Characteristic-Sorted Portfolios: Estimation and Inference*

Matias D. Cattaneo† Richard K. Crump‡ Max H. Farrell§
Ernst Schaumburg¶

January 25, 2018

Abstract

Portfolio sorting is ubiquitous in the empirical finance literature, where it has been widely used to identify pricing anomalies in different asset classes. Despite its popularity, little attention has been paid to the statistical properties of the procedure or conditions under which it produces valid inference. We develop a general, formal framework for portfolio sorting by casting it as a nonparametric estimator. We give precise conditions under which the portfolio sorting estimator is consistent and asymptotically normal, and also establish consistency of both the Fama-MacBeth variance estimator and a new plug-in estimator. Our framework bridges the gap between portfolio sorting and cross-sectional regressions by allowing for linear conditioning variables when sorting. In addition, we obtain a valid mean square error expansion of the sorting estimator, which we employ to develop optimal choices for the number of portfolios. We show that the choice of the number of portfolios is crucial to draw accurate conclusions from the data and provide a simple, data-driven procedure which balances higher-order bias and variance. In many practical settings the optimal number of portfolios varies substantially across applications and subsamples and is, in many cases, much larger than the standard choices of 5 or 10 portfolios used in the literature. We give formal and intuitive justifications for this finding based on the bias-variance trade-off underlying the portfolio sorting estimator. To illustrate the relevance of our results, we revisit the size and momentum anomalies.

Keywords: Portfolio sorts, stock market anomalies, firm characteristics, nonparametric estimation, partitioning, cross-sectional regressions

JEL Classification: C12, C14

*First draft: March 28, 2015. The authors would like to thank Tobias Adrian, Francisco Barillas, Nina Boyarchenko, Jules van Binsbergen, John Campbell, Kent Daniel, Fernando Duarte, Eric Ghysels, Stefano Giglio, Peter Hansen, Ralph Koijen, Gabriele La Spada, Jonathan Lewellen, Erik Loualiche, Stefan Nagel, Andrew Patton, Karen Shen, Stijn Van Nieuwerburgh, Peter Van Tassel, S. Viswanathan, Erik Vogt, Jonathan Wright and participants at the 2015 Interactions Conference and the 2016 MFA Annual Meetings for helpful comments and discussions. Skanda Amarnath, Evan Friedman and Rui Yu provided excellent research assistance. The views expressed in this paper are those of the authors and do not necessarily represent those of the Federal Reserve Bank of New York, the Federal Reserve System or AQR Capital Management LLC. Cattaneo gratefully acknowledges financial support from the National Science Foundation through grant SES 1459931.

†Department of Economics and Department of Statistics, University of Michigan; cattaneo@umich.edu

‡Federal Reserve Bank of New York; richard.crump@ny.frb.org

§Booth School of Business, University of Chicago; max.farrell@chicagobooth.edu

¶AQR Capital Management; ernst.schaumburg@gmail.com

1 Introduction

Portfolio sorting is an important tool of modern empirical finance. It has been used to test fundamental theories in asset pricing, to establish a number of different pricing anomalies, and to identify profitable investment strategies. However, despite its ubiquitous presence in the empirical finance literature, little attention has been paid to the statistical properties of the procedure. We endeavor to fill this gap by formalizing and investigating the properties of so-called characteristic-sorted portfolios—where portfolios of assets are constructed based on similar values for one or more idiosyncratic characteristics and the cross-section of portfolio returns is of primary interest. The empirical applications of characteristic-sorted portfolios are too numerous to list, but some of the seminal work applied to the cross-section of equity returns includes Basu (1977), Stattman (1980), Banz (1981), Bondt and Thaler (1985), Rosenberg et al. (1985), Jegadeesh (1990), Fama and French (1992), and Jegadeesh and Titman (1993). More recently, the procedure has been applied to other asset classes such as currencies (e.g., Lustig and Verdelhan (2007), Lustig et al. (2011)), and across different assets (e.g., Koijen et al. (2015)). Furthermore, portfolio sorting remains a highly popular tool in empirical finance.[1]

We develop a general, formal framework for portfolio sorting by casting the procedure as a nonparametric estimator. Sorting into portfolios has been informally recognized in the literature as a nonparametric alternative to imposing linearity on the relationship between returns and characteristics in recent years (e.g. Fama and French, 2008; Cochrane, 2011), but no formal framework is at present available in the literature. We impose sampling assumptions which are very general and can accommodate momentum and reversal effects, conditional heteroskedasticity in both the cross section and the time series, and idiosyncratic characteristics with a factor structure. Furthermore, our proposed framework allows for both estimated quantiles when forming the portfolios and additive linear-in-parameters conditioning variables entering the underlying model governing the relationship between returns and sorting characteristics. This latter feature of our proposed framework bridges the gap between portfolio sorts and cross-sectional regressions and will allow empirical researchers to investigate new candidate variables while controlling for existing anomalies already identified. More generally, our framework captures and formalizes the main aspects of common empirical work in finance employing portfolio sorts, and therefore gives the basis for a thorough analysis of the statistical properties of popular estimators and test statistics.

Employing our proposed framework, we study the asymptotic properties of the portfolio-sorting estimator and related test statistics in settings with "large" cross-sectional and times-series sample sizes, as this is the most usual situation encountered in applied work. We

[1] Barrot et al. (2016), Bouchaud et al. (2016), Linnainmaa and Roberts (2016), Weber (2016) among others are examples of current working papers utilizing the procedure.

first establish consistency and asymptotic normality of the estimator, explicitly allowing for estimated quantile-spaced portfolios, which reflects standard practice in empirical finance. In addition, we prove the validity of two distinct standard error estimators. The first is a "plug-in" variance estimator which is new to the literature. The second is the omnipresent Fama and MacBeth (1973)-style variance estimator which treats the average portfolio returns as if they were draws from a single, uncorrelated time series. Despite its widespread use, we are unaware of an existing proof of its validity for inference in this setting, although this finding is presaged by the results in Ibragimov and Müller (2010, 2016). Altogether, our first-order asymptotic results provide theory-based guidance to empirical researchers, previously unavailable, highlighting when the portfolio-sorting estimator may be expected to perform well.

Once the portfolio sorting estimator is viewed through the lens of nonparametric estimation, it is clear that the choice of number of portfolios acts as the tuning parameter for the procedure and that an appropriate choice is paramount for drawing valid empirical conclusions. To address this issue, we also obtain higher-order asymptotic mean square error expansions for the estimator which we employ to develop several optimal choices of the total number of portfolios for applications. These optimal choices balance bias and variance and will change depending on the prevalence of many common features of panel data in finance such as unbalanced panels, the relative number of cross-sectional observations versus time-series observations and the presence of conditional heteroskedasticity. In practice, the common approach in the empirical finance literature is to treat the choice of the number of portfolios as invariant to the data at hand—often following historical norms, such as 10 portfolios when sorting on a single characteristic. This is summarized succinctly in Cochrane (2011, p. 1061): "Following Fama and French, a standard methodology has developed: Sort assets into portfolios based on a characteristic, look at the portfolio means (especially the *1–10* portfolio alpha, information ratio, and t-statistic)..." (emphasis added). Thus, another contribution of our paper is to provide a simple, data-driven procedure which is optimal in an objective sense to choose the appropriate number of portfolios. Employing this data-driven procedure provides more power to discern a significant return differential in the data. The optimal choice will vary across time with the cross-sectional sample size and, all else equal, be larger when the number of time-series observations is larger.

We demonstrate the empirical relevance of our theoretical results by revisiting the size anomaly—where smaller firms earn higher returns than larger firms on average, and the momentum anomaly—where firms which have had better relative returns in the recent past also have higher future relative returns on average. We find that in the universe of US stocks the size anomaly is represented by a monotonically decreasing and convex relationship between returns and size, is highly significant, and is robust to different sub-periods including the period from 1980–2015. However, as pointed out in the existing literature, the size anomaly is not robust in sub-samples which exclude "smaller" small firms (i.e., considering only firms listed on the NYSE). We also find that in the universe of US stocks the momentum anomaly is represented by a monotonically increasing and concave relationship between returns and past returns and is highly significant with the "short" side of the trade becoming more profitable in later sub-periods. We also show that the momentum anomaly is distinct from industry momentum by including the latter measure (along with its square and cube) as linear control variables in a portfolio sorting exercise. In both empirical applications we find that the optimal number of portfolios varies substantially over time and is much larger than the standard choice of ten routinely used in the empirical finance literature. In the case of the size anomaly, the optimal number of portfolios can be as small as about 50 in the 1920s and can rise to above 200 in the late 1990s. However, for the momentum anomaly, the optimal number of portfolios can be as small as about 10 in the 1920s and around 50 in the late 1990s.

The financial econometrics literature has primarily focused on the study of estimation and inference in (restricted) factor models featuring common risk factors and idiosyncratic loadings.[2] In contrast, to our knowledge, we are the first to provide a formal framework and to analyze the standard empirical approach of (characteristic-based) portfolio sorting. A few authors have investigated specific aspects of sorted portfolios. Lo and MacKinlay (1990) and Conrad et al. (2003) have studied the effects of data-snooping bias on empirical conclusions drawn from sorted portfolios and argue that they can be quite large. Berk (2000) investigates the power of testing asset pricing models using only the assets within a particular portfolio and argues that this approach biases results in favor of rejecting the model being studied. More recently, Patton and Timmermann (2010) and Romano and Wolf (2013) have proposed tests of monotonicity in the average cross-section of returns taking the sorted portfolios themselves as *given*. Finally, there is a large literature attempting to discriminate between factor-based and characteristic-based explanations for return anomalies. The empirical implementations in this literature often use characteristic-sorted portfolios as test assets although this approach is not universally advocated (see, for example, Fan and Liu (2008), Lewellen et al. (2010), Daniel and Titman (2012), and Kleibergen and Zhan (2013)).

The paper is organized as follows. Section 2 describes our framework and provides a brief overview of our new results. The more general framework is presented in Section 3. Then Sections 4 and 5 treat first-order asymptotic theory and mean square error expansions, re-

[2]See, for example, Gibbons (1982), Kandel (1984), Roll (1985), Shanken (1985), Shanken (1986), Jagannathan and Wang (1998), Kan and Chen (2005), Shanken and Zhou (2007), Kleibergen (2009), Nagel and Singleton (2011), Connor et al. (2012), Adrian et al. (2015), Gospodinov et al. (2014), Ang et al. (2017), Gospodinov et al. (2017) among many others.

spectively; the latter provides guidance on implementation. Section 6 provides our empirical results and Section 7 concludes and discusses further work.

2 Motivation and Overview of Results

This section provides motivation for our study of portfolio sorting and a simplified overview of our results. The premise behind portfolio sorting is to discover whether expected returns of an asset are related to a certain characteristic. A natural, and popular, way to investigate this is to sort observed returns by the characteristic value, divide the assets into portfolios according to the characteristic, and then compare differences in average returns across the portfolios. This methodological approach has found wide popularity in the empirical finance literature not least because it utilizes a basic building block of modern finance, a portfolio of assets, which produces an intuitive estimator of the relationship between asset returns and characteristics.[3] The main goal of this paper is to provide a formal framework and develop rigorous inference results for this procedure. All assumptions and technical results are discussed in detail in the following sections, but omitted here for ease of exposition.

To begin, suppose we observe both the return, R, and value of a single continuous characteristic, z, for n assets over T time periods, that are related through a regression-type model of the form

$$R_{it} = \mu(z_{it}) + \varepsilon_{it}, \qquad i = 1, \ldots, n, \quad t = 1, \ldots, T. \tag{1}$$

Here $\mu(\cdot)$ is the unknown object of interest that dictates how expected returns vary with the characteristic, and is assumed to be continuously differentiable. The general results given in the next section cover a wide range of inference targets and extend the model of equation (1) to include multiple sorting characteristics, conditioning variables, and unbalanced panels, among other features commonly encountered in empirical finance.

To understand the relationship between expected returns and the characteristic at hand, characterized by the unknown function $\mu(z)$, we first form portfolios by partitioning the support of z into quantile-spaced bins. While it is possible to form portfolios in other ways, quantile spacing is the standard technique in empirical finance; our goal is to develop theory that mimics empirical practice as closely as possible (see Remarks 1 and 2 for more discussion). For each period t, it is common practice to form J disjoint portfolios, denoted

[3] For early references see, for example, Black et al. (1972), Fama and MacBeth (1974), Fama (1976), and Jensen (1978).

by P_{jt}, as follows:

$$P_{jt} = \begin{cases} \left[z_{(1)t}, z_{(\lfloor n/J \rfloor)t}\right) & \text{if } j = 1 \\ \left[z_{(\lfloor n(j-1)/J \rfloor)t}, z_{(\lfloor nj/J \rfloor)t}\right) & \text{if } j = 2, \ldots, J-1 \\ \left[z_{(\lfloor n(J-1)/J \rfloor)t}, z_{(n)t}\right] & \text{if } j = J \end{cases} \tag{2}$$

where $z_{(\ell)t}$ denotes the ℓ-th order statistic of the sample of characteristics $\{z_{it} : 1 \leq i \leq n\}$ at each time period $t = 1, 2, \cdots, T$, and $\lfloor \cdot \rfloor$ denotes the floor operator. In other words, each portfolio is a random interval containing roughly $(100/J)$-percent of the observations at each moment in time. This means that the position and length of the portfolios vary over time, but is set automatically, while the number of such portfolios (J) must be chosen by the researcher. A careful (asymptotic) analysis of portfolio-sorting estimators requires accounting for the randomness introduced in the construction of the portfolios, as we do in more detail below.

With the portfolios thus formed, we estimate $\mu(z_*)$ at some fixed point z_* with the average returns within the portfolio containing z_*. Here z_* represents the evaluation point that is of interest to the empirical researcher. For example, one might be interested in expected returns for those individual assets with a very high value of a characteristic. Over time, exactly which portfolio includes assets with characteristic z_* may change. If we let P_{jt}^* represent the appropriate portfolio at each time t then the basic portfolio-sorted estimate is

$$\hat{\mu}(z_*) = \frac{1}{T}\sum_{t=1}^{T} \hat{\mu}_t(z_*), \qquad \hat{\mu}_t(z_*) = \frac{1}{N_{jt}^*}\sum_{i: z_{it} \in P_{jt}^*} R_{it}, \tag{3}$$

where N_{jt}^* is the number of assets in P_{jt}^* at time t. If $J \leq n$, this estimator is well-defined, as there are (roughly) n/J assets in all portfolios. The main motivation for using a sample average of each individual estimator is so that the procedure more closely mimics the actual practice of portfolio choice (where future returns are unknown) and because of the highly unbalanced nature of financial panel data.[4] That said, this estimator (as well as the more general version below) can be simply implemented using ordinary least squares (or weighted least squares in the case of value-weighted portfolios).

The starting point of our formalization is the realization that each $\hat{\mu}_t(z_*)$, $t = 1, \ldots, T$, is a nonparametric estimate of the regression function $\mu(z_*)$, using a technique known as *partitioned regression*. Studied recently by Cattaneo and Farrell (2013), the partition regression estimator estimates $\mu(z_*)$ using observations that are "close" to z_*, which at present means

[4] An alternative interpretation is to allow for time variation in $\mu(\cdot)$ and the estimand becomes the grand mean.

that they are in the same portfolio. A key lesson is that J is the tuning parameter of this nonparametric procedure, akin to the bandwidth in kernel-based estimators or the number of terms in a sieve estimator. For smaller J, the variance of $\hat{\mu}_t(z_*)$ will be low, as a relatively large portion of the sample is in each portfolio, but this also implies that the portfolio includes assets with characteristics quite far from z_*, implying an increased bias; on the other hand, a larger J will decrease bias, but inflate variance. For each cross section, $\hat{\mu}_t(\cdot)$ is a step function with J "rungs", each an average return within a portfolio. While estimation of $\mu(\cdot)$ could be performed with a variety of nonparametric estimators (such as local linear regression, series estimation with spline basis functions etc.), our goal is to explicitly analyze portfolio sorting. From a practitioners perspective, the estimator has the advantage that it has a direct interpretation as a return on a portfolio which is an economically meaningful object.

Moving beyond the cross section, the same structure and lessons holds for the full $\hat{\mu}(z_*)$ of equation (3), but with dramatically different results. Consider Figure 1. Panel (a) shows a single realization of $\hat{\mu}_t(\cdot)$, with $J = 4$, *for a single cross section*. Moving to panel (b), we see that averaging over only two time periods results in a more complex estimator, as the portfolios are formed separately for each cross section. Finally, panel (c) shows the result with $T = 50$ (though a typical application may have T in the hundreds). Throughout, J is fixed, but the increase in T acts to smooth the fit; this point appears to be poorly recognized in practice, and makes clear that the choice of J must depend on T. Next, for the same choices of n and T, Panels (d)–(f) repeat the exercise but with $J = 10$. Comparing panels in the top row to the bottom of Figure 1 shows the bias-variance tradeoff discussed above. Collectively, Figure 1 makes clear that J must depend on the features of the data at hand. Indeed, we show that consistency of $\hat{\mu}(\cdot)$ requires that J diverge with n and T fast enough to remove bias but not so quickly that the variance explodes. We detail explicit, practicable choices of J later in the paper.

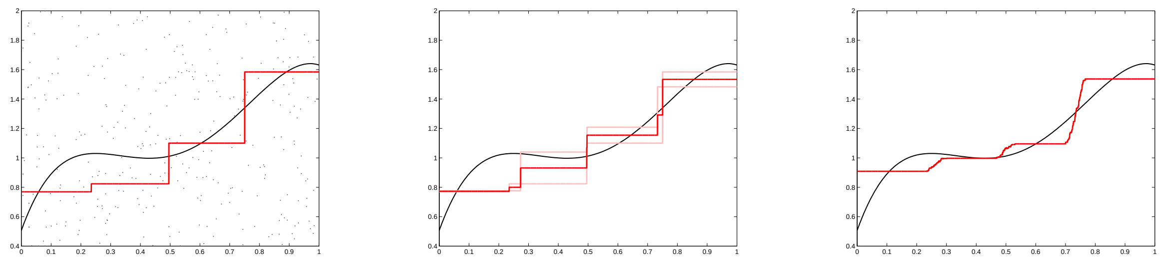
With the portfolios and estimator defined, by far the most common object of interest in the empirical finance literature is the expected returns in the highest portfolio less those in the lowest, which is then either (informally) interpreted as a test of monotonicity of the function $\mu(z)$ or used to construct factors based on the characteristic z. These are different goals (inference and point estimation, respectively), and in particular, require different choices of J (with different rates).

First, consider the test of monotonicity, which is also interpreted as the return from a strategy of buying the spread portfolio: long one dollar of the higher expected return portfolio and short one dollar of the lower expected return portfolio. Formally, we wish to

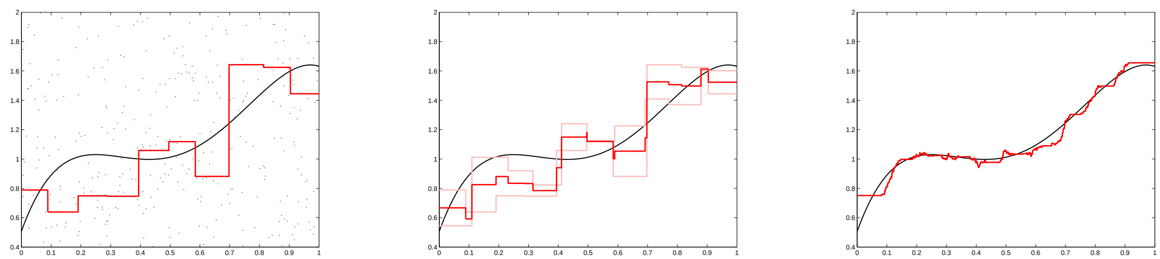
Figure 1: **Introductory Example**

This figure shows the true function $\mu(z) = .45(2.25 + (z-1/2) + 8(z-1/2).^2 + 6(z-1/2)^3 - 30(z-1/2)^4)$ (black line) and the estimated function $\hat{\mu}(z)$ (red line). The left panels show the observed $n = 500$ data points (gray dots) and the middle panels display the estimated function for each time period (pink line). Portfolio breakpoints are chosen as the estimated quantiles of the distribution of z where $z \sim \mathcal{B}eta\,(1,1)$ and $z \sim \mathcal{B}eta\,(1.2, 1.2)$ for odd and even time periods, respectively.

(a) $J = 4$, $T = 1$ (b) $J = 4$, $T = 2$ (c) $J = 4$, $T = 50$



(d) $J = 10$, $T = 1$ (e) $J = 10$, $T = 2$ (f) $J = 10$, $T = 50$



conduct the hypothesis test:

$$\mathsf{H}_0 : \mu(z_H) - \mu(z_L) = 0 \qquad \text{vs.} \qquad \mathsf{H}_1 : \mu(z_H) - \mu(z_L) \neq 0, \tag{4}$$

where $z_L < z_H$ denote "low" and "high" evaluation points. (In practice, z_L and z_H are usually far apart and never within the same portfolio.) Statistical significance in this context is intimately related to the economic significance of the trading strategy, as measured by the Sharpe ratio (Sharpe (1966)). Our general framework allows for a richer class of estimands (see Remark 6) but this estimand will remain our focus throughout the paper because it is the most relevant to empirical researchers.

The following result (or, the more general Theorem 1 below) establishes first-order asymptotic validity for testing (4) using portfolio sorting with estimated quantiles.

Corollary 1. *Under the conditions of Theorem 1 below, and in particular if* $J \log(\max(J,T))/n \to 0$ *and* $nT/J^3 \to 0$, *then*

$$\mathcal{T} = \frac{\left[\hat{\mu}(z_H) - \hat{\mu}(z_L)\right] - \left[\mu(z_H) - \mu(z_L)\right]}{\sqrt{\hat{V}(z_H) + \hat{V}(z_L)}} \to_d \mathcal{N}(0,1),$$

where $\hat{V}(z) \asymp J/(nT)$ is defined in equation (10) below.

The conditions shown formalize the bias-variance trade off restriction on the growth of J. The structure of the estimator implies that the variance of $\hat{\mu}(z_H) - \hat{\mu}(z_L)$ is the sum of each pointwise variance. Consistent variance estimation can be done in several ways, but in particular, we show that the commonly-used Fama and MacBeth (1973) variance estimator, given by

$$\hat{V}_{\mathtt{FM}}(z) = \frac{1}{T^2} \sum_{t=1}^{T} (\hat{\mu}_t(z) - \hat{\mu}(z))^2,$$

is indeed valid for Studentization. See Theorem 2 and the accompanying discussion below. To the best of our knowledge, these results are all new to the literature.

Beyond first-order validity, we also provide explicit, practicable guidance for choice of J via higher-order mean square error (MSE) expansions. To our knowledge, this represents the first theory-founded choice of J for implementing portfolio-sorted based inference: the literature employs *ad hoc* choices, and often $J = 10$, regardless of the data at hand (see quotation from Cochrane (2011) in the Introduction). In contrast, our results provide an objective and data-driven way of choosing the number of portfolios J in applications. For a typical application of portfolio sorting to US equities, we might have $n = 2000$ (on average) and $T = 600$, which our results below imply a value of J that is on the order of 200—suggesting that much larger choices of the number of portfolios may be needed than are currently considered in the literature.

Furthermore, it is important to emphasize that the optimal number of portfolios varies over time, reflecting the fact that cross-sectional sample sizes are very different across the sample period and so the number of portfolios needs to adjust in a systematic way. In fact, in Section 6 we show that the optimal number of portfolios for standard applications in empirical finance range from approximately 10 portfolios early in the sample and over 200 when the cross-sectional sample size is at its largest in the late 1990s. To make this clear notationally, we will write J_t for the number of portfolios in period t. In the context of hypothesis testing (HT), as in equation (4), we find that the optimal number of portfolios is

$$J_t^{\mathtt{HT}} = K^{\mathtt{HT}} n_t^{1/2} T^{1/4}, \qquad t = 1, 2, \cdots, T,$$

where the constant $K^{\mathtt{HT}}$ depends on the data generating process.[5] It is easy to check that $J_t^{\mathtt{HT}}$ satisfies the conditions of Corollary 1. In Section 5 we detail the constant terms and

[5]To simplify the calculations, and give relatively tractable expansions, we assume that the quantiles are known (as opposed to being estimated in each cross-section). This simplification only affects the constants of the higher-order terms in the MSE-expansion, but not the rates, and thus it enables us to provide relatively easy-to-implement plug-in rules for the choice of J_t in applications.

discuss implementation in applications.

Turning to factor construction, we find a different choice of J will be optimal, namely

$$J_t^{\mathtt{PE}} = K^{\mathtt{PE}} n_t^{1/3} T^{1/3}, \qquad t = 1, 2, \cdots, T,$$

where, again, portfolios are chosen separately at each time, $K_t^{\mathtt{PE}}$ depends on the data generating process, and implementation is discussed in Section 5. The major difference here is that for point estimation (PE), the optimal number of portfolios, $J_t^{\mathtt{PE}}$, diverges more slowly than for hypothesis testing, $J_t^{\mathtt{HT}}$ in typical applications where the cross-sectional sample size is much larger than the number of time-series observations. The bias-variance trade off, though still present of course, manifests differently because this is a point estimation problem, rather than one of inference. In particular, the divergence rate will often be slower. This formal choice is a further contribution of our paper, and is new to the literature. However, it does seem that, at least informally, the status quo is to use fewer portfolios for factor construction than for testing. See Remark 10 for further discussion.

Finally, we note that when z_H and z_L are always in the extreme portfolios, the estimator $\hat{\mu}(z_H) - \hat{\mu}(z_L)$, based on (3), is exactly the standard portfolio sorting estimator that enjoys widespread use in empirical finance, but that for estimation and inference about $\mu(z)$ at $z \neq \{z_L, z_H\}$, which again, is uncommon, our approach differs from the standard methods. Our estimator exploits the assumed structure that $\mu(z)$ is constant over time as a function of the characteristic value itself, i.e. as a function of z (see models (1) and (5); Assumption 1). This allows for a richer understanding of the relationship between returns and the underlying characteristics and the ability to investigate more general hypotheses of interest. In these broad terms then, the main contribution of our paper is a formal asymptotic treatment of the standard portfolio-sorts estimator $\hat{\mu}(z_H) - \hat{\mu}(z_L)$, but a further contribution is to show how portfolio sorting can be used for a much wider range of inference targets and correspondingly to allow for inference on additional testable hypotheses generated by theory (e.g., shape restrictions).

Remark 1 (Comparison to the Standard Portfolio Sorting Approach)**.** The standard application of portfolio sorting implicitly assumes that $\mu(\cdot)$ is constant over time as a function of the (random) cross-sectional order statistic of the characteristics, i.e. for any t_1, t_2 and any $q \in [0, 1]$, $\mu(z_{(\lfloor n_{t_1} q \rfloor) t_1}) = \mu(z_{(\lfloor n_{t_2} q \rfloor) t_2})$. We could accommodate this case but with substantial notational complexity. Moreover, the key insights obtained in this paper by formalizing and analyzing the portfolio sorting estimator would not be affected. □

Remark 2 (Analogy to Cross-Sectional Regressions)**.** The assumption that $\mu(z)$ is constant over time as a function of the characteristic value is perfectly aligned with the practice of cross-sectional (or Fama-MacBeth) regressions (Fama and MacBeth (1973)). This approach is motivated by a model of the form,

$$R_{it} = \zeta z_{it} + \varepsilon_{it}, \qquad i = 1, \ldots, n_t, \quad t = 1, \ldots, T,$$

where z_{it} is the value of the characteristic (or a vector of characteristics, more generally). Thus, cross-sectional regressions are then nested in equation (1) under the assumption that $\mu(\cdot)$ is linear in the characteristics (see also Remark 8 below). □

3 General Asset Returns Model and Sorting Estimator

In this section we study a more general model and develop a correspondingly general characteristic-sorted portfolio estimator. We extend beyond the simple case of the previous section in two directions. First, we allow for multiple sorting characteristics, such that z_{it} is replaced by $\mathbf{z}_{it} \in \mathcal{Z} \subset \mathbb{R}^d$. This extension is important because sorting on two variables is quite common in empirical work, and further, we can capture and quantify the empirical reality that sorting is very rarely done on more than two characteristics because this leads to empty portfolios. Intuitively, the nonparametric partitioning estimator, like all others, suffers from the curse of dimensionality, and performance deteriorates rapidly as d increases, as we can make precise (see also Remarks 5 and 8). To address this issue, our second generalization is to allow for other conditioning variables, denoted by $\mathbf{x}_{it} \in \mathbb{R}^{d_x}$, to enter the model in a flexible parametric fashion.

Formally, our model for asset returns is

$$R_{it} = \mu(\mathbf{z}_{it}) + \mathbf{x}_{it}'\boldsymbol{\beta}_t + \varepsilon_{it}, \qquad i = 1, 2, \cdots, n_t, \qquad t = 1, 2, \cdots, T. \tag{5}$$

This model retains the nonparametric structure on $\mu(\mathbf{z})$ as in equation (1), with the same interpretation (though now conditional on $\mathbf{x}_{it}$). Notice that the vector $\mathbf{x}_{it}$ may contain both basic conditioning variables as well as transformations thereof (e.g., interactions and/or power expansions), thus providing a flexible parametric approach to modeling these variables and providing a bridge to cross-sectional regressions from portfolio sorting. Cross-sectional regressions are popular because their linear structure means a larger number of variables can be incorporated compared to the nonparametric nature of portfolio sorting (i.e. cross-sectional regressions do not suffer the curse of dimensionality). Model (5) keeps this property while retaining the nonparametric flexibility and spirit of portfolio sorting. Indeed, the parameters $\boldsymbol{\beta}_t$ are estimable at the parametric rate, in contrast to the nonparametric rate for $\mu(\mathbf{z})$. The additive separability of the conditioning variables, common to both approaches, is the crucial restriction that enables this. Furthermore, due to the linear structure, the sorting estimator can be easily implemented via ordinary least squares, as discussed below.

As in the prior section, the main hypothesis of interest in the empirical finance literature is the presence of a large discrepancy in expected returns between a lower and a higher portfolio. To put (4) into the present, formalized notation, let $\mathbf{z}_L < \mathbf{z}_H$ be two values at or near the lower and upper (observed) boundary points.[6] We are then interested in testing $\mathsf{H}_0 : \mu(\mathbf{z}_H) - \mu(\mathbf{z}_L) = 0$ against the two-sided alternative. Of course, our results also cover other linear transformations such as the "diff-in-diff" approach: e.g., for $d = 2$, the estimand $\mu(\mathbf{z}_{1H}, \mathbf{z}_{2H}) - \mu(\mathbf{z}_{1H}, \mathbf{z}_{2L}) - (\mu(\mathbf{z}_{1L}, \mathbf{z}_{2H}) - \mu(\mathbf{z}_{1L}, \mathbf{z}_{2L}))$. See Nagel (2005) for an example of the latter, and Remark 6 below for further discussion on other potential hypotheses of interest. We will frame much of our discussion around the main hypothesis H_0 for concreteness, while still providing generic results that may be used for other inference targets.

The framework is completed with the following assumption governing the data-generating process. Further conditions required for asymptotic results are stated below.

Assumption 1 (Data-Generating Process)**.** *For the sigma fields $\mathcal{F}_t = \sigma(\mathbf{f}_t)$ generated from a sequence of unobserved (possibly dependent) random vectors $\{\mathbf{f}_t : t = 0, 1, \cdots, T\}$:*

(a) For each $t = 1, 2, \ldots, T$, $\{(R_{it}, \mathbf{z}'_{it}, \mathbf{x}'_{it})' : i = 1, 2, \cdots, n_t\}$ are i.i.d., conditional on $\mathcal{F}_t$;

(b) Model (5) holds with $\mathbb{E}[R_{it}|\mathbf{z}_{it}, \mathbf{x}_{it}, \mathcal{F}_t] = \mu(\mathbf{z}_{it}) + \mathbf{x}'_{it}\boldsymbol{\beta}_t$;

(c) The support of $\mathbf{z}_{it}$, denoted $\mathcal{Z}$, is time-invariant and the product of compact, convex intervals.

These conditions allow for considerable flexibility in the behavior of the time series of returns and the cross-sectional dependence. Indeed, Andrews (2005, p. 1552), using the same condition in a single cross-section, called Assumption 1(a) "surprisingly general".[7] The set up allows for dependence and conditional heteroskedasticity across assets and time. For example, if $\mathbf{f}_t$ were to include a business cycle variable then we could allow for a common business-cycle component in the idiosyncratic variance of returns. As another example, the sampling assumptions allow for a factor structure in the $\mathbf{z}_{it}$ variables. Perhaps most importantly, we do not impose that returns are independent or even uncorrelated over time. Our assumptions accommodate momentum or reversal effects whereby an asset's past *relative* return predicts its future *relative* return, which corresponds to lagged returns entering $\mathbf{z}_{it}$ (see, for example, Bondt and Thaler (1985), Jegadeesh (1990), Lehmann (1990), Jegadeesh and Titman (1993, 2001)). Finally, Assumption 1(c) is needed in order to form the portfolios and allows the density to be bounded away from zero, ensuring they are not empty. The

[6] In practice, the essential ingredient is that the elements of $\mathbf{z}_L$ and $\mathbf{z}_H$ are not in the same portfolio. See Section 4 for more discussion.

[7] Gagliardini and Gourieroux (2014) and Gagliardini et al. (2015) impose similar conditions with applications to panel data in finance.

rectangular structure is without loss of generality for first-order asymptotics but will affect the constants of the mean square error expansions. Finally, the assumption that the support of the characteristics is the same across time-series observations is a common assumption when studying panel data.

In the context of model (5), the portfolio sorting estimator of $\mu(\mathbf{z})$ retains the structure given above in (3), but first the conditioning variables must be projected out. Thus, the cross-sectional estimator $\hat{\mu}_t(\mathbf{z})$ can be constructed by simple ordinary least squares: regressing R_{it} on J_t^d dummies indicating whether $\mathbf{z}_{it}$ is in portfolio j, along with the d_x control variables $\mathbf{x}_{it}$. Note that, in contrast to Section 2, we allow $J = J_t$ to vary over time, in line with having an unbalanced panel. This is particularly important for applications to equities as these data tend to be very unbalanced with cross sections much larger later in the sample as they are at the beginning of the sample. For example, in our empirical applications the largest cross-sectional sample size is approximately fifteen times that of the smallest cross-sectional sample size.

The multiple-characteristic portfolios are formed as the Cartesian products of marginal intervals. That is, we first partition each characteristic into J_t intervals, using its *marginal* quantiles, exactly as in equation (2), and then form J_t^d portfolios by taking the Cartesian products of all such intervals. We retain the notation $P_{jt} \subset \mathbb{R}^d$ for a typical portfolio, where here $j = 1, 2, \ldots, J_t^d$. For $d > 1$, even if $J^d < n$, these portfolios are not uniformly guaranteed to contain any assets, and this concern for "empty" portfolios can be found in the empirical literature (see, for example, Goyal (2012, p. 31)). Our construction mimics empirical practice, and we formalize the constraints on J that ensure nonempty portfolios (a variance condition), while simultaneously controlling bias. While the problem of a large J implying empty portfolios has been recognized (though never studied), the idea of controlling bias appears to be poorly understood. However, in our framework the nonparametric bias arises naturally and is amenable to study. Conditional sorts have been used to "overcome" the empty portfolio issue, but these are problematic for many reasons (see Remark 5).

With the portfolios thus formed, we can define the final portfolio sorting estimator of $\mu(\mathbf{z})$, for a point of interest $\mathbf{z} \in \mathcal{Z}$. First, with an eye to reinforcing the estimated portfolio breakpoints, for a given portfolio P_{jt}, $j = 1, 2, \ldots, J_t^d, t = 1, \ldots, T$, let $\hat{\mathbb{1}}_{jt}(\mathbf{z}) = \mathbb{1}\{\mathbf{z} \in P_{jt}\}$ indicate that the point $\mathbf{z}$ is in P_{jt}, and let $N_{jt} = \sum_{i=1}^{n_t} \hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})$ denote its (random) sample size. The portfolio sorting estimator is then defined as

$$\hat{\mu}(\mathbf{z}) = \frac{1}{T}\sum_{t=1}^{T} \hat{\mu}_t(\mathbf{z}), \qquad \hat{\mu}_t(\mathbf{z}) = \sum_{j=1}^{J_t^d} \frac{1}{N_{jt}} \sum_{i=1}^{n_t} \hat{\mathbf{1}}_{jt} \hat{\mathbb{1}}_{jt}(\mathbf{z}) \hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})(R_{it} - \mathbf{x}_{it}' \hat{\boldsymbol{\beta}}_t), \tag{6}$$

where

$$\hat{\boldsymbol{\beta}}_t = (\mathbf{X}_t'\mathbf{M}_t\mathbf{X}_t)^{-1}\mathbf{X}_t'\mathbf{M}_t\mathbf{R}_t, \qquad \mathbf{R}_t = [R_{1t}, \ldots, R_{n_t t}]',$$
$$\mathbf{X}_t = [\mathbf{x}_{1t}, \mathbf{x}_{2t}, \cdots, \mathbf{x}_{n_t t}]', \qquad \mathbf{M}_t = \mathbf{I}_{n_t} - \hat{\mathbf{B}}_t(\hat{\mathbf{B}}_t'\hat{\mathbf{B}}_t)^{-1}\hat{\mathbf{B}}_t', \tag{7}$$

and $\hat{\mathbf{B}}_t = \hat{\mathbf{B}}_t(\mathbf{z}_t)$ with $\mathbf{z}_t = [\mathbf{z}_{1t}, \mathbf{z}_{2t}, \cdots, \mathbf{z}_{n_t t}]'$ is the $n_t \times J_t^d$ matrix with (i, j) element equal to $\hat{\mathbb{1}}_{jt}(z_{it})$, characterizing the portfolios for the characteristics z_{it}. The indicator function $\hat{\mathbf{1}}_{jt}$ ensures that all necessary inverses exist, and thus takes the value one if P_{jt} is nonempty and $(\mathbf{X}_t'\mathbf{M}_t\mathbf{X}_t/n_t)^{-1}$ is invertible. Both events occur with probability approaching one (see the Supplementary Appendix). It is established there that $N_{jt} \asymp n_t/J_t^d$ with probability approaching one, for all j and t.

Remark 3 (Implementation and Weighted Portfolios)**.** Despite the notational complexity, the estimator $\hat{\mu}_t(\mathbf{z})$ is implemented as a standard linear regression of the outcome R_{it} on the $J_t^d + d_x$ covariates $\hat{\mathbf{B}}_t$ and $\mathbf{x}_{it}$. It is the product of the indicator functions $\hat{\mathbb{1}}_{jt}(\mathbf{z})\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})$ that enforces the nonparametric nature of the estimator: only $\mathbf{z}_{it}$ in the same portfolio as $\mathbf{z}$, and hence "close", are used. The estimator can easily accommodate weighting schemes, such as weighting assets by market capitalization or inversely by their estimated (conditional) heteroskedasticity. For notational simplicity we present all theoretical results without portfolio weights, but all empirical results in Section 6 are based on the value-weighted portfolio estimator. □

It worth emphasizing that the nonparametric estimator $\hat{\mu}_t(\mathbf{z})$ of (6) is nonstandard. At first glance, it appears to be the nonparametric portion of the usual partially linear model, using the partitioning regression estimator as the first stage ($\hat{\boldsymbol{\beta}}_t$ would be the parametric part). However, the partitioning estimator here is formed using estimated quantiles, which makes the "basis" functions of our nonparametric estimator nonstandard, and renders prior results from the literature inapplicable.

Remark 4 (Connection to Other Anomalies Adjustments)**.** A number of authors have attempted to control for existing anomalies by first regressing their proposed anomaly variable on existing variables, and sorting on the residuals (Chen et al., 2002; Hong et al., 2000; Hou and Moskowitz, 2005; Nagel, 2005; Fama and French, 2008; Han and Lesmond, 2011; Wahal and Yavuz, 2013). This is fundamentally (and analytically) different from the estimator we propose in this paper because it implicitly assumes that returns are generated by a model of the form

$$R_{it} = \mu(\mathbf{z}_{it} - \mathbf{x}_{it}'\boldsymbol{\gamma}) + \epsilon_{it},$$

for some unknown parameter $\boldsymbol{\gamma}$ and unobserved disturbance ϵ_{it}. Therefore, this alternative approach assumes an arguably less flexible functional form for the relationship between returns and characteristics. □

Remark 5 (Conditional sorts)**.** A common practice in empirical finance is to perform what are called "conditional" portfolio sorts. These are done by first sorting on one characteristic, and then within each portfolio separately, sorting on a second characteristic, and so forth (usually only two characteristics are considered). In each successive sort, quantile-spaced portfolios are used. In this way, conditional sorts "solve" the empty portfolios problem by construction. However, in this case the characteristic values being compared are not guaranteed to be similar.

To fix ideas, consider sorting first on firm size and then, conditionally, on credit rating. Small firms are less likely to have high credit ratings, and large firms will typically not have low ratings. Thus, even the "top" portfolio for small firms may contain credit ratings lower than the "bottom" portfolio among the large firms.

In this formulation of portfolio sorting, it is implicitly assumed that the function $\mu(\mathbf{z})$ is constant over time as a function of the conditional order statistics, within each portfolio. This is difficult to treat theoretically, as the (population) assumption on $\mu(\mathbf{z})$ must hold for *each* conditional sort for the (estimated) portfolios already constructed. Moreover, it is not clear that this approach can be extended to other interesting estimands. Finally, it would likely be challenging for an economic theory to generate such a constrained (conditional) return generating process.

However, an alternative, and arguably more transparent approach would be to assume additive separability of the function $\mu(\cdot)$ so that

$$R_{it} = \mu_1(z_{1,it}) + \cdots + \mu_d(z_{d,it}) + \varepsilon_{it} \qquad i = 1, \ldots, n_t, \quad t = 1, \ldots, T. \tag{8}$$

and so each characteristic affects returns via their own unknown function, $\mu_\ell(\cdot)$, for $\ell = 1, \ldots, d$. The resulting estimator is always defined for any value $\mathbf{z}$ in the support and so too avoids the problem of empty portfolios (see also Remark 8).

□

4 First-order Asymptotic Theory

With the estimator fully described we now present consistency and asymptotic normality results, and several valid standard error estimators. To our knowledge, these results are all new to the literature. As discussed in Section 2 the empirical literature contains numerous studies that implement exactly the tests validated by the results below, but such validation has heretofore been absent.

Beyond the definition of the model (5) and the conditions placed upon it by Assumption 1, we will require certain regularity conditions and rate restrictions for our asymptotic results. We now make these precise, grouped into the following three assumptions.

Assumption 2 (Regularity Conditions)**.**

(a) The function $\mu(\mathbf{z})$ is continuously differentiable.

(b) For each $t = 1, 2, \ldots, T$, $(R_{it}, \mathbf{z}'_{it}, \mathbf{x}'_{it})'$ has a density $g_t(R, \mathbf{z}, \mathbf{x}; \mathcal{F}_t)$ that is bounded and bounded away from zero.

(c) Uniformly in i and t, $\sigma^2_{it} := \mathbb{E}[|R_{it}|^2 | \mathbf{z}_{it}, \mathbf{x}_{it}, \mathcal{F}_t]$ is bounded and bounded away from zero and $\mathbb{E}[|R_{it}|^{2+\phi} | \mathbf{z}_{it}, \mathbf{x}_{it}, \mathcal{F}_t]$ is bounded for some $\phi > 0$.

(d) For all $\mathbf{a} \in \mathbb{R}^{d_x}$, $\mathbf{a}'\mathbf{x}_{it}$ is sub-Gaussian conditional on $\mathbf{z}_{it}$, $\mathcal{F}_t$ and the conditional expectation of $\mathbf{x}_{it}$ given $\mathbf{z}_{it}$, $\mathcal{F}_t$ is Lipschitz continuous.

(e) The minimum eigenvalue of $\Omega_{\mathrm{uu},t} := \mathbb{E}[\mathbb{V}(\mathbf{x}_{it} | \mathbf{z}_{it}, \mathcal{F}_t)]$ is bounded away from zero.

Assumption 2 collects regularity conditions which are standard in the (cross-sectional) semi- and nonparametric literature. These conditions are not materially stronger than typically imposed, despite the complex nature of the estimation and the use of an estimated set of basis functions in the nonparametric step (due to the estimated quantiles). Assumption 2(a) is a mild smoothness restriction which yields a nonparametric smoothing bias of order $1/J$, while Assumption 2(b) is more related to variance, ensuring that all portfolios are (asymptotically) nonempty. Assumption 2(d) allows for continuous conditioning variables. The remainder collect standard moment conditions.

Assumption 3 (Panel Structure)**.** *The cross-sectional sample sizes diverge proportionally: for a sequence $n \to \infty$, $n_t = \kappa_t n$, with $\kappa_t \leq 1$ and bounded away from zero uniformly in $t \leq T$.*

Assumption 3 requires that the cross-sectional sample sizes grow proportionally. This ensures that each $\hat{\mu}_t(\cdot)$ contributes to the final estimate, and at the same rate. We will also restrict attention to $J_t = J_t(n_t, n, T)$, which implies there is a sequence $J \to \infty$ such at $J_t \propto J$ for all t. Neither of these are likely to be limiting in practice: our optimal choices depend on n_t by design, and there is little conceptual point in letting J_t vary over time beyond accounting for panel imbalance. The notation n and J for common growth rates enables us to present compact and simplified regularity conditions, such as the following assumption, which formalizes the bias-variance requirements on the nonparametric estimator. All limits are taken as $n, T \to \infty$, unless otherwise noted.

Assumption 4 (Rate Restrictions)**.** *The sequences n, T, and J obey: (a) $n^{-1}J^d \log(\max(J^d, T))^2 \to 0$, (b) $\sqrt{nT}J^{-(d/2+1)} \to 0$, and, if $d_x \geq 1$, (c) $T/n \to 0$.*

Assumption 4(a) ensures that all J_t grow slowly enough that the variance of the nonparametric estimator is well-controlled and all portfolios are nonempty, while, 4(b) ensures the nonparametric smoothing bias is negligible. Finally, Assumption 4(c) restricts the rate at which T can grow. This additional assumption is necessary for standard inference when linear conditioning variables are included in the model and $d = 1$. When $d > 1$ then it is implied by Assumptions 4(a) and 4(b).

In general, the performance of the portfolio sorting estimator may be severely compromised if the number of time series observations is large relative to the cross section and/or d is large. To illustrate, suppose for the moment that $J \asymp n^A$ and $T \asymp n^B$. Assumptions 4(a) and (b) require that $A \in ((1+B)/(2+d)\ ,\ 1/d)$, which amounts to requiring $Bd < 2$. If the time series dimension is large, then the number of allowable sorting characteristics is limited. For example, if B is near one, at most two sorting characteristics are allowed, and even then just barely, and may lead to a very poor distributional approximation. Thus, some caution should be taken when applying the estimator to applications with relatively few underlying assets.

Before stating the asymptotic normality result, it is useful to first give an explicit (conditional) variance formula:

$$V(\mathbf{z}) := \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t^d}\frac{1}{N_{jt}}\sum_{i=1}^{n_t}\hat{\mathbf{1}}_{jt}\hat{\mathbb{1}}_{jt}(\mathbf{z})\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\sigma_{it}^2. \tag{9}$$

This formula, and the distributional result below, are stated for a single point $\mathbf{z}$. It is rare that a single $\mu(\mathbf{z})$ would be of interest, but these results will serve as building blocks for more general parameters of interest, such as the leading case of testing (4) treated explicitly below. An important consideration in any such analysis is the covariance between point estimators. The special structure of the portfolio sorting estimator (or partition regression estimator) is useful here: as long as $\mathbf{z}$ and $\mathbf{z}'$ are in different portfolios (which is the only interesting case), $\hat{\mu}(\mathbf{z})$ and $\hat{\mu}(\mathbf{z}')$ are uncorrelated because $\hat{\mathbb{1}}_{jt}(\mathbf{z})\hat{\mathbb{1}}_{jt}(\mathbf{z}') \equiv 0$. The partitioning estimator is, in this sense, a local nonparametric estimator as opposed to a global smoother.

We have the following result.

Theorem 1 (Asymptotic Distribution)**.** *Suppose Assumptions 1–4 hold. Then, for each $\mathbf{z} \in \mathcal{Z}$,*

$$V^{-1/2}(\mathbf{z})(\hat{\mu}(\mathbf{z}) - \mu(\mathbf{z})) = \sum_{t=1}^{T}\sum_{i=1}^{n_t}\hat{w}_{it}(\mathbf{z})\varepsilon_{it} + o_{\mathbb{P}}(1) \to_d \mathcal{N}(0,1),$$

where

$$V(\mathbf{z}) \asymp \frac{J^d}{nT} \qquad \text{and} \qquad \hat{w}_{it}(\mathbf{z}) = V^{-1/2}(\mathbf{z}) \sum_{j=1}^{J_t^d} \frac{1}{TN_{jt}} \hat{\mathbf{1}}_{jt} \hat{\mathbb{1}}_{jt}(\mathbf{z}) \hat{\mathbb{1}}_{jt}(\mathbf{z}_{it}).$$

Theorem 1 shows that the properly normalized and centered estimator $\hat{\mu}(z)$ has a limiting normal distribution. The cost of the flexibility of the nonparametric specification between returns and (some) characteristics comes at the expense at slower convergence — the factor $J^{-d/2}$. Theorem 1 also makes clear why Assumption 4(b) is necessary: the bias of the estimator is of the order J^{-1} and thus, once the rate $J^{-d/2}\sqrt{nT}$ is applied, Assumption 4(b) must hold to ensure that the bias can be ignored for the limiting normal distribution. This undersmoothing approach is typical for bias removal. The statement of the theorem includes a weighted average asymptotic representation for the estimator, which is useful for treatment of estimands beyond point-by-point $\mu(\mathbf{z})$, including linear functionals such as partial means, as discussed in Remark 6.

The final missing piece of the pointwise first-order asymptotic theory is a valid standard error estimator. To this end, we consider two options. The first, due in this context to Fama and MacBeth (1973), makes use of the fact that $\hat{\mu}(\mathbf{z})$ is an average over T "observations", while the second is a plug-in estimator based on an asymptotic approximation to the large sample variability of the portfolio estimator. Define

$$\hat{V}_{\texttt{FM}}(\mathbf{z}) = \frac{1}{T^2} \sum_{t=1}^{T} (\hat{\mu}_t(\mathbf{z}) - \hat{\mu}(\mathbf{z}))^2 \quad \text{and} \quad \hat{V}_{\texttt{PI}}(\mathbf{z}) = \frac{1}{T^2} \sum_{t=1}^{T} \sum_{j=1}^{J_t^d} \sum_{i=1}^{n_t} \hat{\mathbf{1}}_{jt} \frac{1}{N_{jt}^2} \hat{\mathbb{1}}_{jt}(\mathbf{z}) \hat{\mathbb{1}}_{jt}(\mathbf{z}_{it}) \hat{\varepsilon}_{it}^2 \tag{10}$$

with $\hat{\varepsilon}_{it} = R_{it} - \hat{\mu}(\mathbf{z}) - \mathbf{x}_{it}' \hat{\boldsymbol{\beta}}_t$. The following result establishes the validity of both these choices.

Theorem 2 (Standard Errors)**.** *Suppose the assumptions of Theorem 1 hold with $\phi = 2 + \varrho$ for some $\varrho > 0$. Then for fixed $\mathbf{z}$,*

$$\frac{nT}{J^d}(\hat{V}_{\texttt{FM}}(\mathbf{z}) - V(\mathbf{z})) \to_{\mathbb{P}} 0, \qquad \text{and} \qquad \frac{nT}{J^d}(\hat{V}_{\texttt{PI}}(\mathbf{z}) - V(\mathbf{z})) \to_{\mathbb{P}} 0.$$

The Fama and MacBeth (1973) variance estimator is commonly used in empirical work, but this is the first proof of its validity. In contrast, $\hat{V}_{\texttt{PI}}$ is the "plug-in" variance estimator based on the results in Theorem 1. Theorem 2 shows that these variance estimators are asymptotically equivalent. In a fixed sample, it is unclear which of the two estimators is preferred. $\hat{V}_{\texttt{FM}}$ is simple to implement and very popular, while $\hat{V}_{\texttt{PI}}$ is based on estimated residuals and may need a large cross-section. On the other hand, while we assume T diverges, in line with common applications of sorting, it may be established that $\hat{V}_{\texttt{PI}}$ is valid for fixed

T, whereas $\hat{V}_{\texttt{FM}}$ is only valid for large-T panels. However, a related result is due to Ibragimov and Müller (2010), who provided conditions under which the Fama and MacBeth (1973) approach applied to cross-sectional regressions produces inference on a scalar parameter that is valid or conservative, depending on the assumptions imposed.[8] Our empirical results in Section 6 use $\hat{V}_{\texttt{FM}}$ to form test statistics so as to be comparable to existing results in the literature. In general, a consistent message of our results is that caution is warranted in cases applying portfolio sorting to applications with a very modest number of time periods or, as discussed above, when the number of time periods is "large" relative to the cross-sectional sample sizes.

Theorems 1 and 2 lead directly to the following result, which treats the main case of interest under simple and easy-to-interpret conditions.

Corollary 2. *Let the conditions of Theorem 1 hold. Then*

$$\frac{\big[\hat{\mu}(\mathbf{z}_H)-\hat{\mu}(\mathbf{z}_L)\big]-\big[\mu(\mathbf{z}_H)-\mu(\mathbf{z}_L)\big]}{\sqrt{\hat{V}(\mathbf{z}_H)+\hat{V}(\mathbf{z}_L)}}\to_d \mathcal{N}(0,1),$$

where $\hat{V}(\mathbf{z})$ may be $\hat{V}_{\texttt{FM}}$ or $\hat{V}_{\texttt{PI}}$ as defined in Equation (10).

Corollary 1 is the same result, simplified to the model (1). This result shows that testing $\mathsf{H}_0 : \mu(\mathbf{z}_H)-\mu(\mathbf{z}_L)=0$ against the two-sided alternative can proceed as standard: by rejecting H_0 if $|\hat{\mu}(\mathbf{z}_H)-\hat{\mu}(\mathbf{z}_L)|$ greater than $1.96\times\sqrt{\hat{V}(\mathbf{z}_H)+\hat{V}(\mathbf{z}_L)}$. In this way, our work shows under precisely what conditions the standard portfolio sorting approach is valid, and perhaps more importantly, under what conditions it may fail.

Remark 6 (Other Estimands)**.** As we have discussed above, our general framework allows for other estimands aside from the "high minus low" return. For example, a popular estimand in the literature that may be easily treated by our results is the case of partial means, which arises when $d>1$. If we denote the d components of $\mathbf{z}$ by $z^{(1)},z^{(2)},\dots,z^{(d)}$, then for some subset of these of size $\delta<d$, the object of interest is

$$\int_{\times_{\ell=1}^{\delta}}\mu(\mathbf{z})w\big(z^{(1)},z^{(2)},\dots,z^{(\delta)}\big)dz^{(1)}dz^{(2)}\cdots dz^{(\delta)},$$

where the components of $\mathbf{z}$ that are not integrated over are held fixed at some value, or linear combinations for different initial $\mathbf{z}$ points. Prominent examples are the SMB and HML factors of the Fama/French 3 Factors.[9] For recent examples see Fama and French

[8]Specifically, Ibragimov and Müller (2010), in the context of cross-sectional regressions, show that for fixed T and a specific range of size-α tests, the Fama and MacBeth (1973) approach is valid, but potentially conservative.

[9]Available at `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`

(2015) or Hou et al. (2014). The weighting function $w(\cdots)$ is often taken to be the uniform density (based on value-weighted portfolios), but this need not be the case. For example, if $d = 2$, one component may be integrated over before testing the analogous hypothesis to (4):

$$\mathsf{H}_0 : \int_{z^{(1)}} \mu\big(z^{(1)}, z_H^{(2)}\big) w\big(z^{(1)}\big) dz^{(1)} - \int_{z^{(1)}} \mu\big(z^{(1)}, z_L^{(2)}\big) w\big(z^{(1)}\big) dz^{(1)} = 0.$$

In the case of factor construction this corresponds to a test of whether a factor is priced unconditionally. Theorems 1 and 2 can be applied to this case to provide valid inference. $\square$

Remark 7 (Strong Approximations)**.** Our asymptotic results apply to hypothesis tests that can be written as pointwise transformations of $\mu(\mathbf{z})$, with the leading case being (4): $\mathsf{H}_0 : \mu(\mathbf{z}_H) - \mu(\mathbf{z}_L) = 0$. However, there are other hypotheses of interest in this context of portfolio sorting that require moving beyond pointwise results. Chief among these is directly testing monotonicity of $\mu(\cdot)$, rather than using $\mu(\mathbf{z}_H) - \mu(\mathbf{z}_L)$ as a proxy (see discussion in Section 2). Employing coupling results as in Eggermont and LaRiccia (2009, Chapter 22) or Belloni, Chernozhukov, Chetverikov, and Kato (2015), it is possible to establish a valid strong approximation to the suitable centered and scaled stochastic process $\{\hat{\mu}(\mathbf{z}) : \mathbf{z} \in \mathcal{Z}\}$. Such a result would require non-trivial additional technical work, but would allow us to conduct valid asymptotic hypothesis testing for may other non-standard hypothesis of interest such as testing for a "U-shaped" relationship as in Hong et al. (2000) or for the existence of any profitable trading strategy via $\mathsf{H}_0 : |\max_z \mu(\mathbf{z}) - \min_z \mu(\mathbf{z})| = 0$. $\square$

Remark 8 (Analogy to Cross-Sectional Regressions (cont'd))**.** As we have discussed in Remark 2, cross-sectional regressions are the "parametric alternative" to portfolio sorting. In practice, however, the more natural parametric alternative to portfolio sorts with more than one sorting variable—interaction effects in the linear specification—are rarely utilized. Thus the more exact "nonparametric counterpart" to the common implementation of cross-sectional regressions is the additively separable model introduced in equation (8) of Remark 5. The assumption of additive separability would have the effect of ameliorating the "curse of dimensionality"; in fact, it can be shown that in this model the rate restrictions $J \log(\max(J,T))/n \to 0$ and $nT/J^3 \to 0$ (i.e., Assumption 4 when $d = 1$) are sufficient to ensure consistency and asymptotic normality of the estimators, $\hat{\mu}_\ell(z)$, based on the additively separable model with $d \geq 1$ characteristics. $\square$

5 Mean Square Error Expansions and Practical Guidance

With the first-order theoretical properties of the portfolio sorting estimator established, we now turn to issues of implementation. Chief among these is choice of the number of portfolios: with the estimator defined as in equation (6), all that remains for the practitioner is to choose J_t. The results in the previous two sections have emphasized the key role played by choice of J_t in obtaining valid inference. In contrast, the choice of J_t in empirical studies has been *ad hoc*, and almost always set to either 5 or 10 portfolios, without a well-grounded justification. Here we will provide simple, data-driven rules to guide the choice of the number of parameters. To aid in this, we will consider a mean square error expansion for the portfolio estimator, with a particular eye toward testing the central hypothesis of interest: $\mathsf{H}_0 : \mu(\mathbf{z}_H) - \mu(\mathbf{z}_L) = 0$, as the starting point for constructing a plug-in optimal choice.

Our main result for this section is the following characterization of the mean square error of the portfolio sorting estimator. Recall that n and J represent the common growth rates of the $\{n_t\}$ and $\{J_t\}$, respectively.

Theorem 3. *Suppose Assumptions 1, 2, and 3 hold, $J \to \infty$, $n^{-1} J^d \log(\max(J^d, T)) \to 0$, and, if $d_x \geq 1$, then $T/n \to 0$. Then*

$$\mathbb{E}\left[\left(\left[\hat{\mu}(\mathbf{z}_H) - \hat{\mu}(\mathbf{z}_L)\right] - \left[\mu(\mathbf{z}_H) - \mu(\mathbf{z}_L)\right] \right)^2 \Big| \mathfrak{Z}, \mathfrak{X} \right] = \mathcal{V}^{(1)} \frac{J^d}{nT} + \mathcal{V}^{(2)} \frac{J^{2d}}{n^2 T} + \mathcal{B}^2 \frac{1}{J^2} + O_p\left(\frac{1}{nT}\right) + o_p\left(J^{-2} + \frac{J^{2d}}{n^2 T}\right),$$

where $\mathfrak{Z} = (z_{11}, \ldots, z_{n_T T})$, $\mathfrak{X} = (x_{11}, \ldots, x_{n_T T})$ and $\mathcal{B} = \sum_{t=1}^T \mathcal{B}_t(\mathbf{z}_H) - \sum_{t=1}^T \mathcal{B}_t(\mathbf{z}_L)$ and $\mathcal{V}^{(\ell)} = \sum_{t=1}^T \mathcal{V}_t^{(\ell)}(\mathbf{z}_L) + \sum_{t=1}^T \mathcal{V}_t^{(\ell)}(\mathbf{z}_H)$, $\ell \in \{1, 2\}$, and $\mathcal{B}_t(\mathbf{z})$, $\mathcal{V}_t^{(1)}(\mathbf{z})$, and $\mathcal{V}_t^{(2)}(\mathbf{z})$ are defined in the Supplementary Appendix. The term of order $1/(nT)$ captures the limiting variability of $\sqrt{n/T} \sum_{t=1}^T (\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t)$, and does not depend on J.

Under the conditions in Theorem 3, and imposing different possible regularity conditions on the time series structure (e.g., mixing conditions), it is easy to show that

$$\bar{\mathcal{B}} = \mathrm{plim}_{n,T\to\infty} \mathcal{B}, \qquad \bar{\mathcal{V}}^{(1)} = \mathrm{plim}_{n,T\to\infty} \mathcal{V}^{(1)}, \qquad \bar{\mathcal{V}}^{(2)} = \mathrm{plim}_{n,T\to\infty} \mathcal{V}^{(2)},$$

where $\bar{\mathcal{B}}$, $\bar{\mathcal{V}}^{(1)}$ and $\bar{\mathcal{V}}^{(2)}$ are non-random and non-zero quantities. In this paper, however, we remain agnostic about the specific regularity conditions for convergence in probability to occur because our methods do not rely on them.

To obtain an optimal choice for the number of portfolios, note that the first variance term of the expansion will match the first-order asymptotic variance of Theorem 1, which suggests choosing J to jointly minimize the next two terms of the expansion: the bias and higher order variance. This approach is optimal in an inference-targeted MSE sense because it minimizes the two leading terms not accounted for by the large sample approximation of Theorem 1. For testing $\mathsf{H}_0 : \mu(\mathbf{z}_H) - \mu(\mathbf{z}_L) = 0$ we find the optimal number of portfolios to be

$$J_t^\star = \left\lfloor \left(\frac{\bar{\mathcal{B}}^2}{d\bar{\mathcal{V}}^{(2)}} \left(n_t^2 T\right) \right)^{\frac{1}{2d+2}} \right\rfloor, \tag{11}$$

where $\lfloor \cdot \rfloor$ is the integer part of the expression. A simple choice for enforcing the same number of portfolios in all periods is to simply replace n_t with n in this expression. It is straightforward to verify that this choice of $J_t^\star$ satisfies Assumption 4: the condition required remains that $Bd < 2$, for $T \asymp n^B$, which limits the number of sorting characteristics and/or the length of time series allowed (see discussion of Assumption 4).

To make this choice practicable we can select J to minimize a sample version of the MSE expansion underlying equation (11),

$$\widehat{\mathrm{MSE}}\big(\hat{\mu}(\mathbf{z}_H) - \hat{\mu}(\mathbf{z}_L); J\big) = \hat{\mathcal{V}}^{(2)} \frac{J^{2d}}{n^2 T} + \hat{\mathcal{B}}^2 \frac{1}{J^2} \tag{12}$$

where the estimators, $\hat{\mathcal{V}}^{(2)}$ and $\hat{\mathcal{B}}$, will themselves be a function of J. Thus, it is straightforward to search over a grid of values of J and choose based on the minimum value of the expression in equation (12).[10] Alternatively, if we had pilot estimates of $\mathcal{V}^{(2)}$ and $\mathcal{B}$, then we could directly utilize the formula in equation (11) to obtain a choice for each J_t.

Remark 9 (Undersmoothing). (a) A common practice throughout semi- and non-parametric analyses is to select a tuning parameter by undersmoothing a mean square error optimal choice. In theory, this is feasible, but it is necessarily *ad hoc*. In contrast, the choice of $J_t^\star$ of equation (11) has the advantage of being optimal in an objective sense and appropriate for conducting inference. A possible alternative to $J_t^\star$ would be to choose J by balancing $\left|\bar{\mathcal{B}}\right|$ against $\bar{\mathcal{V}}^{(1)}$; however, this would lead to a choice of $J_t \propto (n_t T)^{\frac{1}{d+1}}$ which would tend to result in a larger number of portfolios chosen as compared to $J_t^\star$.

(b) An additional advantage of $J_t^\star$ is that for $d \leq 2$ (the most common case in empirical applications) inference on the parametric component is also valid for this choice of J.

[10] The full details of the implementation of the optimal choice of number of portfolios for our empirical applications are given in the Supplementary Appendix.

It can be shown that for any real, nonzero vector $\mathbf{a} \in \mathbb{R}^{d_x}$,

$$\frac{\frac{1}{T}\sum_{t=1}^{T}\mathbf{a}'(\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t)}{\sqrt{\frac{1}{T^2}\sum_{t=1}^{T}(\mathbf{a}'(\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t))^2}} \to_{\mathrm{d}} \mathcal{N}(0,1) \tag{13}$$

An advantage of the Fama and MacBeth (1973) variance estimator over a "plug-in" alternative in this context is that inference on $\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\beta}_t$ may be conducted without having to estimate the conditional expectation of $\mathbf{x}$ given $\mathbf{z}$ nonparametrically.

□

Remark 10 (Constructing factors)**.** When point estimation rather than inference is of interest, Theorem 3 can be applied to give an optimal choice of J_t, but here using the leading variance and bias terms. Instead of $J_t^\star$ as in equation (11), the optimal choice here will be

$$J_t^{\star\star} = \left\lfloor \left(\frac{2\bar{\mathcal{B}}^2}{d\bar{\mathcal{V}}^{(1)}}(n_t T) \right)^{\frac{1}{d+2}} \right\rfloor,$$

which is different in the constants but more importantly also the rate of divergence: for example, when $d = 1$ then $J_t^{\star\star} \propto n_t^{1/3}T^{1/3}$ whereas $J_t^\star \propto n_t^{1/2}T^{1/4}$. In applications such as for equities where the cross sectional sample size is much larger than the number of time periods then it will be the case that $J_t^{\star\star} = o(J_t^\star)$, i.e., that the optimal number of portfolios is smaller when constructing factors than when conducting inference on whether expected returns vary significantly with characteristics. This point has, at least informally, been recognized in the empirical literature as the number of portfolios used to construct factors has been relatively small (see, for example, Fama and French (1993)). □

6 Empirical Applications

In this section we revisit some notable equity anomaly variables that have been considered in the literature and demonstrate the empirical relevance of the theoretical discussion of the previous sections. We focus on the size anomaly (e.g., Banz (1981), Reinganum (1981)) and the momentum anomaly (e.g., Jegadeesh and Titman (1993)).

6.1 Data and Variable Construction

We use monthly data from the Center for Research in Security Prices (CRSP) over the sample period January 1926 to December 2015. We restrict these data to those firms listed on the New York Stock Exchange (NYSE), American Stock Exchange (AMEX), or Nasdaq and use only returns on common shares (i.e., CRSP share code 10 or 11). To deal with delisting returns we follow the procedure described in Bali et al. (2016) based on Shumway (1997). When forming market equity we use quotes when closing prices are not available and set to missing all observations with 0 shares outstanding. When forming the momentum variable we follow the popular convention of defining momentum by the cumulative return from 12 months ago (i.e, $t-12$) until one month prior to the current month (i.e., $t-2$).[11] We set to missing this variable if any monthly returns are missing over the period. We also construct an industry momentum variable. To do so we use the definitions of the 38 industry portfolios used in Ken French's data library which are based on four digit SIC codes.[12] To construct the industry momentum variable we form a value weighted average of each individual firm's momentum variable within the industry. We use 13-month lagged market capitalization to form weights so they are unaffected by any subsequent changes in price.

We implement the estimator introduced in Section 3 as follows. Since the underlying data are monthly, then portfolios are always formed and then rebalanced at the end of each month. All portfolios, including those based on the standard implementation approach, are value weighted using lagged market equity. We implement the estimators based on the number of portfolios which minimizes our higher-order MSE criterion, described in equation (12) (further details are available in the Supplementary Appendix) since our objective in this section is inference.

Finally, it is important to fully characterize the nature of these data. In particular, the equity return data represent a highly unbalanced panel over our sample period. In Figure 2 we show the cross-sectional sample size over time of both the total CRSP universe and then those firms who are listed on the NYSE. At the beginning of the sample the CRSP universe includes approximately 500 firms, increases to nearly $8,000$ firms in the late 1990s, and is currently at approximately $4,000$ firms. The sharp jumps in cross-sectional sample sizes that occur in 1962 and 1972 represent the addition of firms listed on the AMEX and Nasdaq to the sample. In the bottom panel of Figure 2, the time series of cross-sectional sample sizes is presented for only stocks listed on the NYSE. Even for this subset of firms, the panel is still highly unbalanced. At the beginning of the sample, there are about 500 firms before rising to a high of approximately $2,000$ firms, and is currently slightly below $1,500$ firms.

6.2 Size Anomaly

We first consider the size anomaly—where smaller firms earn higher returns than larger firms on average. To investigate the size anomaly we use market capitalization as our measure of

[11] The one-month gap is to avoid confounding the momentum anomaly variable with the short-term reversal anomaly (Jegadeesh (1990), Lehmann (1990)).

[12] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html. Industry definitions may be found at http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_38_ind_port.html.

size of the firm. Thus, following the notation of Section 3 we have,

$$R_{it} = \mu(\text{ME}_{i(t-1)}) + \varepsilon_{it}, \qquad i = 1, \dots, n_t, \quad t = 1, \dots, T. \tag{14}$$

Here, ME_{it}, represents the market equity of firm i at time t transformed in the following way: (i) the natural logarithm of market equity of firm i at time t is taken; (ii) at each cross section $t = 1, \dots, T$, the natural logarithm of market equity is demeaned and normalized by the inverse of the cross-sectional standard deviation (i.e., a zscore is applied). This latter transformation is necessary in light of Assumption 1(c) and ensures that the measure of the size of a firm is comparable over time.

Figure 3 provides the estimates of the relationship between returns and firm size. The left column shows the estimate, $\{\hat{\mu}(z) : z \in \mathcal{Z}\}$, based on equation (6) whereas the right column plots the average return in each of ten portfolios formed based on the conventional approach currently used in the literature.[13] To ensure comparability both estimates have been placed on the same scale. As is clear from the figure, the conventional approach produces an attenuated return differential between average returns and size. One important reason for this is that the standard approach relies on the same number of portfolios regardless of changes in the cross-sectional sample size. As we have shown in Sections 3 and 4, it is imperative that the choice of the number of portfolios is data-driven, respecting the appropriate rate conditions, in order to deliver valid inference. The standard approach will tend to produce a biased estimate of the return differential and will compromise power to discern a significant differential in the data. This issue will always arise in any unbalanced panel, but is exacerbated by the highly unbalanced nature of these data where the number of firms has been trending strongly over time (see Figure 2).

The estimate, $\{\hat{\mu}(z) : z \in \mathcal{Z}\}$, is shown for three different subsamples in Figure 3, namely, 1926–2015, 1967–2015, and 1980–2015. The estimated shape between returns and size is generally very similar across the three sub-periods with a relatively flat relationship except for small firms where there is a sharp monotonic rise in average returns as size decreases. The peak average return for the smallest firms appears to have risen over time, at approximately 5% over the full sample, 5.5% over the sample from 1967–2015 and slightly above 6% over the sample 1980–2015.

Table A1 shows the associated point estimates and test statistics corresponding to the graphs in Figure 3. We display results for a number of different choices of the pairs (z_h, z_L), namely, $(\Phi^{-1}(.975), \Phi^{-1}(.025))$, $(\Phi^{-1}(.95), \Phi^{-1}(.05))$, and $(\Phi^{-1}(.9), \Phi^{-1}(.1))$.[14] These evalu-

[13] The portfolio breakpoints for the standard approach are commonly chosen using either deciles of the sub-sample of firms listed on the NYSE or deciles based on the entire sample. Here we choose deciles based on the latter as the ensure better comparability across estimators.

[14] Recall that $\Phi^{-1}(.975) = 1.96$, $\Phi^{-1}(.95) = 1.65$, and $\Phi^{-1}(.9) = 1.28$.

ation points correspond to the vertical lines shown in Figure 3. In addition, the table shows the point estimates and corresponding test statistics from the conventional approach using ten portfolios. Over all three sub-periods, the difference between the function evaluated at the two most extreme evaluation points, $(\Phi^{-1}(.975), \Phi^{-1}(.025))$, is associated with a strongly statistically significant effect of size on returns. Even in the shortest subsample, 1980–2015, the t-statistic is -5.46. This is also the case when the evaluation points are shifted inward to $(\Phi^{-1}(.95), \Phi^{-1}(.05))$. Of course, as shown in Figure 3, this result is driven by very small firms. However, the conventional estimator would suggest that the size effect is no longer statistically distinguishable from zero over the last 35 or so years. Instead, what has happened is that "larger" small firms are no longer producing higher returns in the last sub-sample. This pattern can be seen in the innermost set of evaluations points, $(z_H, z_L) = (\Phi^{-1}(.9), \Phi^{-1}(.1))$, where the size effect is estimated to be reversed albeit statistically indistinguishable from zero.

To further investigate the results of Table A1 we reconsider the estimates for the relationship between returns and firm size using only firms listed on the NYSE in Figure 4. In this case, the shape of the estimated relationship changes markedly in the full sample versus the most recent subsamples. In the full sample, the estimated relationship appears very similar to the shape shown in the three charts in Figure 3—a sharp downward slope from smaller firms to larger firms. However, over the samples 1967–2015 and 1980–2015, the estimated shape changes demonstrably toward an upside-down "U" shape. It is important to emphasize that the standard approach implies a very different shape and pattern of the relationship between returns and size for this sample of firms—especially for the sample from 1967–2015 and 1980–2015.

Figure 5 shows time series plots of the optimal number of portfolios in the sample for the size anomaly. The left column displays the optimal number of portfolios chosen based on equation (12), using data for our three sub-periods and based on $z_H = \Phi^{-1}(.975)$, $z_L = \Phi^{-1}(.025)$, and $\Phi(\cdot)$ is the CDF of a standard normal random variable.[15] Notably, the optimal number of portfolios is substantially larger than the standard choice of ten. Instead, the optimal choice is approximately 250 in the largest cross section and around 50 in the smallest cross section. Furthermore, in all three samples, there is substantial variation in the optimal number of portfolios, again, reflecting the strong variation in cross-sectional sample sizes in these data. The right column shows the optimal number of portfolios in the NYSE-only sample. In this restricted sample the cross-sectional sample sizes are lower which, all else equal, will reduce the optimal choice of number of portfolios. However, the bias-variance trade-off also changes in the NYSE-only sample and so it is not always the case that the

[15] We also considered choosing J based on a plug-in version of equation (11). However, this resulted in considerably larger values for J.

restricted sample has a smaller value for the optimal number of portfolios. In the 1980–2015 sample, the optimal choice of portfolios is slightly larger (at its peak) than the case using all stocks reinforcing the point that the appropriate choice of number of portfolios will be strongly affected by the features of the data being used.

6.3 Momentum Anomaly

We next consider the momentum anomaly—where firms which have had better relative returns in the nearby past also have higher relative returns on average. As discussed in Section 3, the generality of our sampling assumptions means that our results apply to anomalies such as momentum where lagged returns enter in the unknown function of interest. Specifically, we have

$$R_{it} = \mu(\textsc{mom}_{it}) + \varepsilon_{it}, \qquad i = 1, \dots, n_t, \quad t = 1, \dots, T. \tag{15}$$

Here, $\textsc{mom}_{it}$, represents the 12-2 momentum measure of firm i at time t transformed in the following way: at each cross section $t = 1, \dots, T$, 12-2 momentum is demeaned and normalized by the inverse of the cross-sectional standard deviation (i.e., a zscore is applied).[16]

Figure 6 shows the estimates of the relationship between returns and momentum. Even more so than in the case of the size anomaly, we observe that $\{\hat{\mu}(z) : z \in \mathcal{Z}\}$ is very similar across subsamples. The relationship is concave with past "winners" (i.e., those with high 12-2 momentum values) earning about 2% in returns, on average. The strategy of investing in past "losers" (i.e., those with low 12-2 momentum values) on the other hand, has resulted in increasing losses in the later subsamples. The nadir in the estimated relationship occurs at approximately -0.8% in the full sample, slightly less than that in the 1967–2015 subsample, and -1.5% in the 1980–2015 subsamples. This suggests that the short side of buying the spread portfolio appears to have become more profitable in recent years.[17] The right column of Figure 6 shows that this insight could not be gleaned by using the conventional estimator. Furthermore, the conventional estimator suggests an approximately linear relationship between returns and momentum with a distinctly compressed differential between the average returns of winners versus losers. This underscores how our more general approach leads to richer conclusions about the underlying data generating process.

The bottom panel of Table A1 shows the corresponding point estimates and test statistics for the momentum anomaly. The results strongly confirm that momentum is a ro-

[16]Unlike in the case of the size anomaly, no transformation is necessary to satisfy Assumption 1(c). We chose to normalize each cross-section in this way as it is the natural counterpart in our setting to the standard portfolio sorting approach to the momentum anomaly. Moreover, the results based directly on 12-2 momentum are similar.

[17]This conclusion is robust to excluding the financial crisis and its aftermath.

bust anomaly. Across all three pairs of evaluation points and the three different samples, the spread is highly statistically significant (last column). Focusing separately on $\mu(z_H)$ and $\mu(z_L)$ we find that the point estimates are positive and negative, respectively, across all our specifications. In fact, the short end of the spread trade, represented by $\mu(z_L)$, appears to have become stronger in the latter samples (see also Figure 6), producing t-statistics which have the largest magnitude in the 1980–2015 sample when evaluated at either $(\Phi^{-1}(.975), \Phi^{-1}(.025))$ or $(\Phi^{-1}(.95), \Phi^{-1}(.05))$. In contrast, the conventional implementation finds that the short side of the trade is never significant across any of the subsamples and a t-statistic of only -0.36 in the 1980–2015 sample.

Cross-sectional regressions are, by far, the most popular empirical alternative to portfolio sorting (see discussion in Remarks 2 and 8). Arguably the most appealing feature of cross-sectional regressions to the empirical researcher is the ability to include a large number of control variables. Given that we have combined the two approaches in a unified framework it is natural to consider an example. Here we will consider the nonparametric relationship between returns and momentum while controlling for industry momentum. This empirical exercise is similar in spirit to Moskowitz and Grinblatt (1999). The model then becomes,

$$R_{it} = \mu(\textsc{mom}_{it}) + \beta_1 \cdot \textsc{Imom}_{it} + \beta_2 \cdot \textsc{Imom}_{it}^2 + \beta_3 \cdot \textsc{Imom}_{it}^3 + \varepsilon_{it}, \qquad i = 1, \ldots, n_t, \quad t = 1, \ldots, T, \tag{16}$$

where $\textsc{Imom}_{it}$ is the industry momentum of firm i at time t. We also include the square and cube of industry momentum as a flexible way to allow for nonlinear effects of this control variable.

Figure 7 shows the estimates of the relationship between returns and momentum controlling for industry momentum as in equation (16) (solid line). For reference the plots in the left column also include $\{\hat{\mu}(z) : z \in \mathcal{Z}\}$ (dash-dotted line) with no control variables (i.e., based on equation (15)) for the same choice of the number of portfolios at each time t.[18] The difference between the two estimated functions tends to be larger for larger values of 12-2 momentum and accounts for, at most, approximately 0.5 percentage point of momentum returns in the full sample. In the two more recent subsamples the differences are smaller but economically meaningful. That said, the broad shape of the relationship between returns and stock momentum is unchanged by controlling for industry momentum. This suggests that, for this choice of specification, momentum of individual firms is generally distinct from momentum within an industry (Moskowitz and Grinblatt (1999), Grundy and Martin (2001)).

[18] To improve comparability, the estimated function without control variables uses the same sequence of $\{J_t : t = 1 \ldots T\}$ as in the case with control variables. Thus, this estimated function differs from that presented in Figure 6.

The bottom panel of Table A1 provides point estimates and associated test statistics based on equation (16) in the rows labelled "w/ controls". First, it is clear that the inclusion of industry momentum does have a noticeable effect on inference. In general, the magnitude of the t-statistics for the high evaluation point, low evaluation point, and difference are shrunk toward zero. For both the high evaluation point and the difference this is uniformly true and, in all cases, results in t-statistics with substantially larger associated p-values. That said, for all subsamples the difference at the high and low evaluation points results in statistically significant return differential at the 5% level. This exercise illustrates the usefulness of our unified framework as it allows for the additional of control variables in a simple and straightforward manner.

Finally, Figure 8 shows time series plots of the optimal number of portfolios in the sample. Just as in the case of the size anomaly, the optimal number of portfolios is well above ten. However, a number of specifications result in a maximum number of portfolios of approximately 55. This is much smaller, in general, than for the size anomaly (Figure 5). In the right column we show the optimal number of portfolios across time when controlling for industry momentum. These are much larger than the corresponding row in the left column. Intuitively, the inclusion of controls soaks up some of the variation in returns previously explained only by 12-2 momentum. This lower variance results in a higher choice of J (see equation (11)). This example makes clear that the appropriate choice of the number of portfolios reflects a diverse set of characteristics of the data such as cross-sectional sample size, the number of time series observations, the shape of the relationship, and the variability of the innovations.

7 Conclusion

This paper has developed a framework formalizing portfolio-sorting based estimation and inference. Despite decades of use in empirical finance, portfolio sorting has received little to no formal treatment. By formalizing portfolio sorting as a nonparametric procedure, this paper made a first step in developing the econometric properties of this widely used technique. We have developed first-order asymptotic theory as well as mean square error based optimal choices for the number of portfolios, treating the most common application, testing high vs. low returns based on empirical quantiles. We have shown that the choice of the number of portfolios is crucial to draw accurate conclusions from the data and, in standard empirical finance applications, should vary over time and be guided by other aspects of the data at hand. We provide practical guidance on how to implement this choice. In addition, we show that once the number of portfolios is chosen in the appropriate, data-driven way, inference based on the "Fama-MacBeth" variance estimator is asymptotically valid.

One of the key challenges in the empirical finance literature is sorting in a multi-characteristic setting where the number of characteristics is quickly limited by the presence of empty portfolios. Instead, researchers often resort to cross-sectional regressions thereby imposing a restrictive parametric assumption. Here, we bridge the gap between the two approaches proposing a novel portfolio sorting estimator which allows for linear conditioning variables.

We have demonstrated the empirical relevance of our theoretical results by revisiting two notable stock-return anomalies identified in the literature—the size anomaly and the momentum anomaly. We find that the estimated relationship between returns and size is characterized by a monotonically decreasing and convex functional form with a significant return differential between the function evaluated at extreme values of the size variable. However, the statistical significance is generated by very small firms and the results are no longer robust once the smallest firms have been removed from the sample. We also find that the estimated relationship between returns and past returns is characterized by a monotonically increasing and concave functional form with a significant and robust return differential. We find that the "short" side of the momentum spread trade has become more profitable in later sub-periods. In both empirical applications we find that the optimal number of portfolios varies substantially over time and is much larger than the standard choice of ten routinely used in the empirical finance literature.

The results presented herein will be useful to empirical researchers in future applications of characteristic-based portfolio sorts, as well as in (re)examining past findings. However, questions remain, including developing similar results for beta-sorted portfolios. We are currently pursuing research into this area.

References

Adrian, T., Crump, R. K., Moench, E., 2015. Regression-based estimation of dynamic asset pricing models. Journal of Financial Economics 118, 211–244.

Andrews, D. W. K., 2005. Cross-section regression with common shocks. Econometrica 73, 1551–1585.

Ang, A., Liu, J., Schwarz, K., 2017. Using stocks or portfolios in tests of factor models, working paper.

Bali, T. G., Engle, R. F., Murray, S., 2016. Empirical Asset Pricing: The Cross Section of Stock Returns. John Wiley & Sons, Inc., Hoboken, NJ.

Banz, R. W., 1981. The relationship between return and market value of common stocks. Journal of Financial Economics 9, 3–18.

Barrot, J.-N., Loualiche, E., Sauvagnat, J., 2016. The globalization risk premium, working paper.

Basu, S., 1977. Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. Journal of Finance 32, 663–682.

Belloni, A., Chernozhukov, V., Chetverikov, D., Kato, K., 2015. On the asymptotic theory for least squares series: Pointwise and uniform results. Journal of Econometrics 186, 345–366.

Berk, J. B., 2000. Sorting out sorts. Journal of Finance 55, 407–427.

Black, F., Jensen, M. C., Scholes, M. S., 1972. The capital asset pricing model: Some empirical tests. In: Jensen, M. C. (ed.), *Studies in Theory of Capital Markets*, Praeger, New York, NY.

Bondt, W. F. M. D., Thaler, R., 1985. Does the stock market overreact? Journal of Finance 40, 793–805.

Bouchaud, J.-P., Krüger, P., Landier, A., Thesmar, D., 2016. Sticky expectations and stock market anomalies, working paper.

Cattaneo, M. D., Farrell, M. H., 2013. Optimal convergence rates, Bahadur representation, and asymptotic normality of partitioning estimators. Journal of Econometrics 174, 127–143.

Chen, J., Hong, H., Stein, J. C., 2002. Breadth of ownership and stock returns. Journal of Financial Economics 66, 171–205.

Cochrane, J. H., 2011. Discount rates. Journal of Finance 66, 1047–1108.

Connor, G., Hagmann, M., Linton, O., 2012. Efficient semiparametric estimation of the Fama-French model and extensions. Econometrica 80, 713–754.

Conrad, J. S., Cooper, M. J., Kaul, G., 2003. Value versus glamour. Journal of Finance 58, 1969–1996.

Daniel, K., Titman, S., 2012. Testing factor-model explanations of market anomalies. Critical Finance Review 1, 103–139.

Eggermont, P. P. B., LaRiccia, V. N., 2009. Maximum Penalized Likelihood Estimation: Regression. Springer, New York, NY.

Fama, E. F., 1976. Foundations of Finance: Portfolio Decisions and Securities Prices. Basic Books, New York, NY.

Fama, E. F., French, K. R., 1992. The cross-section of expected stock returns. Journal of Finance 47, 427–465.

Fama, E. F., French, K. R., 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33, 3–56.

Fama, E. F., French, K. R., 2008. Dissecting anomalies. Journal of Finance 63, 1653–1678.

Fama, E. F., French, K. R., 2015. A five-factor asset pricing model. Journal of Financial Economics 116, 1–22.

Fama, E. F., MacBeth, J. D., 1973. Risk, return, and equilibrium: Empirical tests. Journal of Political Economy 81, 607–636.

Fama, E. F., MacBeth, J. D., 1974. Tests of multiperiod two parameter model. Journal of Financial Economics 1, 43–66.

Fan, X., Liu, M., 2008. Sorting, firm characteristics, and time-varying risk: An econometric analysis. Journal of Financial Econometrics 6, 49–86.

Gagliardini, P., Gourieroux, C., 2014. Efficiency in large dynamic panel models with common factors. Econometric Theory 30, 961–1020.

Gagliardini, P., Ossola, E., Scaillet, O., 2015. Time-varying risk premium in large cross-sectional equity datasets. Econometrica (forthcoming).

Gibbons, M. R., 1982. Multivariate tests of financial models: a new approach. Journal of Financial Economics 10, 3–27.

Gospodinov, N., Kan, R., Robotti, C., 2014. Misspecification-robust inference in linear asset-pricing models with irrelevant risk factors. Review of Financial Studies 27, 2139–2170.

Gospodinov, N., Kan, R., Robotti, C., 2017. Spurious inference in reduced-rank asset-pricing models, working paper.

Goyal, A., 2012. Empirical cross-sectional asset pricing: a survey. Financial Markets and Portfolio Management 26, 3–38.

Grundy, B. D., Martin, J. S., 2001. Understanding the nature of the risks and the source of rewards to momentum investing. Review of Financial Studies 14, 29–78.

Hall, P., Heyde, C. C., 1980. Martingale Limit Theory and its Application. Academic Press, New York.

Han, Y., Lesmond, D. A., 2011. Liquidity biases and the pricing of cross-sectional idiosyncratic volatility. Review of Financial Studies 24, 1590–1629.

Hong, H., Lim, T., Stein, J. C., 2000. Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies. Journal of Finance 55, 265–295.

Hou, K., Moskowitz, T., 2005. Market frictions, price delay, and the cross-section of expected returns. Review of Financial Studies 18, 981–1020.

Hou, K., Xue, C., Zhang, L., 2014. Digesting anomalies: An investment approach. Review of Financial Studies Forthcoming.

Ibragimov, R., Müller, U. K., 2010. t-Statistic based correlation and heterogeneity robust inference. Journal of Business and Economic Statistics 28, 453–468.

Ibragimov, R., Müller, U. K., 2016. Inference with few heterogenous clusters. Review of Economics and Statistics 98, 83–96.

Jagannathan, R., Wang, Z., 1998. An asymptotic theory for estimating beta-pricing models using cross-sectional regression. Journal of Finance 53, 1285–1309.

Jegadeesh, N., 1990. Evidence of predictable behavior of security returns. Journal of Finance 45, 881–898.

Jegadeesh, N., Titman, S., 1993. Returns to buying winners and selling losers: implications for stock market. Journal of Finance 48, 65–92.

Jegadeesh, N., Titman, S., 2001. Profitability of momentum strategies: an evaluation of alternative explanations. Journal of Finance 56, 699–720.

Jensen, M. C., 1978. Some anomalous evidence regarding market efficiency. Journal of Financial Economics 6, 95–101.

Kan, R., Chen, R., 2005. Finite sample analysis of two-pass cross-sectional regressions, working paper.

Kandel, S., 1984. The likelihood ratio test statistic of mean-variance efficiency without a riskless asset. Journal of Financial Economics 13, 575–592.

Kleibergen, F., 2009. Tests of risk premia in linear factor models. Journal of Econometrics 149, 149–173.

Kleibergen, F., Zhan, Z., 2013. Unexplained factors and their effects on second pass R-squared's and t-tests. Working paper, Brown University, Providence, RI.

Koijen, R. S., Moskowitz, T. J., Pedersen, L. H., Vrugt, E. B., 2015. Carry, working paper.

Lehmann, B. N., 1990. Fads, martingales, and market efficiency. Quarterly Journal of Economics 105, 1–28.

Lewellen, J., Nagel, S., Shanken, J., 2010. A skeptical appraisal of asset pricing tests. Journal of Financial Economics 96, 175–194.

Linnainmaa, J. T., Roberts, M., 2016. The history of the cross section of stock returns, working Paper.

Lo, A. W., MacKinlay, A. C., 1990. Data-snooping biases in tests of financial asset pricing models. Review of Financial Studies 3, 431–467.

Lustig, H., Roussanov, N., Verdelhan, A., 2011. Common risk factors in currency markets. Review of Financial Studies 24, 3731–3777.

Lustig, H., Verdelhan, A., 2007. The cross-section of foreign currency risk premia and us consumption growth risk. Review of Financial Studies 97, 89–117.

Moskowitz, T. J., Grinblatt, M., 1999. Do industries explain momentum? Journal of Finance 54, 1249–1290.

Nagel, S., 2005. Short sales, institutional investors and the cross-section of stock returns. Journal of Financial Economics 78, 277–309.

Nagel, S., Singleton, K. J., 2011. Estimation and evaluation of conditional asset pricing models. Journal of Finance 66, 873–909.

Patton, A. J., Timmermann, A., 2010. Monotonicity in asset returns: New tests with applications to the term structure, the CAPM, and portfolio sorts. Journal of Financial Economics 98, 605–625.

Reinganum, M. R., 1981. Misspecification of asset pricing: Empirical anomalies based on earnings' yields and market values. Journal of Financial Economics 9, 19–46.

Roll, R., 1985. A note on the geometry of Shanken's CSR T^2 test for mean/variance efficiency. Journal of Financial Economics 14, 349–357.

Romano, J. P., Wolf, M., 2013. Testing for monotonicity in expected asset returns. Journal of Empirical Finance 23, 93–116.

Rosenberg, B., Reid, K., Lanstein, R., 1985. Persuasive evidence of market inefficiency. Journal of Portfolio Management 11, 9–16.

Shanken, J., 1985. Multivariate tests of the zero-beta CAPM. Journal of Financial Economics 14, 327–348.

Shanken, J., 1986. Testing portfolio efficiency when the zero-beta rate is unknown: a note. Journal of Finance 41, 269–276.

Shanken, J., Zhou, G., 2007. Estimating and testing beta pricing models: Alternative methods and their performance in simulations. Journal of Financial Economics 84, 40–86.

Sharpe, W. F., 1966. Mutual fund performance. Journal of Business 39, 119–138.

Shumway, T., 1997. The delisting bias in crsp data. Journal of Finance 52, 327–340.

Stattman, D., 1980. Book values and stock returns. The Chicago MBA: A Journal of Selected Papers 4, 25–45.

Wahal, S., Yavuz, M. D., 2013. Style investing, comovement and return predictability. Journal of Financial Economics 107, 136–154.

Weber, M., 2016. Cash flow duration and the term structure of equity returns, working paper.

A.1 Implementation

As we discussed in Section 5 we base our choice of the optimal number of portfolios in our empirical applications based on equation (12). To do so let $t_{\max} = \arg\max_{1\leq t\leq T} n_t$, $n = n_{t_{\max}}$ and $J = J_{t_{\max}}$. For all other time periods we scale J_t as $J_t = J\,(n_t/\,n)^{\frac{1}{d+1}}$ (see discussion in Section 4). We then choose a grid of values for J as $J = ((n_{t_{\min}}/\,n)^{\frac{1}{d+1}}, \ldots, J_{\max})$ where $t_{\min} = \arg\min_{1\leq t\leq T} n_t$. In our empirical applications we set $J_{\max} = 400$.

To estimate the MSE in practice we have the following estimator,

$$
\begin{aligned}
&\widehat{\mathrm{MSE}}\big(\hat{\mu}(\mathbf{z}_H) - \hat{\mu}(\mathbf{z}_L);J_1,\ldots,J_T\big) \\
&\quad= \Bigg(\hat{\mu}'\left(\mathbf{z}_H\right)\cdot T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J_t^d}\sum_{i=1}^{n_t}\omega_{it}\hat{\mathbf{1}}_{jt}\hat{\mathbb{1}}_{jt}(\mathbf{z}_H)\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\left(\mathbf{z}_{it}-\mathbf{z}_H\right) \\
&\qquad -\hat{\mu}'\left(\mathbf{z}_L\right)\cdot T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J_t^d}\sum_{i=1}^{n_t}\omega_{it}\hat{\mathbf{1}}_{jt}\hat{\mathbb{1}}_{jt}(\mathbf{z}_L)\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\left(\mathbf{z}_{it}-\mathbf{z}_L\right)\Bigg)^2 \\
&\qquad + T^{-2}\sum_{t=1}^{T}\left(\hat{m}_t\left(\mathbf{z}_H\right)-\hat{m}_t\left(\mathbf{z}_L\right)-\left(\hat{m}\left(\mathbf{z}_H\right)-\hat{m}\left(\mathbf{z}_L\right)\right)\right)^2
\end{aligned}
$$

where

$$
\hat{m}_t\left(\mathbf{z}\right) = \sum_{j=1}^{J_t^d} N_{jt}^{-1/2}\sum_{i=1}^{n_t}\omega_{it}\hat{\mathbf{1}}_{jt}\hat{\mathbb{1}}_{jt}(\mathbf{z})\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})(R_{it}-\mathbf{x}_{it}'\hat{\boldsymbol{\beta}}_t), \qquad \hat{m}\left(\mathbf{z}\right) = T^{-1}\sum_{t=1}^{T}\hat{m}_t\left(\mathbf{z}\right).
$$

Here ω_{it} is the weight applied to the returns in each portfolio which satisfies $\sum_{i=1}^{n_t}\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\omega_{it} = 1$ for each $j = 1,\ldots,J_t^d$ and at each time t. As is common, we use lagged market equity to weight the returns in each portfolio in our empirical applications. The plug-in estimate of $\hat{\mathcal{V}}^{(2)}\frac{J^{2d}}{n^2T}$ implicit in the above expression utilizes the logic of the Fama-MacBeth variance estimator applied to the higher-order variance term. As a plug-in estimator of $\mu'\left(\mathbf{z}\right)$ we use the time-series average of the estimated slope coefficient from a local regression using the 40 closest points to $\mathbf{z}$ (ties included) at each point in time.

A.2 Preliminary Lemmas

Before proceeding to the lemmas it is useful to introduce some additional notation. Define $\hat{q}_{jt} = N_{jt}/\,n_t$ and its population counterpart $q_{jt} = \mathbb{P}\left(\mathbf{z}\in P_{jt}|\,\mathcal{F}_t\right)$.

Lemma 1. *Under our assumptions, we have that*

$$
\max_{1\leq t\leq T}\max_{1\leq j\leq J_t^d}\sup_{\mathbf{z}}\left|\hat{\mathbb{1}}_{jt}\left(\mathbf{z}\right)\mu\left(\mathbf{z}\right)-\hat{\mathbb{1}}_{jt}\left(\mathbf{z}\right)\gamma_{jt}^0\right| = O_p\left(J^{-1}\right),
$$

for a nonrandom γ_{jt}^0 only dependent on j and t and if we define $h_{t,\ell}\left(\mathbf{z}\right) = h_{t,\ell}\left(\mathbf{z},\mathcal{F}_t\right) = \mathbb{E}\left[x_{it,\ell}|\mathcal{F}_t,\mathbf{z}_{it}=\mathbf{z}\right]$ where $x_{it,\ell}$ is the ℓth element of $\mathbf{x}_{it}$,

$$
\max_{1\leq t\leq T}\max_{1\leq j\leq J_t^d}\max_{1\leq\ell\leq d_x}\sup_{\mathbf{z}}\left|\hat{\mathbb{1}}_{jt}\left(\mathbf{z}\right)h_{t,\ell}\left(\mathbf{z}\right)-\hat{\mathbb{1}}_{jt}\left(\mathbf{z}\right)\pi_{jt,\ell}^0\right| = O_p\left(J^{-1}\right),
$$

for a nonrandom $\pi_{jt,\ell}^0$ only dependent on j, t and ℓ.

Stack the $h_{t,\ell}(\cdot)$'s as $h_t(\cdot) = (h_{t,1}(\cdot), \ldots, h_{t,d_x}(\cdot))'$ and then stack again as the $n_t \times d_x$ matrix $\mathbf{H}_t = (h_t(\mathbf{z}_{1t}), \ldots, h_t(\mathbf{z}_{n_t t}))'$. Finally, define $\mathbf{U}_t = \mathbf{X}_t - \mathbf{H}_t$. Recall $\Omega_{\mathrm{uu},t} = \mathrm{plim}_{n\to\infty} U_t U_t'/n_t = \mathbb{E}\left[\mathbb{V}\left(\mathbf{x}_{it} | \mathbf{z}_{it}, \mathcal{F}_t\right) | \mathcal{F}_t\right]$ and by our assumptions we have that $\min_{1\le t\le T} \lambda_{\min}(\Omega_{\mathrm{uu},t})$ is bounded away from zero. Note that under our assumptions $\hat{q}_{jt} \asymp J^{-d}$ and $q_{jt} \asymp J^{-d}$. Let $\hat{\Omega}_{\mathrm{uu},t} = \mathbf{X}_t' M_{B_t} \mathbf{X}_t / n_t$. Define $\hat{\mathbf{1}}_{jt} = \hat{\mathbf{1}}_{q,jt} \hat{\mathbf{1}}_{\beta,t}$ where $\hat{\mathbf{1}}_{q,jt} = \mathbf{1}(q_{jt}/2 \le \hat{q}_{jt} \le 2q_{jt})$ and $\hat{\mathbf{1}}_{\beta,t} = \mathbf{1}\left(\lambda_{\min}\left(\hat{\Omega}_{\mathrm{uu},t}\right) \ge C_{\mathrm{uu}}/2\right)$ where $C_{\mathrm{uu}} := \min_{1\le t\le T} \lambda_{\min}(\Omega_{\mathrm{uu},t})$ which is bounded away from zero by our assumptions.

Lemma 2. *Under our assumptions,*

$$\max_{1\le t\le T} \max_{1\le j\le J_t^d} |\hat{q}_{jt} - q_{jt}|^2 = O_p\left(\frac{\log\left(\max\left(J^d, T\right)\right)}{J^d n}\right).$$

Lemma 3. *Under our assumptions,*

$$T^{-1} \sum_{t=1}^{T} \left\| \hat{\Omega}_{\mathrm{uu},t} - \Omega_{\mathrm{uu},t} \right\|^2 = O_p\left(n^{-1}\right) + O\left(J^{-4}\right) + O_p\left(n^{-2} J^{2d}\right),$$

and

$$\max_{1\le t\le T} \left\| \hat{\Omega}_{\mathrm{uu},t} - \Omega_{\mathrm{uu},t} \right\| = O_p\left(\log(T)\, n^{-1/2}\right) + O_p\left(J^{-2}\right) + O_p\left(n^{-1} J^d\right).$$

Lemma 4. *Under our assumptions* $J^{ad} n^b T^c \max_{1\le t\le T} \max_{1\le j\le J_t^d} |\mathbf{1}_{jt} - 1| = o_p(1)$ *for any fixed* $a, b, c \in \mathbb{R}$

Lemma 5. *Under our assumptions,* $V(\mathbf{z}) = C n^{-1} T^{-1} J^d + o\left(n^{-1} T^{-1} J^d\right)$ *where the constant is bounded and bounded away from zero.*

Lemma 6. *Under our assumptions,*

$$V(\mathbf{z})^{-1} T^{-2} \sum_{t=1}^{T} \sum_{j=1}^{J_t^d} N_{jt}^{-2} \sum_{i=1}^{n_t} \hat{\mathbf{1}}_{jt} \hat{\mathbb{1}}_{jt}(\mathbf{z}) \hat{\mathbb{1}}_{jt}(\mathbf{z}_{it}) \left(\varepsilon_{it}^2 - \sigma_{it}^2\right) = o_p(1).$$

Lemma 7. *Under our assumptions,*

$$\left| T^{-1} \sum_{t=1}^{T} \hat{\mathbf{1}}_{\beta,t} s_t' \left(\hat{\beta}_t - \beta_t\right) \right|^2 = O_p\left(n^{-1} T^{-1}\right) + O\left(J^{-4}\right) + O_p\left(n^{-2} J^{-2}\right) + O_p\left(n^{-1} J^{-6}\right) + O_p\left(J^{2d-2} n^{-3}\right)$$

and

$$T^{-1} \sum_{t=1}^{T} \hat{\mathbf{1}}_{\beta,t} \left(s_t' \left(\hat{\beta}_t - \beta_t\right)\right)^2 = O_p\left(n^{-1}\right) + O\left(J^{-4}\right),$$

where $\|s_t\| \le C$ *a.s. and is nonrandom conditional on* $\mathbf{z}_t$ *and* $\mathcal{F}_t$*.*

A.3 Proofs

Proof of Theorem 1: Our estimator may be written as

$$\hat{\mu}(\mathbf{z}) = T^{-1} \sum_{t=1}^{T} n_t^{-1} \sum_{i=1}^{n_t} \sum_{j=1}^{J_t^d} \hat{\mathbf{1}}_{jt} \hat{\mathbb{1}}_{jt}(\mathbf{z}) \hat{q}_{jt}^{-1} \hat{\mathbb{1}}_{jt}(\mathbf{z}_{it}) \left(R_{it} - \mathbf{x}_{it}' \hat{\beta}_t\right)$$

We can then decompose the estimator as $\hat{\mu}(\mathbf{z}) - \mu(\mathbf{z}) = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4$,where

$$\begin{aligned}
\mathcal{L}_1 &= T^{-1}\sum\nolimits_{t=1}^{T} n_t^{-1}\sum\nolimits_{i=1}^{n_t}\sum\nolimits_{j=1}^{J_t^d} \hat{\mathbf{1}}_{jt}\hat{\mathbb{1}}_{jt}(\mathbf{z})\,\hat{q}_{jt}^{-1}\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\,(\mu(\mathbf{z}_{it}) - \mu(\mathbf{z}))\\
\mathcal{L}_2 &= T^{-1}\sum\nolimits_{t=1}^{T} n_t^{-1}\sum\nolimits_{i=1}^{n_t}\sum\nolimits_{j=1}^{J_t^d} \hat{\mathbf{1}}_{jt}\hat{\mathbb{1}}_{jt}(\mathbf{z})\,\hat{q}_{jt}^{-1}\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\,\varepsilon_{it}\\
\mathcal{L}_3 &= -T^{-1}\sum\nolimits_{t=1}^{T} n_t^{-1}\sum\nolimits_{i=1}^{n_t}\sum\nolimits_{j=1}^{J_t^d} \hat{\mathbf{1}}_{jt}\hat{\mathbb{1}}_{jt}(\mathbf{z})\,\hat{q}_{jt}^{-1}\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\,\mathbf{x}_{it}'\left(\hat{\beta}_t - \beta_t\right)\\
\mathcal{L}_4 &= T^{-1}\sum\nolimits_{t=1}^{T}\sum\nolimits_{j=1}^{J_t^d}\left(\hat{\mathbf{1}}_{jt} - 1\right)\hat{\mathbb{1}}_{jt}(\mathbf{z})\,\mu(\mathbf{z}).
\end{aligned}$$

Define

$$\theta_t(\mathbf{z}) = V(\mathbf{z})^{-1/2}T^{-1}n_t^{-1}\sum\nolimits_{i=1}^{n_t}\sum\nolimits_{j=1}^{J_t^d}\hat{\mathbf{1}}_{jt}\hat{\mathbb{1}}_{jt}(\mathbf{z})\,\hat{q}_{jt}^{-1}\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\,\varepsilon_{it}.$$

We will work with the rescaled version of our centered estimator as

$$V(\mathbf{z})^{-1/2}(\hat{\mu}(\mathbf{z}) - \mu(\mathbf{z})) = V(\mathbf{z})^{-1/2}\mathcal{L}_1 + V(\mathbf{z})^{-1/2}\mathcal{L}_2 + V(\mathbf{z})^{-1/2}\mathcal{L}_3 + V(\mathbf{z})^{-1/2}\mathcal{L}_4.$$

By Lemma 5, $V(\mathbf{z}) \asymp Cn^{-1}T^{-1}J^d + o\left(n^{-1}T^{-1}J^d\right)$ so we need to show that, under our assumptions, $\mathcal{L}_\ell = o_p(J^{d/2}n^{-1/2}T^{-1/2})$ for $\ell \in \{1,3,4\}$ and $V(\mathbf{z})^{-1/2}\mathcal{L}_2 \to_d \mathcal{N}(0,1)$.

First consider $\mathcal{L}_1$. Then,

$$\begin{aligned}
|\mathcal{L}_1| &= \left|T^{-1}\sum\nolimits_{t=1}^{T} n_t^{-1}\sum\nolimits_{i=1}^{n_t}\sum\nolimits_{j=1}^{J_t^d}\hat{\mathbf{1}}_{jt}\hat{\mathbb{1}}_{jt}(\mathbf{z})\,\hat{q}_{jt}^{-1}\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\,(\mu(\mathbf{z}_{it}) - \mu(\mathbf{z}))\right|\\
&\leq T^{-1}\sum\nolimits_{t=1}^{T} n_t^{-1}\sum\nolimits_{i=1}^{n_t}\sum\nolimits_{j=1}^{J_t^d}\hat{\mathbf{1}}_{jt}\hat{\mathbb{1}}_{jt}(\mathbf{z})\,\hat{q}_{jt}^{-1}\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\left|\mu(\mathbf{z}_{it}) - \gamma_{jt}^0\right|\\
&\quad + T^{-1}\sum\nolimits_{t=1}^{T} n_t^{-1}\sum\nolimits_{i=1}^{n_t}\sum\nolimits_{j=1}^{J_t^d}\hat{\mathbf{1}}_{jt}\hat{\mathbb{1}}_{jt}(\mathbf{z})\,\hat{q}_{jt}^{-1}\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\left|\mu(\mathbf{z}) - \gamma_{jt}^0\right|.
\end{aligned}$$

The first term is

$$\begin{aligned}
&T^{-1}\sum\nolimits_{t=1}^{T} n_t^{-1}\sum\nolimits_{i=1}^{n_t}\sum\nolimits_{j=1}^{J_t^d}\hat{\mathbf{1}}_{jt}\hat{\mathbb{1}}_{jt}(\mathbf{z})\,\hat{q}_{jt}^{-1}\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\left|\mu(\mathbf{z}_{it}) - \gamma_{jt}^0\right|\\
\leq\;& \max_{1\leq t\leq T}\max_{1\leq j\leq J_t^d}\sup_{\mathbf{z}}\left|\hat{\mathbb{1}}_{jt}(\mathbf{z})\,\mu(\mathbf{z}) - \hat{\mathbb{1}}_{jt}(\mathbf{z})\,\gamma_{jt}^0\right| \times T^{-1}\sum\nolimits_{t=1}^{T}\sum\nolimits_{j=1}^{J_t^d}\hat{\mathbf{1}}_{jt}\hat{\mathbb{1}}_{jt}(\mathbf{z})
\end{aligned}$$

which is $O_p\left(J^{-1}\right)$. The second term follows by the same steps. Thus, $\mathcal{L}_1 = o\left(J^{d/2}n^{-1/2}T^{-1/2}\right)$ if Assumption 4(b) holds.

Now consider $\mathcal{L}_2$. Define the sigma field, $\mathcal{G}_s = \sigma(z_1, \ldots, z_T, x_1, \ldots, x_T, \mathcal{F}_1, \ldots, \mathcal{F}_T, \varepsilon_1, \ldots \varepsilon_s)$. Then we have that $(\theta_t(\mathbf{z}), \mathcal{G}_t)$ is a martingale difference sequence with $\sum_{t=1}^{T}\left[\,\theta_t(\mathbf{z})^2\middle|\,\mathcal{G}_{t-1}\right] = 1$. By Hall and Heyde (1980, Corollary 3.1) we have that $\sum_{t=1}^{T}\theta_t(\mathbf{z}) \to_d \mathcal{N}(0,1)$ if $\sum_{t=1}^{T}\mathbb{E}\left[|\theta_t(\mathbf{z})|^{2+\chi}\middle|\,\mathcal{G}_{t-1}\right] = o_p(1)$ for some $\chi > 0$. To show this note that

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\left[|\theta_t(\mathbf{z})|^{2+\chi}\middle|\,\mathcal{G}_{t-1}\right]\\
&= V(\mathbf{z})^{-(1+\chi/2)}T^{-(2+\chi)}\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\left|n_t^{-1}\sum_{i=1}^{n_t}\sum_{j=1}^{J_t^d}\hat{\mathbf{1}}_{jt}\hat{\mathbb{1}}_{jt}(\mathbf{z})\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\hat{q}_{jt}\varepsilon_{it}\right|^{2+\chi}\middle|\,\mathcal{F}_t, x_t, z_t, \mathcal{G}_{t-1}\right]\middle|\,\mathcal{G}_{t-1}\right],
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\left|n_t^{-1}\sum_{i=1}^{n_t}\sum_{j=1}^{J_t^d}\hat{\mathbf{1}}_{jt}\hat{\mathbb{1}}_{jt}(\mathbf{z})\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\hat{q}_{jt}\varepsilon_{it}\right|^{2+\chi}\middle|\mathcal{F}_t,x_t,z_t,\mathcal{G}_{t-1}\right]\\
&\leq C\sum_{i=1}^{n_t}\mathbb{E}\left[\left|n_t^{-1}\sum_{j=1}^{J_t^d}\hat{\mathbf{1}}_{jt}\hat{\mathbb{1}}_{jt}(\mathbf{z})\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\hat{q}_{jt}\varepsilon_{it}\right|^{2+\chi}\middle|\mathcal{F}_t,x_t,z_t,\mathcal{G}_{t-1}\right]\vee\\
&\left(\sum_{i=1}^{n_t}\mathbb{E}\left[\left|n_t^{-1}\sum_{j=1}^{J_t^d}\hat{\mathbf{1}}_{jt}\hat{\mathbb{1}}_{jt}(\mathbf{z})\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\hat{q}_{jt}\varepsilon_{it}\right|^{2}\middle|\mathcal{F}_t,x_t,z_t,\mathcal{G}_{t-1}\right]\right)^{1+\chi/2}.
\end{aligned}$$

The first term is

$$\begin{aligned}
&\sum_{i=1}^{n_t}\mathbb{E}\left[\left|n_t^{-1}\sum_{j=1}^{J_t^d}\hat{\mathbf{1}}_{jt}\hat{\mathbb{1}}_{jt}(\mathbf{z})\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\hat{q}_{jt}\varepsilon_{it}\right|^{2+\chi}\middle|\mathcal{F}_t,x_t,z_t,\mathcal{G}_{t-1}\right]\\
&=\sum_{i=1}^{n_t}\left|n_t^{-1}\sum_{j=1}^{J_t^d}\hat{\mathbf{1}}_{jt}\hat{\mathbb{1}}_{jt}(\mathbf{z})\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\hat{q}_{jt}\right|^{2+\chi}\mathbb{E}\left[|\varepsilon_{it}|^{2+\chi}\middle|\mathcal{F}_t,x_t,z_t\right]\\
&\leq Cn^{-(2+\chi)}J^{d(2+\chi)}\sum_{i=1}^{n_t}\left|\sum_{j=1}^{J_t^d}\hat{\mathbb{1}}_{jt}(\mathbf{z})\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\right|^{2+\chi}\\
&=Cn^{-(1+\chi)}J^{d(1+\chi)}.
\end{aligned}$$

By similar steps, the second term is

$$\left(\sum_{i=1}^{n_t}\mathbb{E}\left[\left|n_t^{-1}\sum_{j=1}^{J_t^d}\hat{\mathbf{1}}_{jt}\hat{\mathbb{1}}_{jt}(\mathbf{z})\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\hat{q}_{jt}\varepsilon_{it}\right|^{2}\middle|\mathcal{F}_t,x_t,z_t,\mathcal{G}_{t-1}\right]\right)^{1+\chi/2}\leq C\left(J^dn^{-1}\right)^{1+\chi/2}.$$

Thus,

$$\sum_{t=1}^{T}\mathbb{E}[|\theta_t(\mathbf{z})|^{2+\chi}\big|\mathcal{G}_{t-1}]\leq C\left(J^dn^{-1}T^{-1}\right)^{-(1+\chi/2)}T^{-(1+\chi)}\left(n^{-(1+\chi)}J^{d(1+\chi)}\vee(J^dn^{-1})^{1+\chi/2}\right)\leq CT^{-\chi/2},$$

and so the result follows. Thus, we have that $V(\mathbf{z})^{-1/2}\mathcal{L}_2\to_d\mathcal{N}(0,1)$.

Next consider $\mathcal{L}_3$. If we define

$$\hat{h}_t(\mathbf{z})=\sum\nolimits_{j=1}^{J_t^d}\hat{\mathbb{1}}_{jt}(\mathbf{z})\hat{\pi}_{jt},\qquad\hat{\pi}_{jt}=\hat{\mathbf{1}}_{q,jt}\hat{q}_{jt}^{-1}n_t^{-1}\sum\nolimits_{i=1}^{n_t}\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\mathbf{x}_{it},$$

then

$$\begin{aligned}
-\mathrm{V}_{n,T}^{1/2}\mathcal{L}_3 &= T^{-1}\sum\nolimits_{t=1}^{T}\hat{h}_t(\mathbf{z})'\hat{\mathbf{1}}_{\beta,t}\left(\hat{\beta}_t-\beta_t\right)\\
&= T^{-1}\sum\nolimits_{t=1}^{T}h_t(\mathbf{z})'\mathbf{1}_{\beta,t}\left(\hat{\beta}_t-\beta_t\right)+T^{-1}\sum\nolimits_{t=1}^{T}\left(\hat{h}_t(\mathbf{z})-h_t(\mathbf{z})\right)'\hat{\mathbf{1}}_{\beta,t}\left(\hat{\beta}_t-\beta_t\right)\\
&= \mathcal{L}_{31}+\mathcal{L}_{32}.
\end{aligned}$$

First consider equation $\mathcal{L}_{31}$. By Lemma 7,

$$
\begin{aligned}
|\mathcal{L}_{31}|^2 &= \left|T^{-1}\sum\nolimits_{t=1}^{T} h_t(\mathbf{z})' \mathbf{1}_{\beta,t}\left(\hat{\beta}_t-\beta_t\right)\right|^2 \\
&= O_p\left(n^{-1}T^{-1}\right)+O\left(J^{-4}\right)+O_p\left(n^{-2}J^{-2}\right)+O_p\left(n^{-1}J^{-6}\right)+O_p\left(J^{2d-2}n^{-3}\right)
\end{aligned}
$$

which is $o_p\left(J^d n^{-1}T^{-1}\right)$ under Assumptions 4(a) and Assumption 4(b). Next consider $\mathcal{L}_{32}$. By the Cauchy-Schwarz inequality we have that

$$
\begin{aligned}
|\mathcal{L}_{32}|^2 &= \left|\frac{1}{T}\sum\nolimits_{t=1}^{T}\left(\hat{h}_t(\mathbf{z})-h_t(\mathbf{z})\right)' \hat{\mathbf{1}}_{\beta,t}\left(\hat{\beta}_t-\beta_t\right)\right|^2 \\
&\leq \left[\frac{1}{T}\sum\nolimits_{t=1}^{T}\left(\hat{h}_t(\mathbf{z})-h_t(\mathbf{z})\right)'\left(\hat{h}_t(\mathbf{z})-h_t(\mathbf{z})\right)\right]\left[\frac{1}{T}\sum\nolimits_{t=1}^{T}\hat{\mathbf{1}}_{\beta,t}\left(\hat{\beta}_t-\beta_t\right)'\left(\hat{\beta}_t-\beta_t\right)\right],
\end{aligned}
$$

The first factor, following the proof for the the consistency of $\hat{\mathrm{V}}_{n,T}^{\mathrm{FM}}$ (all terms but $\mathcal{S}_{12}^{\mathrm{FM}}$) is $O_p\left(J^d n^{-1}\right)$. By Lemma 7, the second factor is $O_p\left(n^{-1}T^{-1}\right)+O\left(T^{-1}J^{-4}\right)$. Thus, $|\mathcal{L}_{32}|^2 = O_p\left(J^d n^{-2}\right)+O_p\left(J^{d-4}n^{-1}\right)$ which is $o_p\left(J^d n^{-1}T^{-1}\right)$ under Assumptions 4(a), 4(b) and 4(c).

Finally, consider $\mathcal{L}_4$

$$
|\mathcal{L}_4| \leq T^{-1}\sum\nolimits_{t=1}^{T}\sum\nolimits_{j=1}^{J_t^d}\left|\hat{\mathbf{1}}_{jt}-1\right|\hat{\mathbb{1}}_{jt}(\mathbf{z})\left|\mu(\mathbf{z})\right| = C\cdot\left(\max_{1\leq t\leq T}\max_{1\leq j\leq J_t^d}\left|\hat{\mathbf{1}}_{jt}-1\right|\right),
$$

which is $o_p\left(J^{d/2}n^{-1/2}T^{-1/2}\right)$ by Lemma 4.

Proof of Theorem 2 We have,

$$
\hat{\mathrm{V}}_{n,T}^{\mathrm{FM}} = T^{-2}\sum\nolimits_{t=1}^{T}\left(\hat{\mu}_t(\mathbf{z})-\hat{\mu}(\mathbf{z})\right)^2 = T^{-2}\sum\nolimits_{t=1}^{T}\left(\hat{\mu}_t(\mathbf{z})-\mu(\mathbf{z})\right)^2 - T^{-1}\left(\hat{\mu}(\mathbf{z})-\mu(\mathbf{z})\right)^2.
$$

Recall that,

$$
\begin{aligned}
\hat{\mu}_{t_1}(\mathbf{z})-\mu(\mathbf{z}) &= n_{t_1}^{-1}\sum\nolimits_{j_1}\hat{q}_{j_1t_1}^{-1}\sum\nolimits_{i_1}\hat{\mathbf{1}}_{j_1t_1}\hat{\mathbb{1}}_{j_1t_1}(\mathbf{z})\,\hat{\mathbb{1}}_{j_1t_1}(\mathbf{z}_{i_1t_1})\left(R_{i_1t_1}-x_{i_1t_1}'\hat{\beta}_{t_1}\right)-\mu(\mathbf{z}) \\
&= n_{t_1}^{-1}\sum\nolimits_{j_1}\hat{q}_{j_1t_1}^{-1}\sum\nolimits_{i_1}\hat{\mathbf{1}}_{j_1t_1}\hat{\mathbb{1}}_{j_1t_1}(\mathbf{z})\,\hat{\mathbb{1}}_{j_1t_1}(\mathbf{z}_{i_1t_1})\,\varepsilon_{i_1t_1} \\
&\quad + n_{t_1}^{-1}\sum\nolimits_{j_1}\hat{q}_{j_1t_1}^{-1}\sum\nolimits_{i_1}\hat{\mathbf{1}}_{j_1t_1}\hat{\mathbb{1}}_{j_1t_1}(\mathbf{z})\,\hat{\mathbb{1}}_{j_1t_1}(\mathbf{z}_{i_1t_1})\left(\mu(\mathbf{z}_{i_1t_1})-\mu(\mathbf{z})\right) \\
&\quad - n_{t_1}^{-1}\sum\nolimits_{j_1}\hat{q}_{j_1t_1}^{-1}\sum\nolimits_{i_1}\hat{\mathbf{1}}_{j_1t_1}\hat{\mathbb{1}}_{j_1t_1}(\mathbf{z})\,\hat{\mathbb{1}}_{j_1t_1}(\mathbf{z}_{i_1t_1})\,\mathbf{x}_{i_1t_1}'\left(\hat{\beta}_{t_1}-\beta_{t_1}\right) \\
&\quad + \sum\nolimits_{j_1}\left(\hat{\mathbf{1}}_{j_1t_1}-1\right)\hat{\mathbb{1}}_{j_1t_1}(\mathbf{z})\,\mu(\mathbf{z}).
\end{aligned}
$$

Thus, by Lemma 6 and the Cauchy-Schwarz inequality it is sufficient to show that $\left|\mathcal{S}_{11}^{\mathrm{FM}}\right| = o_p(1)$, $\left|\mathcal{S}_{12}^{\mathrm{FM}}\right| = o_p(1)$, $\left|\mathcal{S}_{13}^{\mathrm{FM}}\right| = o_p(1)$, and $\left|\mathcal{S}_{2}^{\mathrm{FM}}\right| = o_p(1)$ where

$$
\begin{aligned}
\mathcal{S}_{11}^{\mathrm{FM}} &= \frac{n}{TJ^d}\sum\nolimits_{t_1}\left[n_{t_1}^{-1}\sum\nolimits_{j_1}\hat{q}_{j_1t_1}^{-1}\sum\nolimits_{i_1}\hat{\mathbf{1}}_{j_1t_1}\hat{\mathbb{1}}_{j_1t_1}(\mathbf{z})\,\hat{\mathbb{1}}_{j_1t_1}(\mathbf{z}_{i_1t_1})\left(\mu(\mathbf{z}_{i_1t_1})-\mu(\mathbf{z})\right)\right]^2 \\
\mathcal{S}_{12}^{\mathrm{FM}} &= \frac{n}{TJ^d}\sum\nolimits_{t_1}\left[n_{t_1}^{-1}\sum\nolimits_{j_1}\hat{q}_{j_1t_1}^{-1}\sum\nolimits_{i_1}\hat{\mathbf{1}}_{j_1t_1}\hat{\mathbb{1}}_{j_1t_1}(\mathbf{z})\,\hat{\mathbb{1}}_{j_1t_1}(\mathbf{z}_{i_1t_1})\,\mathbf{x}_{i_1t_1}'\left(\hat{\beta}_{t_1}-\beta_{t_1}\right)\right]^2 \\
\mathcal{S}_{13}^{\mathrm{FM}} &= \frac{n}{TJ^d}\sum\nolimits_{t_1}\left[\sum\nolimits_{j_1}\left(\hat{\mathbf{1}}_{j_1t_1}-1\right)\hat{\mathbb{1}}_{j_1t_1}(\mathbf{z})\,\mu(\mathbf{z})\right]^2 \\
\mathcal{S}_{2}^{\mathrm{FM}} &= \frac{n}{J^d}\left(\hat{\mu}(\mathbf{z})-\mu(\mathbf{z})\right)^2.
\end{aligned}
$$

First consider, $\mathcal{S}_2^{\mathrm{FM}}$. We have already shown that $\hat{\mu}(\mathbf{z}) - \mu(\mathbf{z}) = O_p\left(\sqrt{J^d n^{-1} T^{-1}}\right) + o_p\left(\sqrt{J^d n^{-1} T^{-1}}\right)$, so that $\mathcal{S}_2$ satisfies

$$\mathcal{S}_2^{\mathrm{FM}} = \frac{n}{J^d}\left(\hat{\mu}(\mathbf{z}) - \mu(\mathbf{z})\right)^2 = O_p\left(T^{-1}\right) = o_p(1).$$

Next, consider $\mathcal{S}_{11}^{\mathrm{FM}}$,

$$\begin{aligned}
\mathcal{S}_{11}^{\mathrm{FM}} &\leq \frac{n}{TJ^d}\sum\nolimits_{t_1}\left[n_{t_1}^{-1}\sum\nolimits_{j_1}\hat{q}_{j_1t_1}^{-1}\sum\nolimits_{i_1}\hat{\mathbf{1}}_{j_1t_1}\hat{\mathbb{1}}_{j_1t_1}(\mathbf{z})\,\hat{\mathbb{1}}_{j_1t_1}(\mathbf{z}_{i_1t_1})\left(\left|\mu(\mathbf{z}_{i_1t_1}) - \gamma_{j_1t_1}^0\right| + \left|\mu(\mathbf{z}) - \gamma_{j_1t_1}^0\right|\right)\right]^2 \\
&= \frac{n}{TJ^d}\sum\nolimits_{t_1}\left[n_{t_1}^{-1}\sum\nolimits_{j_1}\hat{q}_{j_1t_1}^{-1}\sum\nolimits_{i_1}\hat{\mathbf{1}}_{j_1t_1}\hat{\mathbb{1}}_{j_1t_1}(\mathbf{z})\,\hat{\mathbb{1}}_{j_1t_1}(\mathbf{z}_{i_1t_1})\right. \\
&\qquad \left.\left(\left|\hat{\mathbb{1}}_{j_1t_1}(\mathbf{z}_{i_1t_1})\,\mu(\mathbf{z}_{i_1t_1}) - \hat{\mathbb{1}}_{j_1t_1}(\mathbf{z}_{i_1t_1})\,\gamma_{j_1t_1}^0\right| + \left|\hat{\mathbb{1}}_{j_1t_1}(\mathbf{z})\,\mu(\mathbf{z}) - \hat{\mathbb{1}}_{j_1t_1}(\mathbf{z})\,\gamma_{j_1t_1}^0\right|\right)\right]^2 \\
&\leq C\max_{1\leq t\leq T}\max_{1\leq j\leq J_t^d}\sup_{\mathbf{z}}\left|\hat{\mathbb{1}}_{jt}(\mathbf{z})\,\mu(\mathbf{z}) - \hat{\mathbb{1}}_{jt}(\mathbf{z})\,\gamma_{j_1t_1}^0\right|^2 \times \frac{n}{J^d} \\
&= O\left(nJ^{-(d+2)}\right),
\end{aligned}$$

which is $o(1)$ under Assumption 4(b).

Next consider $\mathcal{S}_{12}^{\mathrm{FM}}$,

$$\begin{aligned}
\mathcal{S}_{12}^{\mathrm{FM}} &= \frac{n}{TJ^d}\sum\nolimits_{t_1}\left[n_{t_1}^{-1}\sum\nolimits_{j_1}\hat{q}_{j_1t_1}^{-1}\sum\nolimits_{i_1}\hat{\mathbf{1}}_{j_1t_1}\hat{\mathbb{1}}_{j_1t_1}(\mathbf{z})\,\hat{\mathbb{1}}_{j_1t_1}(\mathbf{z}_{i_1t_1})\,\mathbf{x}_{i_1t_1}'\left(\hat{\beta}_{t_1} - \beta_{t_1}\right)\right]^2 \\
&= \frac{n}{TJ^d}\sum\nolimits_{t_1}\left[\hat{\mathbf{1}}_{\beta,t_1}\left(\hat{h}_{t_1}(\mathbf{z}) - h_{t_1}(\mathbf{z})\right)'\left(\hat{\beta}_{t_1} - \beta_{t_1}\right) + \mathbf{1}_{\beta,t_1}h_{t_1}(\mathbf{z})'\left(\hat{\beta}_{t_1} - \beta_{t_1}\right)\right]^2 \\
&\leq C\frac{n}{TJ^d}\sum\nolimits_{t_1}\hat{\mathbf{1}}_{\beta,t_1}\left[h_{t_1}(\mathbf{z})'\left(\hat{\beta}_{t_1} - \beta_{t_1}\right)\right]^2 + C\frac{n}{TJ^d}\sum\nolimits_{t_1}\hat{\mathbf{1}}_{\beta,t_1}\left[\left(\hat{h}_{t_1}(\mathbf{z}) - h_{t_1}(\mathbf{z})\right)'\left(\hat{\beta}_{t_1} - \beta_{t_1}\right)\right]^2 \\
&= \mathcal{S}_{121}^{\mathrm{FM}} + \mathcal{S}_{122}^{\mathrm{FM}}.
\end{aligned}$$

$\mathcal{S}_{121}^{\mathrm{FM}}$ follows by the same steps as in the proof for $\mathcal{L}_3$ and so is $o_p(1)$ under Assumptions 4(a) and 4(b). Next consider $\mathcal{S}_{122}^{\mathrm{FM}}$,

$$\begin{aligned}
&\frac{n}{TJ^d}\sum\nolimits_{t_1}\hat{\mathbf{1}}_{\beta,t_1}\left|\left(\hat{h}_t(\mathbf{z}) - h_t(\mathbf{z})\right)'\left(\hat{\beta}_{t_1} - \beta_{t_1}\right)\right|^2 \\
\leq\ & \frac{n}{TJ^d}\left(\sum\nolimits_{t_1}\hat{\mathbf{1}}_{\beta,t_1}\left|\left(\hat{h}_t(\mathbf{z}) - h_t(\mathbf{z})\right)'\left(\hat{\beta}_{t_1} - \beta_{t_1}\right)\right|\right)^2 \\
\leq\ & \frac{n}{TJ^d}\left(\sum\nolimits_{t_1}\hat{\mathbf{1}}_{\beta,t_1}\left\|\hat{h}_t(\mathbf{z}) - h_t(\mathbf{z})\right\|\left\|\hat{\beta}_{t_1} - \beta_{t_1}\right\|\right)^2 \\
\leq\ & \left(\left(\frac{n}{J^d}\right)T^{-1}\sum\nolimits_{t_1}\left\|\hat{h}_t(\mathbf{z}) - h_t(\mathbf{z})\right\|^2\right)\left(\sum\nolimits_{t_1}\hat{\mathbf{1}}_{\beta,t_1}\left\|\hat{\beta}_{t_1} - \beta_{t_1}\right\|^2\right).
\end{aligned}$$

The first factor is $O_p(1)$. To see this note that we can show that $\left(J^d n^{-1}\right)T^{-1}\sum_{t_1}\left(\hat{\mu}_t(\mathbf{z}) - \mu(\mathbf{z})\right)^2 = O_p(1)$ by showing $\left|\mathcal{S}_{11}^{\mathrm{FM}}\right| = o_p(1)$, $\left|\mathcal{S}_{12}^{\mathrm{FM}}\right| = o_p(1)$, and $\left|\mathcal{S}_2^{\mathrm{FM}}\right| = o_p(1)$. We can then follow the same steps to show that $\left(J^d n^{-1}\right)T^{-1}\sum_{t_1}\left\|\hat{h}_t(\mathbf{z}) - h_t(\mathbf{z})\right\|^2 = O_p(1)$. The second factor is $O_p\left(Tn^{-1}\right) + O_p\left(TJ^{-4}\right)$ which is $o_p(1)$ by Assumptions 4(a), 4(b), and 4(c).

Finally consider $\mathcal{S}_{13}^{\mathrm{FM}}$,

$$\mathcal{S}_{13}^{\mathrm{FM}} = \frac{n}{TJ^d}\sum\nolimits_{t_1}\left[\sum\nolimits_{j_1}\left(\hat{\mathbf{1}}_{j_1t_1} - 1\right)\hat{\mathbb{1}}_{j_1t_1}(\mathbf{z})\,\mu(\mathbf{z})\right]^2 \leq C\left(\frac{n}{J^d}\max_{1\leq t\leq T}\max_{1\leq j\leq J_t^d}\left|\hat{\mathbf{1}}_{jt} - 1\right|\right),$$

which is $o_p(1)$ by Lemma 4. Thus, $\mathcal{S}_{13}^{\mathrm{FM}} = o_p(1)$.

Next we need to show that $\frac{nT}{J^d}\left(\hat{\mathrm{V}}_{n,T}^{\mathrm{PI}} - \mathrm{V}_{n,T}\right) = o_p(1)$. First note that

$$\frac{nT}{J^d}\left(\hat{\mathrm{V}}_{n,T}^{\mathrm{PI}} - \mathrm{V}_{n,T}\right) = \frac{nT}{J^d}\left(\tilde{\mathrm{V}}_{n,T} - \mathrm{V}_{n,T}\right) + \mathcal{S}_1^{\mathrm{PI}} + \mathcal{S}_2^{\mathrm{PI}} + \mathcal{S}_3^{\mathrm{PI}},$$

where

$$\begin{aligned}
\mathcal{S}_1^{\mathrm{PI}} &= -\frac{n}{J^d T}\sum\nolimits_{t=1}^{T} n_t^{-2}\sum\nolimits_{i_1\neq i_2}\sum\nolimits_{j=1}^{J_t^d} \hat{\mathbf{1}}_{jt}\hat{\mathbb{1}}_{jt}(\mathbf{z})\,\hat{q}_{jt}^{-2}\hat{\mathbb{1}}_{jt}(\mathbf{z}_{i_1t})\,\hat{\mathbb{1}}_{jt}(\mathbf{z}_{i_2t})\,\varepsilon_{i_1t}\varepsilon_{i_2t} \\
\mathcal{S}_2^{\mathrm{PI}} &= \frac{2n}{J^d T}\sum\nolimits_{t=1}^{T} n_t^{-2}\sum\nolimits_{i=1}^{n_t}\sum\nolimits_{j=1}^{J_t^d} \hat{\mathbf{1}}_{jt}\hat{\mathbb{1}}_{jt}(\mathbf{z})\,\hat{q}_{jt}^{-2}\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\,\varepsilon_{it}\,(\hat{\varepsilon}_{it}-\varepsilon_{it}) \\
\mathcal{S}_3^{\mathrm{PI}} &= \frac{n}{J^d T}\sum\nolimits_{t=1}^{T} n_t^{-2}\sum\nolimits_{i=1}^{n_t}\sum\nolimits_{j=1}^{J_t^d} \hat{\mathbf{1}}_{jt}\hat{\mathbb{1}}_{jt}(\mathbf{z})\,\hat{q}_{jt}^{-2}\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\,(\hat{\varepsilon}_{it}-\varepsilon_{it})^2 .
\end{aligned}$$

Note that

$$\begin{aligned}
&(\hat{\varepsilon}_{i_1t_1} - \varepsilon_{i_1t_1}) \\
= \;& -T^{-1}\sum\nolimits_{t_2} n_{t_2}^{-1}\sum\nolimits_{j_2}\hat{q}_{j_2t_2}^{-1}\sum\nolimits_{i_2}\hat{\mathbf{1}}_{j_2t_2}\hat{\mathbb{1}}_{j_2t_2}(\mathbf{z}_{i_1t_1})\,\hat{\mathbb{1}}_{j_2t_2}(\mathbf{z}_{i_2t_2})\,\varepsilon_{i_2t_2} \\
& -T^{-1}\sum\nolimits_{t_2} n_{t_2}^{-1}\sum\nolimits_{j_2}\hat{q}_{j_2t_2}^{-1}\sum\nolimits_{i_2}\hat{\mathbf{1}}_{j_2t_2}\hat{\mathbb{1}}_{j_2t_2}(\mathbf{z}_{i_1t_1})\,\hat{\mathbb{1}}_{j_2t_2}(\mathbf{z}_{i_2t_2})\,(\mu(\mathbf{z}_{i_2t_2}) - \mu(\mathbf{z}_{i_1t_1})) \\
& +T^{-1}\sum\nolimits_{t_2} n_{t_2}^{-1}\sum\nolimits_{j_2}\hat{q}_{j_2t_2}^{-1}\sum\nolimits_{i_2}\hat{\mathbf{1}}_{j_2t_2}\hat{\mathbb{1}}_{j_2t_2}(\mathbf{z}_{i_1t_1})\,\hat{\mathbb{1}}_{j_2t_2}(\mathbf{z}_{i_2t_2})\left[\mathbf{x}_{i_2t_2}'\left(\hat{\beta}_{t_2}-\beta_{t_2}\right) - \mathbf{x}_{i_1t_1}'\left(\hat{\beta}_{t_1}-\beta_{t_1}\right)\right] \\
& -T^{-1}\sum\nolimits_{t_2}\sum\nolimits_{j_2}\left(\hat{\mathbf{1}}_{j_2t_2}-1\right)\hat{\mathbb{1}}_{j_2t_2}(\mathbf{z}_{i_1t_1})\,\mu(\mathbf{z}_{i_1t_1}) .
\end{aligned}$$

First consider $\mathcal{S}_1^{\mathrm{PI}}$:

$$\begin{aligned}
\mathbb{E}\left|\mathcal{S}_1^{\mathrm{PI}}\right|^2 &= \mathbb{E}\left|\frac{n}{J^d T}\sum\nolimits_{t=1}^{T} n_t^{-2}\sum\nolimits_{i_1\neq i_2}\sum\nolimits_{j=1}^{J_t^d}\hat{\mathbf{1}}_{jt}\hat{\mathbb{1}}_{jt}(\mathbf{z})\,\hat{q}_{jt}^{-2}\hat{\mathbb{1}}_{jt}(\mathbf{z}_{i_1t})\,\hat{\mathbb{1}}_{jt}(\mathbf{z}_{i_2t})\,\varepsilon_{i_1t}\varepsilon_{i_2t}\right|^2 \\
&= \left(\frac{n}{J^d T}\right)^2\sum\nolimits_{t_1,t_2} n_{t_1}^{-2}n_{t_2}^{-2}\sum\nolimits_{\substack{i_1\neq i_2,\\ i_3\neq i_4}}\sum\nolimits_{j_1,j_2}\mathbb{E}\left[\hat{\mathbf{1}}_{j_1t_1}\hat{\mathbf{1}}_{j_2t_2}\hat{\mathbb{1}}_{j_1t_1}(\mathbf{z})\,\hat{\mathbb{1}}_{j_2t_2}(\mathbf{z})\,\hat{q}_{j_1t_1}^{-2}\hat{q}_{j_2t_2}^{-2}\times\right. \\
&\qquad \left.\hat{\mathbb{1}}_{j_1t_1}(\mathbf{z}_{i_1t_1})\,\hat{\mathbb{1}}_{j_1t_1}(\mathbf{z}_{i_2t_1})\,\hat{\mathbb{1}}_{j_2t_2}(\mathbf{z}_{i_3t_2})\,\hat{\mathbb{1}}_{j_2t_2}(\mathbf{z}_{i_4t_2})\,\varepsilon_{i_1t_1}\varepsilon_{i_2t_1}\varepsilon_{i_3t_2}\varepsilon_{i_4t_2}\right] .
\end{aligned}$$

The expectation is nonzero only when $(t_1 = t_2)$ and either $(i_1 = i_3),(i_2 = i_4)$ or $(i_1 = i_4),(i_2 = i_3)$. This yields

$$\begin{aligned}
\mathbb{E}\left|\mathcal{S}_1^{\mathrm{PI}}\right|^2 &\leq C\left(\frac{n}{J^d T}\right)^2\sum\nolimits_t n_t^{-4}\sum\nolimits_{i_1,i_2}\sum\nolimits_{j_1,j_2}\mathbb{E}\left[\hat{\mathbf{1}}_{j_1t}\hat{\mathbf{1}}_{j_2t}\hat{\mathbb{1}}_{j_1t}(\mathbf{z})\,\hat{\mathbb{1}}_{j_2t}(\mathbf{z})\,\hat{q}_{j_1t}^{-2}\hat{q}_{j_2t}^{-2}\times\right. \\
&\qquad \left.\hat{\mathbb{1}}_{j_1t}(\mathbf{z}_{i_1t})\,\hat{\mathbb{1}}_{j_2t}(\mathbf{z}_{i_1t})\,\hat{\mathbb{1}}_{j_1t}(\mathbf{z}_{i_2t})\,\hat{\mathbb{1}}_{j_2t}(\mathbf{z}_{i_2t})\,\varepsilon_{i_1t}^2\varepsilon_{i_2t}^2\right] \\
&\quad +C\left(\frac{n}{J^d T}\right)^2\sum\nolimits_t n_t^{-4}\sum\nolimits_{i_1,i_2}\sum\nolimits_{j_1,j_2}\mathbb{E}\left[\hat{\mathbf{1}}_{j_1t}\hat{\mathbf{1}}_{j_2t}\hat{\mathbb{1}}_{j_1t}(\mathbf{z})\,\hat{\mathbb{1}}_{j_2t}(\mathbf{z})\,\hat{q}_{j_1t}^{-2}\hat{q}_{j_2t}^{-2}\times\right. \\
&\qquad \left.\hat{\mathbb{1}}_{j_1t}(\mathbf{z}_{i_1t})\,\hat{\mathbb{1}}_{j_2t}(\mathbf{z}_{i_1t})\,\hat{\mathbb{1}}_{j_1t}(\mathbf{z}_{i_2t})\,\hat{\mathbb{1}}_{j_2t}(\mathbf{z}_{i_2t})\,\varepsilon_{i_1t}^2\varepsilon_{i_2t}^2\right] \\
&= C\left(\frac{n}{J^d T}\right)^2\sum\nolimits_t n_t^{-4}\sum\nolimits_{i_1,i_2}\sum\nolimits_{j_1}\mathbb{E}\left[\hat{\mathbf{1}}_{j_1t}\hat{\mathbb{1}}_{j_1t}(\mathbf{z})\,\hat{q}_{j_1t}^{-4}\hat{\mathbb{1}}_{j_1t}(\mathbf{z}_{i_1t})\,\hat{\mathbb{1}}_{j_1t}(\mathbf{z}_{i_2t})\,\varepsilon_{i_1t}^2\varepsilon_{i_2t}^2\right] \\
&\leq C\left(\frac{n}{J^d T}\right)^2\sum\nolimits_t n_t^{-4}\sum\nolimits_{i_1,i_2}\sum\nolimits_{j_1}\mathbb{E}\left[\hat{\mathbf{1}}_{j_1t}\hat{\mathbb{1}}_{j_1t}(\mathbf{z})\,\hat{q}_{j_1t}^{-4}\hat{\mathbb{1}}_{j_1t}(\mathbf{z}_{i_1t})\,\hat{\mathbb{1}}_{j_1t}(\mathbf{z}_{i_2t})\right] \\
&= C\left(\frac{n}{J^d T}\right)^2\sum\nolimits_t n_t^{-2}\sum\nolimits_{j_1}\mathbb{E}\left[\mathbf{1}_{j_1t}\hat{\mathbb{1}}_{j_1t}(\mathbf{z})\,\hat{q}_{j_1t}^{-2}\right] \\
&\leq CJ^{2d}\left(\frac{n}{J^d T}\right)^2 Tn^{-2} \\
&= CT^{-1},
\end{aligned}$$

so that $\mathcal{S}_1^{\mathrm{PI}} = O_p\left(T^{-1/2}\right)$ by Markov's inequality. $\mathcal{S}_2^{\mathrm{PI}}$ and $\mathcal{S}_3^{\mathrm{PI}}$ follow by similar bounding arguments as above.

Proof of Theorem 3 Note first that

$$\mathcal{L}_1 = \mu'(\mathbf{z}) \cdot T^{-1} \sum_{t=1}^{T} \sum_{j=1}^{J_t^d} N_{jt}^{-1} \sum_{i=1}^{n_t} \hat{\mathbf{1}}_{jt} \hat{\mathbb{1}}_{jt}(\mathbf{z}) \hat{\mathbb{1}}_{jt}(\mathbf{z}_{it}) (\mathbf{z}_{it} - \mathbf{z}) + o_p(1).$$

Thus, the constant associated with the bias is $\mathcal{B} = \lim_{n,T\to\infty} \mathcal{B}_{n,T}$, where

$$\mathcal{B}_{n,T} = J \cdot \mu'(\mathbf{z}) \cdot T^{-1} \sum_{t=1}^{T} \sum_{j=1}^{J_t^d} N_{jt}^{-1} \sum_{i=1}^{n_t} \hat{\mathbf{1}}_{jt} \hat{\mathbb{1}}_{jt}(\mathbf{z}) \hat{\mathbb{1}}_{jt}(\mathbf{z}_{it}) (\mathbf{z}_{it} - \mathbf{z}).$$

For the constants associated with the variance note that $\mathcal{L}_2 = \mathcal{L}_{21} + \mathcal{L}_{22} + \mathcal{L}_{23}$ where

$$\begin{aligned}
\mathcal{L}_{21} &= T^{-1} \sum\nolimits_{t=1}^{T} n_t^{-1} \sum\nolimits_{i=1}^{n_t} \sum\nolimits_{j=1}^{J_t^d} \hat{\mathbf{1}}_{jt} \hat{\mathbb{1}}_{jt}(\mathbf{z})\, q_{jt}^{-1} \hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\, \varepsilon_{it} \\
\mathcal{L}_{22} &= T^{-1} \sum\nolimits_{t=1}^{T} n_t^{-1} \sum\nolimits_{i=1}^{n_t} \sum\nolimits_{j=1}^{J_t^d} \hat{\mathbf{1}}_{jt} \hat{\mathbb{1}}_{jt}(\mathbf{z})\, q_{jt}^{-2} (\hat{q}_{jt} - q_{jt})\, \hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\, \varepsilon_{it} \\
\mathcal{L}_{23} &= T^{-1} \sum\nolimits_{t=1}^{T} n_t^{-1} \sum\nolimits_{i=1}^{n_t} \sum\nolimits_{j=1}^{J_t^d} \hat{\mathbf{1}}_{jt} \hat{\mathbb{1}}_{jt}(\mathbf{z})\, \hat{q}_{jt}^{-1} q_{jt}^{-2} (\hat{q}_{jt} - q_{jt})^2\, \hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\, \varepsilon_{it}
\end{aligned}$$

By similar bounding arguments as in the proof of Theorem 2 it can be shown that $\mathcal{L}_{23} = O_p(T^{-1/2} n^{-3/2} J^{3d/2})$. For $\mathcal{L}_{22}$ we will assume that the quantiles are known (see footnote 5). Define $\tilde{q}_{jt} = n_t^{-1} \sum_{i=1}^{n_t} \mathbb{1}_{jt}(\mathbf{z}_{it})$ and

$$\tilde{\mathcal{L}}_{22} = T^{-1} \sum\nolimits_{t=1}^{T} n_t^{-1} \sum\nolimits_{i=1}^{n_t} \sum\nolimits_{j=1}^{J_t^d} \hat{\mathbf{1}}_{jt} \hat{\mathbb{1}}_{jt}(\mathbf{z})\, q_{jt}^{-2} (\tilde{q}_{jt} - q_{jt})\, \hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\, \varepsilon_{it}.$$

Then,

$$\begin{aligned}
\mathbb{E}\left[\tilde{\mathcal{L}}_{22}^2 \,\middle|\, \mathfrak{Z}, \mathfrak{X}, \mathcal{F}_1, \ldots, \mathcal{F}_T\right] &= T^{-2} \sum_{t=1}^{T} n_t^{-4} \sum_{j=1}^{J_t^d} \sum_{i=1}^{n_t} \mathbb{1}_{jt}(\mathbf{z}) q_{jt}^{-4} (q_{jt}^2 - 2q_{jt} + 1) \mathbb{1}_{jt}(\mathbf{z}_{it}) \sigma_{it}^2 \\
&\quad + T^{-2} \sum_{t=1}^{T} n_t^{-4} (n_t - 1) \sum_{j=1}^{J_t^d} \sum_{i=1}^{n_t} \mathbb{1}_{jt}(\mathbf{z}) q_{jt}^{-3} (q_{jt}^{-3} + 3q_{jt}^2 - 3q_{jt} + 1) \mathbb{1}_{jt}(\mathbf{z}_{it}) \sigma_{it}^2 \\
&= \mathcal{V}_{n,T}^{(2)} \times T^{-1} n^{-2} J^{2d} + o_p(1)
\end{aligned}$$

where

$$\mathcal{V}_{n,T}^{(2)} = T^{-1} J^{-2d} n^2 \sum_{t=1}^{T} n_t^{-4} (n_t - 1) \sum_{j=1}^{J_t^d} \sum_{i=1}^{n_t} \mathbb{1}_{jt}(\mathbf{z}) q_{jt}^{-3} \mathbb{1}_{jt}(\mathbf{z}_{it}) \sigma_{it}^2$$

Finally, $\mathcal{V}^{(2)} = \lim_{n,T\to\infty} \mathcal{V}_{n,T}^{(2)}$. By similar steps we have that $\mathcal{V}^{(1)} = \lim_{n,T\to\infty} \mathcal{V}_{n,T}^{(1)}$ where

$$\mathcal{V}_{n,T}^{(1)} = T^{-1} J^{-d} n \sum_{t=1}^{T} n_t^{-2} \sum_{j=1}^{J_t^d} \sum_{i=1}^{n_t} \mathbb{1}_{jt}(\mathbf{z}) q_{jt}^{-2} \mathbb{1}_{jt}(\mathbf{z}_{it}) \sigma_{it}^2.$$

Table A1: **Empirical Results**

This table reports point estimate and associated test statistics from the models specified in equation (14) (top panel) and equations (15) and (16) (bottom panel) where $J^\star$ has been chosen based on the estimand $\mu(z_H) - \mu(z_L)$ where (z_H, z_L) are listed in the second column of each panel. The standard estimator refers to the standard implementation based on 10 portfolios as described in Remark 1. Test statistics are formed using $\hat{V}_{\texttt{FM}}$ for the variance estimator. $\Phi(\cdot)$ is the CDF of a standard normal random variable. All returns are in monthly changes and all portfolios are value weighted based on lagged market equity.

Size Anomaly

		Point Estimate			*Test Statistic*		
	(z_H, z_L)	High	Low	Difference	High	Low	Difference
1926–2015	$(\Phi^{-1}(.975), \Phi^{-1}(.025))$	0.0089	0.0407	-0.0317	5.38	8.77	-6.45
	$(\Phi^{-1}(.95), \Phi^{-1}(.05))$	0.0088	0.0232	-0.0144	5.03	5.82	-3.31
	$(\Phi^{-1}(.9), \Phi^{-1}(.1))$	0.0107	0.0147	-0.0039	5.91	4.41	-1.04
	Standard Estimator	0.0089	0.0204	-0.0115	5.81	6.00	-3.09
1967–2015	$(\Phi^{-1}(.975), \Phi^{-1}(.025))$	0.0095	0.0464	-0.0369	4.70	8.68	-6.46
	$(\Phi^{-1}(.95), \Phi^{-1}(.05))$	0.0096	0.0227	-0.0131	4.63	6.36	-3.17
	$(\Phi^{-1}(.9), \Phi^{-1}(.1))$	0.0103	0.0137	-0.0034	4.83	4.32	-0.88
	Standard Estimator	0.0089	0.0183	-0.0094	4.93	5.59	-2.51
1980–2015	$(\Phi^{-1}(.975), \Phi^{-1}(.025))$	0.0107	0.0453	-0.0346	4.62	7.67	-5.46
	$(\Phi^{-1}(.95), \Phi^{-1}(.05))$	0.0111	0.0238	-0.0127	4.63	5.35	-2.51
	$(\Phi^{-1}(.9), \Phi^{-1}(.1))$	0.0108	0.0092	0.0016	4.45	2.58	0.36
	Standard Estimator	0.0101	0.0163	-0.0062	4.79	4.52	-1.49

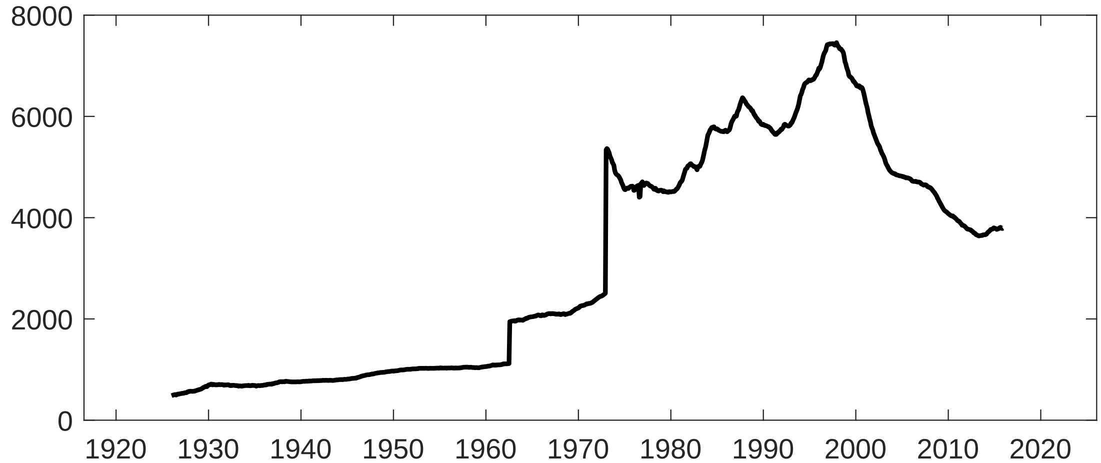
Momentum Anomaly

		Point Estimate			*Test Statistic*		
	(z_H, z_L)	High	Low	Difference	High	Low	Difference
1926–2015	$(\Phi^{-1}(.975), \Phi^{-1}(.025))$	0.0170	-0.0074	0.0244	7.39	-1.83	5.25
	w/ controls	0.0136	-0.0102	0.0238	3.57	-1.75	3.42
	$(\Phi^{-1}(.95), \Phi^{-1}(.05))$	0.0172	-0.0062	0.0234	7.74	-1.46	4.87
	w/ controls	0.0138	-0.0041	0.0179	3.31	-0.62	2.32
	$(\Phi^{-1}(.9), \Phi^{-1}(.1))$	0.0143	-0.0000	0.0152	6.64	-0.23	3.37
	w/ controls	0.0115	-0.0021	0.0136	3.03	-0.42	2.13
	Standard Estimator	0.0159	0.0000	0.0155	7.70	0.13	4.05
1967–2015	$(\Phi^{-1}(.975), \Phi^{-1}(.025))$	0.0175	-0.0082	0.0257	5.60	-1.76	4.58
	w/ controls	0.0146	-0.0168	0.0314	3.44	-2.01	3.35
	$(\Phi^{-1}(.95), \Phi^{-1}(.05))$	0.0163	-0.0047	0.0210	5.48	-1.07	3.94
	w/ controls	0.0131	-0.0125	0.0255	3.28	-1.77	3.16
	$(\Phi^{-1}(.9), \Phi^{-1}(.1))$	0.0157	-0.0063	0.0220	5.65	-1.41	4.20
	w/ controls	0.0083	-0.0131	0.0214	2.02	-2.35	3.09
	Standard Estimator	0.0156	-0.0023	0.0180	5.62	-0.58	3.66
1980–2015	$(\Phi^{-1}(.975), \Phi^{-1}(.025))$	0.0150	-0.0159	0.0309	4.13	-2.45	4.15
	w/ controls	0.0128	-0.0208	0.0336	2.35	-1.90	2.74
	$(\Phi^{-1}(.95), \Phi^{-1}(.05))$	0.0143	-0.0127	0.0270	4.09	-2.06	3.82
	w/ controls	0.0117	-0.0152	0.0269	2.20	-1.47	2.31
	$(\Phi^{-1}(.9), \Phi^{-1}(.1))$	0.0144	-0.0073	0.0216	4.47	-1.32	3.40
	w/ controls	0.0093	-0.0098	0.0191	1.67	-1.35	2.08
	Standard Estimator	0.0150	-0.0018	0.0168	4.59	-0.36	2.84

Figure 2: **Cross-Sectional Sample Sizes**

The top chart shows the monthly cross-section sample sizes over time, n_t, for the primary data set from the Center for Research in Security Prices (CRSP). The bottom chart shows the cross-section sample sizes over time for those stocks listed on the New York Stock Exchange (NYSE).

All



NYSE Only

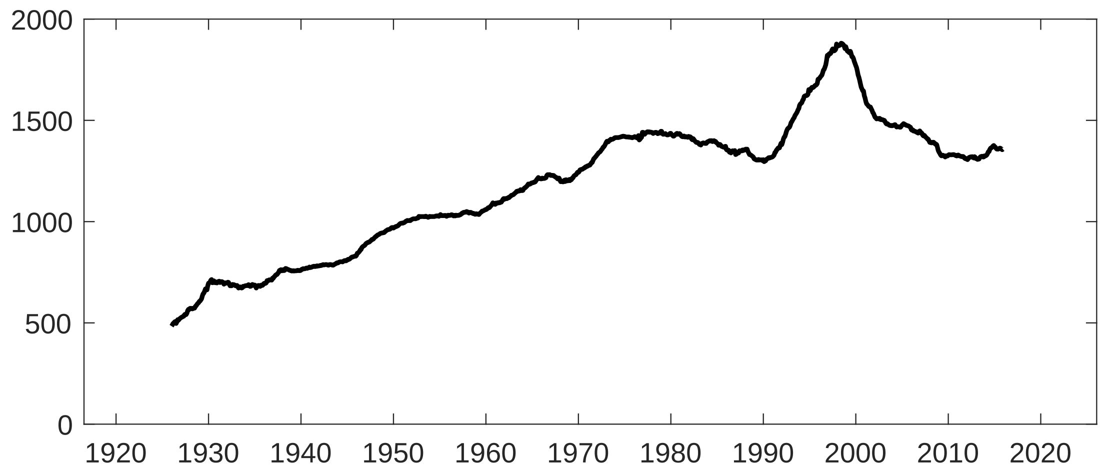
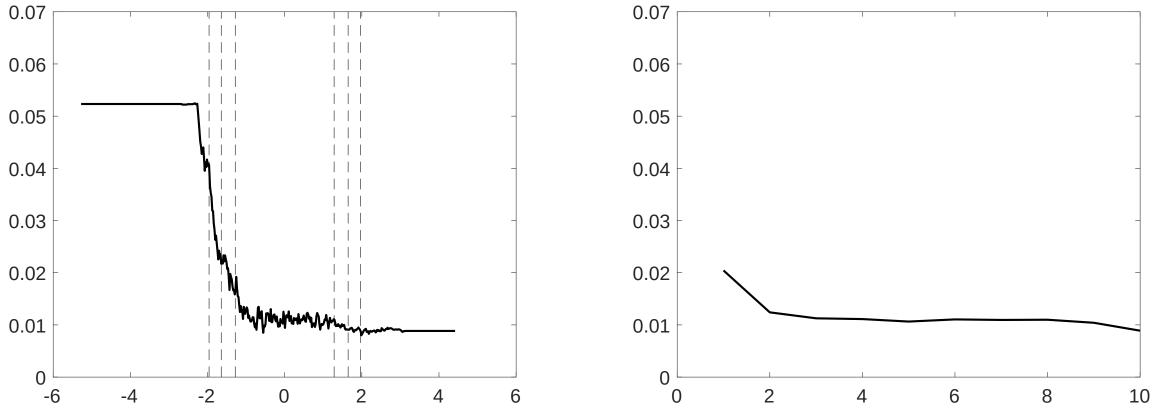


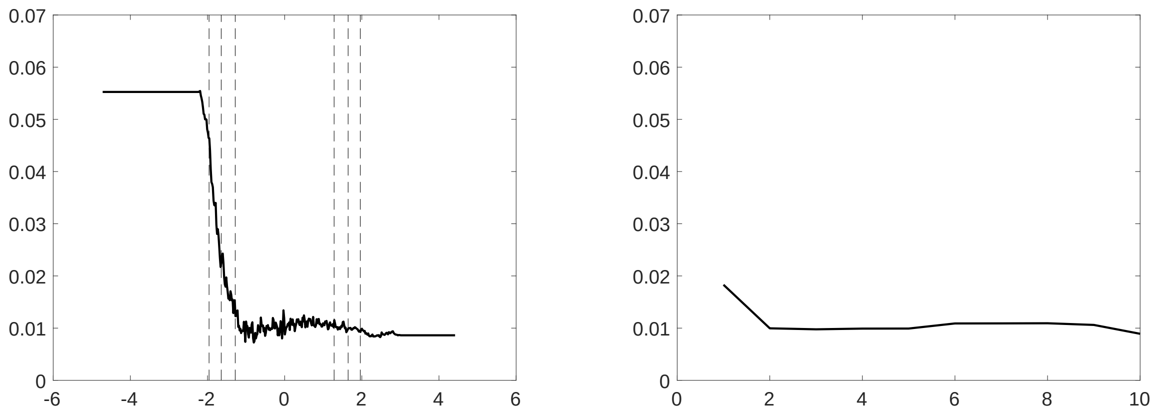
Figure 3: **Size Anomaly: All Stocks**

This figure shows the estimated relationship between the cross section of equity returns, R_{it} and lagged market equity, $\textsc{me}_{i(t-1)}$ as specified by equation (14). The left column displays $\{\hat{\mu}(z) : z \in \mathcal{Z}\}$ where $\{J_t : t = 1 \ldots T\}$ has been chosen based on equation (12), $z_H = \Phi^{-1}(.975)$, $z_L = \Phi^{-1}(.025)$, and $\Phi(\cdot)$ is the CDF of a standard normal random variable. The right column displays the estimated relationship using the standard portfolio sorting implementation based on 10 portfolios. Dotted lines designate $(\Phi^{-1}(.025), \Phi^{-1}(.975))$, $(\Phi^{-1}(.05), \Phi^{-1}(.95))$, and $(\Phi^{-1}(.1), \Phi^{-1}(.9))$, respectively. All returns are in monthly changes and all portfolios are value weighted based on lagged market equity.

1926–2015



1967–2015



1980–2015

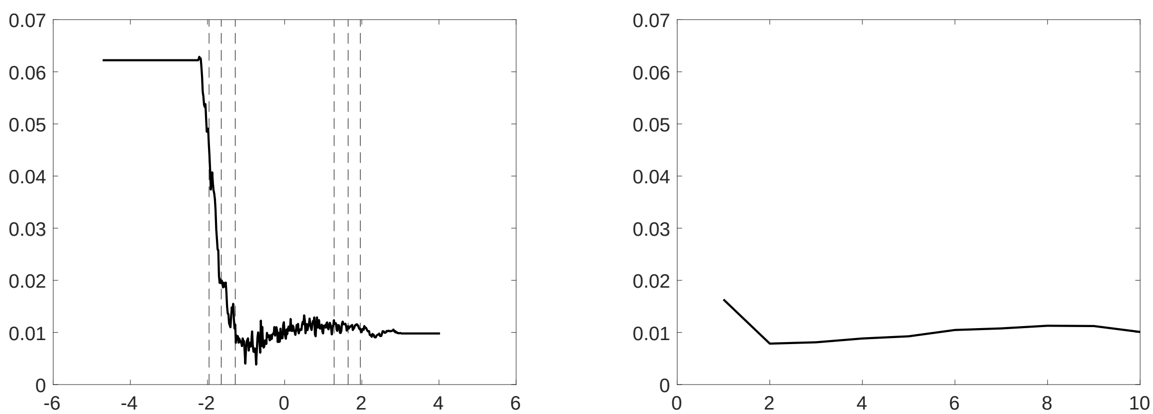
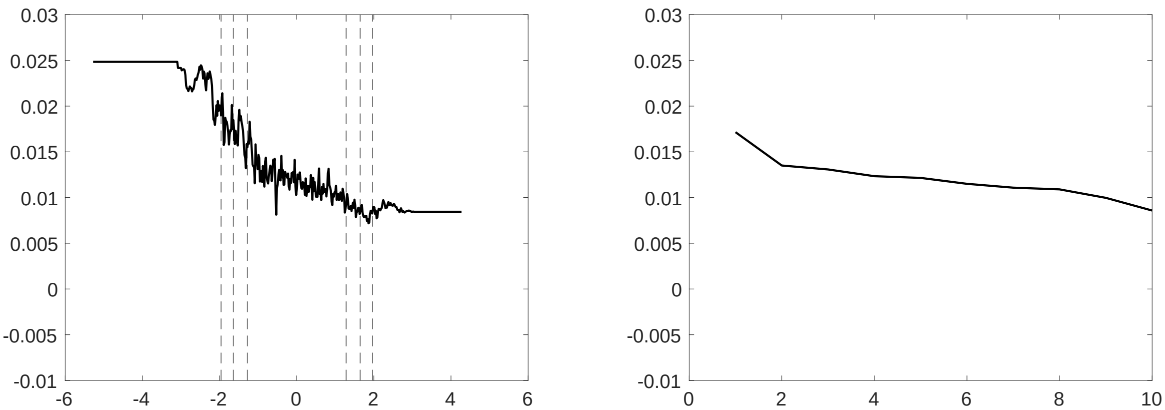


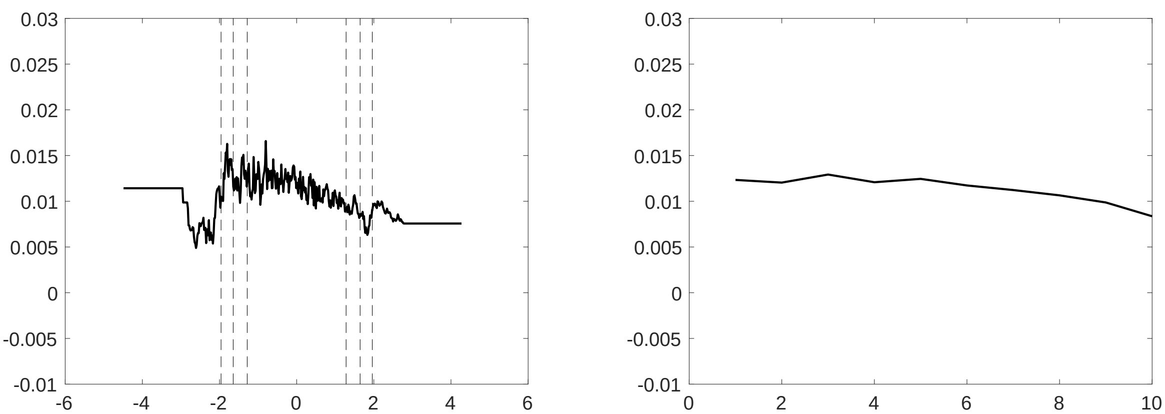
Figure 4: **Size Anomaly: NYSE Only**

This figure shows the estimated relationship between the cross section of equity returns, R_{it} and lagged market equity, $\mathrm{ME}_{i(t-1)}$ as specified by equation (14) restricted to firms listed on the New York Stock Exchange (NYSE). The left column displays $\{\hat{\mu}(z) : z \in \mathcal{Z}\}$ where $\{J_t : t = 1, \ldots, T\}$ has been chosen based on equation (12), $z_H = \Phi^{-1}(.975)$, $z_L = \Phi^{-1}(.025)$, and $\Phi(\cdot)$ is the CDF of a standard normal random variable. The right column displays the estimated relationship using the standard portfolio sorting implementation based on 10 portfolios. Dotted lines designate $(\Phi^{-1}(.025), \Phi^{-1}(.975))$, $(\Phi^{-1}(.05), \Phi^{-1}(.95))$, and $(\Phi^{-1}(.1), \Phi^{-1}(.9))$, respectively. All returns are in monthly changes and all portfolios are value weighted based on lagged market equity.

1926–2015



1967–2015



1980–2015

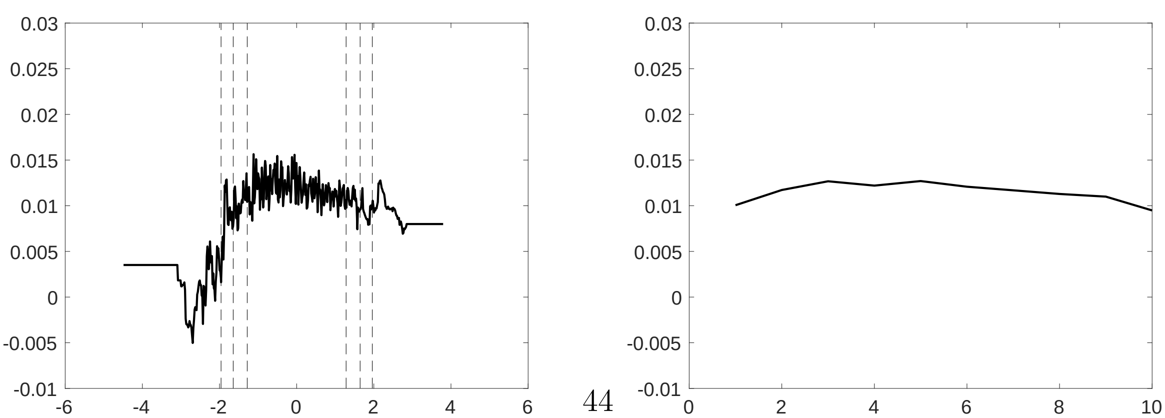
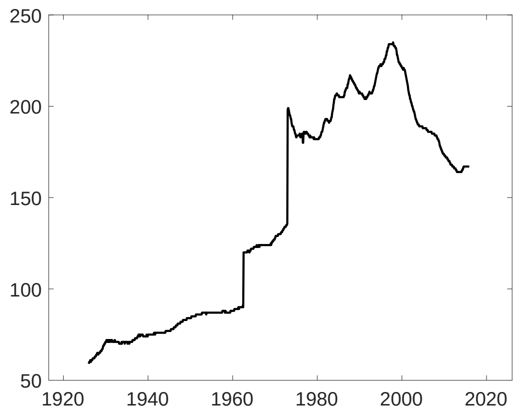


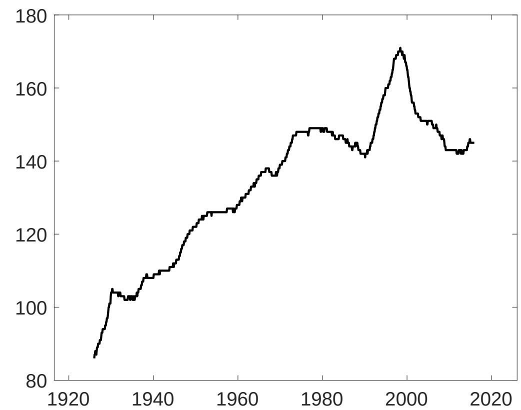
Figure 5: **Size Anomaly: Optimal Portfolios Counts**

This figure shows the optimal number of portfolios for the estimated relationship between the cross section of equity returns, R_{it} and lagged market equity, $\text{ME}_{i(t-1)}$ as specified by equation (14). $\{J_t : t = 1 \dots T\}$ has been chosen based on equation (12) where $z_H = \Phi^{-1}(.975)$, $z_L = \Phi^{-1}(.025)$, and $\Phi(\cdot)$ is the CDF of a standard normal random variable.

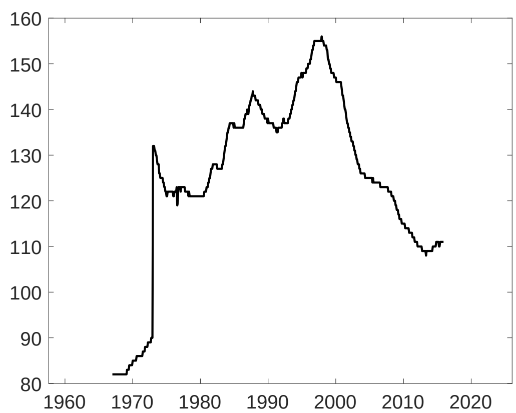
1926–2015, All Stocks



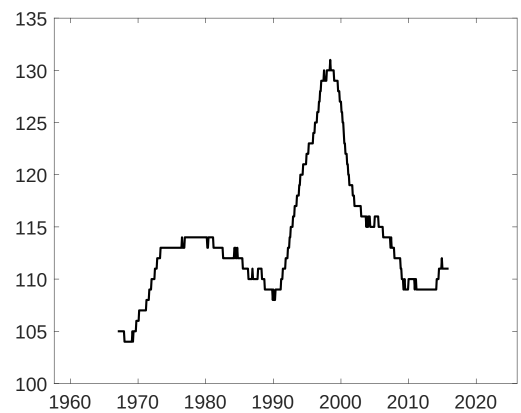
1926–2015, NYSE Only



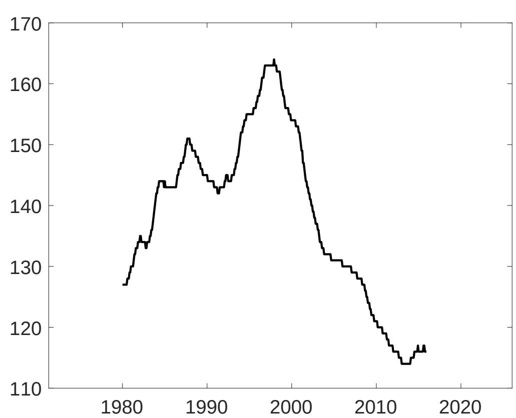
1967–2015, All Stocks



1967–2015, NYSE Only



1980–2015, All Stocks



1980–2015, NYSE Only

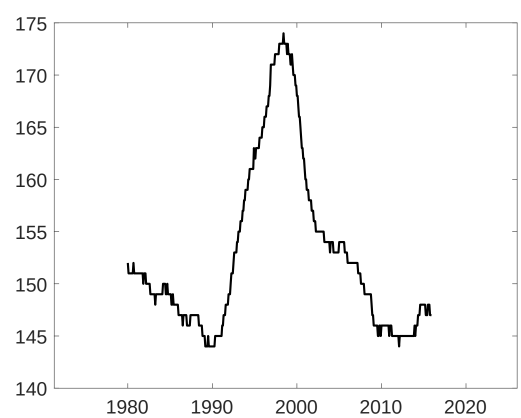
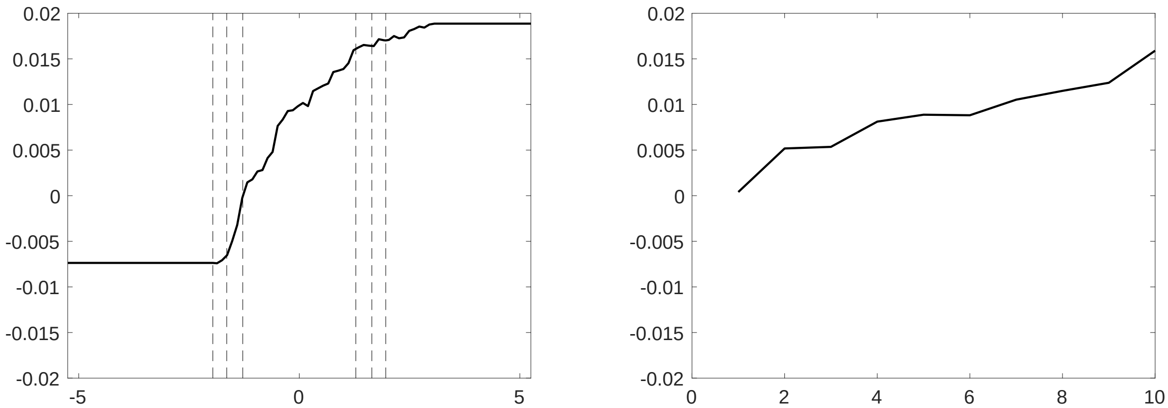


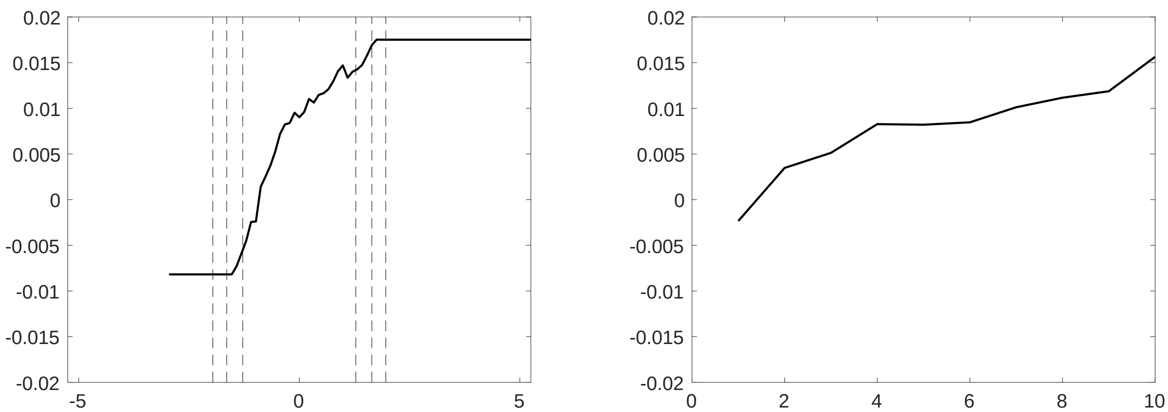
Figure 6: **Momentum Anomaly**

This figure shows the estimated relationship between the cross section of equity returns, R_{it} and 12-2 momentum, $\textsc{mom}_{it}$ as specified by equation (15). The left column displays $\{\hat{\mu}(z) : z \in \mathcal{Z}\}$ where $\{J_t : t = 1 \dots T\}$ has been chosen based on equation (12), $z_H = \Phi^{-1}(.975)$, $z_L = \Phi^{-1}(.025)$, and $\Phi(\cdot)$ is the CDF of a standard normal random variable. The right column displays the estimated relationship using the standard portfolio sorting implementation based on 10 portfolios. Dotted lines designate $(\Phi^{-1}(.025), \Phi^{-1}(.975))$, $(\Phi^{-1}(.05), \Phi^{-1}(.95))$, and $(\Phi^{-1}(.1), \Phi^{-1}(.9))$, respectively. All returns are in monthly changes and all portfolios are value weighted based on lagged market equity.

1927–2015



1967–2015



1980–2015

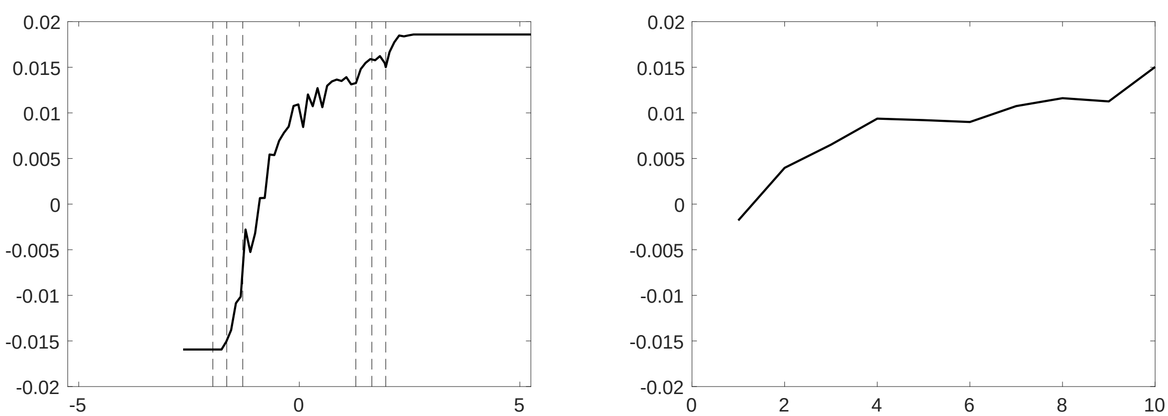
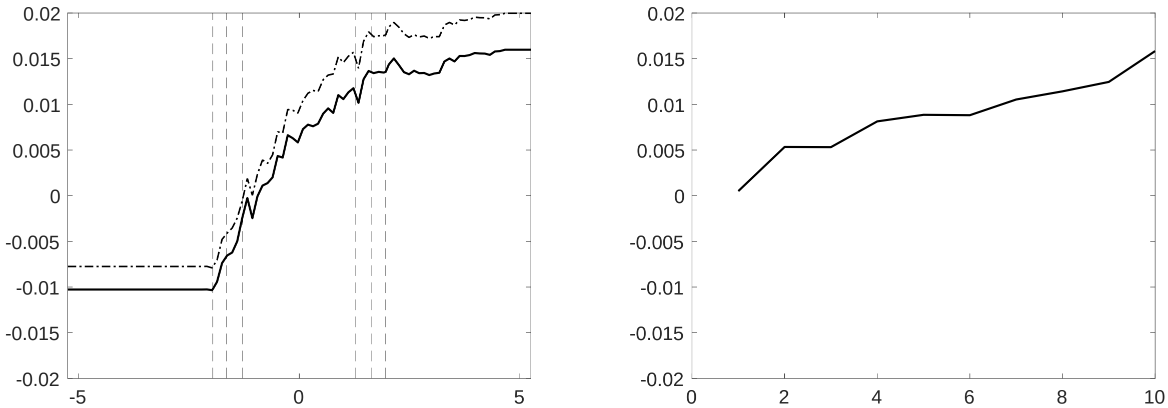


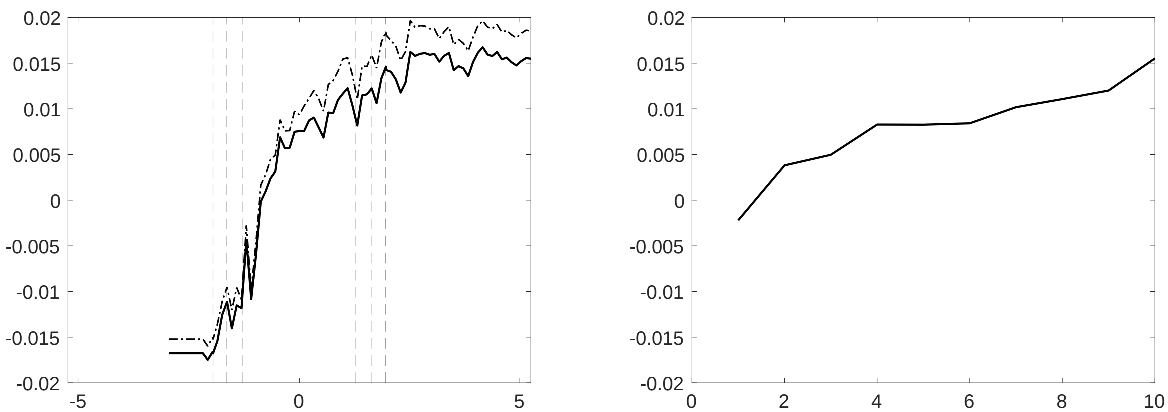
Figure 7: **Momentum Anomaly: Controlling for Industry Momentum**

This figure shows the estimated relationship between the cross section of equity returns, R_{it} and 12-2 momentum, $\textsc{mom}_{it}$. The left column displays $\{\hat{\mu}(z) : z \in \mathcal{Z}\}$ controlling for $\textsc{Imom}_{it}$, $\textsc{Imom}_{it}^2$ and $\textsc{Imom}_{it}^3$ (solid line) as specified by equation (16) where $\{J_t : t = 1 \ldots T\}$ has been chosen based on equation (12), $z_H = \Phi^{-1}(.975)$, $z_L = \Phi^{-1}(.025)$, and $\Phi(\cdot)$ is the CDF of a standard normal random variable. The dash-dotted line shows $\{\hat{\mu}(z) : z \in \mathcal{Z}\}$ without control variables as specified by equation (15) for the same $\{J_t : t = 1 \ldots T\}$. The right column displays the estimated relationship using the standard portfolio sorting implementation based on 10 portfolios (with no control variables). Dotted lines designate $(\Phi^{-1}(.025), \Phi^{-1}(.975))$, $(\Phi^{-1}(.05), \Phi^{-1}(.95))$, and $(\Phi^{-1}(.1), \Phi^{-1}(.9))$, respectively. All returns are in monthly changes and all portfolios are value weighted based on lagged market equity.

1927–2015



1967–2015



1980–2015

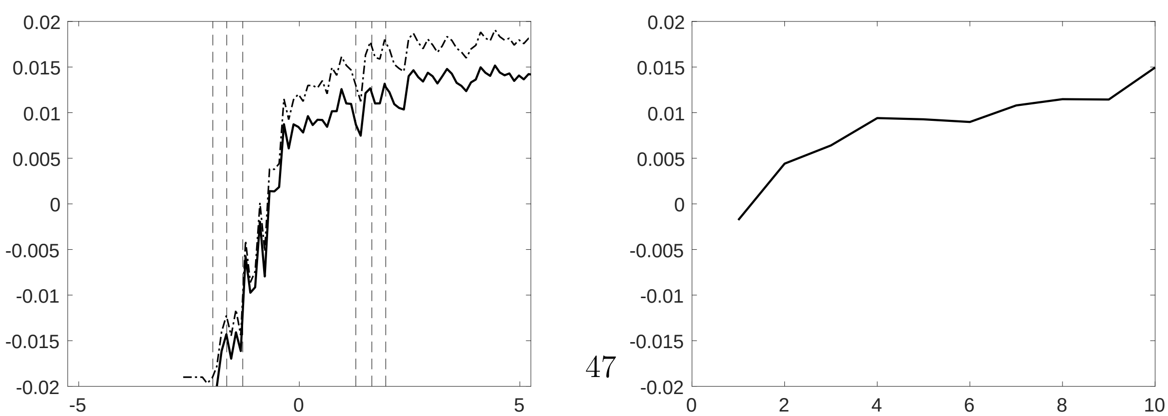
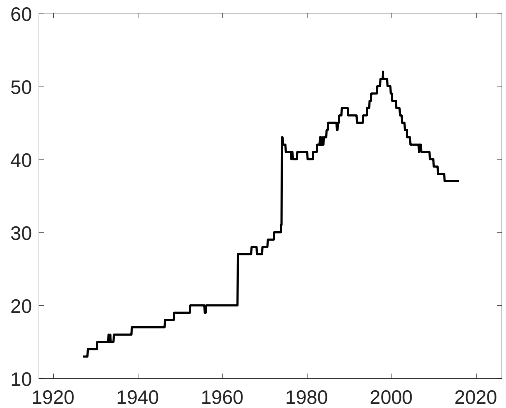


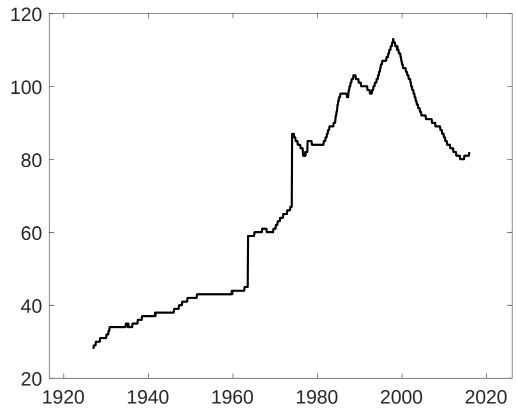
Figure 8: **Momentum Anomaly: Optimal Portfolios Counts**

This figure shows the optimal number of portfolios, $J_t^{\star}$, for the estimated relationship between the cross section of equity returns, R_{it} and 12-2 momentum, $\textsc{mom}_{it}$ as specified by equation (15) (left column) and 12-2 momentum, $\textsc{mom}_{it}$ controlling for $\textsc{Imom}_{it}$, $\textsc{Imom}_{it}^2$ and $\textsc{Imom}_{it}^3$ as specified by equation (16) (right column). $\{J_t : t = 1 \dots T\}$ has been chosen based on equation (12) where $z_H = \Phi^{-1}(.975)$, $z_L = \Phi^{-1}(.025)$, and $\Phi(\cdot)$ is the CDF of a standard normal random variable.

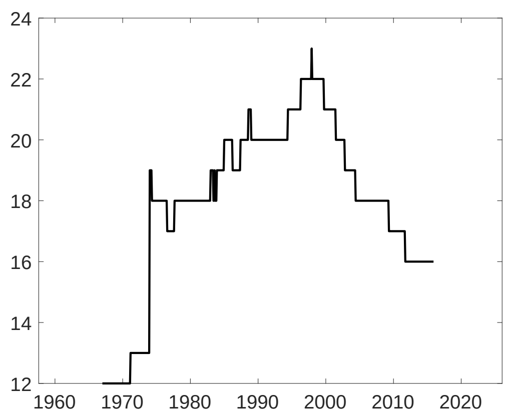
1927–2015, No Controls



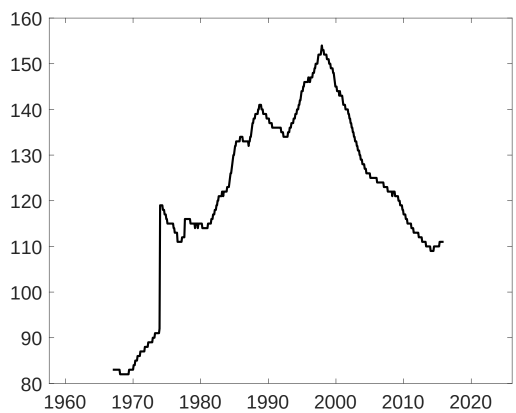
1927–2015, Controlling for Industry Momentum



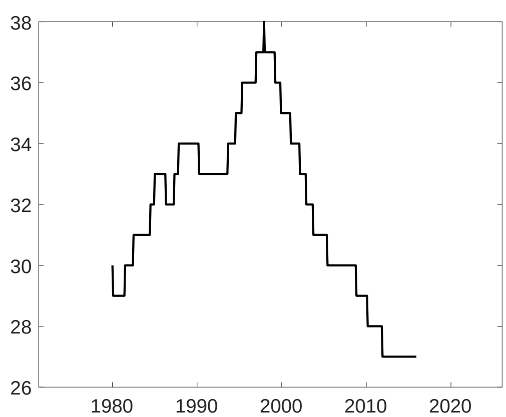
1967–2015, No Controls



1967-2015, Controlling for Industry Momentum



1980–2015, No Controls



1980-2015, Controlling for Industry Momentum