sectional comparison, the autocovariances of small-cap stocks tend to persist for a longer period than the large caps, perhaps due to limits to arbitrage or for liquidity reasons.

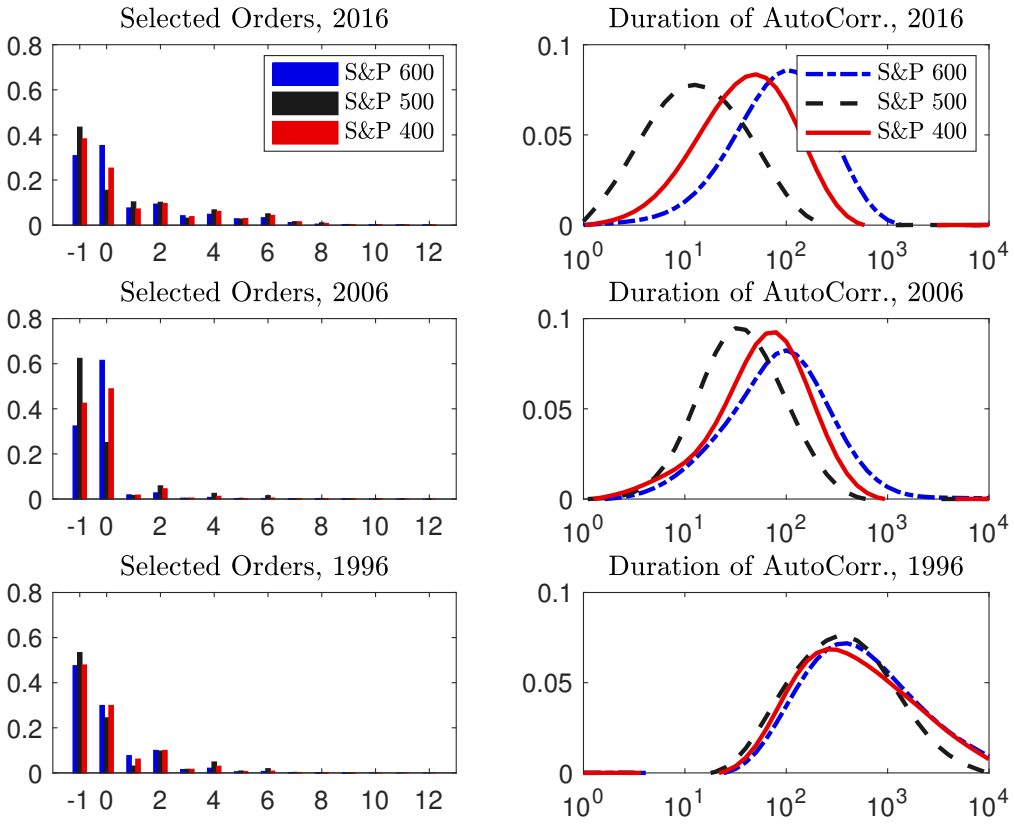


Figure 3.4: Selected Orders and Durations of Autocorrelations

Note: Left panels provide the frequencies of selected orders using BIC for each stock-day pair in 1996, 2006, and 2016, respectively. "-1" represents the case of small noise, i.e., the stock-day pair for which no reliable estimate of noise variance exists. "0" represents the case of i.i.d. noise, whereas other values are the selected orders of MA processes. Panels on the right provide the corresponding (fitted) histograms of the durations of autocorrelations in the case of dependent noise. Duration in terms of seconds is defined as the product of the selected order and the average trading frequency for each stock-day pair. The x-axis is transformed to a logarithmic scale for the sake of presentation.

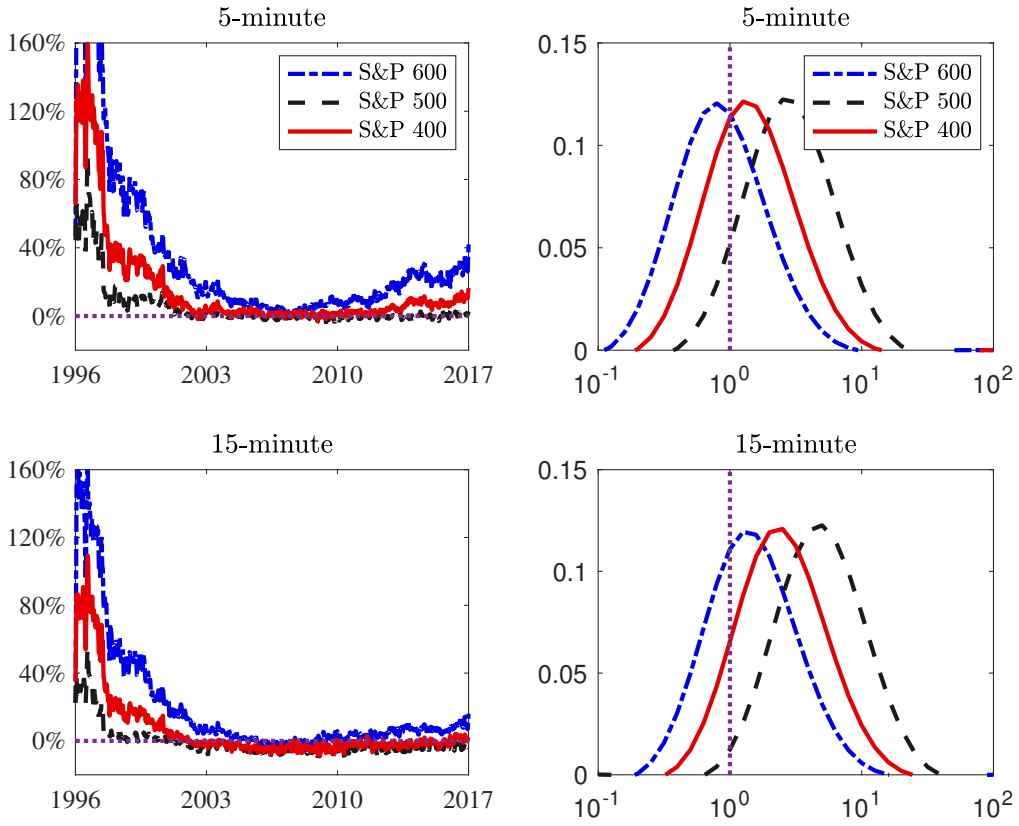


Figure 3.5: Relative Biases and Standard Errors of the Realized Volatility against QMLE

Note: The right panels plot percentage biases in the cross-sectional medians of 5-minute and 15-minute realized volatility estimates, respectively, relative to their corresponding QMLEs using the entire sample. Time series are smoothed with equal weights over a moving window of 21 days. The right panels provide the histograms of the ratios of standard errors between the 5-minute (resp. 15-minute) realized volatility estimator and the QMLE, for each stock-day pair in 2016. The x-axes on the right panels are transformed to the a logarithmic scale for the sake of presentation.

Chapter 4

The Statistical Limit of Arbitrage[1]

4.1 Introduction

It is a fundamental underlying principle of most asset pricing theories, including the Arbitrage Pricing Theory (APT), that investment opportunities with extremely high ratios of reward to risk do not exist in financial markets. Implicitly, these theories rest on the premise that such near-arbitrage opportunities would attract arbitrageurs who exploit and thereby eliminate these opportunities. An important assumption in these theories is that parameters in the data-generating process (DGP) of returns are known to arbitrageurs. Therefore, near-arbitrage opportunities in the DGP of returns are ruled out.

In practice, however, sophisticated investors searching for near-arbitrage opportunities do not know the true parameters. Instead, they commonly conduct statistical analyses to learn about the existence of such opportunities from historical returns data. As a consequence, they face statistical uncertainty. In some settings, such as in some derivatives pricing applications, for instance, the statistical uncertainty may be sufficiently small that it is not a significant impediment to arbitrageur activity. But in noisy, high-dimensional settings such as the cross-section of stock returns, statistical uncertainty can be substantial and it can constitute a statistical limit to arbitrage.

1. This Chapter is a joint work with Stefan Nagel and Dacheng Xiu.

To analyze the effects of arbitrageur learning, we consider a setting in which returns follow a statistical linear factor model. Near-arbitrage opportunities are characterized by high Sharpe ratios achieved by factor-neural trading strategies. To exploit such opportunities, arbitrageurs need knowledge of factor model alphas, but they must learn about these from historical realizations of returns. We derive the optimal Sharpe ratio achievable by any feasible arbitrage trading strategies, which is strictly dominated by the infeasible optimal Sharpe ratio that arbitrageurs could achieve if they were endowed with perfect knowledge of alphas. This, in turn, provides a new no-near-feasible-arbitrage bound on the Sharpe ratio that accounts for the statistical limit to arbitrage.

The difficulty of the learning problem hinges on the DGP of alpha signals. While our theory generally does not rely on specific cross-sectional distributions of alpha signals, we use simple special cases to demonstrate how the optimal Sharpe ratio varies with the strength and sparsity of alphas. When alphas are strong and not too rare relative to the dimensionality of the cross-section and the sample size, arbitrageurs can learn the distribution of alpha perfectly in the limit. But when alpha is weaker and more rare, its inference becomes more challenging and a gap arises between the optimal feasible Sharpe ratio and the infeasible Sharpe ratio that requires perfect knowledge of alphas. For instance, the infeasible Sharpe ratio may explode asymptotically, while the feasible Sharpe ratio stays bounded.

The existence of this statistical limit to arbitrage implies a widening of the bounds in which mispricing can survive in equilibrium compared with a situation in which arbitrageurs know the DGP and its parameters. Some mispricing may survive because it is clouded by too much statistical uncertainty. Empirically therefore, the feasible, not the infeasible, Sharpe ratio tells us about the minimum reward-to-risk compensation that arbitrageurs require.

We further demonstrate how arbitrageurs can construct a feasible trading strategy that achieves the theoretically optimal feasible Sharpe ratio, uniformly over DGPs of alphas, regardless of the strength and sparsity of alphas. This means that the feasible Sharpe ratio

bound is in fact sharp. A uniformly valid trading strategy is desirable because in reality arbitrageurs do not know which DGP is a correct description of the observed data. The optimal strategy estimates the empirical distribution of alpha signals and assigns weights based on the relative magnitudes and associated uncertainty of the alpha estimates. Assets with high alpha t-statistics get portfolio weights proportional to their signal strength. Weaker alphas are more difficult to exploit, yet simply ignoring them would lead to a suboptimal trading strategy. The optimal strategy constructs portfolio weights for weak signals by locally smoothing alpha signals cross-sectionally.

To empirically contrast feasible and infeasible Sharpe ratios, we also propose an estimator of the infeasible Sharpe ratio that a hypothetical arbitrageur endowed with perfect knowledge of DGP parameters would perceive. While this Sharpe ratio can be estimated consistently, it cannot be realized by any feasible portfolio with weights constructed using historical data. The infeasible Sharpe ratio often serves as the building block for tests of APT, see, e.g., Gibbons, Ross, and Shanken (1989), Gagliardini, Ossola, and Scaillet (2016), Fan, Liao, and Yao (2015), and Pesaran and Yamagata (2017). While such tests are powerful and may lead to discoveries of alpha signals, they are not relevant for arbitrageurs who are confined to feasible trading strategies. Our effort in constructing the optimal feasible arbitrage portfolio and evaluating its economic performance directly responds to Shanken's call (Shanken (1992)): "... practical content is given to the notion of 'approximate arbitrage,' by characterizing the investment opportunities that are available as a consequence of the observed expected return deviation ... Far more will be learned, I believe, by examining the extent to which we can approximate an arbitrage with existing assets."

While the optimal strategy outlined in our paper may not be adopted by investors, exploring alternative strategies that are commonly employed by practitioners is both relevant and insightful. Specifically, we investigate the potential of alternative strategies that leverage techniques such as multiple testing, shrinkage, and selection to construct arbitrage portfolios,

assessing their ability to achieve the optimal feasible Sharpe ratio. With alphas estimated from cross-sectional regressions, one strategy adopts a multiple-testing (BH) procedure as in Benjamini and Hochberg (1995) on the individual p-values of t-statistics for alpha, in order to guard against potential false discoveries among significant alphas, before building the optimal portfolio weights using selected alphas. Other strategies use either LASSO or Ridge penalties to regularize portfolio weights based on alpha estimates. Such strategies amount to imposing a prior distribution on the alphas. We illustrate with a simple example that these strategies can achieve optimal Sharpe ratio under distinct alpha assumptions. In particular, BH procedure achieves optimal performance only when few true alpha signals are substantially strong. Its failure to achieve optimality is precisely due to its conservativeness nature against the less potent alphas. In contrast, the ridge-based portfolio is equivalent to that constructed by alpha estimates from plain cross-sectional regressions. This approach can achieve optimality when almost all true alphas are either uniformly strong or uniformly weak. The LASSO approach attempts to strike a balance between the aforementioned two methods, with a small gap to achieving the theoretically optimal Sharpe ratio, provided an optimal tuning parameter.

Finally, we demonstrate the empirical implications of the statistical limits of arbitrage by examining 56 years of monthly individual equity returns in US stock market from 1965 to 2020. The average number of stocks over this period exceeds 4000. We construct residuals via cross-sectional regressions from a multi-factor model that directly uses observed characteristics as risk exposures. These characteristics include market beta (Fama and MacBeth (1973)), size (Banz (1981)), operating profits/book equity (Fama and French (2006)), book equity/market equity (Fama and French (2006)), asset growth (Cooper, Gulen, and Schill (2008)), momentum (Jegadeesh and Titman (1993)), short-term reversal (Jegadeesh (1990)), industry momentum (Moskowitz and Grinblatt (1999)), illiquidity (Amihud (2002)), leverage (Bhandari (1988)), return seasonality (Heston and Sadka (2008)), sales growth (Lakonishok,

Shleifer, and Vishny (1994)), accruals (Sloan (1996)), dividend yield (Litzenberger and Ramaswamy (1979)), tangibility (Hahn and Lee (2009)), and idiosyncratic risk (Ang, Hodrick, Xing, and Zhang (2006)), as well as 11 Global industry Classification Standard (GICS) sectors. These characteristics and industry dummies capture similar equity factors in the MSCI Barra model widely-used among practitioners.

Our paper builds on a large literature on the arbitrage pricing theory (APT) developed by Ross (1976) and later refined by Huberman (1982), Chamberlain and Rothschild (1983), and Ingersoll (1984). As in these papers, we rely on asymptotic arguments that do not rely on assumptions about investor preferences, but these results should be seen as an asymptotic approximation for a more realistic setting with a finite number of assets in which weak preference restrictions rule out Sharpe ratios far above the Sharpe ratios of diversified factor portfolios. The statistical limit to arbitrage that we highlight in this paper relax this Sharpe ratio bound compared with an economy in which arbitrageurs are endowed with perfect knowledge of DGP parameters. In this regard, our paper is also related to another large strand of literature on the limit of arbitrage, see Gromb and Vayanos (2010) for a comprehensive review. Complementary to the existing literature, the arbitrage limit in our setting stems from statistical uncertainty, instead of being induced from risk, costs, frictions, and other constraints rational expectation investors are facing.

Kozak, Nagel, and Santosh (2018a) argue that the absence of near-arbitrage opportunities enforces the expected returns to approximately line up linearly with common factor covariances, even in a world in which belief distortions affect asset prices. Our study focuses on the deviations of expected returns from this approximate linear relation and how statistical limits to arbitrage allow bigger deviations. A closely related paper to ours is Kim, Korajczyk, and Neuhierl (2020), which proposes a characteristics-based factor model to construct feasible arbitrage portfolios. Their asymptotic theory does not preclude arbitrage opportunities with a theoretically infinite Sharpe ratio, which implies a rather strong signal-to-noise ratio in their

alpha signals. Relatedly, Uppal and Zaffaroni (2018) propose a methodology to construct robust portfolios that can be decomposed into alpha (arbitrage) portfolios and beta (factor) portfolios. Our setting is considerably different from both papers in that the premise of our framework rules out infinite feasible Sharpe ratios, which enforces weak and rare signals. In our setting, alphas cannot possibly be recovered with certainty even when the sample size is large. On the empirical side, Guijarro-Ordonez, Pelger, and Zanotti (2022) propose a deep learning approach to statistical arbitrage that achieves a sizable out-of-sample Sharpe ratio. The profits of their trading strategy stem from generalized return reversals at daily to weekly frequencies, potentially due to liquidity provision and other microstructure channels. Our empirical analysis is not targeted towards characterizing the reward-to-risk ratios for high frequency traders, nor for traders that turnover a large portion of their portfolios daily.

Our paper also contributes to the evolving literature on applications of statistical and machine learning in asset pricing, and in particular on the topic of testing the APT, e.g., Gibbons, Ross, and Shanken (1989), Gagliardini, Ossola, and Scaillet (2016), and Fan, Liao, and Yao (2015), as well as on testing for alphas, e.g., Barras, Scaillet, and Wermers (2010), Harvey and Liu (2020), and Giglio, Liao, and Xiu (2021). The first literature focus on testing a null that all alphas are equal to zero. This is certainly an interesting null hypothesis, but as we emphasize in this paper, the APT does allow for alphas as long as they do not induce an explosive feasible Sharpe ratio. The second literature focuses on detecting strong alphas, in which widely used multiple testing methods, such as the BH method by Benjamini and Hochberg (1995), or its extensions can be applied to control the false discovery rate (FDR). In contrast, we allow for rare and weak alpha signals such that any procedure aiming to control the FDR is too conservative with too few or no discoveries.[2] Our objective here is not on model testing or signal detection. Rather, we strive for the optimal economic performance of arbitrage portfolios. We show that even if signals were so weak that they are

2. Donoho and Jin (2004) adopt the so-called higher criticism approach, dating back to Tukey (1976), to detect rare and weak signals in a stylized multiple testing problem.

undetectable by multiple testing methods, they may lead to a portfolio with a considerable Sharpe ratio.

There has been a long-standing critique of rational expectation models in macroeconomics and finance in which economic agents are not confronted with statistical uncertainty over structure parameters, see Hansen (2007). Bayesian learning is one way to expose model agents to statistical uncertainty. Pastor and Veronesi (2009) survey the literature on learning in financial markets. In many settings, e.g., Collin-Dufresne, Johannes, and Lochstoer (2016), learning can be sufficiently slow such that its effects persist in empirically realistic sample sizes, even though convergence to rational expectations takes place in the long-run. An exception is Martin and Nagel (2021) where learning effects persist because investors face a high-dimensional inference problem about the process generating firm cash flows. Similarly, arbitrageurs in our model attempt to make inference on a high-dimensional parameter vector with a potentially insufficient sample size, but they learn about returns, not firms' underlying cash flows. We examine different sequences of DGPs and in most scenarios, our learning system does not converge to a rational expectations limit.[3]

Our paper is also related to Chen, Hansen, and Hansen (2021b) and Chen, Hansen, and Hansen (2021a) in that they also account for the distinction between beliefs of economic agents and the DGP revealed by empirical evidence. They model belief distortions as a change of measure in moment conditions, use statistical measures of divergence relative to rational expectation to bound the set of subjective probabilities, and seek robust inference with this form of misspecification. In the spirit of Hansen (2014), we develop an optimal feasible Sharpe ratio for arbitrageurs inside the economic model, which is in contrast with the (infeasible) one from an outside econometrician's point of view. In our setting, the deviation from rational expectations stems naturally from the statistical obstacles economic agents

3. Our analysis is related to a large literature in econometrics and statistics that discuss uniform validity of asymptotic approximations, see, e.g., Staiger and Stock (1997), Imbens and Manski (2004), Leeb and Pötscher (2005), Andrews, Cheng, and Guggenberger (2020).

are facing. A subtle and important point we strive to make here is that economic agents embracing machine learning methods in a high dimensional environment could achieve a distinct outcome as opposed to what rational expectation agents could asymptotically.

From a methodological perspective, the optimal portfolio weights are proportional to the posterior mean of alpha, which resembles the classical normal mean problem in empirical Bayes, dating back to Robbins (1956), where the unknown parameters, alpha, are regarded as random draws from some common distribution, and only a noisy version of alpha (in the form of ex-factor returns) is observed. Our nonparametric approach thereby shares the same spirit of nonparametric empirical Bayes, see, e.g., Johns (1957), Zhang (1997), and Brown and Greenshtein (2009). Yet unlike the classical empirical Bayes inference, our analysis allows for weak and rare alphas as motivated from economic restrictions, and digs further into the Sharpe ratios above and beyond the posterior mean of alphas.

Our paper proceeds as follows. Section 4.2 develops our main result on statistical limit to arbitrage. Specifically, Section 4.2.1 sets up the model, Section 4.2.2 motivates and then defines the feasibility constraint facing arbitrageurs, Sections 4.2.3 - 4.2.4 specify arbitrageurs' decision problem, derive the optimal strategy, illustrate the Bayes correction for alpha, and demonstrate the gap between feasible and infeasible Sharpe ratios, Section 4.2.5 constructs a feasible trading strategy that achieves the bound, Section 4.2.6 proposes an estimator of the infeasible Sharpe ratio, and finally Section 4.2.7 analyzes alternative strategies. Section 4.3 provides simulation evidence, followed by an empirical analysis in Section 4.4. Section 4.5 concludes. The appendix provides technical details.

4.2 Main Theoretical Results

We start by revisiting the arbitrage pricing theory framework developed by Ross (1976). This theory is primarily based on a reduced-form statistical model for asset returns, which, despite its stylized nature, offers significant theoretical insights and remains relevant for guiding empirical investment decisions.

4.2.1 Factor Model Setup

To be more concrete, the factor economy has N assets in the investment universe. The $N \times 1$ vector of excess returns r_t follows a reduced-form linear factor model:

$$r_t = \alpha + \beta\gamma + \beta v_t + u_t, \tag{4.2.1}$$

where β is an $N \times K$ matrix of factor exposures (with the first column being a vector of 1s), α is an $N \times 1$ vector of pricing errors, v_t is a $K \times 1$ vector of zero-mean factor innovations with covariance matrix Σ_v, γ is a $K \times 1$ vector of risk premia (first entry corresponding to the column of 1s is the zero beta rate), and u_t is a vector of zero-mean idiosyncratic returns, independent of v_t, with a diagonal covariance matrix Σ_u.[4]

We assume at any given time t, arbitrageurs examine a sample of size T, derived from Equation (4.2.1), spanning from $t - T + 1$ to t. Throughout we will consider asymptotic limits as N and T increase while K and t are fixed.[5] To facilitate our asymptotic analysis along the cross-sectional dimension, N, we regard high dimensional objects such as α, β, and Σ_u as random variables drawn from some cross-sectional distributions, whereas γ and Σ_v are regarded as deterministic parameters. This distinction is made because γ and Σ_v serve as nuisance parameters in our analysis and their dimensions remain fixed. We assume that α has mean zero, and is cross-sectionally independent of β, and that β has full column rank and is pervasive. These conditions are essential for identification of γ in a model that

4. While approximate factor models become more prevalent following Chamberlain and Rothschild (1983), allowing for off-diagonal entries in the covariance matrix Σ_u would introduce additional statistical obstacles due to the estimation of large covariance matrix for inference on alpha and for building optimal portfolios. For simplicity, we illustrate the economic insight of limits to arbitrage using a strict factor model, leaving discussions on violations of model assumptions later.

5. This framework presents a slight deviation from the conventional scenario in which arbitrageurs observe a sample spanning $t = 1, \ldots, T$ and make their investment decision at time $T + 1$. Essentially, our approach highlights the continuous process of making investment decisions, rather than conceiving these decisions as occurring at a single, distant future point, T. That said, this conceptual difference does not result in any tangible difference in our theoretical and empirical results.

allows for pricing errors. We formalize these conditions below.

To allow for a more general DGP, we may consider a conditional version of (4.2.1):

$$r_t = \alpha_{t-1} + \beta_{t-1}\gamma_{t-1} + \beta_{t-1}v_t + u_t, \tag{4.2.2}$$

where β_t is a vector of time-varying factor loadings and γ_t is a vector of time-varying risk premia.[6] Despite the conditional model's capacity to encapsulate a broader DGP, it does not provide additional economic insights, particularly concerning the theoretical limits of arbitrage, compared to the unconditional model. While our primary theoretical findings, such as those provided in Theorems 7 and 8, remain applicable when β is substituted with β_t, we concentrate on the more stylized model (4.2.1) to illustrate our theory for clarity.

There are at least three variations of the factor model, depending on what econometricians assume to be observable. The most common setup in academic finance literature imposes that factors are observable as in e.g., Fama and French (1993).[7] The second setting, which has gained more popularity recently since its debut in Connor and Korajczyk (1986), assumes that factors are latent. The third setting, arguably most prevalent among practitioners, is the MSCI Barra model originally proposed by Rosenberg (1974), where factor exposures (e.g., characteristics) are assumed observable. One notable advantage of this model is its avoidance of the cumbersome task of estimating a large number of potentially time-varying stock-level factor exposures. Such estimation processes can prove both statistically inefficient and computationally demanding. By explicitly specifying risk exposures as linear functions of readily observable characteristics, the MSCI Barra model simplifies the estimation procedure

6. This model is overly parametrized that parameters are not identifiable without additional restrictions. Some examples of parsimonious conditional factor models include Connor, Hagmann, and Linton (2012), Gagliardini, Ossola, and Scaillet (2016), and Kelly, Pruitt, and Su (2019a).

7. This is different from saying factor innovations, v_t, are observable. The setting of observable factors typically involves another equation that $f_t = \mu + v_t$, where μ are the population means of the observed factors f_t, which are not necessarily identical to the factor risk premia, γ. Since μ is an unknown parameter, v_t is still not observable.

and significantly reduces computational complexity.

The three approaches are intrinsically linked: if we assume that (4.2.2) holds with $\alpha_t = c_t\alpha$ and $\beta_t = c_t\beta$ for certain characteristics c_t, then projecting returns onto lagged characteristics results in:

$$(c_{t-1}^\intercal c_{t-1})^{-1} c_{t-1}^\intercal r_t = \alpha + \beta(\gamma_{t-1} + v_t) + (c_{t-1}^\intercal c_{t-1})^{-1} c_{t-1}^\intercal u_t. \tag{4.2.3}$$

This transformation converts a conditional model for individual stocks, r_t, into an unconditional latent factor model for managed portfolios, $(c_{t-1}^\intercal c_{t-1})^{-1} c_{t-1}^\intercal r_t$.

In our empirical work, we prefer to use the third model for its practicality in analyzing individual stock returns, adhering to (4.2.2) but with a constant alpha.[8] This approach is particularly pertinent for practitioners. We extend our examination to include portfolios as test assets, employing a static factor model, (4.2.1), with latent factors. In light of our discussion about (4.2.3), this approach effectively accommodates for individual stocks' alpha that may vary over time.

4.2.2 Feasible Near-Arbitrage Opportunities

Building upon the insight of Ross (1976), Huberman (1982) and Ingersoll (1984) established the concept of near-arbitrage, which can be formalized in a more general setting as below:

Definition 1: A portfolio strategy w at time t is said to generate a near-arbitrage under a sequence of data-generating processes, such as (4.2.1), defined in a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_s\}_{s\leq t}, \mathrm{P})$, if it satisfies $w \in \mathcal{F}_t$, and along some diverging subsequence,[9] with

8. Empirically, we use a moving window method for alpha estimation, that effectively addresses alphas that vary slowly over time. Given that accurately inferring alpha typically demands a large sample size, opting for a constant alpha model serves as a reasonable compromise. This assumption sidesteps the theoretical complexities associated with alpha's temporal variability without losing empirical relevance. Meanwhile, our analysis based on managed portfolios incorporates time-varying alphas effectively.

9. We adopt the same subsequence definition as that used in Ingersoll (1984). The subsequence typically depends on the count of investment opportunities, i.e., N, though we do not need make this explicit in this definition. For simplicity of notation and without ambiguity, we omit the dependence of w on N and t.

probability approaching one,

$$\mathrm{Var}(w^\intercal r_{t+1}|\mathcal{F}_t) \to 0, \quad \mathrm{E}(w^\intercal r_{t+1}|\mathcal{F}_t) \geq \delta > 0.$$

Intuitively, no near-arbitrage means there exists no sequence of portfolios that earn positive expected returns with vanishing risks. Ingersoll (1984) established that a sufficient and necessary condition for the absence of near-arbitrage is that[10]

$$S^\star = \sqrt{\alpha^\intercal \Sigma_u^{-1} \alpha} \lesssim_{\mathrm{P}} 1. \tag{4.2.4}$$

Here, $S^\star$ is the theoretically optimal Sharpe ratio arbitrageurs can achieve in this economy using a portfolio strategy that has zero exposure to factor risks, namely, a "statistical arbitrage" strategy in the jargon of practitioners. This result suggests that moderate mispricing in the form of nonzero alphas is permitted in an economy without near-arbitrage opportunities, but there cannot be too many alphas that are too large, to the extent that $S^\star$ explodes.[11]

To achieve this optimal Sharpe ratio, arbitrageurs should hold a portfolio with weights given by $w^\star \propto \Sigma_u^{-1}\alpha$, according to Ingersoll (1984).[12] Under the rational expectation assumption, arbitrageurs (agents in this model) know the true (population) parameters: α and Σ_u. In reality, however, the true parameters are blind to arbitrageurs as they can only learn these parameters from a finite sample of data. This learning effect is sometimes harmless

10. We use the notation $a \lesssim b$ to denote $a \leq Cb$ for some constant $C > 0$, $a \lesssim_{\mathrm{P}} b$ to denote $a = O_{\mathrm{P}}(b)$, $a \eqsim b$ if $a \lesssim b$ and $b \lesssim a$, and use $a \eqsim_{\mathrm{P}} b$ accordingly.

11. Assuming α_i is *i.i.d.* and $\lambda_{\max}(\Sigma_u) \lesssim_{\mathrm{P}} 1$, by equation (4.2.4), we have $\alpha^\intercal\alpha \lesssim_{\mathrm{P}} \alpha^\intercal \lambda_{\min}(\Sigma_u^{-1})\alpha \lesssim_{\mathrm{P}} \alpha^\intercal\Sigma_u^{-1}\alpha \lesssim_{\mathrm{P}} 1$, so that $\mathrm{E}(\alpha_i^2) = o(1)$ by the law of large numbers.

12. In Ingersoll (1984), α is defined to be the cross-sectional projection of the expected returns onto β in the population model such that $\alpha^\intercal\Sigma_u^{-1}\beta = 0$. Contrary to this, our paper sets forth a predetermined DGP as shown in (4.2.1), where we introduce α as a random variable with the property $\mathrm{E}(\alpha^\intercal\beta) = 0$. Consequently, the optimal strategy, $w^\star$, as derived in Theorem 7 and illustrated by (4.2.10), differs from the formula presented by Ingersoll (1984). That said, we can show that our $w^\star$ achieves the same Sharpe ratio $S^\star$ (asymptotically as N increases).

since it can be expected that when the sample size is large enough, the true parameters are (asymptotically) revealed, and hence the predictions under rational expectation hold approximately. Fundamentally, this phenomenon is due to the assumption that the learning problem in the limiting experiment becomes increasingly simpler as the sample size increases.

In the current context, the difficulty of the learning problem also hinges on the number of investment opportunities, N. As N increases, it becomes increasingly difficult for arbitrageurs to determine which among all assets truly have nonzero alphas for a given sample size, T. If the learning problem remains difficult as N and T increase, the learning effect persists, which could lead to distinct limiting implications as opposed to the rational expectation case. It turns out that the rational expectation limit $S^\star$ is only relevant for rather restrictive scenarios. In more realistic settings, e.g., N is much larger than T, the optimal Sharpe ratio arbitrageurs can achieve without factor exposures is far smaller than $S^\star$ because of their inability to make error-free inference. Therefore, the condition (4.2.4) could be excessively restrictive in such scenarios.

To illustrate this intuition, we consider a simple and specific example.

EXAMPLE 3: Suppose the cross-section of alphas is drawn from the following distribution:

$$\alpha_i \overset{i.i.d.}{\sim} \begin{cases} \mu & \text{with prob. } \rho/2 \\ -\mu & \text{with prob. } \rho/2 \\ 0 & \text{with prob. } 1-\rho \end{cases}, \quad 1 \le i \le N, \tag{4.2.5}$$

where $\mu \ge 0$ and $0 \le \rho \le 1$, and they potentially vary with N and T. In addition, we also assume $\beta = 0$, $\Sigma_u = \sigma^2 \mathbb{I}_N$, for some $\sigma > 0$. Consequently, Equation (4.2.1) becomes $r_t = \alpha + u_t$.

In this example, μ dictates the strength of alphas, ρ describes how rare alphas are, whereas σ is a nuisance parameter. By modeling parameters μ and ρ as functions of the sample size and dimensions of the investment set, we introduce greater flexibility in depicting

the challenges arbitrageurs encounter in finite sample situations.[13] To emphasize the role of signal strength and count, we impose in this example that all assets share the same alpha distribution and the same idiosyncratic variance.

Now suppose, more specifically, that the magnitude of (μ, ρ) satisfies

$$\mu \asymp T^{-1/2} \quad \text{and} \quad \rho \asymp N^{-1/2}. \tag{4.2.6}$$

This condition (4.2.6) implies that the signal strength μ vanishes as the sample size increases $(T \to \infty)$ and the signal percentage count ρ decays as the investment universe expands $(N \to \infty)$. This setup is used to approximate a reality with only a small portion of assets having a nonzero yet small alpha. On the other hand, σ is assumed a fixed constant, since in reality idiosyncratic risks never vanish, whereas alphas can be small driven by competition among arbitrageurs. This model rests on an uncommon territory in the existing literature of asset pricing: weak and rare alphas. In fact, the classical no near-arbitrage condition (4.2.4) imposes, implicitly, weakness or rareness on alphas; otherwise, if alphas are strong and dense, $\alpha^\intercal\alpha$ would explode rather rapidly. Even in the current setting, in light of the fact that $\mathrm{E}(\alpha^\intercal\alpha) = \rho\mu^2 N$, we still have $\alpha^\intercal\alpha \xrightarrow{\mathrm{P}} \infty$ as long as $N^{1/2}/T \to \infty$. In other words, a near-arbitrage opportunity arises according to (4.2.4), with a strategy $w = \sigma^{-2}\alpha$.

However, the statistical obstacle prevents arbitrageurs from having this "free lunch." In general, it is only possible to recover any element of alpha up to some estimation error of magnitude $T^{-1/2}$.[14] Since by design the true alpha is of the same order of magnitude as its

13. Adopting a drifting sequence for parameters is a common trick in econometrics to provide more accurate finite sample approximations. As Bekker (1994) put, "in evaluating the results, it is important to keep in mind that the parameter sequence is designed to make the asymptotic distribution fit the finite sample distribution better. It is completely irrelevant whether or not further sampling will lead to samples conforming to this sequence or not."

14. Giglio, Liao, and Xiu (2021) develop the asymptotic normality result for alpha estimates via a Fama-MacBeth procedure in various scenarios, in which factors are (partially) observable or latent whereas β is unknown. The CLTs in these scenarios share the same form: for any $1 \le i \le N$,

$$\sqrt{T}(\widehat{\alpha}_i - \alpha_i) \xrightarrow{d} \mathcal{N}(0, \sigma_i^2(1 + \gamma^\intercal(\Sigma_v)^{-1}\gamma)), \tag{4.2.7}$$

level of statistical uncertainty, i.e., $\mu \approx T^{-1/2}$, it is impossible for arbitrageurs to determine precisely which assets among all have nonzero alpha.

For illustration purpose, suppose that arbitrageurs adopt the strategy $\widehat{w} = \sigma^{-2}\widehat{\alpha}$,[15] replacing α in w with $\widehat{\alpha} = \bar{r} = \alpha + \bar{u}$.[16] Out of sample, this portfolio's conditional expected return and conditional variance can be written as:

$$\begin{aligned}
\mathrm{E}\left(\sigma^{-2}\left(\alpha+\bar{u}\right)^{\intercal}\left(\alpha+u_{t+1}\right)|\mathcal{F}_t\right) &= \sigma^{-2}(\alpha^{\intercal}\alpha+\bar{u}^{\intercal}\alpha),\\
\mathrm{Var}\left(\sigma^{-2}\left(\alpha+\bar{u}\right)^{\intercal}\left(\alpha+u_{t+1}\right)|\mathcal{F}_t\right) &= \sigma^{-2}(\alpha^{\intercal}\alpha+2\alpha^{\intercal}\bar{u}+\bar{u}^{\intercal}\bar{u}),
\end{aligned}$$

where u_{t+1} denotes a future de-meaned return at $t+1$, that shares the same distribution as $\{u_s\}_{s\leq t}$, but is independent of $\bar{u}$ which belongs to the information set up to t, $\mathcal{F}_t$. The resulting squared conditional Sharpe ratio is given by:

$$S^2 = \frac{\sigma^{-4}(\alpha^{\intercal}\alpha+\bar{u}^{\intercal}\alpha)^2}{\sigma^{-2}(\alpha^{\intercal}\alpha+2\alpha^{\intercal}\bar{u}+\bar{u}^{\intercal}\bar{u})} \lesssim_{\mathrm{P}} T^{-1} \to 0, \tag{4.2.8}$$

where we use the fact that $\bar{u}^{\intercal}\bar{u} \approx_{\mathrm{P}} N/T$ when u_t is i.i.d.. In other words, this portfolio achieves a Sharpe ratio equal to zero asymptotically.

Is there a superior trading strategy capable of maintaining a non-vanishing Sharpe ratio? The straightforward answer is no. Our discussion below will elucidate that, within this context, namely, (4.2.6) holds true, the highest achievable Sharpe ratio for all *feasible* trading strategies employed by arbitrageurs, represented as S^{OPT}, vanishes asymptotically as $N, T \to \infty$. Conversely, the *infeasible* optimal Sharpe ratio, denoted $S^{\star}$, diverges under the

---

where σ_i^2 is the ith entry of Σ_u. In the case that β is observable (but factors are not), we can show that the CLT has a similar form except that the scalar $(1+\gamma^{\intercal}(\Sigma_v)^{-1}\gamma)$ disappears.

15. The knowledge of σ is ultimately inconsequential for our purposes, as we will demonstrate subsequently for a more general setting. Despite σ being known, this strategy fails to yield any positive Sharpe ratio.

16. For any time series of random vector a_t, we use $\bar{a}$ to denote its sample average. As we will point out later in the paper, this strategy $\widehat{w}$, which we will denote by $\widehat{w}^{\mathrm{CSR}}$, fails to achieve the optimal Sharpe ratio in all scenarios. We will discuss the optimal strategy in Section 4.2.5.

condition that $N^{1/2}/T \to \infty$. There is a vast disparity, as shown by this example, between S^{OPT} and $S^{\star}$.

The difference between feasible and infeasible strategies is primarily driven by the information set accessible to arbitrageurs when they implement their trading strategies. In this example, the infeasible strategy assumes that arbitrageurs have access to a comprehensive information set, $\mathcal{F}_t$, that encompasses the knowledge of α. This knowledge proves to be extremely valuable when α is both rare and weak, thereby establishing a significant gap between this strategy and others lacking access to such information.

To clarify, we formalize the definitions as follows:

DEFINITION 2: A strategy is deemed *feasible* when it relies solely on observable data, in stark contrast to an *infeasible* strategy that presupposes complete knowledge of the DGP, encapsulated by the full information set, $\mathcal{F}_t$.

Recognizing the distinction between feasible and infeasible strategies, we proceed to investigate the maximum Sharpe ratio attainable by feasible strategies. This exploration necessitates defining the decision-making problem faced by arbitrageurs, which we turn to next.

4.2.3 Arbitrageurs' Decision Problem and Feasible Sharpe Ratio Bound

In the broader context of the DGP, given by equation (4.2.1), the information set $\mathcal{F}_t$ not only includes α but also other variables such as Σ_u, β, and v_t, alongside unknown parameters like Σ_v and γ. For a strategy to be considered feasible, it must refrain from leveraging these unknown variables and parameters. This constraint introduces complexity to the arbitrageurs' decision-making process by intertwining optimization with statistical techniques, resulting in solutions that are not uniquely defined.

To navigate this complexity, we introduce a wider set of trading strategies that specifically exclude reliance on information about α, recognizing that the Sharpe ratio wedge between

feasible and feasible strategies mainly stems from the process of learning about α. Given this constraint, we derive an achievable upper bound for the Sharpe ratio that applies to all such strategies, and by extension, to all feasible strategies, since they similarly eschew information on α. Yet, this upper bound is not sharp, because the restriction does not prevent strategies that may leverage information on other variables, like Σ_u. In Section 4.2.5, we further show that it is possible to attain this upper bound with a well-crafted feasible trading strategy. Therefore, even lacking knowledge of other variables and parameters, arbitrageurs can effectively use statistical inference to estimate these unknowns, with the statistical error on such estimates becoming negligible asymptotically. This highlights that the main hurdle for arbitrageurs is not in learning about factors, volatilities, or risk premia, but in acquiring knowledge about α.

We are now ready to characterize the decision problem arbitrageurs confront. Operating under an information set, $\mathcal{G}$, a subset of $\mathcal{F}_t$, arbitrageurs are tasked with solving a mean-variance optimization problem expressed as:

$$\widetilde{w} = \arg \max_{w \in \mathcal{G}: w^\intercal \beta = 0} U(w), \quad \text{where} \quad U(w) = \mathrm{E}\left(w^\intercal r_{t+1} \middle| \mathcal{G}\right) - \frac{\kappa}{2} \mathrm{Var}\left(w^\intercal r_{t+1} \middle| \mathcal{G}\right),$$

and κ measures the degree of risk aversion. Our objective is to identify the optimal strategy subject to the measurability constraint: $w \in \mathcal{G}$. Given the necessity for statistical arbitrage to be factor-neutral, the strategy w must be orthogonal to β, with β being predetermined, i.e., $\beta \in \mathcal{G}$.

To streamline the portfolio formation process, we exclude transaction cost considerations, transforming it into a static, single-period optimization problem. Thus, we omit the subscript t wherever possible. This approach aligns with the APT framework introduced by Ross (1976), where the arbitrageurs' chosen strategy aims to eliminate exposure to systematic factors, concentrating instead on capitalizing on alphas and balancing idiosyncratic risks. Their utility is modeled as a Bayesian posterior expected loss, a concept discussed in Berger

(1985). The next theorem presents a solution with minimal conditions on $\mathcal{G}$.

Theorem 7. *Suppose that r_t follows (4.2.1) and that β is $\mathcal{G}$-measurable. It holds that*

$$\widetilde{w} = \frac{1}{\kappa}\widetilde{\Sigma}_u^{-1/2}\mathbb{M}_{\widetilde{\Sigma}_u^{-1/2}\beta}\widetilde{\Sigma}_u^{-1/2}\widetilde{\alpha}, \tag{4.2.9}$$

where $\widetilde{\alpha} := \mathrm{E}(\alpha|\mathcal{G})$ and $\widetilde{\Sigma}_u := \mathrm{E}(\Sigma_u|\mathcal{G}) + \mathrm{Var}(\alpha|\mathcal{G})$.[17]

This result associates the optimal strategy with the posterior summary statistics of α and Σ_u. Essentially, the choice of the optimal $\mathcal{G}$-measurable strategy is determined by arbitrageurs' best parameter estimates based on their information set, rather than the true values of the unknown parameters. In the special scenario where $\mathcal{G} = \mathcal{F}_t$, meaning arbitrageurs possess complete information up to time t, we have $\widetilde{\alpha} = \alpha$, and $\widetilde{\Sigma}_u = \Sigma_u$. Under these conditions, the optimal infeasible strategy, denoted by $w^\star$, is given by

$$w^\star = \frac{1}{\kappa}\Sigma_u^{-1/2}\mathbb{M}_{\Sigma_u^{-1/2}\beta}\Sigma_u^{-1/2}\alpha. \tag{4.2.10}$$

Moving forward, our attention is directed towards a specific choice of $\mathcal{G}$ as generated by $\{(r_s, \beta, v_s, \Sigma_u) : t - T + 1 \leq s \leq t\}$.[18] Accordingly, $\widetilde{w}$ is assumed associated with this particular choice by default hereafter.

Next, we examine the investment performance of arbitrageurs' optimal strategy, $\widetilde{w}$. It is important to note that in evaluating a strategy, the conditional Sharpe ratio is calculated based on the comprehensive information set, $\mathcal{F}_t$, rather than the arbitrageurs' specific information set, $\mathcal{G}$. For any given strategy w, its conditional Sharpe ratio is defined as:

$$S(w) \coloneqq \mathrm{E}(w^\intercal r_{t+1}|\mathcal{F}_t)/\mathrm{Var}(w^\intercal r_{t+1}|\mathcal{F}_t)^{1/2}.$$

17. $\mathbb{M}_A = \mathbb{I} - A(A^\intercal A)^{-1}A^\intercal$ for any matrix A.

18. The nuisance parameters γ and Σ_v are deterministic constants and, as such, are encompassed within any information set. Although they are part of $\mathcal{G}$, their presence does not influence the optimal strategy of arbitrageurs.

To describe the asymptotic behavior of a strategy, we need impose assumptions on the return generating process:

Assumption 11. *For each $N \geq 1$, the following conditions hold:*

(a) (α_i, u_i) is i.i.d. across i, and satisfies $\mathrm{E}(\alpha_i|(\Sigma_u)_{i,i}) = 0$ and $\mathrm{E}\|\alpha\|^2_{\mathrm{MAX}} = o(1)$. Moreover, it holds that $1 \lesssim_{\mathrm{P}} \lambda_{\min}(\Sigma_u) \leq \lambda_{\max}(\Sigma_u) \lesssim_{\mathrm{P}} 1$.[19]

(b) The pricing errors α, factors v_t, factor loadings β, and idiosyncratic errors u_t are, conditionally on Σ_u, mutually independent.

Condition (a) suggests that the alphas in our model are inherently weak; as the number of assets, N, increases, their magnitudes diminish.[20] Moreover, this condition ensures that volatilities remain within upper and lower bounds. These two conditions imply that learning about alpha is a more arduous task than learning about volatilities. Also, condition (b) is imperative for the model's identification. For instance, assuming independence between α and β is key to identify the risk premia, γ, see, e.g., Giglio, Kellly, and Xiu (2022).

Given this assumption, the subsequent theorem illustrates that the optimal strategy $\widetilde{w}$ yields a Sharpe ratio that is nearly at its maximum.

Theorem 8. *Suppose that r_t follows (4.2.1) and that Assumption 11 holds. Let $\mathcal{G}$ represent the information set generated by $\{(r_s, \beta, v_s, \Sigma_u) : t - T + 1 \leq s \leq t\}$. For any strategy w that is both $\mathcal{G}$-measurable and factor-neutral, if follows that as $N \to \infty$,*

$$S(w) \leq S(\mathcal{G}) + o_{\mathrm{P}}(1 + S(\mathcal{G})), \tag{4.2.11}$$

19. For a matrix A, we use $\|A\|$ and $\|A\|_{\mathrm{MAX}} = \max_{i,j} |a_{ij}|$ to denote the operator norm (or ℓ_2 norm) and the ℓ_∞ norm of A on the vector space. We use $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to denote the minimum and maximum eigenvalues of A.

20. By Assumption 13(a), $\mathrm{Var}(\alpha_i) = \mathrm{E}(\alpha_i^2) = o(1)$. Referencing our earlier discussion (footnote 11), a diminishing variance in α is essential for precluding near-arbitrage opportunities within Ross' APT framework.

where $S(\mathcal{G})^2 = \widetilde{\alpha}^\intercal \Sigma_u^{-1} \widetilde{\alpha}$*, and the equality in* (4.2.11) *holds if* $w = \widetilde{w}$*. Moreover, the optimal strategy for arbitrageurs,* $\widetilde{w}$*, is approximately equal to*

$$\left\| \widetilde{w} - \frac{1}{\kappa} \mathbb{M}_\beta \Sigma_u^{-1} \widetilde{\alpha} \right\| = o_{\mathrm{P}}(1 + S(\mathcal{G})). \tag{4.2.12}$$

Accordingly, their optimal utility satisfies $\sqrt{2\kappa U(\widetilde{w})} = S(\mathcal{G}) + o_{\mathrm{P}}(1 + S(\mathcal{G}))$.

Theorem 8 derives an upper bound, denoted as $S(\mathcal{G})$, on the Sharpe ratio that any $\mathcal{G}$-measurable strategy can attain. This upper bound also connects with arbitrageurs' optimal utility. By the definition of $S(\mathcal{G})$, this upper bound satisfies that

$$\mathrm{E}\big(S(\mathcal{G})^2\big) \leq \mathrm{E}\big(\alpha^\intercal \Sigma_u^{-1} \alpha\big), \tag{4.2.13}$$

with the equality holds only when $\widetilde{\alpha} = \alpha$ almost surely, where the right-hand side corresponds to the infeasible scenario in which arbitrageurs have perfect knowledge of α, which echoes (4.2.4), the result given by Huberman (1982).

Furthermore, the result demonstrates that the optimal strategy, as articulated in Theorem 7, can reach this upper bound with a negligible approximation error. In scenarios where $S(\mathcal{G})$ is finite, that is $O_{\mathrm{P}}(1)$, the error term $o_{\mathrm{P}}(1+S(\mathcal{G}))$ reduces to $o_{\mathrm{P}}(1)$ and vanishes in the limit. The theorem additionally addresses situations where $S(\mathcal{G})$ may diverge, in which instance the approximation error is reduced to $o_{\mathrm{P}}(S(\mathcal{G}))$, remaining inconsequentially small relative to $S(\mathcal{G})$ itself.

Also, Equation (4.2.12) offers a more straightforward and intuitive formula for approximating the optimal feasible strategy, $\widetilde{w}$. Additionally, as a byproduct of Theorem 8, the optimal infeasible strategy in Equation (4.2.10) can also be approximated in a simpler form:

$$\left\| w^\star - \frac{1}{\kappa} \mathbb{M}_\beta \Sigma_u^{-1} \alpha \right\| = o_{\mathrm{P}}(1 + S^\star), \tag{4.2.14}$$

which represents a special case of Equation (4.2.12) where $\mathcal{G}$ is set as $\mathcal{F}_t$. At its core, it is $\widetilde{\alpha}$, the posterior estimate of the pricing errors that dictates the optimal Sharpe ratio achievable by arbitrageurs, rather than α themselves. Intuitively, part of the construction in (4.2.12), $\Sigma_u^{-1}\widetilde{\alpha}$, is the optimal allocation to the ex-factor returns, $\alpha + u_t = r_t - \beta(\gamma + v_t)$, based on a simple mean-variance analysis, except for the use of $\widetilde{\alpha}$, that addresses the problem that arbitragers do not observe true alphas in the DGP. Multiplying by $\mathbb{M}_\beta$ to $\Sigma_u^{-1}\widetilde{\alpha}$ simply eliminates its factor exposures, because $(\mathbb{M}_\beta\Sigma_u^{-1}\widetilde{\alpha})^\intercal\beta = 0$.

Theorems 7 and 8 are concerned with factor-neutral strategies, as our main results focus on the limits of arbitrage. Our appendix broadens this scope to encompass any strategy, revealing that any $\mathcal{G}$-measurable strategy, w, adheres to a similar upper limit:

$$S(w) \leq \left(S(\mathcal{G})^2 + \gamma^\intercal\Sigma_v^{-1}\gamma\right)^{1/2} + o_{\mathrm{P}}(1 + S(\mathcal{G})), \tag{4.2.15}$$

where $\gamma^\intercal\Sigma_v^{-1}\gamma$ is the squared optimal Sharpe ratio earned from factor portfolios or SDF.

This result provides a solution to a long standing problem in optimal portfolio allocation amidst parameter uncertainty. Historically, the mean-variance portfolio, derived from sample means and covariance matrices, is noted for its suboptimal performance. When assuming normally distributed returns, Kan and Zhou (2007) studied the expected performance of the plug-in mean-variance portfolio and found its Sharpe ratio is smaller than that of the infeasible Sharpe ratio, $S^\star$. Nevertheless, they did not provide an upper bound of the feasible optimal Sharpe ratio under parameter uncertainty. Our results fill this gap, identifying an upper bound of the feasible optimal Sharpe ratio under the same factor model framework as in APT.

The result in Theorem 8 appears to require that arbitrageurs rely on the information set $\mathcal{G}$, which embodies perfect knowledge of factors, v_t, and their exposures, β, in addition to past asset returns, r_t. Moreover, arbitrageurs appear to have perfect knowledge of the covariance matrix of idiosyncratic errors, Σ_u. In fact, the upper bound in (4.2.15) still holds

if arbitrageurs are endowed with less information, because for any information sets $\mathcal{G}'$ and $\mathcal{G}$ such that $\mathcal{G}' \subseteq \mathcal{G}$, we have $\mathrm{E}(S(\mathcal{G}')^2) \leq \mathrm{E}(S(\mathcal{G})^2)$. Since all feasible strategies are measurable with respect to $\mathcal{G}$, $S(\mathcal{G})$ establishes an upper bound for the Sharpe ratio applicable to this set of strategies, although it may not be sharp. In light of this and Definitions 1 and 2, we immediately obtain a sufficient condition of the absence of near-arbitrage with feasible strategies:

Corollary 3. *Suppose the same assumptions as in Theorem 8 hold. For any given return-generating process satisfying (4.2.1), there exists no feasible strategy that leads to a near-arbitrage, if*

$$S(\mathcal{G}) \lesssim_{\mathrm{P}} 1, \quad \textit{as} \quad N \to \infty. \tag{4.2.16}$$

4.2.4 Bayes Correction for Selection Bias

While the previous discussion underscored the role of the posterior mean of α, denoted as $\widetilde{\alpha} = \mathrm{E}(\alpha|\mathcal{G})$, on the optimal $\mathcal{G}$-measurable portfolio, this expression remains implicit and not directly actionable. Our focus now shifts to reducing the information set $\mathcal{G}$ to its sufficient statistics for α. This step is crucial for understanding why $\widetilde{\alpha}$ leads to enhanced portfolio performance. To facilitate this analysis, we introduce further assumptions:

Assumption 12. *For each $N \geq 1$, the following conditions hold:*

(a) $u_{i,t} = \sigma_i \varepsilon_{i,t}$, *where $\varepsilon_{i,t}$ follows a standard normal distribution, and is i.i.d. across (i,t) and independent of Σ_u.*

(b) $s_i := \alpha_i/\sigma_i$ *is independent of σ_i.*

Based on the DGP described in Equation (4.2.1) and given the information set $\mathcal{G}$, Assumptions 11 and 12(a) together ensure that the key summary statistics for α_i are the volatility σ_i and the sample average of the ex-factor returns, expressed as $\check{\alpha}_i = \bar{r}_i - \beta_i(\gamma + \bar{v}) = \alpha_i + \bar{u}_i$. In other words, $\mathrm{E}(\alpha_i|\mathcal{G}) = \mathrm{E}(\alpha_i|\check{\alpha}_i, \sigma_i)$. Consequently, this assumption simplifies the

conditional information set $\mathcal{G}$ in the posterior distribution of α to merely a two-dimensional vector comprising these conditioning variables. To evaluate this conditional expectation, it becomes necessary to assume a specific form of dependence between α_i and σ_i.

In light of this, we introduce Assumption 12(b), which allows us to further express $\mathrm{E}(\alpha_i|\check{\alpha}_i, \sigma_i) = \sigma_i \mathrm{E}(s_i|\check{\alpha}_i, \sigma_i) = \sigma_i \mathrm{E}(s_i|\check{\alpha}_i/\sigma_i)$.[21] This leads to that $\widetilde{s}_i := \mathrm{E}(s_i|\mathcal{G}) = \mathrm{E}(s_i|\check{s}_i)$, where $\check{s}_i := \check{\alpha}_i/\sigma_i$. Consequently, in terms of the scaled version of α, namely, s, the conditional information set is now further simplified to a single scalar variable, significantly simplifying the estimation problem later. Further, in light of Theorem 8, $\widetilde{\alpha}$, $\widetilde{w}$, and $S(\mathcal{G})^2$ can all be represented in relation to $\widetilde{s}$:[22]

$$\widetilde{\alpha} = \Sigma_u^{-1/2}\widetilde{s}, \quad \widetilde{w} \propto \mathbb{M}_\beta \Sigma_u^{-1/2}\widetilde{s}, \quad S(\mathcal{G})^2 = \widetilde{s}^\intercal \widetilde{s}. \tag{4.2.17}$$

Finally, note that

$$\check{s}_i = s_i + \bar{\varepsilon}_i \sim \mathcal{N}(s_i, T^{-1}), \quad \text{conditional on } s_i.$$

This formulation casts the original posterior inference problem into the framework of the classical Gaussian sequence model for recovering a high-dimensional mean vector, s, from noisy observations $\check{s}$, which has been extensively studied in the statistics literature, see, e.g., Robbins (1956), Efron (2011), and Efron (2019). Although the assumption that ε_i follows a Gaussian distribution may seem restrictive, this framework is sufficiently versatile to incorporate a wide range of distributions for s_i.

Arbitrageurs face the challenge of identifying the true underlying signal, s, from an observed noisy signal $\check{s}$. This task is complicated by what Efron (2011) described as selection bias or "the winner's curse." This phenomenon suggests that a high observed signal, $\check{s}_i$, may

21. This equality relies on the result that conditional on $\widehat{\alpha}_i/\sigma_i$, α_i/σ_i is independent of σ_i. We impose this condition primarily for clarity of exposition and simplicity of Algorithm 3 below.

22. Given that risk aversion does not impact the out-of-sample Sharpe ratio, we will use $\propto$ to substitute for κ^{-1} in our subsequent discussions on the portfolio strategy.

be reflective of the high s_i, or it could be the result of "lucky" noise, $\bar{\varepsilon}_i$, being unusually large. Consequently, arbitrageurs must carefully adjust their investment strategies to mitigate this potential bias. The correction involves relying on $\widetilde{s}$ that accounts for this bias. To see this, we present an explicit formula for $\widetilde{s}$:

Theorem 9. *Suppose that r_t follows (4.2.1) and Assumptions 11 and 12 hold. We introduce a function $\psi(a) = \mathrm{E}\left(s_i | \check{s}_i = a\right)$, with which we have $\widetilde{\alpha}_i = \sigma_i \widetilde{s}_i$, where $\widetilde{s}_i = \psi(\check{s}_i)$. Moreover, it holds that*

$$\psi(a) = a + \frac{1}{T}\frac{d}{da}\log p(a), \tag{4.2.18}$$

where $p(a) = \mathrm{E}\big(\phi_{1/T}(a - s_i)\big)$ is the probability distribution function of $\check{s}_i$.[23]

The preceding discussion directly leads to the first assertion of the theorem. Equation (4.2.18) is rooted in Tweedie's formula (Robbins (1956)), which establishes a connection between the posterior mean of s, given $\check{s} = a$, denoted $\psi(a)$, and the posterior probability distribution of $\check{s}$. The formula's second component, $T^{-1}d\log p(a)/da$, plays a crucial role in adjusting for the selection bias in the observed signal, $\check{s}_i$, which introduces several intriguing properties. An example of these properties, as shown by Andrews, Arnold, and Krutchkoff (1972) is that $\psi(a)$ is a nondecreasing function of a. This property ensures that the relative magnitude of various signals is preserved post-correction. Moreover, under mild assumptions about the prior distribution of s_i, such as being symmetric and unimodal, the posterior mean $\widetilde{s}$ induces a shrinkage effect towards the prior mean, which, in our context, is zero.

The posterior shrinkage on α, or equivalently on s, consequently induces a "shrinkage" effect on the optimal Sharpe ratio achievable by arbitrageurs, as demonstrated by the inequality (4.2.13). Based on the results of Theorem 9, and under a technical condition concerning the tail behavior of s_i, we can obtain a more explicit formula for $S(\mathcal{G})$:

23. $\phi_x(a)$ is the distribution function for $\mathcal{N}(0, x)$.

Corollary 4. *Under the same assumptions outlined in Theorem 9, and with the additional condition that* $\mathrm{E}(s_i^2 \mathbb{1}_{\{|s_i| \geq c_N\}}) = o(N^{-1})$*, we arrive at the conclusion that:*

$$S(\mathcal{G}) = S^{\mathrm{OPT}} + o_{\mathrm{P}}(1), \quad \textit{with} \quad S^{\mathrm{OPT}} = \left(N \int \psi(a)^2 p(a) da \right)^{1/2}.$$

This result enables us to compute S^{OPT} in various examples. For instance, by utilizing this result, we can compare S^{OPT} with $S^\star$ from Huberman (1982) in Example 3. This comparison illuminates the feasibility of attaining non-vanishing Sharpe ratios , highlighting how different DGP conditions influence their achievability.

Corollary 5. *Suppose that the same assumptions as in Corollary 4 hold. In addition, we assume alpha follows (4.2.5) as in Example 3. Then we have* $S^\star = \sigma^{-1}\mu(\rho N)^{1/2} + o_{\mathrm{P}}(1)$*. Further, assuming that* $\sigma^{-1}\mu(\rho N)^{1/2}$ *does not vanish, then it holds that* $S^{\mathrm{OPT}} \leq (1 - \epsilon)\sigma^{-1}\mu(\rho N)^{1/2}$ *for some* $\epsilon > 0$*, if and only if*

$$T^{1/2}\mu/\sigma - \sqrt{-2\log\rho} \lesssim 1. \tag{4.2.19}$$

Corollary 5 suggests that when $T^{1/2}\mu/\sigma$ is large that the constraint (4.2.19) is violated, $S^\star \eqsim_{\mathrm{P}} S^{\mathrm{OPT}}$, that is, in the limit, learning does not play any role, so that arbitrageurs in this scenario achieve the same optimal Sharpe ratio as in Huberman (1982). Furthermore, the rareness parameter ρ does not make much difference if $T^{1/2}\mu/\sigma$ gets sufficiently large. That said, if ρ approaches to zero so fast to the extent that $\sqrt{-2\log\rho}$ dominates $T^{1/2}\mu/\sigma$, that is, alpha is extremely rare and sufficiently weak, the learning problem becomes rather challenging and hence S^{OPT} is dominated by $S^\star$ in the limit, resulting in a strictly smaller Sharpe ratio than the infeasible Sharpe ratio in the classical case.

To give a concrete example of Corollary 5, consider an alternative DGP assumption as

opposed to (4.2.6):[24]

$$\mu \approx N^{-\eta} \quad \text{and} \quad \rho > 0 \quad \text{is fixed.} \tag{4.2.20}$$

In this scenario, $(S^\star)^2 \approx_{\mathrm{P}} N^{1-2\eta}$, which explodes unless $\eta > 1/2$. If further assuming that $N/T \to \psi > 0$, then the left-hand-side of condition (4.2.19) is of order $N^{1/2-\eta} \vee 1$, so that (4.2.19) holds if and only if $\eta \geq 1/2$. Therefore, $\eta < 1/2$ is not consistent with absence of (feasible) near arbitrage in that $S^\star$ explodes, while in the mean time $S^{\mathrm{OPT}} = S^\star$ (by Corollary 5) and hence explodes. If $\eta > 1/2$, $S^\star$ (and hence S^{OPT}) vanishes, which does not seem like an economically plausible case. If we think that arbitrageur activity is required to prevent substantial mispricing, then a setting where mispricing disappears asymptotically even if the frictions faced by arbitrageurs are very large is not plausible. This suggests that under this DGP (4.2.20), the only economically plausible case with absence of near-arbitrage is $\eta = 1/2$. That is, η can be thought as determined in equilibrium, in which there are substantial asset demand distortions such that mispricing in the absence of arbitrageur action would be non-negligible asymptotically, and arbitrageurs are aggressive enough so that near-arbitrage opportunities do not exist asymptotically.

We now illustrate the behavior of S^{OPT} numerically and verify the theoretical predictions of Corollary 5 using the DGP specified in Example 3. Figure 4.1 reports the Sharpe ratio, S^{OPT}, of optimal feasible arbitrage portfolios for a range of μ/σ and ρ values in the case of $N = 1,000$ and $T = 20$ years. Recall that according to model (4.2.5), a ρ percentage of assets have alphas with a Sharpe ratio μ/σ. That is, ρ characterizes the rareness of the alpha signal, whereas μ/σ captures its strength. We intentionally choose a wide range of μ/σ (with annualized Sharpe ratios from 0.11 to 10.95) and ρ (from 0.12% to 50%) to shed light on the dependence landscape of Sharpe ratios on signal weakness and rareness, despite that some of the resulting portfolio Sharpe ratios (the top left conner of Figure 4.1) are unrealistically

24. It is easy to show that the setup (4.2.20) satisfies all assumptions of Corollary 4 for all fixed $\eta > 0$.

high. Note that when $\mu/\sigma \times \sqrt{12}$ hits 0.44, its corresponding t-statistic based on a 20-year sample exceeds 1.96, the typical t-hurdle for a standard student-t test.

The pattern of Sharpe ratios agrees with our intuition and theoretical predictions. For any fixed ρ, as the alpha signal weakens (i.e., μ/σ decreases), the optimal Sharpe ratio drops. The same is true if we decrease the signal count (i.e., ρ vanishes), for any fixed value of μ/σ. The arbitrageur's learning problem is the easiest when signal is strong and count is large (top left conner), and the most challenging towards the right bottom corner, where the optimal Sharpe ratios drop to near zero.

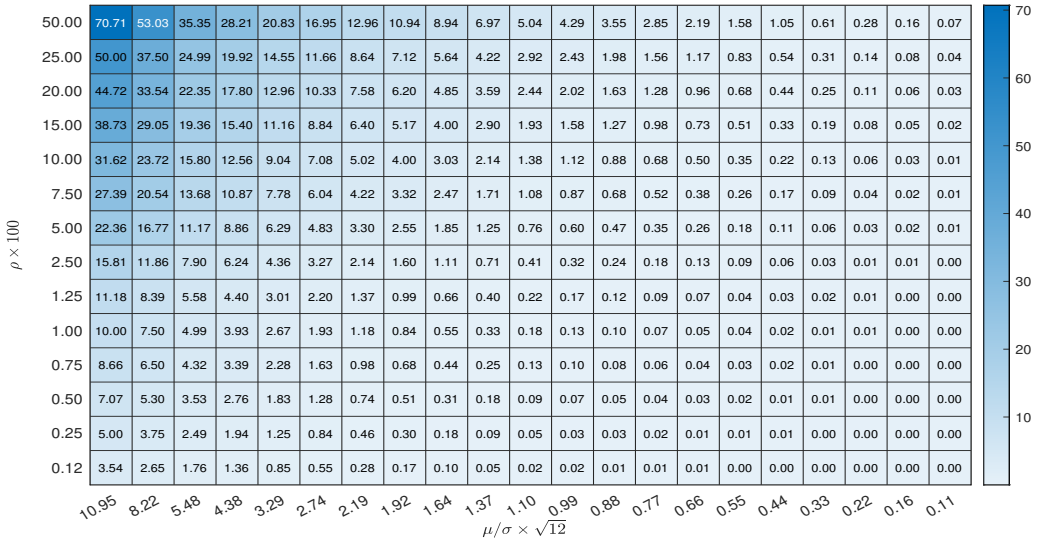


Figure 4.1: Optimal Sharpe Ratios (S^{OPT}) of Feasible Arbitrage Portfolios

Note: The figure reports optimal Sharpe ratios of feasible arbitrage portfolios in model (4.2.5), in which a $100 \times \rho\%$ of assets have alphas that correspond to an annualized Sharpe ratio $\mu/\sigma \times \sqrt{12}$.

The reported Sharpe ratios on Figure 4.1 are only a fraction of the corresponding (infeasible) Sharpe ratios, $S^\star = \sqrt{\alpha^\intercal(\Sigma_u)^{-1}\alpha} = \mu/\sigma\sqrt{\rho N}$, as shown by Figure 4.2. The pattern we see from Figure 4.2 agrees with theoretical predictions of Corollary 5. When the annualized Sharpe ratio $\mu/\sigma \times \sqrt{12}$ is larger than 2.74, regardless of the values of ρ, the signal-to-noise ratio of the learning problem is sufficiently strong that the statistical limit to arbitrage does not matter much, and hence $S^{\text{OPT}}/S^\star$ is close to 1. Nonetheless, this regime is irrelevant

in practice, since it is mostly associated with unrealistically high Sharpe ratios (see Figure 4.1). In contrast, as μ/σ diminishes, the gap between $S^\star$ and S^{OPT} widens. In almost all empirically relevant scenarios, $S^\star$ is largely exaggerated.

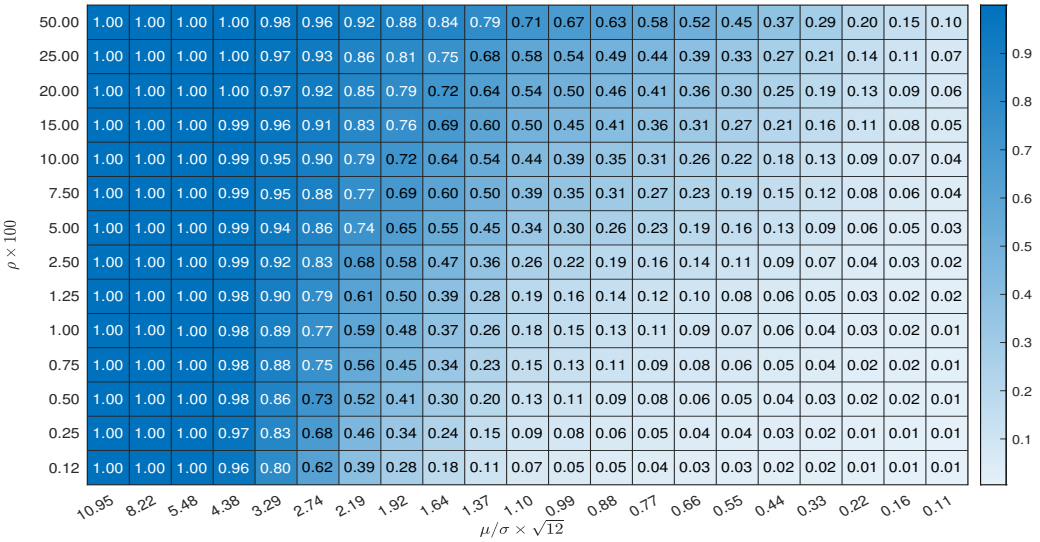


Figure 4.2: Ratios between S^{OPT} and $S^\star$

Note: The figure reports the ratios of optimal Sharpe ratios between feasible and infeasible arbitrage portfolios. The simulation setting is based on model (4.2.5), in which a $100 \times \rho\%$ of assets have alphas that correspond to an annualized Sharpe ratio $\mu/\sigma \times \sqrt{12}$.

4.2.5 Constructing the Optimal Arbitrage Portfolio

In our prior discussion, Theorem 8 established that the optimal Sharpe ratio attainable by any feasible strategy is capped by $S(\mathcal{G})$, which, by Corollary 4, is approximately equal to S^{OPT} under additional conditions. Following this, Corollary 5 illustrates the dependence of S^{OPT} on the unknown DGP parameters. However, these findings presuppose access to the information set $\mathcal{G}$, which not only includes past returns, r_t, but also potentially unknown variables, such as β, v_t, and Σ_u, alongside unknown parameters in the DGP, γ and Σ_v. As a result, the strategy $\widetilde{w}$ is infeasible, given its reliance on Σ_u and $\widetilde{s}$, which in turn depends on the ex-factor returns, $\bar{r}_i - \beta_i(\gamma + \bar{v})$. Therefore, arbitrageurs encounter substantial challenges in this context. They must conduct statistical inference concerning unknown variables and

parameters to shape their investment strategies. Moreover, the devised optimal strategy must be flexible and adapt to the nuances of different DGPs.

Despite the challenges, arbitrageurs can indeed construct a uniformly optimal strategy that attains S^{OPT} across a broad spectrum of DGPs, even without perfect knowledge of the true DGP. In fact, this strategy leverages only observable data, making it entirely feasible. Specifically, we formulate this strategy within a framework where factors remain latent, but factor exposures are observable – a scenario that aligns with our empirical analysis.

We describe this portfolio strategy as "all weather," signifying its adaptability to all considered DGP scenarios. Moreover, the fact that this strategy achieves S^{OPT} implies that the feasible Sharpe ratio upper bound we derive is sharp. We outline the following algorithm to construct this strategy:

Algorithm 3.

Inputs: r_t, $t \in \mathcal{T} = \{t - T + 1, \ldots, t\}$ *and* β.

S1. Construct cross-sectional regression estimates of alpha and idiosyncratic volatilities, for each $i = 1, 2, \ldots, N$*:*

$$\widehat{\alpha} = T^{-1} \sum_{s \in \mathcal{T}} \mathbb{M}_\beta r_s, \qquad \widehat{\sigma}_i^2 = T^{-1} \sum_{s \in \mathcal{T}} \big((\mathbb{M}_\beta r_s)_i - \widehat{\alpha}_i \big)^2, \quad \textit{and} \quad \widehat{s}_i := \widehat{\alpha}_i / \widehat{\sigma}_i.$$

S2. Construct a nonparametric estimate of the marginal density of $\widehat{s}$ *using Gaussian kernel function* $\phi_{1/T}(x)$ *and bandwidth* $k_N \sim (\log N)^{-1}$*:*

$$\widehat{p}(a) = \frac{1}{N k_N} \sum_i \phi_{1/T}\Big(\frac{\widehat{s}_i - a}{k_N} \Big).$$

S3. Estimate $\widetilde{s}$ by Tweedie's formula (4.2.18) and isotonic regression:

$$\breve{\psi}(a) = a + \frac{1+k_N^2}{T}\frac{d}{da}\log \widehat{p}(a).$$
$$\widehat{\psi} = \text{argmin}_{x\in\mathbb{R}^N}\|x-\breve{\psi}\|^2,\ \ s.t.\ \ x_i \le x_j \ \ if\ \widehat{s}_i \le \widehat{s}_j,\ \ for\ 1\le i,j\le N,$$

where $\breve{\psi}_i := \breve{\psi}(\widehat{s}_i)$.

S4. Construct the arbitrage portfolio weights as $\widehat{w}^{\text{OPT}} \propto \mathbb{M}_\beta \widehat{\Sigma}_u^{-1/2}\widehat{\psi}$.

Outputs: $\widehat{w}^{\text{OPT}}$.

Step S1 of Algorithm 3 provides feasible estimates of $\widehat{\alpha}$ and $\widehat{\Sigma}_u = \text{Diag}(\widehat{\sigma}_i^2)$, which in turn leads to the sufficient statistic, $\widehat{s}$. The motivation behind Steps S2 and S3 stems from Tweedie's formula; here, we employ a nonparametric empirical Bayes method for estimating the posterior mean function using kernel density estimation, as suggested by Brown and Greenshtein (2009).[25] The incorporation of the factor $(1+k_N^2)$ serves to adjust for finite sample biases introduced by estimation errors in $\widehat{p}(a)$. To enhance the nonparametric estimator's performance in finite samples, enforcing monotonicity on $\psi(\cdot)$ proves beneficial. For this purpose, isotonic regression is utilized (see Robertson, Dykstra, and Wright (1988)), yielding a monotonic piece-wise linear approximation of $\psi(\cdot)$. Step S4 constructs the optimal portfolio weights, $\widehat{w}^{\text{OPT}}$, following (4.2.17).

An essential step towards achieving optimality involves aggregating information from assets with comparable $\widehat{s}_i$, as done in Step S2. This strategy outperforms the alternatives, some of which directly use estimated $\widehat{s}$ as if these estimates are not susceptible to estimation errors even when they are rather weak, or simply ignore the contribution of all weaker signals. Like any machine learning method, the proposed approach requires a tuning parameter k_N, which can be selected in a validation sample.

25. Step S2 exploits symmetry in the prior distribution of s for the construction of the nonparametric kernel density estimator. This enhances finite sample performance and streamlines the proof.

The following theorem demonstrates the optimality of $\widehat{w}^{\mathrm{OPT}}$:

Theorem 10. *Let $\mathbb{P}$ denote the collection of all data-generating processes under which r_t follows (4.2.1), and Assumptions 11 and 12 hold. Moreover, we assume $\|\beta\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} 1$, $\lambda_{\min}(\beta^{\intercal}\beta) \gtrsim_{\mathrm{P}} N$, and that the distribution of s_i is symmetric. In addition, suppose that $N^d \lesssim T \lesssim N^{d'}$ for fixed constants $d > 1/2$ and $d' < 1$. We denote the Sharpe ratio achieved by $\widehat{w}^{\mathrm{OPT}}$ as $\widehat{S}^{\mathrm{OPT}} := \mathrm{E}(r_{t+1}^{\intercal}\widehat{w}^{\mathrm{OPT}}|\mathcal{F}_t)/\mathrm{Var}(r_{t+1}^{\intercal}\widehat{w}^{\mathrm{OPT}}|\mathcal{F}_t)^{1/2}$. Then it holds that $\widehat{w}^{\mathrm{OPT}}$ achieves, asymptotically, the upper bound S^{OPT}, uniformly over all sequences of data-generating processes. That is, for any $\epsilon > 0$,*

$$\lim_{N,T\to\infty} \sup_{\mathrm{P}\in\mathbb{P}} \mathrm{P}\big(\big|\widehat{S}^{\mathrm{OPT}} - S^{\mathrm{OPT}}\big| \geq \epsilon S^{\mathrm{OPT}} + \epsilon\big) = 0.$$

Theorem 10 imposes a mile assumption on $\lambda_{\min}(\beta^{\intercal}\beta)$ requires that all factors are pervasive.[26] This condition is frequently utilized in factor model literature and is particularly relevant here, given our assumption that the factors within our model are latent. A significant difference between our results and the traditional empirical Bayes literature (e.g., Brown and Greenshtein (2009)) is our lack of direct access to observable signals. To eliminate factor exposure, cross-sectional regressions become essential, inevitably introducing estimation errors into $\widehat{s}_i$. The additional conditions imposed by Theorem 10 ensure that such estimation errors become asymptotically negligible.

Theorem 10 concludes that in the context of a linear factor model, arbitrageurs can construct this strategy, without any knowledge besides past returns and risk exposures (beta), to achieve the maximal Sharpe ratio over all feasible trading strategies that have zero exposure to factor risks. This Sharpe ratio precisely characterizes the limit of feasible arbitrages in economic terms. The rationale behind this conclusion stems from the setting where the idiosyncratic variances represented by Σ_u remain significant as N and T increase, in contrast

26. See, e.g., Assumption I.1 of Giglio and Xiu (2021a). While our theoretical results may extend to certain weak factor settings, this is not our emphasis here.

to alphas. This distinction is empirically grounded; alphas are expected to be small and rare , a result of the competitive nature of arbitrage activities, as opposed to idiosyncratic risks that persistently exist. Therefore, in the context of arbitrage limitations, the importance of learning about Σ_u is minimal and becomes negligible. Moreover, acquiring knowledge about v_t and risk premia γ is considerably more straightforward given their low-dimensional characteristics and the knowledge of β.

The term $\epsilon S^{\text{OPT}} + \epsilon$ accommodates both small and large values of S^{OPT}. If $S^{\text{OPT}} \lesssim 1$, then ϵ dominates and the estimation error inside the probability is characterized by the absolute difference between $\widehat{S}^{\text{OPT}}$ and S^{OPT}. Otherwise, if $S^{\text{OPT}} \to \infty$, the estimation error is described in percentage terms. This is necessary because we simultaneously consider a large class of models.

With Theorem 10, we establish the necessity for the no near-arbitrage condition given by (4.2.16).

Corollary 6. *Suppose the same assumptions as in Theorem 10 hold. The portfolio weights by $\widehat{w}^{\text{OPT}}$ yields a near-arbitrage strategy under any sequences of data-generating processes for which condition (4.2.16) does not hold.*

We have shown that arbitrageurs can construct an optimal strategy that realizes S^{OPT}. Now suppose that the equilibrium "cost" of implementing an arbitrage is C in an economy with statistical limit of arbitrage. In equilibrium, $S^{\text{OPT}} = C$, otherwise arbitrageurs can trade until it is no longer profitable to do so. We can thereby interpret $\widehat{S}^{\text{OPT}}$ as an empirical estimate of the arbitrage cost, which we will estimate empirically.

4.2.6 Estimating Optimal Infeasible Sharpe Ratio

We are also interested in estimating the optimal infeasible Sharpe ratio, $S^\star$, which can be perceived as the optimal Sharpe ratio from an outside econometrician's point of view, and yet cannot be realized by a feasible portfolio. Existing literature on testing APT often construct test statistics in the spirit of Gibbons, Ross, and Shanken (1989), which are effectively based

on $S^\star$, see, e.g., Pesaran and Yamagata (2017) and Fan, Liao, and Yao (2015). While such tests are powerful and may lead to detection of alphas, their relevance to arbitrageurs might be limited. The challenge for arbitrageurs lies in translating the statistical evidence into a feasible portfolio strategy that enables profit realization from these discovered alphas, a dilemma that our proposed S^{OPT} addresses directly.

To provide a benchmark for $\widehat{S}^{\text{OPT}}$, we propose an estimator for $S^\star$ inspired by its sample analog:

$$\widetilde{S}^\star = \left(\bar{r}^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{r}\right)^{1/2}. \tag{4.2.21}$$

Unfortunately, this estimator has a non-vanishing asymptotic bias for certain DGPs we consider, as we will show later. To fix this issue, we propose a new estimator that is uniformly consistent:

$$\widehat{S}^\star = \left(\bar{r}^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{r} - N/T\right)^{1/2}. \tag{4.2.22}$$

The next proposition summarizes the asymptotic properties of both estimators.

Proposition 3. *Suppose that r_t follows (4.2.1) and that Assumption 11 holds. Assume that $\mathrm{E}(\alpha_i^2 \mathbb{1}_{\{|\alpha_i| \geq c_N\}}) \leq c_N N^{-1}$, $T \lesssim N$, $T^{-1} N^{1/2} \log N \leq c_N$, for some sequence $c_N \to 0$, and that $\varepsilon_{i,t}$ has finite eighth moment. Then we have*

$$\begin{aligned} \left|\widehat{S}^\star - S^\star\right| \big/ \left(1 + S^\star\right) &= o_{\mathrm{P}}\left(T^{-1/2} N^{1/4} \sqrt{\log N}\right), \\ \left|\widetilde{S}^\star - \left((S^\star)^2 + NT^{-1}\right)^{1/2}\right| \Big/ \left(1 + S^\star\right) &= o_{\mathrm{P}}\left(T^{-1} N^{1/2} \log N\right). \end{aligned}$$

Similar to Theorem 10, the estimation error is relative when $S^\star$ dominates 1.0 asymptotically, and in absolute terms if $S^\star$ is dominated by 1.0.[27] This accommodates a large class of models, some of which have an exploding or a shrinking $S^\star$. While it is possible to estimate $S^\star$, it is not possible to build a portfolio that realize it, unless the signal-to-noise ratio is

27. Obviously, the threshold 1.0 can be replaced by any fixed constant.

sufficiently large such that $S^\star = S^{\text{OPT}}$. Empirically, the difference between $\widehat{S}^\star$ and $\widehat{S}^{\text{OPT}}$ thereby informs us about the impact of learning from the data.

4.2.7 Alternative Strategies for Arbitrage Portfolios

Algorithm 3 introduces a nuanced approach enabling arbitrageurs to reach a state of feasible optimality. However, it is unclear whether this strategy has been employed by arbitrageurs in real-world settings. Consequently, this section delves into various alternative methodologies. While none of these alternatives guarantee uniform optimality across all DGPs we consider, their relative simplicity and observed prevalence in practical applications warrant examination. Our analysis helps illustrate their pros and cons in different DGP scenarios.

Cross-Sectional Regression

One of the most common strategies involves forming portfolios based directly on the cross-sectional estimates of α, obtained in Step S1 of Algorithm 3. This method is referred to as CSR, with its portfolio represented by $\widehat{w}^{\text{CSR}}$:

$$\widehat{w}^{\text{CSR}} \propto \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \widehat{\alpha}. \tag{4.2.23}$$

The strategy described is effectively a sample analogue of the approximated infeasible strategy given by (4.2.14).

We now exploit Example 3 to illustrate the pros and cons of the CSR strategy. Figure 4.3 illustrates the relationship between S^{CSR}, the theoretical Sharpe ratio achieved by $\widehat{w}^{\text{CSR}}$, and S^{OPT} across a spectrum of DGPs. According to our proposition in the online appendix, S^{CSR} tends to be overshadowed by S^{OPT} in scenarios where alpha signals are sparse ($\sqrt{\rho T}\mu/\sigma$ is not excessively large) and simultaneously strong ($\sqrt{T}\mu/\sigma$ is not exceedingly small). This specific regime of dominance is clearly marked with black numerals within the heatmap in Figure 4.3. As the ratio $\mu/\sigma \times \sqrt{12}$ edges closer to 1.0—either moving towards this vertical threshold from the right hand side or descending from the upper left corner—the gap between

S^{CSR} and S^{OPT} enlarges increasingly.

The CSR approach takes all signal estimates directly, without differentiating the significant ones from the insignificant ones. Consequently, even fake signals (pure noise) are assigned non-zero weights, which, in turn, hurts the portfolio's performance. On the other hand, the CSR strategy can achieve optimality when the strong signals are abundant (so that portfolio weights allocated to noise are inconsequential) or when all signals are weak (so that they do not differ too much from fake ones). The latter case is interesting, as it also suggests that simply ignoring weaker signals is not optimal. That said, Figure 4.1 shows that the DGPs for which the cross-sectional regression approach is strongly dominated by our optimal strategy are associated with realistic Sharpe ratios.

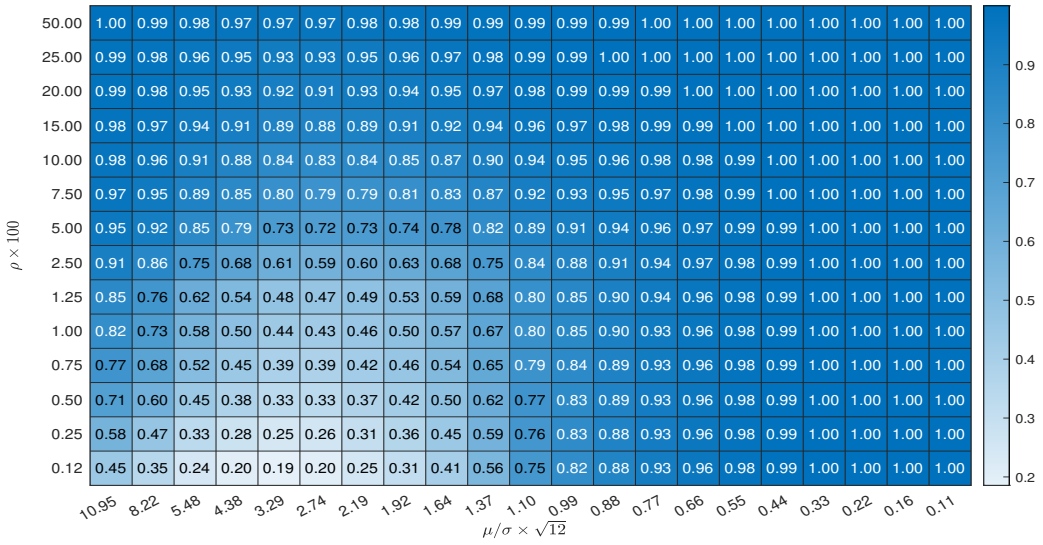


Figure 4.3: Ratios between S^{CSR} and S^{OPT}

Note: The figure reports the ratios between the Sharpe ratios of the OLS based portfolio and the feasible optimal arbitrage portfolio. The theoretical Sharpe ratio achieved by CSR is denoted as S^{CSR}, given explicitly by Proposition **??**. The simulation setting is based on model (4.2.5), in which a $100 \times \rho\%$ of assets have alphas that correspond to an annualized Sharpe ratio $\mu/\sigma \times \sqrt{12}$.

The CSR approach is a simple benchmark as it does not rely on any advanced statistical techniques to detect signals or distinguish their strength. The strategy we discuss next is more advanced, in that it controls false discoveries among selected strong signals using the

B-H procedure proposed by Benjamini and Hochberg (1995).

False Discovery Rate Control

To address the aforementioned selection bias in identifying profitable alpha signals, an alternative methodology conceptualizes the search for alpha as a multiple testing problem. In this case, assuming there are N assets each potentially associated with a nonzero α_i, we can establish for each asset i a null hypothesis $\mathbb{H}_0^i : \alpha_i = 0$. Rejection of this null hypothesis leads to the discovery of a non-zero alpha. Rather than focusing on the significance level of individual tests, a more appropriate strategy involves controlling the FDR, an approach recommended by Barras, Scaillet, and Wermers (2010), Bajgrowicz and Scaillet (2012), and Harvey, Liu, and Zhu (2016) in various asset pricing contexts. Giglio, Liao, and Xiu (2021) have proved the validity of the B-H procedure for FDR control in a general factor model setting for alpha detection. Below we describe the necessary steps to prepare alpha estimates for constructing an arbitrage portfolio.

Begin with a series of p-values, p_i, where each is the result of a t-test on the cross-sectional regression estimate of α_i, $\sqrt{T}\widehat{s}_i$, for $i = 1, 2, \ldots, N$. These p-values assess the significance of each α_i's deviation from zero. Arrange these p-values in ascending order, from the smallest to the largest, resulting in a sorted sequence $p_{(1)} \leq \ldots \leq p_{(N)}$. Identify a critical index, $\widehat{k}$, defined as the maximum i such that $p_{(i)} \leq \tau i/N$ where τ is a predetermined significance level, commonly set at 5%.

The selection of $\widehat{k}$ is strategic, aiming to ensure that, on average, at least a fraction $(1-\tau)$ of the alpha estimates identified as significant (i.e., those with p-values smaller than $p_{(\widehat{k})}$) are truly non-zero. This B-H criterion proves effective irrespective of the overall proportion of non-zero alphas present in the DGP. This method allows for a systematic and statistically robust selection of alpha estimates, minimizing the risk of including false positives, and these selected alpha estimates are then utilized as inputs for constructing an arbitrage portfolio,

as illustrated by the following equation:

$$\widehat{w}^{\mathrm{BH}}(\tau) \propto \mathbb{M}_{\beta}\widehat{\Sigma}_{u}^{-1}\widehat{\alpha}^{\mathrm{BH}}(\tau), \text{ where } \widehat{\alpha}_{i}^{\mathrm{BH}}(\tau) = \widehat{\alpha}_{i}\mathbb{1}_{\{p_i \leq p_{(\widehat{k})}\}}. \tag{4.2.24}$$

This strategy introduces a hard thresholding mechanism to the alpha estimates, effectively nullifying the impact of alphas deemed insignificant.

Controlling the FDR on top of the CSR estimates is intuitively appealing, but doing so incurs a potential loss of power, leading to less investment opportunities. Our focus is on optimal portfolio construction instead of FDR control. Our proposition in the appendix shows that in the context of Example 3, arbitrageurs who adopt the B-H based-trading strategy cannot achieve optimal portfolio for a large class of DGP sequences.

As shown by the proposition in the appendix and illustrated numerically by Figure 4.4, the discrepancy between the optimal Sharpe ratio and that achieved through the B-H method is largely determined by the signal strength. The B-H procedure nears optimality when the signal strength, quantified by $\sqrt{T}\mu/\sigma$, is substantial—exceeding the threshold of $\sqrt{-2\log\rho}$. The instances where the B-H strategy reaches optimality are depicted by the white values on Figure 4.4, with the boundary of this optimal region approximated by the line where $\mu/\sigma\sqrt{12} = 2.19$. This demonstrates that the B-H method excels in identifying strong signals, leading to near-optimal portfolios when signals are strong. Conversely, in the presence of weak signals, the B-H procedure, which amounts to hard-thresholding, tends to underperform compared to CSR.

This point is further elaborated in Figure 4.1, which underscores that even if individual alphas are weak, their aggregated effect on a portfolio's Sharpe ratio can be non-trivial. The B-H approach takes a notably conservative stance towards signal selection, especially in contexts where signals are weak. This cautious approach ensures the reliability of selected alphas by focusing on those that are truly significant. However, this method might not fully capitalize on the potential cumulative impact of weaker signals. In contrast, our optimal

arbitrage portfolio leverages the full spectrum of alpha estimates, including false positives, extending beyond the significant selections made through the B-H procedure.

At its core, this delineates a subtle yet critical divergence between two objectives: alpha testing and portfolio construction. The former prioritizes the identification of statistically significant alphas with FDR under control, whereas the latter focuses on the utilization of all available information to optimize portfolio performance. These objectives do not always align.

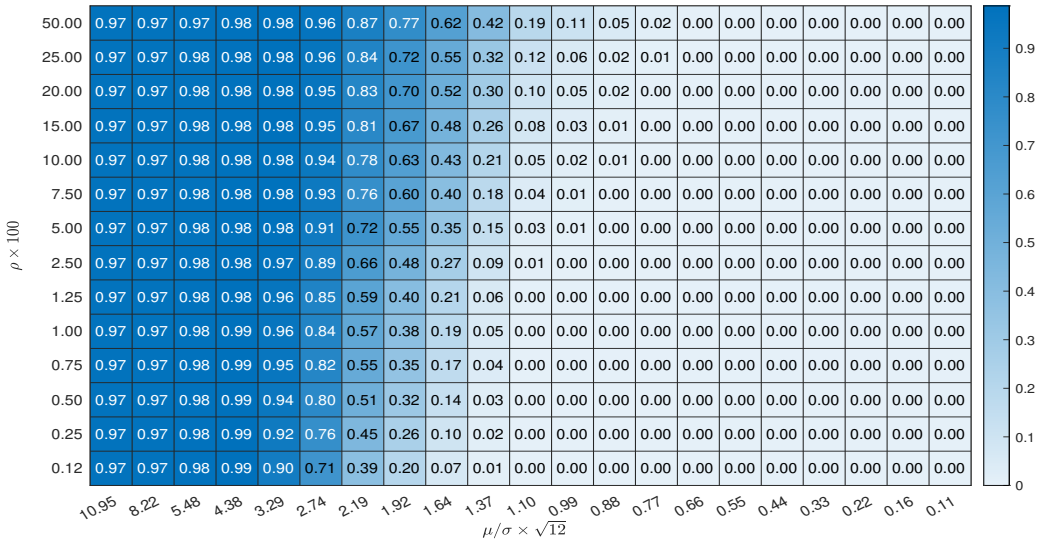


Figure 4.4: Ratios between S^{BH} and S^{OPT}

Note: The figure reports the ratios between the Sharpe ratios of the multiple testing based portfolio (via B-H procedure) and the feasible optimal arbitrage portfolio. The theoretical Sharpe ratio achieved by B-H is denoted as S^{BH}, given explicitly by Proposition **??**. The simulation setting is based on model (4.2.5), in which a $100 \times \rho\%$ of assets have alphas that correspond to an annualized Sharpe ratio $\mu/\sigma \times \sqrt{12}$.

The CSR and the B-H approaches represent two typical strategies in practice. The former trades all signals without distinguishing their strength, whereas the latter only trades the stronger signals. Neither approach always achieves optimality.

Shrinkage Approaches

The analysis above suggests that we can construct the optimal portfolio out of the ex-factor returns, while imposing regularization on portfolio weights, before rewriting the regularized portfolio weights in terms of raw returns (i.e., multiplying the weights by $\mathbb{M}_\beta$). Regularizing portfolio weights amounts to imposing priors directly on the alpha estimates. To see this, we adopt a shrinkage approach, when constructing arbitrage portfolios on residual returns:

$$\max_w \{w^\intercal \widehat{\alpha} - \frac{1}{2} w^\intercal \widehat{\Sigma}_u w - p_\lambda(\widehat{\Sigma}_u^{1/2} w)\},$$

where $p_\lambda(x) = \lambda\|x\|_1$ or $\lambda\|x\|_2^2$, for some $\lambda > 0$. Since $\widehat{\Sigma}_u$ is diagonal, this optimization problem has a closed-form solution: $\breve{w}_{q,i}(\lambda) = \widehat{\sigma}_i^{-1}\psi_q(\widehat{s}_i, \lambda)$, for $i = 1, 2, \ldots, N$, where $q = 1$ corresponds to the Lasso penalty and $q = 2$ the ridge, and $\psi_q(s, \lambda)$ is

$$\psi_1(s, \lambda) = \text{sgn}(s)(|s| - \lambda)_+, \quad \psi_2(s, \lambda) = (1 + 2\lambda)^{-1} s.$$

This leads to the optimal portfolio weight on r_t:[28]

$$\widehat{w}_q(\lambda) \propto \mathbb{M}_\beta \breve{w}_q(\lambda), \quad q = 1, 2.$$

Depending on the magnitude of λ, the Lasso approach replaces all smaller signals (i.e., $\widehat{s}_i$) by zero and shrinks the larger ones by λ in absolute terms. In other words, the Lasso approach is the soft-thresholding alternative to the B-H method. In contrast, the ridge penalty shrinks all signals proportionally, which means ridge is equivalent to CSR! This "embedded" shrinkage effect of CSR explains why it performs well in the case of small signals.

The online appendix, offers explicit formulae for the optimal feasible Sharpe ratios in the

28. An alternative strategy is to impose sparsity directly on the portfolio weights with respect to raw returns. While this approach might be appealing from the transaction cost point of view, it does not associate with an explicit prior on alpha, hence is more difficult to interpret.

shrinkage case. Unlike Ridge, which remains unaffected by the tuning parameter regarding its Sharpe ratio, Lasso's performance is contingent on its tuning parameter. However, even with the optimal choice of tuning parameter, it cannot achieve the optimal feasible Sharpe ratio in all DGPs.

Figure 4.5 compares S_λ^{LASSO} with S^{OPT}, where S_λ^{LASSO} denotes the theoretical Sharpe ratio achieved by Lasso for a given tuning parameter choice λ. Given that the Lasso approach involves this tuning parameter, it necessitates a cross-validation procedure. We adopt an infeasible and theoretically optimal tuning parameter that maximizes S_λ^{LASSO} in Figure 4.5. Although theoretical results in the appendix suggest that Lasso is not uniformly optimal, it performs remarkably well, achieving the optimal Sharpe ratio in almost all regimes. Intuitively, when signals are very strong, Lasso behaves like a hard-thresholding selector, as shrinkage has minimal effect. Conversely, when signals are weak, Lasso behaves akin to Ridge (and hence CSR), since shrinking these signals does not alter the fact that they are nearly indistinguishable from noise.

4.3 Simulation Evidence

This section demonstrates the empirical relevance of our theory via simulations and examines the finite sample performance of the proposed portfolio strategies.

For simplicity and clarity, we simulate a one-factor (CAPM) model of returns given by (4.2.1). We choose the factor risk premium as 5% per year and set the annualized volatility at 25%. We model the cross-section of betas using a normal distribution with mean 1 and variance 1. Since we focus on the arbitrage portfolio, the parameters about the factor component (including the number of factors) are inconsequential, because factors, if any, are eliminated by $\mathbb{M}_\beta$ in the first step when constructing these trading strategies. In addition, we adopt model (4.2.5) in Example 3 for the cross-sectional distribution of alpha, and fix the idiosyncratic volatilities of all assets at σ, since it is α/σ that determines the signal strength and that there is no need of varying both α and σ in the cross section.

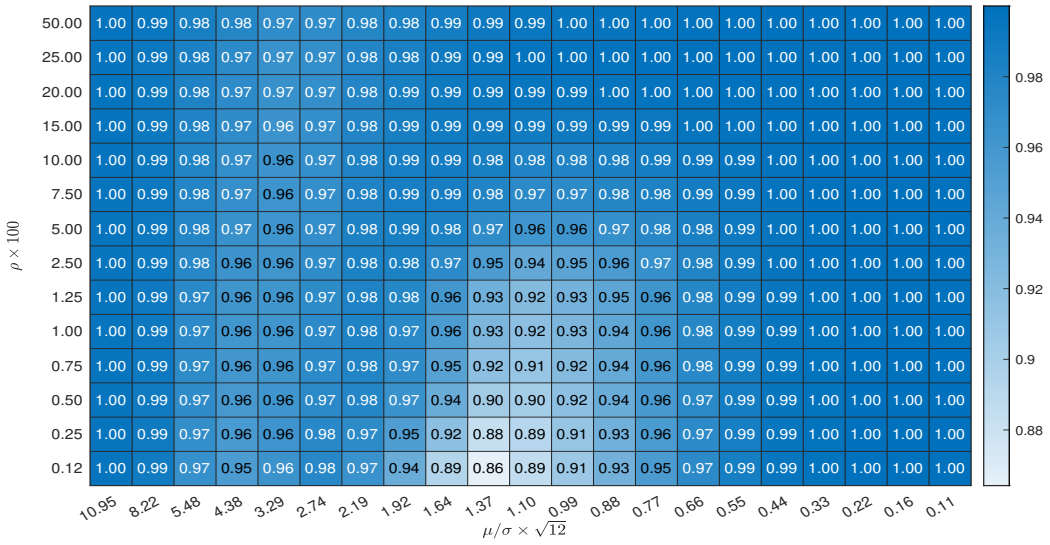


Figure 4.5: Ratios between $S_{\lambda}^{\text{LASSO}}$ and S^{OPT}

Note: The figure reports the ratios between the Sharpe ratios of the Lasso based portfolio and the feasible optimal arbitrage portfolio. The theoretical Sharpe ratio achieved by Lasso is treated as the maximum of $S_{\lambda}^{\text{LASSO}}$ over all λ, given explicitly in the appendix. The simulation setting is based on model (4.2.5), in which a $100 \times \rho\%$ of assets have alphas that correspond to an annualized Sharpe ratio $\mu/\sigma \times \sqrt{12}$. The tuning parameter λ is selected to maximize $S_{\lambda}^{\text{LASSO}}$.

We now compare the finite sample performance of our portfolio estimators over different DGPs. For any given parameter value $(\mu/\sigma, \rho)$ in a DGP, we estimate the portfolio weights, $\widehat{w}^{\text{OPT}}$, using our Algorithm 3, and calculate the resulting (theoretical) Sharpe ratio: $\widehat{w}^{\text{OPT}\intercal}\mu/\sqrt{\widehat{w}^{\text{OPT}\intercal}\Sigma_u^{-1}\widehat{w}^{\text{OPT}}}$. We then calculate the average Sharpe ratio over all Monte Carlo repetitions. Our approach requires a tuning parameter k_n. For robustness, we report results based on three parameter values $(0.5k_n, k_n, 2k_n)$ with $k_n = 0.25$, covering a wide range of tuning parameters. We repeat this exercise for the CSR, B-H, and Lasso methods for comparison. For the Lasso method, we adopt its theoretically optimal yet impractical tuning parameter. This optimal parameter is chosen to maximize $S_{\lambda}^{\text{LASSO}}$, thus ensuring the Lasso method is assessed under the most favorable conditions possible.

In light of Theorem 10, a sensible choice of the estimation error can be written as:

$$\mathrm{Err}^{\mathrm{A}}(\mu/\sigma, \rho) = |\widehat{S}^{\mathrm{A}} - S^{\mathrm{OPT}}|/(1 + S^{\mathrm{OPT}}),$$

where A denotes OPT, CSR, BH, or LASSO, and the dependence of $\widehat{S}^{\mathrm{A}}$ and S^{OPT} on μ/σ and ρ is omitted. When S^{OPT} is large (i.e., $>> 1$), this error is in percentages relative to S^{OPT}; when S^{OPT} is small (i.e., $o_{\mathrm{P}}(1)$), the error is measured in terms of the absolute difference. The error is defined this way because S^{OPT} itself can diverge or diminish depending on different parameters in the simulated DGPs.

Table 4.1 reports the maximal error over all values of μ/σ and ρ. The results show that OPT has a smaller error in almost all cases for all tuning parameters than CSR, BH, or LASSO. As T increases from 10 years to 40 years, the maximum error drops from 0.377 to 0.263 in the case of $N = 1,000$ for $k_n = 0.25$, whereas CSR, BH and LASSO stay above 0.44. The maximal error for CSR is achieved at the lower left conner of Figure 4.1, where signals are strong but rare; for BH, the worst performance occurs around the upper right corner, where many weak signals exist; for LASSO, the worse is near the bottom but in the middle, where signals are neither too strong nor too weak.

	$N = 1,000$, Monthly			$N = 3,000$, Monthly			$N = 1,000$, Daily		
	$T = 10$	$T = 20$	$T = 40$	$T = 10$	$T = 20$	$T = 40$	$T = 10$	$T = 20$	$T = 40$
	0.385	0.332	0.289	0.442	0.367	0.320	0.449	0.440	0.408
OPT	0.377	0.309	0.263	0.437	0.333	0.282	0.411	0.382	0.356
	0.381	0.282	0.233	0.446	0.318	0.247	0.370	0.334	0.303
CSR	0.540	0.489	0.441	0.618	0.570	0.515	0.537	0.485	0.427
BH	0.742	0.703	0.651	0.814	0.789	0.748	0.760	0.715	0.657
LASSO	0.537	0.488	0.440	0.615	0.568	0.512	0.536	0.483	0.426

Table 4.1: Sharpe Ratio Comparison in Simulations

Note: This table reports the maximum error, defined by $\sup_{\mu/\sigma,\rho} \mathrm{Err}^{\mathrm{A}}(\mu/\sigma, \rho)$, where A denotes either OPT, or CSR, or BH, or LASSO, over all values of μ/σ and ρ in Figure 4.1, for several choices of N, T (in years), and data frequencies. The first three rows correspond to the OPT approach with three different values of tuning parameters, $0.5k_n$, k_n, and $2k_n$, respectively, where $k_n = 0.25$. The BH approach controls false discovery rate at a level 5%. The Lasso approach uses the optimal (infeasible) tuning parameter that optimizes S^{LASSO}.

Finally, Figure 4.6 reports the estimation error $\left|\widehat{S}^{\star} - S^{\star}\right| / \left(1 + S^{\star}\right)$ in simulations. The result confirms the consistency result given by Proposition 3. The error is relative when $S^{\star}$ is large or moderate ($>> 1$). We find the relative error is around 1% towards the left top corner. For DGPs near the bottom right corner of Figure 4.6, $S^{\star}$ vanishes as shown by Figures 4.1 and 4.2, the error becomes absolute ($S^{\star} << 1$) and is moderately small given the sample size and the cross-sectional dimension.

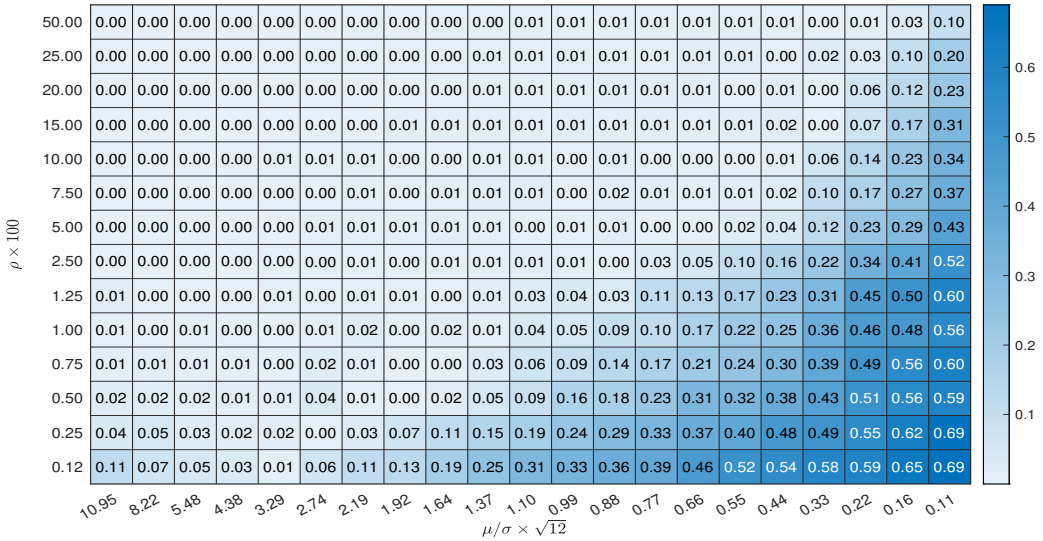


Figure 4.6: Comparison between $\widehat{S}^{\star}$ and $S^{\star}$

Note: The figure reports the error between $\widehat{S}^{\star}$ and $S^{\star}$ defined as $\left|\widehat{S}^{\star} - S^{\star}\right| / \left(1 + S^{\star}\right)$. The simulation setting is based on model (4.2.5), in which a $100 \times \rho\%$ of assets have αs that correspond to an annualized Sharpe ratio $\mu/\sigma \times \sqrt{12}$. In this experiment, $N = 1,000$ and $T = 20$ years.

4.4 Empirical Analysis of US Equities

To demonstrate the empirical relevance of the statistical limit of arbitrage, we study US monthly equity returns from January 1965 to December 2020.

4.4.1 Data Preprocessing

We adopt a multi-factor model with 16 characteristics and 11 GICS sectors, which are selected to incorporate empirical insight from existing asset pricing literature and industry

practice. The selected characteristics include market beta, size, operating profits/book equity, book equity/market equity, asset growth, momentum, short-term reversal, industry momentum, illiquidity, leverage, return seasonality, sales growth, accruals, dividend yield, tangibility, and idiosyncratic risk, which are downloaded directly from the website openassetpricing.com, see Chen and Zimmermann (2020) for construction details.

We download the monthly return data for individual equities from CRSP. We take a number of steps to preprocess the data. First, we single out delisted stocks, and attach delisting returns as their last returns (on the delisting months). Next, we merge the returns data with the aforementioned characteristics database using permnos. The total number of unique permnos on average per month is 6,536. We then apply the usual filters (share codes 10 and 11 and exchange codes 1, 2, and 3) to the database, to eliminate (part of) the sampling periods for stocks that fail to meet these criteria. The remaining average number of stocks per month is 4,756. For stocks whose returns are missing for more than 3 months, we eliminate the missing periods, otherwise we fill the missing returns by zeros.

We now deal with missing characteristics. We start by removing all characteristics data for any stocks since their delisting months. We then fill missing GICS codes with the corresponding stocks' most recent records prior to their missing dates. Stocks without any GICS codes over the entire sample period are eliminated. If the GICS codes become available later in the sample for some stocks, their sample prior to the first dates when GICS become available are eliminated, which mainly occurs prior to 1990. With GICS information, we adopt a two-step procedure to fill in other missing characteristics. For any missing value in a stock's characteristic, we fill it with the sector-wise median of this characteristic each month. If a characteristic's values are not available for an entire sector in a certain month, we fill them with this characteristic's cross-sectional median over all stocks in this month. After data preprocessing, the final average number of stocks per month is reduced to 4,067.

The resulting panel is not balanced, because we do not fill in missing data before a stock's

IPO or after its delisting. Our approach to filling missing data thereby avoids forward-looking bias.

4.4.2 Model Performance

At the end of each month, we run cross-sectional regressions of next month returns onto the 27 cross-sectional predictors (including the intercept). We do so using all stocks in the current month's cross sections. Following Gu, Kellly, and Xiu (2020), the 16 characteristics are rank-normalized within each cross-section, alleviating the impact of extreme outliers in characteristics, though this barely changes any follow-up results.

Figure 4.7 plots the time series of the cross-sectional regression R^2s over time. The R^2 has been on the decline since the beginning of the sample till 1990s. This coincides with the period when the number of stocks in the US equity markets increases. The R^2s are moderately low, with an average of 8.25%. The low R^2s suggest that a substantial portion of cross-sectional variation of individual equity returns is idiosyncratic noise. Therefore, learning alphas from residuals of the factor model is an incredibly difficult statistical task.

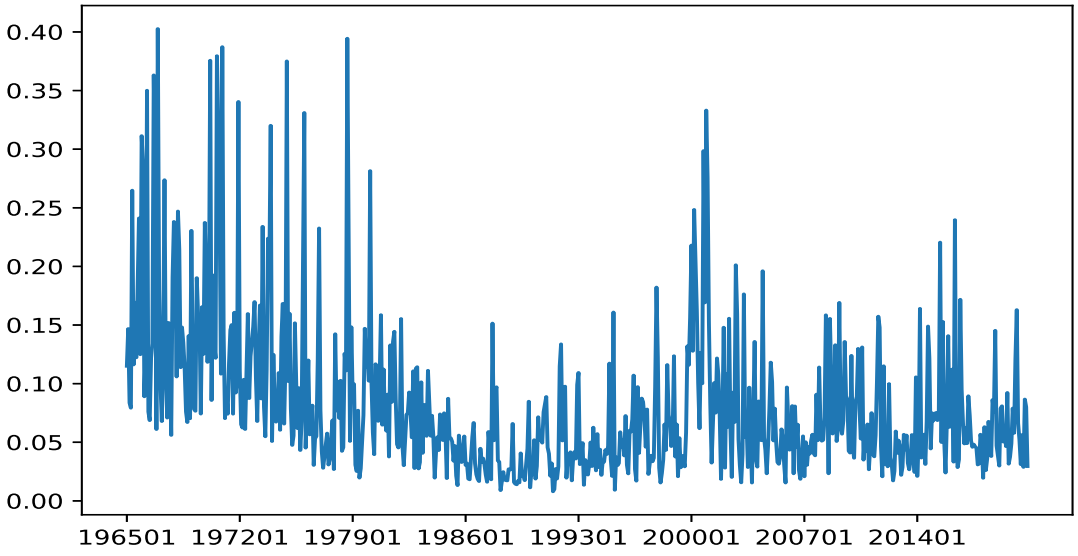


Figure 4.7: Time-series of the Cross-sectional R^2s

4.4.3 Rare and Weak Alphas

We now study the statistical properties of alphas using the full sample data. For each stock, we collect its regression residuals and take their average as an estimate for its alpha. We impose that all residuals have at least 60 observations. This ensures enough sample size for inference on alpha, although the distribution of alphas' t-statistics turns out not sensitive to this requirement. Figure 4.8 provides histograms of the t-statistics and Sharpe ratios for alphas of all 12,415 stocks in our sample that meet this criterion. Because these stocks have different sample sizes, the histograms of the Sharpe ratios are not simply the scaled version of the histogram of the t-statistics.

Only 6.35% of the t-statistics exceed 2.0 in magnitude, and more than 0.63% exceed 3.0. This suggests that truly significant alphas are extremely rare. Moreover, the largest Sharpe ratio of all individual stocks' alphas is rather modest, about 1.699. Only 0.505% of the alphas have a Sharpe ratio greater than 1.0. These summary statistics suggest that rare and weak alpha is perhaps the most relevant scenario in practice.

4.4.4 Performance of Arbitrage Portfolios

Throughout we assume alphas do not vary over time. If alphas are driven by some observable characteristics, then it is possible to construct a factor using these characteristics via cross-sectional regressions, which turns "alpha" into risk premia. In this regard, alphas are meaningless without reference to a specific factor model. Extracting more "factors" out of alphas would lead to even smaller arbitrage profits.

We now compare arbitrage portfolios based on various strategies, including the optimal strategy, the cross-sectional regression (CSR) approach, the multiple-testing based procedure (BH), and LASSO approach. The ridge approach is omitted, since it is equivalent to the CSR.

Specifically, at the end of each month, we build optimal portfolio weights using these strategies. We only invest in stocks with a continuous record for at least 96 months. We

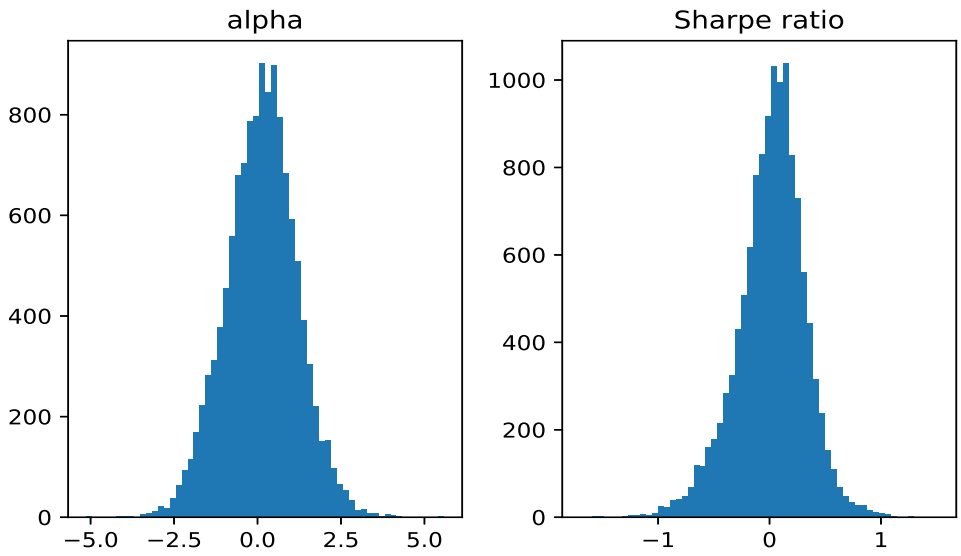


Figure 4.8: Histograms of the t-Statistics and Sharpe Ratios of Estimated Alphas

Note: The figure provides the histograms of the t-statistics (left) and Sharpe ratios (right) of estimated alphas for all tickers in our sample with at least 60 months of data. The total number of tickers available is 12,482.

rebalance these portfolios at the end of each month, with weights recalculated using a 120-month rolling window. Both Lasso and the optimal strategy require a tuning parameter. Out of the 10-year rolling window, we leave the last 2 years as the validation sample for tuning parameter selection. As expected, optimal tuning parameter is difficult to select, which undermines the performance of both strategies.

All these strategies yield similar Sharpe ratios. BH and OPT tie for the top of the chart, yielding 0.497 and 0.496, respectively, followed by CSR that scores 0.450. The LASSO approach only obtains 0.384. The Sharpe ratios of different strategies are not influenced by risk aversion, though the cumulative returns are. To compare cumulative returns, we normalize all strategies to have the same (ex-post) volatility. The resulting time-series of normalized cumulative returns are shown in Figure 4.9.

Closely examining these strategies reveals more insight. BH is highly conservative. Out of 46 years of out-of sample trading months (1975/01 - 2020/12), 289 months have no trading

activities. The largest number of stocks selected for trading in a month is 10, and the average over all non-zero periods is 2.43. In contrast, CSR trade all stocks that meet our trading criteria, with an average of 2,366 stocks per month. OPT almost does so, with an average of 2,359. The number of stocks traded by LASSO is rather volatile, varying between none and all stocks from month to month, with an average of 757.6 per month. This is likely caused by the noise in the tuning process.

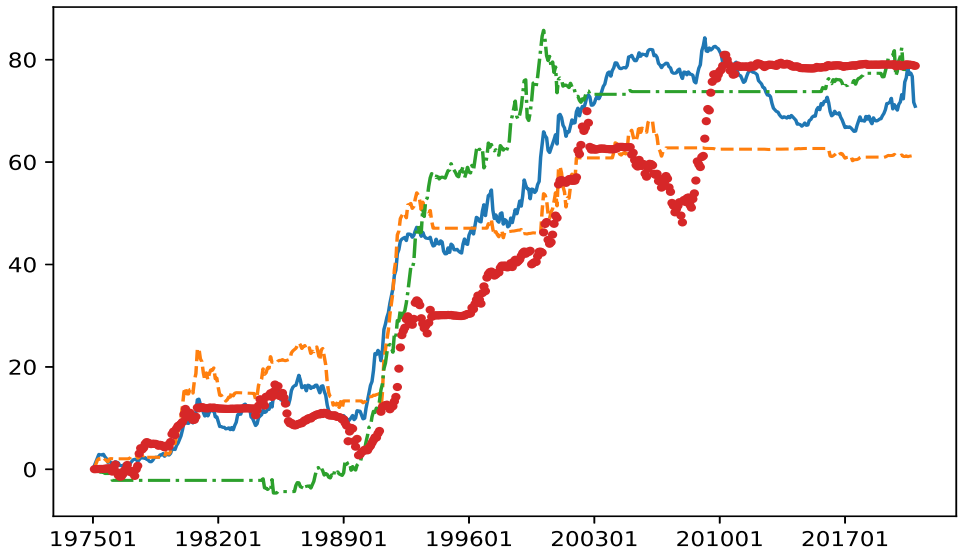


Figure 4.9: Normalized Cumulative Returns of Arbitrage Portfolios

Note: This figure compares the cumulative returns of OPT (red dotted), CSR (blue solid), BH (green dot-dashed), and LASSO (orange dashed) strategies. We normalize all returns by their realized volatilities calculated by the square root of the sum of the squared returns over the entire sample, only for comparison purpose.

We also calculate the perceived Sharpe ratios using (4.2.21), and provide a time-series plot of $\widehat{S}^\star$ in Figure 4.10. We also compare it with the biased estimates $\widetilde{S}^\star$ using (4.2.22). We observe a huge gap between the estimated perceived Sharpe ratios using these formulae. As predicted by Proposition 3, $\widetilde{S}^\star$ overestimates $S^\star$, though it guarantees positive values. Our estimate $\widetilde{S}^\star$ is averaged around 2.55 (we truncate negative estimates by 0), but can sometimes exceed 7.5. These estimates are far greater than the feasible Sharpe ratios we obtain for any of these strategies. That said, even the infeasible Sharpe ratios can be as

low as 0 for certain periods of the sample. The feasible portfolio returns seem in agreement with the prediction. For instance, OPT, LASSO, and BH's cumulative returns are almost flat post 2010, whereas CSR has negative returns.

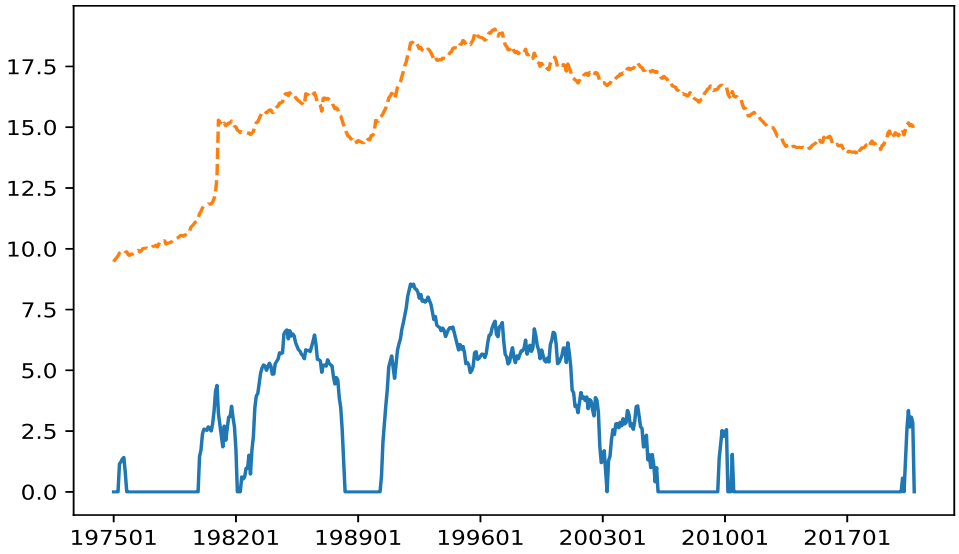


Figure 4.10: Time Series of Sharpe Ratios

Note: The figure compares naive estimates (orange dashed) and their bias-corrected estimates (blue solid) of the infeasible Sharpe ratios based on a rolling window of 120 months.

4.5 Conclusion

Taking stock, our paper provides a new theoretical framework to understanding the implications of statistical learning in asset pricing. In the age of big data, rational expectations assumption often fails to retain its relevance in practice, and hence understanding its limitation and the role of statistical learning is vitally important. We introduce new econometric tools in the spirit of nonparametric empirical Bayes, which could be adopted in other contexts.

The empirical message should be confined within the context of monthly rebalancing strategies via linear factor models. The gap between feasible and infeasible Sharpe ratios will further increase if arbitrageurs face additional statistical challenges, e.g., model misspeci-

fication, omitted factors, weak factors, large non-sparse idiosyncratic covariance matrix, etc. Consequently, the empirical gap should remain for any arbitrageurs, including those who engage in higher frequency trading or use more complex nonlinear models.

More broadly, existing literature have documented impressive Sharpe ratios on various machine learning based trading strategies. Such strategies often rely on ad-hoc model design (e.g., a neural network with a specific architecture) and tuning parameters selection. In this regard, the empirical analysis can at best provide a "lower bound" on the performance of machine learning strategies in investment. Our paper provides a theoretical framework to understand the "upper bound" on the performance of any strategy in the specific context of arbitrage pricing theory, tying together this statistical limit with economic rationale.

On a side note, our theoretical and empirical analyses also have implications on the econometric analysis in asset pricing. Examining the economic performance of asset pricing models is as important as and complementary to statistical tests. The criteria of a good statistical test are primarily statistical in nature, such as Type I and Type II errors, false discovery rate, etc; whereas in practice, it is the economic performance that agents in the economy fundamentally care about. There is often a wedge between these two objectives. For instance, a statistical procedure that guards against false discovery rate may be overly conservative for investment purpose; rejection by a powerful test statistic may not necessarily lead to the practical irrelevance of an economic theory. While the asset pricing literature has seen an explosion of statistical machine learning tools imported from other areas, we caution against their use without guidance of economics.

Chapter 5

Market Efficiency with Many Investors

5.1 Introduction

Tremendous success has been achieved in testing the semi-strong form of the market efficiency hypothesis Fama (1970) by estimating expectations as functions of public information, e.g., prices, firm characteristics, and macro variables. Yet, this common practice does not directly allow for testing the strong version of the market efficiency hypothesis and investigate the important function of financial markets of aggregating private information. Despite the enormous theoretical achievements regarding the role of private information in financial market since seminal Grossman and Stiglitz (1980) and Kyle (1985b), empirically investigating the role of private information is not among the easiest tasks. Indeed, the model of Kyle (1985b) powerfully reveal that, while her private information allows the informed trader to constantly profit, the price process is still a martingale under public information set, which means using data on public information alone can not possibly identify the presence of information asymmetry.

Koijen and Yogo (2019) demonstrates that richer economic implications can be extracted from the joint moments of prices and quantities under equilibrium frameworks. In the context of Kyle (1985b) model, the trading of the informed (noise) trader would positively (negatively) correlate with future price movements, suggesting the potential of joint moments

of prices and quantities in studying the role of private information.

Modern financial markets contain many investors. High-dimensionality of a financial market may have significant implications on both the behaviors of agents within the economy (Martin and Nagel (2022)) and on how econometricians shall analyze the data generated by the economy (Giglio, Kelly, and Xiu (2022)).

In this paper, we investigate, in the presence of many investors, the role of information in affecting investor decision-making, and the efficiency and liquidity of the market, as revealed by the joint moments of prices and investor trading. Building on the symmetric model of Kyle, Obizhaeva, and Wang (2018), we model a continuous-time economy which consists of many (but not a continuum of) investors with heterogeneous information and belief structure, trading a risk-free asset and a risky asset in a centralized market. The risky asset generates a cash flow with unobservable growth rate. Investors receive private signal flows of heterogeneous precision about the growth rate, potentially misperceive the precision of their own signals to different extents, and infer the future price movements utilizing both the private signals and the public information embedded in past price and cash flow. At each moment, investors make optimal consumption-portfolio decisions based on their inference, taking into account the the impact of their actions on prices. However, investors do not observe the actions of each other.

To tractably characterize the equilibrium under heterogeneity, we deviate from the standard equilibrium definition, denoted by "asymptotic equilibrium", by leveraging the presence of many investors. Specifically, in an asymptotic equilibrium, given the price function and others' strategies, an agent is willing to take a suboptimal strategy, as long as the difference in trading under the suboptimal strategy and an exactly optimal one, compared to the magnitude of trading itself, vanishes as the total number of investors grows. In other words, we only require investors to act optimally (under their subjective measures) to the leading order. These leading-order optimal strategies are much simpler than the exactly optimal strategies

in the current context, which then allows for a tractable characterization of an equilibrium. Indeed, one crux of the tractability is that, with many investors, the total precision of public and private information is almost the same across investors with different private information quality and misperception, because the total precision mostly comes from the public part. In this regard, the presence of many investors allows investors in the economy to take simple trading strategies while still acting almost optimally.

In equilibrium, as in Kyle (1985b), investors' trading rates, rather than positions, are proportional to their subjective expectations of future price changes, to avoid incurring large trading cost due to price impact. They also rebalance proportionally to their current position to optimize the risk exposure. An investor's expected price change, on the other hand, is proportional to the difference between her private signal in the recent past and the average of others' private signals, which is reflected through price, multiplied by the misperception-adjusted precision of her own signal. The price dependence of investors' trading in turn determines the equilibrium market liquidity and price impact that every investor takes into account.

The equilibrium has empirical implications in three aspects. In the cross section, because the willingness to trade, i.e., the private signal, is highly idiosyncratic, the trading across investors follows a weak factor model, with the factor, up to a constant scalar, being minus a weighted sum of the idiosyncratic shocks to private signals across investors and capturing the minus pricing error. Unlike standard factor models, the idiosyncratic shocks and the factor negatively correlate, necessary for the aggregate trading to be zero. The loading of an investor's trading on the factor depends on both her trading intensity and the information quality, where the two are not one-to-one mapped because of her misperception. The more intensively the investor trades, the more she contributes to the pricing error, whereas the more informative she is, the more she is able to eliminate the pricing error. In the same direction, the investor's trading positively (negatively) predict future price changes when

she trades relatively conservatively (aggressively) compared to others given her information quality, which is in fact captured by whether her misperception is above or below the average level. Finally, the equilibrium model indicates that the time-series property of an investor's trading is governed by the size of the rebalancing effect and the persistence of predictable price changes.

The empirical implications leads to straightforward empirical strategies. The covariance matrix of trading across investors provide information about the magnitude of investor trading and the distribution of misperception across investors. The regression of future price changes on trading identifies the magnitude of the pricing error and how quickly the predictable price change decays. Using the autocorrelations of investor trading, econometricians can estimate the rebalancing magnitude and compare the trading-implied predictability persistence with the one from the predictive regression.

In the presence of many investors, however, the implications are of high dimension. Directly using the sample covariance matrix of trading and running predictive regression with trading of each institution would incur large estimation errors. Motivated by the factor structure of trading, and given that the factor is weak, our estimation procedures are centered around conducting dimension reduction with a modification of standard principal component analysis (PCA) approach. Specifically, we replace all the diagonal elements of the sample covariance matrix with zeros and conduct eigendecomposition afterwards. When private information is the only trading motive, the eigendecomposition would generate two (one) eigenvectors[1] if there is (is not) misperception heterogeneity, due to the factor-idiosyncratic shock correlation. There could be a few more eigenvectors when there are other motives such as public information driving the trading pattern. Regardless, because the eigenvectors span the pricing error factor, it would be sufficient to run predictive regression with the linear combinations of all the investors' trading using those small number of eigenvectors

1. Precisely, eigenvectors with nonzero eigenvalues.

and achieve dimension reduction for this part as well. We further provide the procedure and conditions to identify misperception.

The theory part of our paper extends the symmetric model of Kyle, Obizhaeva, and Wang (2018) to allow for flexible heterogeneity in information quality and investor misperception in a dynamic economy, and contribute to the broad literature of asset pricing models with information asymmetry: Kyle (1989), Wang (1993), Wang (1994), He and Wang (1995), Vayanos (1999), Vives (2011), and Du and Zhu (2017). The empirically relevant many-investor setup indicates that investors take simple equilibrium strategies. It also indicates that, although in equilibrium there could be sizable pricing errors, each investor can at best profit out of a tiny portion of it. These results highlights the impact of high-dimensionality on agent behaviors and in this regard the paper is also related to Martin and Nagel (2022).

The econometric procedures in the paper build on and contribute to the evolving literature on the applications of statistical and machine learning in asset pricing, in particular on the topic of factor models and PCA approaches, e.g., Kelly and Pruitt (2013), Kozak, Nagel, and Santosh (2018b), Kelly, Pruitt, and Su (2019b), Kozak, Nagel, and Santosh (2020), Giglio and Xiu (2021b), and Giglio, Xiu, and Zhang (2021). Complementary to this literature, the factors in our paper are weak yet pervasive and we propose a modification of the standard PCA accordingly. On the other hand, the way we conduct predictive regression shares the same spirit of Da, Nagel, and Xiu (2022), which demonstrates that weak predictors, if efficiently combined, can collectively generate significant predictive power. In our paper, however, the relative predictive power of each institution's trading is revealed by the trading correlations across institutions, which directly guides how we combine the predictors.

Moreover, our paper relates to the enormous literature on mutual funds, e.g., Berk and Green (2004), Fama and French (2010), Pástor and Stambaugh (2012), Kacperczyk, Van Nieuwerburgh, and Veldkamp (2014, 2016), and Song (2020).[2] We demonstrate that

2. See also Wermers (2000), Berk and Van Binsbergen (2015), Pástor, Stambaugh, and Taylor (2015), Pástor, Stambaugh, and Taylor (2017), and Pástor, Stambaugh, and Taylor (2020).

how much predictive power econometricians can extract from investors' positions hinges on the dispersion of their misperception and that investors' portfolio returns and their contributions to market efficiency are not one-to-one mapped. For instance, in the fully symmetric case, even though investors are injecting their private information into the price, their trading has zero correlation with price movements and the alphas of their portfolios are all zero. We further provide estimates of the contributions to price informativeness by each type of institutions.

Finally, our paper is also connected to the literature that empirically estimates price elasticities of demand in financial market under various scenarios that are orders of magnitude smaller than what the standard models would imply,[3] including Harris and Gurel (1986), Shleifer (1986), Chang, Hong, and Liskovich (2015), Koijen and Yogo (2019), and Gabaix and Koijen (2021).

Our paper proceeds as follows. Sections 5.2 – 5.4 characterize the equilibrium model with many investors. Specifically, Section 5.2 sets up the economy, Section 5.3 characterizes the equilibrium under symmetry , and Section 5.4 generalizes the characterization to the scenario with heterogeneous information quality and misperception. Section 5.5 presents the empirical implications of the equilibrium model and develops econometric procedures to combine the implications with price and quantity data. Section 5.6 concludes. The appendix provides technical details.

5.2 Setup of the Economy

5.2.1 Assets

There is a risky financial asset in the economy with zero total supply and its price at time t is denoted by $P_{i,t}$. In addition there exists a zero-return risk-free asset with risk-free rate r.

3. See, e.g., Petajisto (2009).

The risky asset generates cash flow at rate $D_t + D_t^o$. It evolves according to

$$dD_t = -\xi^D D_t dt + G_t dt + \sigma^D dZ_t^D, \quad \text{and} \quad dD_t^o = -\xi^D D_t^o dt + G_t^o dt + \sigma^D dZ_t^{D,o} \tag{5.2.1}$$

Here ξ^D is the mean-reversion parameter, σ^D is the volatility parameter, and Z_t^D and $Z_t^{D,o}$ are mutually independent standard Brownian motions. The change rates of D_t and D_t^o are also affected by the dividend growth rates G_t and G_t^o, which both follow mean-reversion processes:

$$dG_t = -\xi^G G_t dt + \sigma^G dZ_t^G \quad \text{and} \quad dG_t^o = -\xi^o G_t^o dt + \sigma^o dZ_t^o, \tag{5.2.2}$$

where Z_t^G and Z_t^o are two other mutually independent standard Brownian motions, which do not depend on Z_t^D and $Z_t^{D,o}$ as well.

5.2.2 Preference

We index the many investors participating the economy by $j \in \mathcal{J}$. Investor j trades both risk-free and risky assets and we use $x_{j,t}^f$ to stand for the value of her time-t holding of risk-free asset, and denote by $x_{j,t}$ her time-t holding of the risky asset. Their time derivatives are denoted by by $\dot{x}_{j,t}^f$ and $\dot{x}_{j,t}$. The objective of investor j is

$$\max_{\{\dot{x}_{j,s}, c_{j,s}\}_{s \geq t}} \mathrm{E}_{j,t} \int_t^\infty e^{-\rho(s-t)} u(c_{j,s}) ds, \quad \text{with} \quad u(c_{j,s}) = -e^{-\gamma c_{j,s}}, \tag{5.2.3}$$

where the change of the risk-free asset value satisfies

$$\dot{x}_{j,t}^f = r x_{j,t}^f - c_{j,t} + x_{j,t}\big(D_t + D_t^o\big) - \dot{x}_{j,t} P_t \tag{5.2.4}$$

We use $\mathrm{E}_{j,t}(\cdot)$ to represent expectation under investor j' subjective measure, conditional on her time-t information set, which we will further discuss shortly. The trading rate $\dot{x}_{j,t}^f$ and $\dot{x}_{j,t}$ must also be measurable to the time-t information set. Market-clearing price P_t can depend on the investor's current and past trading, and the investor takes her price impact

into account when making decisions.

5.2.3 Information and Belief

For the risky asset, we assume investors directly observe its price and cash flow components D_t and D_t^o. The cash-flow growth rate component G_t^o is also observable to each investor. However, investors misperceive how G_t^o drives cash-flow growth. Concretely, under investor j's subjective measure, the cash-flow rate follows

$$dD_t^o = -\xi^D D_t^o dt + \big(G_t + (1+\eta_j)G_t^o\big)dt + \sigma^D d\widehat{Z}_{j,t}^{D,o}, \tag{5.2.5}$$

where investor j considers $\widehat{Z}_{j,t}^{D,o} := Z_t^{D,o} - (\sigma^D)^{-1}\eta_j \int^t G_s^o ds$ a standard Brownian motion. On the other hand, investors do not observe the growth rate component G_t. Investor j instead receives a noisy signal flow $S_{j,t}$ about G_t, which satisfies

$$dS_{j,t} = G_t dt + \sigma_j^S dZ_{j,t}^S. \tag{5.2.6}$$

Here $Z_{j,t}^S$ is a standard Brownian motion independent of $(Z_t^D, Z_t^G$, and is also independent across assets. The volatility parameter σ_j^S controls the signal-to-noise ratio of $S_{j,t}$. The investor, however, perceive the drift of her signal as $\omega_j G_t$ in constructing her subjective expectation $\mathrm{E}_{j,t}(\cdot)$:

$$dS_{j,t} = \omega_j G_t dt + \sigma_j^S d\widehat{Z}_{j,t}^S.$$

where $\widehat{Z}_{j,t}^S := Z_{j,t}^S + (\sigma_j^S)^{-1}(1-\omega_j)\int^t G_s ds$ is considered by investor j as a standard Brownian motion. In other words, she has an incorrect understanding of the informativeness of her signals. This matches the econometric fact that it is hard to measure drift precisely without long time span, whereas volatility is perfectly measurable in continuous time. We assume that investors correctly perceive the dynamics of both asset fundamentals and others' private signals. Investors do not observe each other's private signals or trading actions. Therefore,

investor j's subjective expectation $\mathrm{E}_{j,t}(\cdot)$ is only measurable to $\mathcal{F}_{j,t}$, the information set generated by $\{S_{j,s}, P_s, D_s, D_s^o, G_s^o\}_{s\leq t}$, i.e., her private signals, the price, the cash flow, and the observable growth rate component, up to time t.

5.3 Equilibrium with Symmetric Information Structure

In this section, we conduct equilibrium analysis under the symmetric information structure.

5.3.1 Setup

For results in this section, we impose the following assumption:

Assumption 13. *The cash flow component D_t^o and the observable growth rate G_t^o are both zero. Moreover, signal noise parameter σ_j^S and the misperception parameter ω_j stay invariant across all investor $j \in \mathcal{J}$:*

$$\sigma_j^S = \sigma^S, \quad \textit{and} \quad \omega_j = \omega, \quad \forall j \in \mathcal{J}. \tag{5.3.7}$$

Under Assumption 13, our setup matches that of Kyle, Obizhaeva, and Wang (2018), where an equilibrium with linear flow-strategies is elegantly characterized. We conduct an asymptotic exercise to understand the properties of the equilibrium of Kyle, Obizhaeva, and Wang (2018) as the total number of investors gets large. This matches our empirical goal as, for instance, in stock markets typical stocks are being traded by a large number of investors. Moreover, in this case we are able to obtain provide closed-form expressions of all the endogenous parameters, whereas in general numerical tools to required to solve key endogenous parameters. Specifically, we consider the following drifting sequence of exogenous parameters:

Assumption 14. *As J, i.e., the size of $\mathcal{J}$, increases,*

$$\left(\sigma^S\right)^2 = J\iota,$$

whereas ι and all the other exogenous parameters $(\xi^D, \xi^G, \sigma^D, \sigma^G, \omega, \rho, \gamma)$ stay unchanged.

The assumption is to prevent total precision of all investors' signals from exploding, which would lead to that asset price converges to the exact fundamental value and that price impact vanishes. It is notable that we keep γ constant. Arguably, absolute risk aversion changes with level of wealth. That the market consists of a lot of investors is connected with that each investor is relatively small and potentially has large absolute risk aversion. Fortunately, under the current setup, equilibria with different values of γ are isomorphic, and characterizing an equilibrium with any value of γ would be sufficient.[4]

Given the information structure, investors form their expected returns, which could depend on both their private signals and price history. On the other hand, however, price is endogenous and is affected by how investors trade based on the subjective expected returns. The equilibrium concept is formally defined as:

DEFINITION 3: An equilibrium is a set of investor trading strategies $\{\dot{x}^f_{j,t}(\cdot), \dot{x}_{j,t}(\cdot)\}_{j\in\mathcal{J}}$ and price function $P_t\big(\{\dot{x}_{j,s}(\cdot), D_s, S_{j,s}, P_{s'}\}_{s\leq t, s'<t, j\in\mathcal{J}}\big)$ such that

(i) For each investor $j \in \mathcal{J}$, given the price functions and the strategies of all the other investors, the trading strategy $\big(\dot{x}^f_{j,t}(\cdot), \dot{x}_{j,t}(\cdot)\big)$ solves the optimization problem specified by (5.2.3) and (5.2.4), subject to that $\dot{x}^f_{j,t}(\cdot)$ and $\dot{x}_{j,t}(\cdot)$ are measurable to the information set $\mathcal{F}_{j,t}$;

(ii) The risky asset market clears:

$$\sum_{j\in\mathcal{J}} x_{j,t} = 0.$$

Notably, in Definition 3 trading strategies can depend on contemporaneous prices. In other words, investors can trade in the form of submitting demand schedules, which allows the market to clear in the absence of market makers.

4. If we double the risk aversion, then all the price variables would not change and all the quantity variable would be halved. See Theorem 4 of Kyle, Obizhaeva, and Wang (2018).

5.3.2 Equilibrium

We conjecture that in equilibrium all investors submit symmetric linear demand schedules as follows:

$$\dot{x}_{j,t} = \psi^D G_t^D + \psi^C D_t + \psi^S G_{j,t} - \psi^P P_t - \psi^H x_{j,t}, \tag{5.3.8}$$

where processes G_t^D and $G_{j,t}$ are constructed as follows:

$$G_t^D = (\xi^P - \xi^G) \int_{-\infty}^{t} e^{-\xi^P (t-s)} (dD_s - \xi^G D_s ds), \tag{5.3.9}$$

$$G_{j,t} = (\xi^P - \xi^G) \int_{-\infty}^{t} e^{-\xi^P (t-s)} dS_{j,s}, \tag{5.3.10}$$

with the parameter ξ^P defined by

$$\xi^P = \sqrt{(\xi^G)^2 + (\sigma^G)^2 \big((\sigma^D)^{-2} + \iota^{-1}\big)}. \tag{5.3.11}$$

In other words, G_t^D and $G_{j,t}$ are constructed using cash flow and private signals up to t, which both contain information about the growth rate G_t. The reason behind the specific way of construction will be explained shortly.

Market clearing forces both aggregate position and aggregate trading to be zero, which leads to

$$\dot{x}_{j,t} + \psi^H x_{j,t} = - \sum_{j \in \mathcal{J}: j' \neq j} (\dot{x}_{j',t} + \psi^H x_{j',t}). \tag{5.3.12}$$

Then, suppose all investors $j' \in \mathcal{J}$ with $j' \neq j$ submit demand schedule (5.3.8), the supply curve faced by investor j, which is a function of investor trading, would be

$$P_t\big(\dot{x}_{j,t}\big) = \big(\psi^P\big)^{-1} \bigg(\psi^D G_t^D + \psi^C D_t + \psi^S \frac{1}{J-1} \sum_{j \in \mathcal{J}: j' \neq j} G_{j',t} + \frac{1}{J-1} \big(\dot{x}_{j,t} + \psi^H x_{j,t}\big) \bigg). \tag{5.3.13}$$

Investor j's problem is to solve her optimization problem specified by (5.2.3) and (5.2.4)

under (5.3.13). The equilibrium would be established if her optimal strategy also satisfy the conjecture (5.3.8).

The next proposition presents the implications of large J on the equilibrium characterized by Kyle, Obizhaeva, and Wang (2018):

Proposition 4. *Suppose Assumptions 13 and 14 hold. Then there exists an equilibrium as in Definition 3 if J is sufficiently large, $\xi^G > r$, and $\omega > 2$.*[5] *The equilibrium has the following properties:*

(i) In equilibrium the price satisfies

$$P_t = \frac{D_t}{\xi^D + r} + \frac{\mathrm{E}(G_t|\{D_s, \bar{S}_s\}_{s\le t})}{(\xi^D + r)(\xi^G + r)} + O_{\mathrm{P}}(J^{-1}), \quad \textit{with} \quad \bar{S}_s = \frac{1}{J}\sum_{j\in\mathcal{J}} S_{j,s}, \tag{5.3.14}$$

where $\mathrm{E}(\cdot)$ is the expectation under the objective measure. Moreover, it holds that

$$\mathrm{E}(G_t|\{D_s, \bar{S}_s\}_{s\le t}) = \frac{(\sigma^D)^{-2} G_t^D + \iota^{-1}\bar{G}_t}{(\sigma^D)^{-2} + \iota^{-1}}, \quad \textit{with} \quad \bar{G}_t := \frac{1}{J}\sum_{j\in\mathcal{J}} G_{j,t}, \tag{5.3.15}$$

(ii) The optimal trading strategy follows (5.3.8), with $\left((J-1)\psi^P\right)^{-1} = \zeta + O_{\mathrm{P}}(J^{-1})$ and $\psi^H = b + O_{\mathrm{P}}(J^{-1})$, and the trading satisfies

$$\dot{x}_{j,t} = a\frac{\xi^P + r}{(\xi^D + r)(\xi^G + r)}\frac{\omega(\sigma^S)^{-2}}{(\sigma^D)^{-2} + \iota^{-1}}(G_{j,t} - \bar{G}_t) - b x_{j,t} + O_{\mathrm{P}}(J^{-3/2}). \tag{5.3.16}$$

Here the endogenous parameters a, b, and ζ are given by

$$a = \frac{1}{\zeta\omega(\xi^P + r)}, \quad b = \frac{1}{2}(\omega - 2)(\xi^P + r), \quad \zeta = \frac{1}{2}\frac{r\gamma(\sigma^P)^2}{\left(b + \frac{1}{4}r\right)^2 - \frac{1}{16}r^2}, \tag{5.3.17}$$

5. This shows that the existence condition conjectured by Kyle, Obizhaeva, and Wang (2018), at least when the market is large, is correct.

where $\sigma^P = (\xi^D + r)^{-1}\sqrt{\left(\sigma^D\right)^2 + \left(\sigma^G\right)^2(\xi^G + r)^{-2}}$.

As revealed by property (i), when the market becomes large, the gap between the equilibrium price and what best reflects the present value of future cash flows converges to zero. In general, as demonstrated by Kyle, Obizhaeva, and Wang (2018), the equilibrium price is

$$P_t = \frac{D_t}{\xi^D + r} + \frac{\phi}{(\xi^D + r)(\xi^G + r)}\frac{1}{J}\sum_{j\in\mathcal{J}}\widetilde{\mathrm{E}}_{j,t}(G_t),$$

where ϕ is a endogenous parameter conjectured to be smaller than one, and $\widetilde{\mathrm{E}}_{j,t}(\cdot)$ is the investor j's expectation if she hypothetically observes everyone's private signal up to t. Therefore, the gap comes from two sources: the "price-dampening" parameter ϕ and that $\widetilde{\mathrm{E}}_{j,t}\big(G_t\big)$ differs from $\mathrm{E}(G_t|\{D_s, \bar{S}_s\}_{s\leq t})$. The latter one naturally shrinks, because the difference between the two expectations only comes from investor j's misperception of her own signal, which plays a vanishing role in forming the expectations as J increases. The price dampening effect originates from that investor j understand (correctly) that other investors misperceive the precision of their signals in forming their growth rate estimates $\widetilde{\mathrm{E}}_{j,t}\big(G_t\big)$. If $\phi = 1$, the average expected return across investors would be negative (positive) when G_t is positive (negative). So would the aggregate demand. As a result, $\phi < 1$ is needed to clear the market. Hence, when the market is large, ϕ is pushed towards one as $\widetilde{\mathrm{E}}_{j,t}\big(G_t\big)$ converges to $\mathrm{E}(G_t|\{D_s, \bar{S}_s\}_{s\leq t})$.

Equation (5.3.15) is a direct result of standard Kalman-Bucy filtering. $G_{j,t}$ is constructed using investor j's own signal, whereas G_t^D is intended to capture the information about the growth rate contained in the cash flow. The expectation of G_t is a weighted average of them, with the weights determined by their relative informativeness. The parameter ξ^P determines the relative weights on signals from recent and distant past. Indeed, when the information on G_t contained in the cash flow or the signal is of high quality, ξ^P increases and $G_{j,t}$ and G_t^D is mostly composed of very recent signals.

The equilibrium trading is quite interpretable as well. The gap between average growth rate estimate $\bar{G}_t$ to the true growth rate G_t mean-reverts to zero at rate ξ^P.[6] With G_t unobservable, investor j measure the gap using $G_{j,t} - \bar{G}_t$ and shrink it by the factor $\omega(\sigma^S)^{-2}/\big((\sigma^D)^{-2} + \iota^{-1}\big)$, as she is aware of the noise of her own signal. The factor $\big(\xi^D + r\big)^{-1}\big(\xi^G + r\big)^{-1}$ reflects how price is connected to the average growth rate estimate. Finally, the investor chooses parameters a and b to balance capturing expected returns and alleviating trading costs, with the price impact $\big((J-1)\psi^P\big)^{-1} = \zeta + O_{\mathrm{P}}(J^{-1})$ being endogenously determined. The equilibrium trading resembles the one obtained by Gârleanu and Pedersen (2013, 2016) in partial equilibrium models with exogenous trading cost function, where investors directly derive utility from after-cost investment performance rather than from consumption.

5.4 Equilibrium with Heterogeneous Information Structure

This section is devoted to the equilibrium analysis under general heterogeneous information structure.

5.4.1 Setup

Motivated by the simplification of the equilibrium under large market demonstrated by Section 5.3.2, we restrict our analysis to the large market scenario, in order to generate a tractable characterization of the equilibrium in the presence of heterogeneity in belief and information structure. To regulate the asymptotic behaviors of various parameters, we impose the following assumption, which accommodates Assumption 14 as a special case.

Assumption 15. *As J, i.e., the size of $\mathcal{J}$, increases,*

$$\big(\sigma_j^S\big)^2 = J\iota_j,$$

6. Precisely speaking, when investor j makes her trading decisions, what matters should be the average growth rate estimate among all the *other* investors, but the difference is of second order when the market is large.

whereas ι_j *and exogenous parameters* $(\xi^D, \xi^G, \sigma^D, \sigma^G, \omega_j, \rho, \gamma)$ *stay unchanged. The belief parameters* η_j *satisfies*

$$\eta_j = \bar{\eta} + \kappa_j, \quad \textit{with} \quad \sum_{j \in \mathcal{J}} \kappa_j = 0.$$

where $\bar{\eta}$ *can either stay constant or vary with* J*, and satisfies* $\bar{\eta} = O(1)$*. On the other hand,* κ_j *satisfies* $\max_{j \in \mathcal{J}} |\kappa_j| = O(J^{-1/2})$.

In contrast to Assumption 14, here we explicitly allow for incorrect beliefs on G^o_t represented by η_j. The requirement that belief dispersion must not be larger than $\sim J^{-1/2}$ might sound restrictive, but it actually already allows the belief dispersion to generate almost arbitrarily strong correlations between investors' trading.

The main message from Proposition **??** is that, even though investors solve their optimization problems exactly, as the market becomes large, the equilibrium price and investor trading are dominated by leading-order terms of simple forms. Now we take one step further and only require the investors to approximately solve their optimization problems, that is, intuitively, as long as a strategy leads to trading that is sufficiently close to trading under an exactly optimal one, they are willing to take it. To formalize this idea, we update the equilibrium concept as follows:

DEFINITION 4: An asymptotic equilibrium is a set of trading strategies $\{\dot{x}^f_{j,t}(\cdot), \dot{x}_{j,t}(\cdot)\}_{j \in \mathcal{J}}$ and price functions $P_t\big(\{\dot{x}_{j,s}(\cdot), D_s, D^o_s, G^o_s, S_{j,s}, P_{s'}\}_{s \le t, s' < t, j \in \mathcal{J}}\big)$ such that

(i) For each investor $j \in \mathcal{J}$, given the price function and the strategies of all the other investors, $\big(\dot{x}^f_{j,t}(\cdot), \dot{x}_{j,t}(\cdot)\big)$ are measurable to the information set $\mathcal{F}_{j,t}$ and satisfy

$$\mathrm{E}_j\big(|\dot{x}_{j,t} - \dot{x}^*_{j,t}|^2\big) \lesssim J^{-2} \quad \text{and} \quad \mathrm{E}_j\big(|x_{j,t} - x^*_{j,t}|^2\big) \lesssim J^{-2},$$

for some strategy $\big(\dot{x}^{f*}_{j,t}(\cdot), \dot{x}^*_{j,t}(\cdot)\big)$ that solves the optimization problem specified by (5.2.3) – (5.2.4), subject to that $\dot{x}^{f*}_{j,t}(\cdot)$ and $\dot{x}^*_{j,t}(\cdot)$ are measurable to $\mathcal{F}_{j,t}$;

(ii) The risky asset market clears:

$$\sum_{j\in\mathcal{J}} x_{j,t} = 0.$$

We note that in Proposition 4 the magnitude of the leading terms for both trading rate $\dot{x}^*_{j,t}$ and position $x^*_{j,t}$ are $\approx J^{-1/2}$, which turns out to be also true in the asymptotic equilibrium studied in this section. The small magnitude comes from that investors' signals are highly noisy, under Assumptions 14 or 15. These leading terms, however, dominate the deviations from exactly optimal trading and position allowed by requirement (i) in Definition 4.

5.4.2 Equilibrium

As in Section 5.3.2, we conjecture that in equilibrium all investors submit linear demand schedules

$$\dot{x}_{j,t} = \psi^D_j G^D_t + \psi^C_j \big(D_t + D^o_t\big) + \psi^S_j G_{j,t} + \psi^o_j G^o_t - \psi^P_j P_t - \psi^H x_{j,t}. \tag{5.4.18}$$

All coefficients but ψ_H are j-dependent. The processes G^D_t and $G_{j,t}$ are also defined by (5.3.9) and (5.3.10), where ξ^P is now defined by

$$\xi^P = \sqrt{(\xi^G)^2 + (\sigma^G)^2\big((\sigma^D)^{-2} + \widetilde{\iota}^{-1}\big)}, \tag{5.4.19}$$

with

$$\widetilde{\iota} = \sum_{j\in\mathcal{J}} \pi_j^2 (\sigma^S_j)^2 \quad \text{and} \quad \pi_j = \omega_j (\sigma^S_j)^{-2} \Big/ \sum_{j'\in\mathcal{J}} \omega_{j'} (\sigma^S_{j'})^{-2}.$$

Under symmetric information structure (Assumption 14), weight π_j would reduce to equal weight, $\widetilde{\iota}$ would be equal to ι, and ξ^P would be the same as the one in (5.3.11). We will explain why the weight π_j appears later.

Suppose all investors $j' \in \mathcal{J}$ with $j' \neq j$ submit demand schedule (5.4.18). Similar to the symmetric case, the market clearing condition would lead to that the supply curve $P_t\big(\dot{x}_{j,t}\big)$

investor j faces is:

$$\left(\widetilde{\psi}_j^P\right)^{-1}\left(\widetilde{\psi}_j^D G_t^D + \widetilde{\psi}_j^C\left(D_t + D_t^o\right) + \widetilde{\psi}_j^o G_t^o + \frac{1}{J-1}\sum_{j\in\mathcal{J}:j'\neq j}\psi_{j'}^S G_{j',t} + \frac{1}{J-1}\left(\dot{x}_{j,t} + \psi^H x_{j,t}\right)\right), \tag{5.4.20}$$

where $\left(\widetilde{\psi}_j^P, \widetilde{\psi}_j^D, \widetilde{\psi}_j^C, \widetilde{\psi}_j^o\right) = (J-1)^{-1}\sum_{j\in\mathcal{J}:j'\neq j}\left(\psi_{j'}^P, \psi_{j'}^D, \psi_{j'}^C, \psi_{j'}^o\right)$. According to Definition 4, an asymptotic equilibrium exists if we find a set of parameters $\left\{\psi_j^P, \psi_j^D, \psi_j^C, \psi_j^o, \psi_j^P, \psi^H\right\}_{j\in\mathcal{J}}$ such that, for each investor $j \in \mathcal{J}$, (5.4.18) "almost" (as in requirement (i) of Definition 4) solves the optimization problem (5.2.3) and (5.2.4), under the supply curve (5.4.20). The following theorem presents the result.

Theorem 11. *Suppose Assumption 15 holds. Then there exists an asymptotic equilibrium as in Definition 4 if* $\xi^G > r$ *and* $\widetilde{\omega} > 2$, *where* $\widetilde{\omega} = \sum_{j\in\mathcal{J}}\pi_j\omega_j$. *The equilibrium has the following properties:*

(i) In equilibrium the price satisfies

$$P_t = \frac{D_t + D_t^o}{\xi^D + r} + \frac{\mathrm{E}(G_t|\{D_s, \bar{S}_s\}_{s\leq t})}{(\xi^D + r)(\xi^G + r)} + \frac{(1+\bar{\eta})G_t^o}{(\xi^D + r)(\xi^o + r)}, \quad \textit{with} \quad \bar{S}_s = \sum_{j\in\mathcal{J}}\pi_j S_{j,s}, \tag{5.4.21}$$

and $\mathrm{E}(\cdot)$ *is the expectation under the objective measure. Moreover, it holds that*

$$\mathrm{E}(G_t|\{D_s, \bar{S}_s\}_{s\leq t}) = \frac{(\sigma^D)^{-2}G_t^D + \widetilde{\iota}^{-1}\bar{G}_t}{(\sigma^D)^{-2} + \widetilde{\iota}^{-1}}, \quad \textit{with} \quad \bar{G}_t = \sum_{j\in\mathcal{J}}\pi_j G_{j,t}. \tag{5.4.22}$$

(ii) Investor j's trading strategy follows (5.4.18) with $\psi_j^P = \zeta^{-1}\pi_j$ *and* $\psi_j^H = b$, *and the trading satisfies*

$$\dot{x}_{j,t} = a\frac{\xi^P + r}{(\xi^D + r)(\xi^G + r)}\frac{\omega_j(\sigma_j^S)^{-2}}{(\sigma^D)^{-2} + \widetilde{\iota}^{-1}}(G_{j,t} - \bar{G}_t) + a^o\frac{\kappa_j}{\xi^D + r}G_t^o - bx_{j,t}. \tag{5.4.23}$$

Here the endogenous parameters a, a^o, b, and ζ are given by

$$a = \frac{1}{\zeta\widetilde{\omega}(\xi^P + r)}, \quad a^o = \frac{1}{2\zeta}\frac{1}{b + \xi^o + r}, \quad b = \frac{\widetilde{\omega} - 2}{2}(\xi^P + r), \quad \zeta = \frac{1}{2}\frac{r\gamma(\sigma^P)^2}{\left(b + \frac{1}{4}r\right)^2 - \frac{1}{16}r^2}, \tag{5.4.24}$$

where $\sigma^P = (\xi^D + r)^{-1}\sqrt{2(\sigma^D)^2 + (\sigma^G)^2\left(\xi^G + r\right)^{-2} + (1 + \bar{\eta})^2(\sigma^o)^2\left(\xi^o + r\right)^{-2}}$ *is the volatility of the equilibrium price.*

The equilibrium properties resembles those in Proposition 4. The equilibrium price (5.4.21) takes the same form of its counterpart (5.3.14), where $\bar{S}_{j,t}$ is an average of $S_{j,t}$ weighted by π_j. Parameter $\widetilde{\iota}$ appearing in (5.4.22) and (5.4.19) is the squared volatility of $\bar{S}_{j,t}$, which reflects its noise level. The homogeneous misperception parameter ω that affects trading and price impact in (5.3.17) is replaced with $\widetilde{\omega}$ as in (5.4.24), which is the average of ω_j weighted by π_j as well. The weight π_j reflects that investors with higher signal precision and larger upward bias in perceiving it would play more important roles in determining equilibrium price and market liquidity. Its specific form originates from the $\omega_j(\sigma_j^S)^{-2}$ factor appearing in the trading (5.4.23). Comparing (5.4.23) with the equilibrium price (5.4.21) and (5.4.22), we clearly see that the trading can be generated using a linear demand schedule of form (5.4.18).

5.5 Econometric Analysis of the Equilibrium Model

This section studies the identification and estimation of the following parameters: price mean-reversion coefficient ξ^P, price impact parameter ζ, misperception parameters $\{\omega_j\}_{j\in\mathcal{J}}$, and information quality parameters $\{\sigma_j^S\}_{j\in\mathcal{J}}$. As we can not rule out other types of equilibria, we impose the following assumption:

Assumption 16. *Equilibrium price and investor trading are the ones characterized in Theorem 11.*

5.5.1 Empirical Content of the Equilibrium Model

The equilibrium has two major implications: how trading depends on private and public information (5.4.23), and the relation of prices to private information (5.4.21). They connect the parameters of interest to moments of holdings and prices. The equilibrium implication on trading can be written more compactly as

$$\dot{x}_{j,t} + bx_{j,t} = y_{j,t} := \phi^S \pi_j (G_{j,t} - \bar{G}_t) + \phi^o \kappa_j G^o_t,$$

where the definitions of constants ϕ^S and ϕ^o is clear from (5.4.23). Moreover, we introduce notation

$$\varepsilon_{j,t} = \phi^S (\xi^P - \xi^G) \int_{-\infty}^{t} e^{-\xi^P (t-s)} \sigma^S_j dZ^S_{j,s}, \quad f_t = -\sum_{j \in \mathcal{J}} \pi_j \varepsilon_{j,t}, \quad \text{and} \quad g_t = \phi^o G^o_t.$$

Then the definition of y_t can be further simplified into:

$$y_{j,t} = \pi_j f_t + \kappa_j g_t + \pi_j \varepsilon_{j,t}.$$

In other words, trading across investors follows a simple factor structure. Investor trades on f_t, which is the aggregation of the noise in each investor's private signal entering the price. Because observe f_t is not observable to investors, to load more on the factor f_t, an investor will have to load more on the noise of her own signal as well, which more intensively moves the price against herself. The factor g_t originates from the observable growth rate component G^o_t, on which there is a belief dispersion. The next proposition provides statistical moments regarding the above factor model.

Proposition 5. *Suppose Assumptions 15 and 16 hold. Then we have, with some constant*

ϕ that only depends on (ϕ^S, ξ^P, ξ^G) and some constant $\bar{\phi}$ that only depends on $(\phi^o, \xi^o, \sigma^o)$,

$$\mathrm{E}(f_t^2) = \phi\widetilde{\iota}, \quad \mathrm{E}(\varepsilon_{j,t}^2) = \phi(\sigma_j^S)^2, \quad \mathrm{E}(f_t\varepsilon_{j,t}) = -\phi\pi_j(\sigma_j^S)^2, \quad \mathrm{E}(g_t^2) = \bar{\phi}.$$

The covariance matrix of rebalancing-adjusted trading rate across investors is

$$\mathrm{Cov}\big(y_t\big) = \phi\widetilde{\iota} \cdot \Big(\beta\beta^\intercal - \nu\nu^\intercal + \mathrm{diag}\big(\nu\big)\Big) + \bar{\phi} \cdot \kappa\kappa^\intercal. \tag{5.5.25}$$

Here $\nu = (\nu_1, \nu_2, ..., \nu_J)^\intercal$ with $\nu_j = \pi_j\widetilde{\omega}^{-1}\omega_j$ and $\beta = \pi - \nu$, whereas y_t, π, and κ are all J-dimensional column vectors whose entries are clear from the context. The average misperception $\widetilde{\omega}$ is defined in Theorem 11.

Because weight π_j is of order $\sim J^{-1}$, and the standard deviation of idiosyncratic shock $\pi_j\varepsilon_{j,t}$ is of order $\sim J^{-1/2}$, factor f is a weak one. As mentioned earlier, even though Assumption 15 impose a bound on the size of κ_j, it is not highly restrictive. Given the magnitude of the idiosyncratic shocks, Assumption 15 allows g_t to be a standard strong factor and the dominant driver of trading correlation patterns. On the other hand, the factors f_t and g_t in fact carry all the predictive power of trading on future price changes. To be concrete, we introduce $\Pi_t = P_t + \int^t (D_s + D_s^o - rP_s)ds$, the excess gain process of holding one unit of the risky asset.

Proposition 6. *Suppose Assumptions 15 and 16 hold and denote by $\widetilde{\mathcal{F}}_t$ the information set generated by $\big\{x_{j,s}, f_s, g_s\big\}_{s\le t, j\in\mathcal{J}}$. Then it holds that, for all $\tau \ge 0$,*

$$\mathrm{E}\big(d\Pi_{t+\tau}\big|\widetilde{\mathcal{F}}_t\big)/dt = e^{-\xi^P\tau}\big(a\widetilde{\omega}\big)^{-1} \cdot f_t - e^{-\xi^o\tau}(a^o)^{-1} \cdot \bar{\eta}g_t,$$

where a is introduced in (5.4.24).

Notably, even though g_t could dominate f_t in generating the cross-sectional comovement of trading, it possesses similar or smaller predictive power compared to f_t, depending on

the size of average belief distortion $\bar{\eta}$. The reason is that investors do not observe f_t, which appears in the conditioning information set $\widetilde{\mathcal{F}}_t$, and they are unwilling to expeditiously trade on it because of the low quality of their private signals. Econometricians, on the other hand, can efficiently aggregate their private signals implied by their trading and obtain a much more precise estimate of f_t, as long as the misperception parameter ω_j differs across investors. Indeed, as a result of Proposition 6, whenever $\tau \geq 0$,

$$\mathrm{E}\big(y_t d\Pi_{t+\tau}\big)/dt = e^{-\xi^P \tau}\big(a\widetilde{\omega}\big)^{-1} \cdot \phi\widetilde{\imath}\beta - e^{-\xi^o \tau}(a^o)^{-1} \cdot \bar{\phi}\bar{\eta}\kappa. \tag{5.5.26}$$

If there is no heterogeneity in ω_j, then $\beta = 0$ and every investor's trading has zero correlation with future price movements, making it impossible for econometricians to extract any predictive power.

However, in reality econometricians do not observe y_t. They only observe investor positions at discrete time $\Delta x_t = x_t - x_{t-1}$. The following proposition connects their statistical moments.

Proposition 7. *Suppose Assumptions 15 and 16 hold. Suppose $\xi^P = \xi^o$. Then we have, for all $\tau \geq 1$,*

$$\Sigma := \mathrm{Cov}(\Delta x_t) = \lambda \mathrm{Cov}(y_t), \quad R_\tau := \mathrm{E}(\Delta x_t \Delta \Pi_{t+\tau}) = \bar{\lambda}\mathrm{E}(y_t \Delta \Pi_{t+\tau}), \tag{5.5.27}$$

where $\Delta\Pi_t = \Pi_t - \Pi_{t-1}$, $\bar{\lambda} = (\xi^P + b)^{-1}(1 - e^{-\xi^P})$, and λ also only depends on ξ^P and b. Moreover, it holds that, for all $\tau \geq 1$,

$$\rho_\tau := \mathrm{Corr}(x_{j,t}, x_{j,t+\tau}) = \frac{\xi^P e^{-b\tau} - b e^{-\xi^P \tau}}{\xi^P - b}. \tag{5.5.28}$$

Therefore, given ξ^P, by looking at the autocorrelation of position change $\Delta x_{j,t}$, the econometrician can identify b, which allows her to impute the moments involving y_t from

those based on Δx_t.

Finally, in equilibrium a and $\widetilde{\omega}$ are connected to b, ξ^P, ζ, and r as in (5.4.24).

5.5.2 Identification and Estimation Procedure

Proposition 8. *Suppose Assumptions 15 and 16 hold and ω_j is not invariant across j. Also suppose $\bar{\eta} = 0$ and $\xi^P = \xi^o$. If the econometrician has access to risk-free rate r, the covariance matrix Σ, the expected return vector R_τ, and the trading autocorrelation ρ_τ, then using (5.5.25), (5.5.26), (5.5.27), (5.5.28) and (5.4.24), she can identify, for all $j \in \mathcal{J}$,*

$$\xi^P, \quad \pi_j, \quad \omega_j, \quad \zeta, \quad \text{and} \quad \widetilde{\phi\iota}.$$

Given π_j and ω_j, we can only obtain σ_j^S up to a constant common across j. In other words, only the relative magnitude of information quality is identified, because we do not know parameters such as ξ^D and ξ^G that connect growth rate to price. The quantity $\widetilde{\phi\iota}$ affects the amount of predictable return and is a combination of how much the average estimate of growth rate deviate from the true value and how much the growth rate affects price.

The identification is achieved through the following procedure:

Algorithm 4. *Inputs: risk-free rate r, covariance matrix Σ, portfolio price change vector R_τ, and trading autocorrelation ρ_τ.*

S1. Given Σ, we can utilize (5.5.25) and (5.5.27) to obtain

$$\widetilde{\nu} \coloneqq (\lambda\widetilde{\phi\iota})\nu \quad \text{and} \quad \widetilde{\Sigma} \coloneqq \lambda\widetilde{\phi\iota}\big(\beta\beta^\intercal - \nu\nu^\intercal\big) + \lambda\bar{\phi}\kappa\kappa^\intercal.$$

S2. Using $\widetilde{\nu}$ and $\widetilde{\Sigma}$, we calculate $\lambda\widetilde{\phi\iota}$ using $\lambda\widetilde{\phi\iota} = -\widetilde{\nu}^\intercal\widetilde{\Sigma}^{-1}\widetilde{\nu}$.[7] We then obtain ν from $\widetilde{\nu}$ and $\lambda\widetilde{\phi\iota}$.

7. $\widetilde{\Sigma}$ is singular and $\widetilde{\Sigma}^{-1}$ stands for its pseudo inverse.

S3. Further, using $\lambda\widetilde{\phi\iota}$, $\widetilde{\Sigma}$, *and* R_τ, *we obtain, with any* $\tau \geq 1$,

$$\beta = \frac{R_\tau}{\sqrt{\lambda\widetilde{\phi\iota}R_\tau^\intercal\widetilde{\Sigma}^{-1}R_\tau}}.$$

S4. From how R_τ *and* ρ_τ *change with* τ *specified by* (5.5.26) *and* (5.5.28), *we obtain* ξ^P *and* b, *and thereby* λ *and* $\bar{\lambda}$.

S5. From β *and* ν, *we obtain* π *and* $\widetilde{\omega}^{-1}\omega$. *From* β, $\lambda\widetilde{\phi\iota}$, λ, $\bar{\lambda}$, *and* R_τ, *we obtain* $\widetilde{\phi\iota}$ *directly and obtain* $a\widetilde{\omega}$ *using* (5.5.26) *and* (5.5.27).

S6. Utilizing (5.4.24), *we obtain* $\widetilde{\omega}$ *from* b, ξ^P, *and* r. *Then we obtain* a *from* $a\widetilde{\omega}$. *Using* (5.4.24) *again, we obtain* ζ.

Outputs: ξ^P, π_j, ω_j, ζ, *and* $\widetilde{\phi\iota}$.

The above procedure does not rely on that the econometrician observe all the market participants.[8] In the case where we do observe every investor's trading, we do not need step 2 thanks to that $\sum_{j\in\mathcal{J}}\nu_j = 1$ by definition.

The natural implementation of Algorithm 4 is to construct empirical counterparts of the population moments. Given the large dimension of Σ and R_τ, using the sample covariance matrix and sample mean vector directly would incur large estimation erros. Motivated by the factor structure of trading manifested by (5.5.25), we propose an estimation method by modifying the standard principal component analysis (PCA) approach, which we call truncated PCA. Indeed, as discussed after Proposition 5, the eigenvalue generated by factor f_t is at the same order of magnitude as that from the idiosyncratic component $\varepsilon_{j,t}$. To bypass this issue, unlike the standard PCA that conducts eigendecomposition directly on covariance or correlation matrices, we replace all the diagonal elements of the sample version of Σ with

8. The dimension of matrices and vectors involved in the procedure then apparently have dimensions being the number of observed investors, and we of course only identify π_j and ω_j for those observed investors.

zero and conduct eigendecomposition afterwards. The diagonal elements themselves can be used to estimate $\widetilde{\nu}$ directly, which is needed in in step 2 of Algorithm 4, because $\lambda\phi\widetilde{\iota}\nu$ is the dominating part of the diagonal elements of Σ.[9] The eigendecomposition would generate eigenvectors that span β, ν, and κ. As a result, as demonstrated by (5.5.26), the expected return vector is also spanned by those eigenvectors. Therefore, we only need to estimate the projection of R_τ on a small number of eigenvectors, which is therefore of low dimension. The next Algorithm presents the details.

Algorithm 5. *Inputs: position change Δx_t and gain change $\Delta\Pi_t$.*

S1. Construct sample covariance matrix $\widehat{\Sigma} = \widehat{\mathrm{Cov}}(\Delta x_t)$. Then estimate $\widetilde{\nu}$ using

$$\widehat{\widetilde{\nu}}_j = \widehat{\Sigma}_{j,j}.$$

S2. Replace diagonal elements of $\widehat{\Sigma}$ with zero. Conduct eigendecomposition and take d eigenvectors with large eigenvalues in absolute value. The eigenvectors and eigenvalues can be written as a $J \times d$ matrix $\widehat{\Lambda}$ and $d \times d$ diagonal matrix $\widehat{D}$,[10] where Λ satisfies $\Lambda^\intercal\Lambda = I_d$. Estimate $\widetilde{\Sigma}$ as

$$\widehat{\widetilde{\Sigma}} = \widehat{\Lambda}\widehat{D}\widehat{\Lambda}^\intercal.$$

S3. Estimate R_τ as $\widehat{R}_\tau = \widehat{\Lambda}\widehat{\mathrm{E}}\big(\widehat{\Lambda}^\intercal\Delta x_t\Delta\Pi_{t+\tau}\big)$.

Outputs: $\widehat{\widetilde{\nu}}$, $\widehat{\widetilde{\Sigma}}$, $\widehat{R}_\tau$, $\widehat{\Lambda}$, and $\widehat{D}$.

5.6 Conclusion

Taking stock, our paper provides a new conceptual framework and appropriate econometric procedures to understanding the role of private information in the financial market utilizing

9. The estimates would be biased if κ is big enough. But in this case g_t becomes a strong factor and the danger would be easily detected.

10. We reuse the letter D which appears in the equilibrium model to represent cash flow rate. The context shall eliminate ambiguity.

price and quantity data together. The presence of many investors allows for a tractable equilibrium with heterogeneity in information and belief structure, and at the same time requires for properly designed econometric methods. Examining the equilibrium implications with the joint moments of price and institution holding data allows us to measure, among others, the magnitude of market inefficiency and the contributions of various investors to the price informativeness. More broadly, it would be interesting to investigate that to what extent the current empirical strategies, that are directly generated by our equilibrium model, can actually apply beyond under the current structural assumptions.

References

Aït-Sahalia, Y., and J. Jacod (2014): *High Frequency Financial Econometrics*. Princeton University Press.

Aït-Sahalia, Y., P. A. Mykland, and L. Zhang (2005): "How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise," *Review of Financial Studies*, 18, 351–416.

——— (2011): "Ultra High Frequency Volatility Estimation with Dependent Microstructure Noise," *Journal of Econometrics*, 160, 160–175.

Aït-Sahalia, Y., and D. Xiu (2016): "Increased Correlation Among Asset Classes: Are Volatility or Jumps to Blame?," *Journal of Econometrics*, 194, 205–219.

——— (2019): "A Hausman Test for the Presence of Market Microstructure Noise in High Frequency Data," *Journal of Econometrics*, 211, 176–205.

Amihud, Y. (2002): "Illiquidity and stock returns: cross-section and time-series effects," *Journal of Financial Markets*, 5(1), 31–56.

Amihud, Y., and H. Mendelson (1980): "Dealership market: Market-making and inventory," *Journal of Financial Economics*, 8, 31–53.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2003): "Modeling and Forecasting Realized Volatility," *Econometrica*, 71, 579–625.

Andrews, D. W. K., and X. Cheng (2012): "Estimation and Inference with Weak, Semi-Strong, and Strong Identification," *Econometrica*, 80(5), 2153–2211.

Andrews, D. W. K., X. Cheng, and P. Guggenberger (2020): "Generic Results for Establishing the Asymptotic Size of Confidence Sets and Tests," *Journal of Econometrics*, 218(2), 496–531.

Andrews, R. L., J. C. Arnold, and R. G. Krutchkoff (1972): "Shrinkage of the Posterior Mean in the Normal Case," *Biometrika*, 59(3), 693–695.

Ang, A., R. J. Hodrick, Y. Xing, and X. Zhang (2006): "The cross-section of volatility and expected returns," *Journal of Finance*, 61(1), 259–299.

Bajgrowicz, P., and O. Scaillet (2012): "Technical trading revisited False discoveries, persistence tests, and transaction costs," *Journal of Financial Economics*, 106(3), 473–491.

Bandi, F. M., and J. R. Russell (2008): "Microstructure Noise, Realized Volatility and Optimal Sampling," *Review of Economic Studies*, 75, 339–369.

Banz, R. W. (1981): "The relationship between return and market value of common stocks," *Journal of Financial Economics*, 9(1), 3–18.

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2008): "Designing realized kernels to measure ex-post variation of equity prices in the presence of noise," *Econometrica*, 76, 1481–1536.

——— (2009): "Realized Kernels in Practice: Trades and Quotes," *Econometrics Journal*, 12, 1–32.

——— (2011): "Subsampling Realised Kernels," *Journal of Econometrics*, 160(1), 204–219.

Barras, L., O. Scaillet, and R. Wermers (2010): "False discoveries in mutual fund performance: Measuring luck in estimated alphas," *Journal of Finance*, 65(1), 179–216.

Bekker, P. A. (1994): "Alternative Approximations to the Distributions of Instrumental Variable Estimators," *Econometrica*, 62(3), 657–681.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *Review of Economic Studies*, 81, 608–650.

Benjamini, Y., and Y. Hochberg (1995): "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), 289–300.

Berger, J. O. (1985): *Statistical Decision Theory and Bayesian Analysis*. Springer.

Berk, J. B., and R. C. Green (2004): "Mutual fund flows and performance in rational markets," *Journal of political economy*, 112(6), 1269–1295.

Berk, J. B., and J. H. Van Binsbergen (2015): "Measuring skill in the mutual fund industry," *Journal of financial economics*, 118(1), 1–20.

Bertsimas, D., and A. W. Lo (1998): "Optimal control of execution costs," *Journal of Financial Markets*, 1, 1–50.

Bhandari, L. C. (1988): "Debt/equity ratio and expected common stock returns: Empirical evidence," *Journal of Finance*, 43(2), 507–528.

Bibinger, M., N. Hautsch, P. Malec, and M. Reiss (2019): "Estimating the Spot Covariation of Asset Prices – Statistical Theory and Empirical Evidence," *Journal of Business & Economic Statistics*, 37(3), 419–435.

Box, G. E. P., G. M. Jenkins, and G. C. Reinsel (2007): *Time Series Analysis: Forecasting and Control*. Wiley, 4th edn.

Brockwell, P. J., and R. A. Davis (1991): *Time Series: Theory and Methods*. Springer-Verlag, New York, second edn.

Brogaard, J. A., T. Hendershott, and R. Riordan (2014): "High Frequency Trading and Price Discovery," *Review of Financial Studies*, 27, 2267–2306.

Brown, L. D., and E. Greenshtein (2009): "Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means," *The Annals of Statistics*, 37(4), 1685 – 1704.

Chamberlain, G., and M. Rothschild (1983): "Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets," *Econometrica*, 51, 1281–1304.

Chang, J., A. Delaigle, P. Hall, and C. Tang (2018): "A frequency domain analysis of the error distribution from noisy high-frequency data," *Biometrika*, 105(2), 353–369.

Chang, Y.-C., H. Hong, and I. Liskovich (2015): "Regression discontinuity and the price effects of stock market indexing," *The Review of Financial Studies*, 28(1), 212–246.

Chen, A. Y., and T. Zimmermann (2020): "Open source cross-sectional asset pricing," *Available at SSRN*.

Chen, X., L. P. Hansen, and P. G. Hansen (2021a): "Robust Identification of Investor Beliefs," *Proceedings of the National Academy of Sciences*, 117(52), 33130–33140.

——— (2021b): "Robust Inference for Moment Condition Models without Rational Expectations," *Journal of Econometrics, forthcoming*.

Collin-Dufresne, P., M. Johannes, and L. A. Lochstoer (2016): "Parameter Learning in General Equilibrium: The Asset Pricing Implications," *American Economic Review*, 106(3), 664–698.

Comte, F., and E. Renault (1996): "Long memory continuous time models," *Journal of Econometrics*, 73, 101–149.

——— (1998): "Long memory in continuous-time stochastic volatility models," *Mathematical Finance*, 8, 291–323.

Connor, G., M. Hagmann, and O. Linton (2012): "Efficient Semiparametric Estimation of the Fama-French Model and Extensions," *Econometrica*, 80(2), 713–754.

Connor, G., and R. A. Korajczyk (1986): "Performance Measurement with the Arbitrage Pricing Theory: A New Framework for Analysis," *Journal of Financial Economics*, 15(3), 373–394.

Cooper, M. J., H. Gulen, and M. J. Schill (2008): "Asset Growth and the Cross-Section of Stock Returns," *Journal of Finance*, 63(4), 1609–1651.

Da, R., S. Nagel, and D. Xiu (2022): "The Statistical Limit of Arbitrage," Discussion paper, Technical Report, Chicago Booth.

Da, R., and D. Xiu (2021a): "Disentangling Autocorrelated Intraday Returns," Discussion paper, University of Chicago.

——— (2021b): "When Moving-Average Models Meet High-Frequency Data: Uniform Inference on Volatility," Discussion paper, University of Chicago.

Donoho, D., and J. Jin (2004): "Higher Criticism for Detecting Sparse Heterogeneous Mixtures," *Annals of Statistics*, 32(3), 962–994.

Du, S., and H. Zhu (2017): "What is the optimal trading frequency in financial markets?," *The Review of Economic Studies*, 84(4), 1606–1651.

Efron, B. (2011): "Tweedie's formula and selection bias," *Journal of the American Statistical Association*, 106(496), 1602–1614.

——— (2019): "Bayes, Oracle Bayes and Empirical Bayes," *Statistical Science*, 34(2), 177 – 201.

Fama, E. F. (1970): "Efficient capital markets: A review of theory and empirical work," *The journal of Finance*, 25(2), 383–417.

Fama, E. F., and K. R. French (1993): "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, 33(1), 3–56.

——— (2006): "Profitability, investment and average returns," *Journal of Financial Economics*, 82(3), 491–518.

Fama, E. F., and K. R. French (2010): "Luck versus skill in the cross-section of mutual fund returns," *The journal of finance*, 65(5), 1915–1947.

Fama, E. F., and J. D. MacBeth (1973): "Risk, Return, and Equilibrium: Empirical Tests," *Journal of Political Economy*, 81, 607–636.

Fan, J., Y. Liao, and J. Yao (2015): "Power Enhancement in High-Dimensional Cross-Sectional Tests," *Econometrica*, 83(4), 1497–1541.

Fukasawa, M. (2010): "Realized volatility with stochastic sampling," *Stochastic Processes and Applications*, 120, 829–852.

Gabaix, X., and R. S. Koijen (2021): "In search of the origins of financial fluctuations: The inelastic markets hypothesis," Discussion paper, National Bureau of Economic Research.

Gagliardini, P., E. Ossola, and O. Scaillet (2016): "Time-varying risk premium in large cross-sectional equity datasets," *Econometrica*, 84(3), 985–1046.

Garbade, K., and Z. Lieber (1977): "On the Independence of Transactions on the New York Stock Exchange," *Journal of Banking and Finance*, 1(151-172).

Gardner, G., A. C. Harvey, and G. D. A. Phillips (1980): "Algorithm AS 154: An Algorithm for Exact Maximum Likelihood Estimation of Autoregressive-Moving Average Models by Means of Kalman Filtering," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(3), 311–322.

Gârleanu, N., and L. H. Pedersen (2013): "Dynamic trading with predictable returns and transaction costs," *The Journal of Finance*, 68(6), 2309–2340.

——— (2016): "Dynamic portfolio choice with frictions," *Journal of Economic Theory*, 165, 487–516.

Gibbons, M. R., S. A. Ross, and J. Shanken (1989): "A Test of the Efficiency of a Given Portfolio," *Econometrica*, 29, 1121–1152.

Giglio, S., B. Kellly, and D. Xiu (2022): "Factor Models, Machine Learning, and Asset Pricing," *Annual Review of Financial Economics*, 14, 337–368.

Giglio, S., B. Kelly, and D. Xiu (2022): "Factor models, machine learning, and asset pricing," *Annual Review of Financial Economics*, 14, 337–368.

Giglio, S., Y. Liao, and D. Xiu (2021): "Thousands of Alpha Tests," *Review of Financial Studies*, 34(7), 3456–3496.

Giglio, S., and D. Xiu (2021a): "Asset pricing with omitted factors," *Journal of Political Economy*, 129(7), 1947–1990.

——— (2021b): "Asset pricing with omitted factors," *Journal of Political Economy*, 129(7), 1947–1990.

Giglio, S., D. Xiu, and D. Zhang (2021): "Test assets and weak factors," Discussion paper, National Bureau of Economic Research.

Griffin, J. E., and R. C. Oomen (2008): "Sampling returns for realized variance calculations: Tick time or transaction time?," *Econometric Reviews*, 27, 230–253.

Gromb, D., and D. Vayanos (2010): "Limits of Arbitrage," *Annual Review of Financial Economics*, 2, 251–275.

Grossman, S. J., and J. E. Stiglitz (1980): "On the impossibility of informationally efficient markets," *The American economic review*, 70(3), 393–408.

Gu, S., B. Kellly, and D. Xiu (2020): "Empirical Asset Pricing via Machine Learning," *Review of Financial Studies*, 33(5), 2223–2273.

Guijarro-Ordonez, J., M. Pelger, and G. Zanotti (2022): "Deep Learning Statistical Arbitrage," Discussion paper, Stanford University.

Hahn, J., and H. Lee (2009): "Financial Constraints, Debt Capacity, and the Cross-Section of Stock Returns," *Journal of Finance*, 64(2), 891–921.

Hannan, E. J. (1980): "The Estimation of the Order of an ARMA Process," *The Annals of Statistics*, pp. 1071–1081.

Hansen, L. P. (2007): "Beliefs, Doubts, and Learning: Valuing Macroeconomic Risk," *American Economic Review*, 97(2), 1–30.

——— (2014): "Nobel Lecture: Uncertainty Outside and Inside Economic Models," *Journal of Political Economy*, 122(51), 945–987.

Hansen, P. R., J. Large, and A. Lunde (2008): "Moving average-based estimators of integrated variance," *Econometric Reviews*, 27, 79–111.

Hansen, P. R., and A. Lunde (2006): "Realized variance and market microstructure noise," *Journal of Business & Economic Statistics*, 24(2), 127–161.

Harris, L., and E. Gurel (1986): "Price and volume effects associated with changes in the S&P 500 list: New evidence for the existence of price pressures," *the Journal of Finance*, 41(4), 815–829.

Harvey, C. R., and Y. Liu (2020): "False (and Missed) Discoveries in Financial Economics," *Journal of Finance, forthcoming*.

Harvey, C. R., Y. Liu, and H. Zhu (2016): "... And the cross-section of expected returns," *Review of Financial Studies*, 29(1), 5–68.

Hasbrouck, J. (2007): *Empirical Market Microstructure*. Oxford University Press, New York, NY.

Hasbrouck, J., and T. S. Y. Ho (1987): "Order Arrival, Quote Behavior, and the Return-Generating Process," *Journal of Finance*, 42(4), 1035–1048.

Hautsch, N., and M. Podolskij (2013): "Pre-Averaging Based Estimation of Quadratic Variation in the Presence of Noise and Jumps: Theory, Implementation, and Empirical Evidence," *Journal of Business & Economic Statistics*, 31(2), 165–183.

He, H., and J. Wang (1995): "Differential information and dynamic behavior of stock trading volume," *The Review of Financial Studies*, 8(4), 919–972.

Heston, S. L., and R. Sadka (2008): "Seasonality in the cross-section of stock returns," *Journal of Financial Economics*, 87(2), 418–445.

Huberman, G. (1982): "A Simple Approach to Arbitrage Pricing Theory," *Journal of Economic Thoery*, 28(1), 183–191.

Imbens, G. W., and C. F. Manski (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72(6), 1845–1857.

Ingersoll, J. E. (1984): "Some Results in the Theory of Arbitrage Pricing," *Journal of Finance*, 39(4), 1021–1039.

Jacod, J., Y. Li, P. A. Mykland, M. Podolskij, and M. Vetter (2009): "Microstructure Noise in the Continuous Case: The Pre-Averaging Approach," *Stochastic Processes and Their Applications*, 119, 2249–2276.

Jacod, J., Y. Li, and X. Zheng (2017): "Statistical Properties of Microstructure Noise," *Econometrica*, 85(4), 1133–1174.

——— (2019): "Estimating the Integrated Volatility with Tick Observations," *Journal of Econometrics*, 208, 80–100.

Jacod, J., M. Podolskij, and M. Vetter (2010): "Limit Theorems for Moving Averages of Discretized Processes Plus Noise," *Annals of Statistics*, 38, 1478–1545.

Jacod, J., and P. Protter (2011): *Discretization of Processes*. Springer-Verlag.

Jegadeesh, N. (1990): "Evidence of predictable behavior of security returns," *Journal of Finance*, 45(3), 881–898.

Jegadeesh, N., and S. Titman (1993): "Returns to buying winners and selling losers: Implications for stock market efficiency," *Journal of Finance*, 48(1), 65–91.

Johns, M. V. (1957): "Non-parametric empirical Bayes procedures," *Annals of Mathematical Statistics*, 28, 649–669.

Kabaila, P. (1995): "The Effect of Model Selection on Confidence Regions and Prediction Regions," *Econometic Theory*, 11(3), 537–549.

Kacperczyk, M., S. Van Nieuwerburgh, and L. Veldkamp (2014): "Time-varying fund manager skill," *The Journal of Finance*, 69(4), 1455–1484.

——— (2016): "A rational theory of mutual funds' attention allocation," *Econometrica*, 84(2), 571–626.

Kalnina, I., and O. Linton (2008): "Estimating quadratic variation consistently in the presence of endogenous and diurnal measurement error," *Journal of Econometrics*, 147, 47–59.

Kan, R., and G. Zhou (2007): "Optimal Portfolio Choice with Parameter Uncertainty," *Journal of Financial and Quantitative Analysis*, 42(3), 621–656.

Kelly, B., and S. Pruitt (2013): "Market expectations in the cross-section of present values," *The Journal of Finance*, 68(5), 1721–1756.

Kelly, B., S. Pruitt, and Y. Su (2019a): "Some Characteristics are Risk Exposures, and the Rest are Irrelevant," *Journal of Financial Economics, forthcoming*.

Kelly, B. T., S. Pruitt, and Y. Su (2019b): "Characteristics are covariances: A unified model of risk and return," *Journal of Financial Economics*, 134(3), 501–524.

Kim, S., R. Korajczyk, and A. Neuhierl (2020): "Arbitrage Portfolios," *Review of Financial Studies, forthcoming*.

Koijen, R. S., and M. Yogo (2019): "A demand system approach to asset pricing," *Journal of Political Economy*, 127(4), 1475–1515.

Kozak, S., S. Nagel, and S. Santosh (2018a): "Interpreting factor models," *Journal of Finance*, 73(3), 1183–1223.

——— (2018b): "Interpreting factor models," *The Journal of Finance*, 73(3), 1183–1223.

——— (2020): "Shrinking the cross-section," *Journal of Financial Economics*, 135(2), 271–292.

Kyle, A. S. (1985a): "Continuous Auctions and Insider Trading," *Econometrica*, 53, 1315–1336.

Kyle, A. S. (1985b): "Continuous auctions and insider trading," *Econometrica: Journal of the Econometric Society*, pp. 1315–1335.

——— (1989): "Informed speculation with imperfect competition," *The Review of Economic Studies*, 56(3), 317–355.

Kyle, A. S., A. A. Obizhaeva, and Y. Wang (2018): "Smooth trading with overconfidence and market power," *The Review of Economic Studies*, 85(1), 611–662.

Lakonishok, J., A. Shleifer, and R. W. Vishny (1994): "Contrarian investment, extrapolation, and risk," *Journal of Finance*, 49(5), 1541–1578.

Leeb, H., and B. M. Pötscher (2005): "Model Selection and Inference: Facts and Fiction," *Econometic Theory*, 21(1), 21–59.

Leeb, H., and B. M. Pötscher (2008): "Can one estimate the unconditional distribution of post-model-selection estimators?," *Econometric Theory*, 24(2), 338–376.

Li, J., and D. Xiu (2016): "Generalized Method of Integrated Moments for High-Frequency Data," *Econometrica*, 84, 1613–1633.

Li, M., and O. Linton (2021): "A ReMeDI for Microstructure Noise," Discussion paper, Cambridge University.

Li, Y., P. A. Mykland, E. Renault, L. Zhang, and X. Zheng (2014): "Realized Volatility When Sampling Times are Possibly Endogenous," *Econometric Theory*, 30, 580–605.

Li, Z. M., R. J. Laeven, and M. H. Vellekoop (2020): "Dependent Microstructure Noise and Integrated Volatility Estimation from High-Frequency Data," *Journal of Econometrics*, 215(2), 536–558.

Litzenberger, R. H., and K. Ramaswamy (1979): "The effects of personal taxes and dividends on capital asset prices: Theory and empirical evidence," *Journal of Financial Economics*, 7, 163–195.

Mancini, C. (2001): "Disentangling the Jumps of the Diffusion in a Geometric Jumping Brownian Motion," *Giornale dell'Istituto Italiano degli Attuari*, 64, 19–47.

Martin, I., and S. Nagel (2021): "Market Efficiency in the Age of Big Data," *Journal of Financial Economics, forthcoming*.

Martin, I. W., and S. Nagel (2022): "Market efficiency in the age of big data," *Journal of financial economics*, 145(1), 154–177.

Mikusheva, A. (2012): "One Dimensional Inference in Autoregressive Models with the Potential Presence of a Unit Root," *Econometrica*, 80(1), 173–212.

Moskowitz, T. J., and M. Grinblatt (1999): "Do industries explain momentum?," *Journal of Finance*, 54(4), 1249–1290.

Niederhoffer, V., and M. Osborne (1966): "Market Making and Reversal on the Stock Exchange," *Journal of American Statistical Association*, 61(316), 897–916.

Pástor, L., and R. F. Stambaugh (2012): "On the size of the active management industry," *Journal of Political Economy*, 120(4), 740–781.

Pástor, L., R. F. Stambaugh, and L. A. Taylor (2015): "Scale and skill in active management," *Journal of Financial Economics*, 116(1), 23–45.

——— (2017): "Do funds make more when they trade more?," *The Journal of Finance*, 72(4), 1483–1528.

——— (2020): "Fund tradeoffs," *Journal of Financial Economics*, 138(3), 614–634.

Pastor, L., and P. Veronesi (2009): "Learning in Financial Markets," *Annual Review of Financial Economics*, 1(1), 361–381.

Pesaran, H., and T. Yamagata (2017): "Testing for Alpha in Linear Factor Pricing Models with a Large Number of Securities," Discussion paper.

Petajisto, A. (2009): "Why do demand curves for stocks slope down?," *Journal of Financial and Quantitative Analysis*, 44(5), 1013–1044.

Pötscher, B. M. (1991): "Effects of Model Selection on Inference," *Econometic Theory*, 7(2), 163–185.

Reiss, M. (2011): "Asymptotic equivalence for inference on the volatility from noisy observations," *Annals of Statistics*, 39, 772–802.

Renault, E., and B. J. Werker (2011): "Causality Effects in Return Volatility Measures with Random Times," *Journal of Econometrics*, 160, 272–279.

Robbins, H. (1956): "An empirical Bayes approach to statistics," *Berkeley Symposium on Mathematical Statistics and Probability*, 3, 157–163.

Robertson, T., R. Dykstra, and F. Wright (1988): "Order restricted statistical inference," *New York: Wiley*.

Rosenberg, B. (1974): "Extra-Market Components of Covariance in Security Returns," *Journal of Financial and Quantitative Analysis*, 9(2), 263–274.

Ross, S. A. (1976): "The arbitrage theory of capital asset pricing," *Journal of Economic Theory*, 13(3), 341–360.

Shanken, J. (1992): "The Current State of the Arbitrage Pricing Theory," *Journal of Finance*, 47(4), 1569–1574.

Shibata, R. (1980): "Asymptotically Efficient Selection of the Order of the Model for Estimating Parameters of a Linear Process," *The Annals of Statistics*, 8(1), 147–164.

——— (1986): "Consistency of Model Selection and Parameter Estimation," *Journal of Applied Probability*, 23, 127–141.

Shleifer, A. (1986): "Do demand curves for stocks slope down?," *The Journal of Finance*, 41(3), 579–590.

Simmons, D. M. (1971): "Common-Stock Transaction Sequences and the Random-Walk Model," *Operations Research*, 19(4), 845–861.

Sloan, R. G. (1996): "Do Stock Prices Fully Reflect Information in Accruals and Cash Flows about Future Earnings?," *The Accounting Review*, 71(3), 289–315.

Song, Y. (2020): "The mismatch between mutual fund scale and skill," *The Journal of Finance*, 75(5), 2555–2589.

Staiger, D., and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65(3), 557–586.

Tóth, B., I. Palit, F. Lillo, and J. D. Farmer (2015): "Why is Equity Order Flow so Persistent?," *Journal of Economic Dynamics & Control*, 51, 218–239.

Tukey, J. W. (1976): "The Higher Criticism. Course Notes, Statistics 411," Discussion paper, Princeton University.

Uppal, R., and P. Zaffaroni (2018): "Robust Portfolio Choice," Discussion paper, EDHEC Business School and Imperial College London.

Varneskov, R. T. (2016): "Estimating the quadratic variation spectrum of noisy asset prices using generalized flat-top realized kernels," *Econometric Theory*, pp. 1–45.

Vayanos, D. (1999): "Strategic trading and welfare in a dynamic market," *The Review of Economic Studies*, 66(2), 219–254.

Vives, X. (2011): "Strategic supply function competition with private information," *Econometrica*, 79(6), 1919–1966.

Wang, J. (1993): "A model of intertemporal asset prices under asymmetric information," *The Review of Economic Studies*, 60(2), 249–282.

——— (1994): "A model of competitive stock trading volume," *Journal of political Economy*, 102(1), 127–168.

Wermers, R. (2000): "Mutual fund performance: An empirical decomposition into stock-picking talent, style, transactions costs, and expenses," *The Journal of Finance*, 55(4), 1655–1695.

White, H. (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25.

Whittle, P. (1951): *Hypothesis Testing in Time Series Analysis*. New York: Hafner Publishing Company.

Wilson, G. (1969): "Factorization of the Covariance Generating Function of a Pure Moving Average Process," *SIAM Journal on Numerica Analysis*, 6(1), 1–7.

Xiu, D. (2010): "Quasi-Maximum Likelihood Estimation of Volatility with High Frequency Data," *Journal of Econometrics*, 159, 235–250.

Zhang, C.-H. (1997): "EMPIRICAL BAYES AND COMPOUND ESTIMATION OF NORMAL MEANS," *Statistica Sinica*, 7(1), 181–193.

Zhang, L. (2006): "Efficient Estimation of Stochastic Volatility Using Noisy Observations: A Multi-Scale Approach," *Bernoulli*, 12, 1019–1043.

Zhang, L., P. A. Mykland, and Y. Aït-Sahalia (2005): "A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High-Frequency Data," *Journal of the American Statistical Association*, 100, 1394–1411.