CHARACTERISTIC-SORTED PORTFOLIOS: ESTIMATION AND INFERENCE

Matias D. Cattaneo, Richard K. Crump, Max H. Farrell, and Ernst Schaumburg*

Abstract—Portfolio sorting is ubiquitous in the empirical finance literature, where it has been widely used to identify pricing anomalies. Despite its popularity, little attention has been paid to the statistical properties of the procedure. We develop a general framework for portfolio sorting by casting it as a nonparametric estimator. We present valid asymptotic inference methods and a valid mean square error expansion of the estimator leading to an optimal choice for the number of portfolios. In practical settings, the optimal choice may be much larger than the standard choices of five or ten. To illustrate the relevance of our results, we revisit the size and momentum anomalies.

I. Introduction

PORTFOLIO sorting is an important tool of modern empirical finance. It has been used to test fundamental theories in asset pricing, establish a number of different pricing anomalies, and identify profitable investment strategies. However, despite its ubiquity 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), De Bondt and Thaler (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, and across different assets; furthermore, portfolio sorting remains a highly popular tool in empirical finance.

Received for publication October 8, 2018. Revision accepted for publication March 6, 2019. Editor: Olivier Coibion.

*Cattaneo: Princeton University; Crump: Federal Reserve Bank of New York; Farrell: University of Chicago; Schaumburg: AQR Capital Management.

We thank Tobias Adrian, Francisco Barillas, Tim Bollerslev, Nina Boyarchenko, Jules van Binsbergen, John Campbell, Kent Daniel, Fernando Duarte, Eric Ghysels, Stefano Giglio, Peter Hansen, Ralph Koijen, Gabriele La Spada, Jonathan Lewellen, Jia Li, Erik Loualiche, Stefan Nagel, Andrew Patton, Karen Shen, George Tauchen, Stijn Van Nieuwerburgh, Peter Van Tassel, S. Viswanathan, Erik Vogt, Brian Weller, and Jonathan Wright, as well as seminar and conference participants at the 2015 Interactions Conference, the 2016 MFA Annual Meetings, University of Miami, and Duke University for helpful comments and discussions. Skanda Amarnath, Evan Friedman, and Rui Yu provided excellent research assistance. Last but not least, we thank the editor, Olivier Coibion, and three anonymous reviewers for their comments. The views expressed in this paper are our own 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. Farrell gratefully acknowledges financial support from the Richard N. Rosett and John E. Jeuck Fellowships.

A supplemental appendix is available online at http://www.mitpressjournals.org/doi/suppl/10.1162/rest_a_00883.

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 (Fama & French, 2008; Cochrane, 2011), but no formal framework is at present available in the literature. We impose sampling assumptions that 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 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 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 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.

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 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

The Review of Economics and Statistics, July 2020, 102(3): 531–551

https://doi.org/10.1162/rest_a_00883

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 ten portfolios when sorting on a single characteristic. This is summarized succinctly in Cochrane (2011, 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 that 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 for longer time series. Our results thus have a direct impact on empirical practice by providing a transparent, objective, data-driven way to choose the number of portfolios that nonetheless capture intuitive, real-world concerns in data analysis.

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 that have had better relative returns in the recent past also have higher future relative returns on average. We find that in the universe of U.S. stocks, the size anomaly is significant using our methods and is robust to different subperiods including the period from 1980 to 2015. Moreover, this conclusion would not be reached with the ad hoc, yet standard, choice of ten portfolios; our results are thus crucial for data analysis. Our results suggest that the relationship is monotonically decreasing and convex; this is borne out graphically. As pointed out in the existing literature, the size anomaly is not robust in subsamples that exclude "smaller" small firms (i.e., considering only firms listed on the NYSE). We also find that in the universe of U.S. stocks, the momentum anomaly is significant, with the "short" side of the trade becoming more profitable in later subperiods. Graphically, the relationship appears monotonically increasing and concave. 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 and, more important, that substantive conclusions change with the number of portfolios chosen for analysis. 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 is about 10 in the 1920s and about 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. For recent examples, see Shanken and Zhou (2007), Kleibergen and Zhan (2015), Nagel and Singleton (2011), Connor, Hagmann, and Linton (2012), Adrian, Crump, and Moench (2015), Ang, Liu, and Schwarz (forthcoming), and Gospodinov, Kan, and Robotti (2017), among others. In contrast, to our knowledge, we are the first to provide a formal framework and 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, Cooper, and Kaul (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 (Lewellen, Nagel, & Shanken, 2010; Kleibergen & Zhan, 2015).

The paper is organized as follows. Section II describes our framework and provides a brief overview of our new results. The more general framework is presented in section III. Then sections IV and V treat first-order asymptotic theory and mean square error expansions, respectively; the latter provides guidance on implementation. Section VI provides our empirical results, and section VII concludes and discusses further work.

II. 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 uses a basic building block of modern finance, a portfolio of assets, which produces an intuitive estimator of the relationship between asset returns and characteristics. 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

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

Here $\mu(\cdot)$ is the unknown object of interest that dictates how expected returns vary with the characteristic and is assumed to be twice 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. For each period t, it is common practice to form J disjoint portfolios, denoted by P_{jt}, as follows: $P_{jt} = [z_{(\lfloor n(j-1)/J \rfloor)t}, z_{(\lfloor nj/J \rfloor)t})$ if $j = 1, \ldots, J-1$, and $P_{Jt} = [z_{(\lfloor n(J-1)/J \rfloor)t}, z_{(n)t}]$, where $z_{(\ell)t}$ denotes the ℓth order statistic of the sample of characteristics $\{z_{it} : 1 \le i \le n\}$ at each time period $t = 1, 2, \ldots, T$, and $\lfloor \cdot \rfloor$ denotes the floor operator. In other words, each portfolio is a random interval containing roughly $(100/J)\%$ of the observations at each moment in time. This means that the position and size of the portfolios vary over time, but are 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

$$\hat{\mu}(z_*) = \frac{1}{T}\sum_{t=1}^{T} \hat{\mu}_t(z_*), \quad \hat{\mu}_t(z_*) = \frac{1}{N^*_{jt}} \sum_{i : z_{it} \in P^*_{jt}} R_{it}, \tag{2}$$

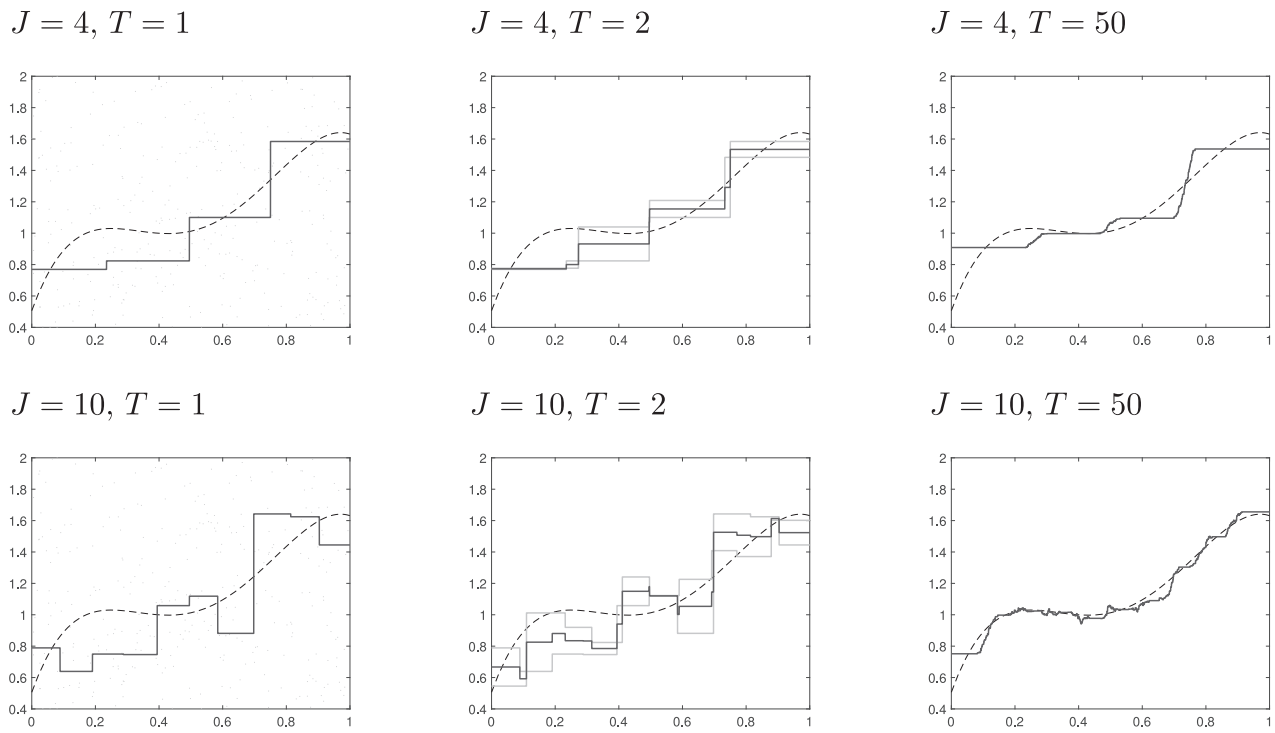
where N^*_{jt} is the number of assets in P^*_{jt} at time t. If $J \le n$, this estimator is well defined, as there are (roughly) n/J assets in all portfolios. The main motivation for using a sample average of each individual estimator is so that the procedure more closely mimics the actual practice of portfolio choice (where future returns are unknown) and because of the highly unbalanced nature of financial panel data. 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 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 (such as knots for spline regression). It is well documented that nonparametric inference is sensitive to tuning parameter choices, and empirical finance is no exception. 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 kernel or series regression), our goal is to explicitly analyze portfolio sorting. Such methods are not immune from tuning parameter choice sensitivity and may require stronger assumptions than portfolio sorting. From a practitioner's 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 (2), but with dramatically different results. Consider figure 1. The top-left plot shows a single realization of $\hat{\mu}_t(\cdot)$, with $J = 4$, for a single cross section. Moving to the top-center plot, 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, the top-right plot 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, the bottom row repeats the exercise but with $J = 10$. Comparing panels in the top row to the bottom of figure 1 shows the bias-variance trade-off. Figure 1 makes clear that J must depend on the features of the data at hand. 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 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

FIGURE 1.—INTRODUCTORY EXAMPLE



This figure shows the true (dashed line) and estimated function (solid line). The left panels show the $n = 500$ data points (gray dots), and the middle panels display the estimated function for each time period (light gray lines). Break points are chosen as estimated quantiles of z where $z \sim beta(1, 1)/z \sim beta(1.2, 1.2)$ for odd or even time periods.

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 thus require different choices of J.

First, consider the test of monotonicity, which is also interpreted as the return from a strategy of buying the spread portfolio: long \$1 of the higher expected return portfolio and short \$1 of the lower expected return portfolio. Formally, we wish to conduct the hypothesis test,

$$\begin{aligned} &\mathsf{H}_0 : \mu(z_H) - \mu(z_L) = 0 \quad \text{versus} \\ &\mathsf{H}_1 : \mu(z_H) - \mu(z_L) \neq 0, \end{aligned} \tag{3}$$

where $z_L < z_H$ denotes "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. Our general framework allows for a richer class of estimands (see remark 4), but this estimand will remain our focus throughout the paper because it is the most relevant to empirical researchers.

Our main result establishes asymptotic validity for testing equation (3) using portfolio sorting with estimated quantiles. Namely, it follows from (the more general) theorem 1 that

$$\mathcal{T} = \frac{\big[\hat{\mu}(z_H) - \hat{\mu}(z_L)\big] - \big[\mu(z_H) - \mu(z_L)\big]}{\sqrt{\hat{V}(z_H) + \hat{V}(z_L)}} \to_{\mathrm{d}} \mathcal{N}(0, 1),$$

provided that $J \log(\max(J, T))/n \to 0$ and $nT/J^3 \to 0$, and other regularity conditions hold. The growth restrictions on J formalize the bias-variance trade-off in this problem.

Consistent variance estimation can be done in several ways. 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 and that $\hat{V}(z) \asymp J/(nT)$. 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, as is a novel plug-in approach. Both are given in equation (9). (See theorem 2 for a complete discussion.) 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-sorting-based inference. The literature typically employs ad hoc choices, and often $J = 10$ (see the quotation from Cochrane, 2011, above). However, given the nonparametric nature of the problem, J should depend on the features of the data and, moreover, should change over time because cross-sectional sample sizes vary substantially. To make this clear notationally, we will write J_t for the number of portfolios in period t. Even if these facts are recognized by empirical

researchers and the need for $J \neq 10$ is clear, a lack of principled tools may be holding back practice. Our results fill this gap by providing a transparent, data-driven method of portfolio choice, so that practitioners who wish to use something other than ten may do so in a replicable, objective way. For example, in our data, n ranges from 500 to nearly 8,000 (see section VI in the text and figure A1 in the supplemental appendix) and the optimal choice of J_t, for example, varies from 13 to 52 for the momentum anomaly (figure 5).

In the context of hypothesis testing, as in equation (3), we find that the optimal number of portfolios obeys

$$J_t^{\star} = K^{\star} n_t^{1/2} T^{1/4}, \quad t = 1, 2, \ldots, T,$$

where the constant $K^{\star}$ depends on the data-generating process. It is easy to check that $J_t^{\star}$ satisfies the conditions above (i.e., those for theorem 1). In section V, we detail the constant terms and discuss implementation in applications. Turning to factor construction, we find a different choice of J will be optimal,

$$J_t^{\star\star} = K^{\star\star} n_t^{1/3} T^{1/3}, \quad t = 1, 2, \ldots, T,$$

where, again, portfolios are chosen separately at each time, $K^{\star}$ depends on the data generating process, and implementation is discussed in section V. The major difference here is that for point estimation, the optimal number of portfolios, $J_t^{\star\star}$, diverges more slowly than for hypothesis testing, $J_t^{\star}$, 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 9 for further discussion.)

We illustrate the use and importance of our results in our empirical applications (section VI). As a preview, consider the momentum anomaly. We find that in the universe of U.S. stocks, the momentum anomaly results in statistically significant average returns, both overall and also individually for the long side and short side of the trade (see table 1). Graphically, the relationship between past relative returns and current returns appears monotonically increasing and concave, shown in figure 2. Alongside we show the results using the standard approach based on 10 portfolios. This makes clear that these same conclusions would not be reached using the conventional estimator.

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 equation (2), is exactly the standard portfolio sorting estimator that enjoys widespread use in empirical finance. We exploit the assumed structure that $\mu(z)$ is constant over time as a function of the characteristic value itself, which allows for intuitive and interpretable estimation and inference about $\mu(z)$ at $z \neq \{z_L, z_H\}$. The analogous assumption implicitly required in standard portfolio sorting is that $\mu(\cdot)$ is constant over time as a function of the (random) cross-sectional order statistic of the characteristics, that is, the ranks. These two overlap in the special case when z_H and z_L are always in the extreme portfolios. 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. In these broad terms, then, the main contribution of our paper is a formal asymptotic treatment of the standard portfolio-sorts test on $\hat{\mu}(z_H) - \hat{\mu}(z_L)$, 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).

An alternative interpretation that unifies the two approaches, which researchers may hold implicitly, is as inference on the grand mean at that point, even if $\mu(\cdot)$ is not constant in z itself or in its rank. That is, recalling equation (2), we interpret the estimand as (the limit of) $\bar{\mu}(z_*) = \sum_{t=1}^{T} \mu_t(z_*)/T$. When z_H and z_L are always in the extreme portfolios, this interpretation may be natural and the quantity $\hat{\mu}(z_H) - \hat{\mu}(z_L)$ directly interpretable. Our method accommodates this interpretation without substantive change.

Remark 1 (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 & MacBeth, 1973). This approach is motivated by a model of the form $R_{it} = \zeta z_{it} + \varepsilon_{it}$, $i = 1, \ldots, n_t$, $t = 1, \ldots, T$, 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 6 below).

III. General Asset Returns Model and Sorting Estimator

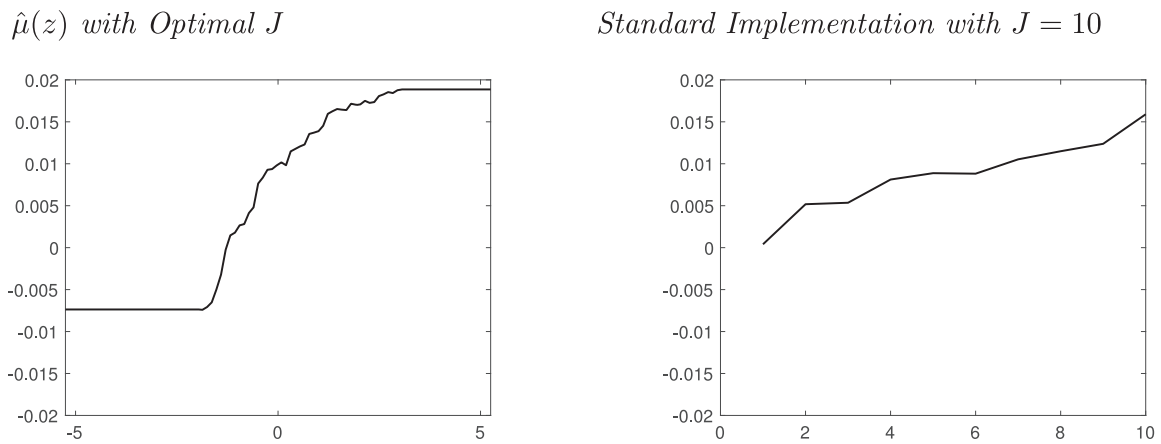
In this section, we study a more general model and develop a correspondingly general characteristic-sorted portfolio estimator. We extend beyond the simple case of the previous section in two directions. First, we allow for multiple sorting characteristics, such that z_{it} is replaced by $\mathbf{z}_{it} \in \mathcal{Z} \subset \mathbb{R}^d$. 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 as d increases, as we can make precise (see also section IIIA and remark 6). 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 parametric fashion.

TABLE 1.—EMPIRICAL RESULTS

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

This table reports point estimate and associated test statistics from the models specified in equation (13) (top panel) and equations (14) and (15) (bottom panel) using $J_t^\star$. The standard estimator refers to the standard implementation with $J = 10$. Test statistics are formed using $\hat{V}_{\text{FM}}$ for the variance estimator. All returns are in monthly changes, and all portfolios are value weighted based on lagged market equity.

FIGURE 2.—MOMENTUM ANOMALY EXAMPLE



This figure shows the estimated relation between equity returns and 12-2 momentum. The left column shows $\hat{\mu}(z)$ using $J_t^\star$; the right column shows the estimated relation using the standard implementation with $J = 10$. All returns are in monthly changes, and all portfolios are value weighted based on lagged market equity. The sample period is 1927 to 2015.

Formally, our model for asset returns is

$$R_{it} = \mu(\mathbf{z}_{it}) + \mathbf{x}_{it}'\boldsymbol{\beta}_t + \varepsilon_{it}, \quad i = 1, 2, \ldots, n_t, \quad t = 1, 2, \ldots, T. \tag{4}$$

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 (4) 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 equation (3) 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. 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—for example, 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 4 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, which also includes regularity conditions for our asymptotic results.

Assumption 1 (Data-Generating Process). Let the sigma fields $\mathcal{F}_t = \sigma(\mathbf{f}_t)$ be generated from a sequence of unobserved (possibly dependent) random vectors $\{\mathbf{f}_t : t = 0, 1, \ldots, T\}$. For $t = 1, 2, \ldots, T$, the following conditions hold.

(a) Conditional on $\mathcal{F}_t$, $\{(R_{it}, \mathbf{z}_{it}', \mathbf{x}_{it}')' : i = 1, 2, \ldots, n_t\}$ are i.i.d. satisfying model (4).

(b) $\mathbb{E}[\varepsilon_{it}|\mathbf{z}_{it}, \mathbf{x}_{it}, \mathcal{F}_t] = 0$; uniformly in t, $\Omega_{\mathrm{uu},t} = \mathbb{E}[\mathbb{V}(\mathbf{x}_{it}|\mathbf{z}_{it}, \mathcal{F}_t)]$ is bounded and its minimum eigenvalue is bounded away from 0, $\sigma_{it}^2 = \mathbb{E}[|\varepsilon_{it}|^2|\mathbf{z}_{it}, \mathbf{x}_{it}, \mathcal{F}_t]$ is bounded and bounded away from 0, $\mathbb{E}[|\varepsilon_{it}|^{2+\phi}|\mathbf{z}_{it}, \mathbf{x}_{it}, \mathcal{F}_t]$ is bounded for some $\phi > 0$, and $\mathbb{E}[\mathbf{a}'\mathbf{x}_{it}|\mathbf{z}_{it}, \mathcal{F}_t]$ is sub-Gaussian for all $\mathbf{a} \in \mathbb{R}^{d_x}$.

(c) Conditional on $\mathcal{F}_t$, $\mathbf{z}_{it}$ has time-invariant support, denoted $\mathcal{Z}$, and continuous Lebesgue density bounded away from 0.

(d) $\mu(\mathbf{z})$ is twice continuously differentiable; $|\mathbb{E}[x_{it,\ell}|\mathbf{z}_{it} = \mathbf{z}, \mathcal{F}_t] - \mathbb{E}[x_{it,\ell}|\mathbf{z}_{it} = \mathbf{z}', \mathcal{F}_t]| \leq C\|\mathbf{z} - \mathbf{z}'\|$ for all $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}$ where $x_{it,\ell}$ is the ℓth element of $\mathbf{x}_{it}$ and the constant C is not a function of t or $\mathcal{F}_t$.

These conditions allow for considerable flexibility in the behavior of the time series of returns and the cross-sectional dependence. Indeed, Andrews (2005, 1552), using the same condition in a single cross-section, called assumption 1a "surprisingly general." The setup allows for dependence and conditional heteroskedasticity across assets and time. For example, if $\mathbf{f}_t$ were to include a business cycle variable, 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 important, 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 $\mathbf{z}_{it}$ (De Bondt & Thaler, 1985; Jegadeesh, 1990; Lehmann, 1990; Jegadeesh & Titman, 1993, 2001).

Assumption 1 requires that the density of $\mathbf{z}_{it}$ be bounded away from 0, for each $t = 1, 2, \ldots, T$, which is useful to form (asymptotically) nonempty portfolios. The assumption that the support of the characteristics is the same across time-series observations is common when studying panel data. The other restrictions are mostly regularity conditions standard in the (cross-sectional) semi-/nonparametric literature, related to boundedness of moments and smoothness conditions of unknown functions. 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).

In the context of model (4), the portfolio-sorting estimator of $\mu(\mathbf{z})$ retains the structure given above in equation (2), 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 II, 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 than they are at the beginning of the sample. For example, in our empirical applications the largest cross-sectional sample size is approximately fifteen times the smallest.

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, 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 (Goyal, 2012). 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 different conceptually, as discussed below.

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 break points, 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

$$\hat{\mu}(\mathbf{z}) = \frac{1}{T}\sum_{t=1}^{T} \hat{\mu}_t(\mathbf{z}),$$

$$\hat{\mu}_t(\mathbf{z}) = \sum_{j=1}^{J_t^d} \frac{1}{N_{jt}} \sum_{i=1}^{n_t} \hat{\mathbf{1}}_{jt}\hat{\mathbb{1}}_{jt}(\mathbf{z})\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})(R_{it} - \mathbf{x}_{it}'\hat{\boldsymbol{\beta}}_t), \qquad (5)$$

where

$$\hat{\boldsymbol{\beta}}_t = (\mathbf{X}_t'\mathbf{M}_t\mathbf{X}_t)^{-1}\mathbf{X}_t'\mathbf{M}_t\mathbf{R}_t, \quad \mathbf{R}_t = [R_{1t}, \ldots, R_{n_t t}]',$$
$$\mathbf{X}_t = [\mathbf{x}_{1t}, \mathbf{x}_{2t}, \ldots, \mathbf{x}_{n_t t}]', \quad \mathbf{M}_t = \mathbf{I}_{n_t} - \hat{\mathbf{B}}_t(\hat{\mathbf{B}}_t'\hat{\mathbf{B}}_t)^{-1}\hat{\mathbf{B}}_t', \quad (6)$$

and $\hat{\mathbf{B}}_t = \hat{\mathbf{B}}_t(\mathbf{z}_t)$ with $\mathbf{z}_t = [\mathbf{z}_{1t}, \mathbf{z}_{2t}, \ldots, \mathbf{z}_{n_t t}]'$ 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 1 if P_{jt} is nonempty and $(\mathbf{X}_t'\mathbf{M}_t\mathbf{X}_t/n_t)$ is invertible. Both events occur with probability approaching 1 (see the supplemental appendix). It is established there that $N_{jt} \asymp n_t/J_t^d$ with probability approaching 1, for all j and t.

Remark 2 (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 $\mathbf{X}_t$. 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 ease, we present our theory without portfolio weights, but all empirical results in section VI are based on the value-weighted portfolio estimator.

It worth emphasizing that the nonparametric estimator $\hat{\mu}_t(\mathbf{z})$ of equation (5) 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 3 (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. This is fundamentally (and analytically) different from what we study in this paper, and this approach does not, in general, enjoy the usual interpretation of estimating the effect of $\mathbf{z}_{it}$ on R_{it} controlling for additional variables. In contrast, our framework retains the standard interpretation through the additive separability assumption as described by model (4).

A. *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 section, we discuss how our framework relates to conditional sorts, based on two distinct interpretations of conditional sorting: first as conditional testing and second as a mechanical solution to empty portfolios.

To fix ideas, consider firm size and credit rating. Small firms are less likely to have high credit ratings, and so in the "high" credit rating portfolio, there may be no truly small firms. Directly applying equation (5) would thus yield empty portfolios. Conditional sorts "solve" the empty portfolios problem by construction: first sorting by rating and then within each rating-based portfolio, by size, but have the issue that the "small firm" portfolio within the highest rating portfolio will typically have larger firms than conditional on lower ratings.

But this may not present a problem if we seek to study whether smaller firms still earn higher average returns if we keep credit rating fixed (section VI finds evidence for the size anomaly marginally). To answer this question, we could test the "high minus low" hypothesis within each credit-based portfolio. Our framework directly applies here, that is, the results and discussion in the following sections, provided one is careful to interpret the results conditionally on the first sort. Further, if $\mu(\cdot)$ is truly monotonic in size, then these

conditional results can be extrapolated to "fill" the empty bins, but our theory does not justify this.

A second interpretation of conditional sorts is that they are designed solely to solve the problem of empty portfolios. This is distinct from the above, and our framework does not apply here because 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 (or interest is in a specific grand mean, as above, though here mixing qualitatively different firms). 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 to empty portfolios would be to assume additive separability of the function $\mu(\cdot)$ so that if we denote the d components of $\mathbf{z}_{it}$ by $z_{it,1}, \ldots, z_{it,d}$, we suppose

$$R_{it} = \mu_1(z_{it,1}) + \cdots + \mu_d(z_{it,d}) + \varepsilon_{it} \quad i = 1, \ldots, n_t, \quad t = 1, \ldots, T, \tag{7}$$

and so each characteristic affects returns via its 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 6).

IV. First-Order Asymptotic Theory

With the estimator fully described, we now present consistency and asymptotic normality results and two valid standard error estimators. To our knowledge, these results are all new to the literature. As discussed in section II, 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 model (4) and the conditions placed on it by assumption 1, we will require certain rate restrictions for our asymptotic results. We now make these precise, grouped into the following two assumptions.

Assumption 2 (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 uniformly bounded away from 0.

Assumption 2 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 a sequence $J \to \infty$ such that $J_t \propto J$ for all t. Neither of these is 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 notations n and J for common growth rates enable 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 3 (Rate Restrictions). The sequences n, T, and J obey *(a)* $n^{-1}J^d \log(\max(J^d, T)) \log(n) \to 0$, *(b)* $\sqrt{nT}J^{-(d/2+1)} \to 0$, and, if $d_x \geq 1$, *(c)* $T/n \to 0$.

Assumption 3a ensures that all J_t grow slowly enough that the variance of the nonparametric estimator is well controlled and all portfolios are nonempty, while assumption 3b ensures the nonparametric smoothing bias is negligible. Finally, assumption 3c 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 3a and 3b.

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 or d is large. To illustrate, suppose for the moment that $J \asymp n^A$ and $T \asymp n^B$. Assumptions 3a and 3b 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 1, 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:

$$V(\mathbf{z}) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t^d}\frac{1}{N_{jt}}\sum_{i=1}^{n_t}\hat{\mathbf{1}}_{jt}\hat{\mathbb{1}}_{jt}(\mathbf{z})\hat{\mathbb{1}}_{jt}(\mathbf{z}_{it})\sigma_{it}^2. \tag{8}$$

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 equation (3) 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 can now state our first main result.

Theorem 1 *(Asymptotic Distribution). Suppose assumptions 1, 2, and 3 hold. Then,*

$$V(\mathbf{z})^{-1/2}(\hat{\mu}(\mathbf{z}) - \mu(\mathbf{z}))$$

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

where

$$V(\mathbf{z}) \asymp \frac{J^d}{nT} \quad \textit{and}$$

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

Theorem 1 shows that the properly normalized and centered estimator $\hat{\mu}(z)$ has a limiting normal distribution. The flexibility of the nonparametric specification between returns and (some) characteristics comes at the expense of slower convergence—the factor $J^{-d/2}$. Theorem 1 also makes clear why assumption 3b is necessary: the bias of the estimator is of the order J^{-1}; thus, once the rate $J^{-d/2}\sqrt{nT}$ is applied, assumption 3b 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 4.

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

$$\hat{V}_{\texttt{FM}}(\mathbf{z}) = \frac{1}{T^2} \sum_{t=1}^{T} (\hat{\mu}_t(\mathbf{z}) - \hat{\mu}(\mathbf{z}))^2 \quad \text{and}$$

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

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 options.

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

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

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

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 valid only 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. 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. Our empirical results in section VI 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 1. *Let the conditions of theorem 2 hold. Then,*

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

where $\hat{V}(\mathbf{z})$ *may be* $\hat{V}_{\texttt{FM}}$ *or* $\hat{V}_{\texttt{PI}}$ *as defined in equation (9).*

Section II states this same result, simplified to 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 important, under what conditions it may fail.

Remark 4 (Other Estimands). As we have discussed, 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 $z^{(1)}, z^{(2)}, \ldots, z^{(d)}$, then for some subset of these of size $\delta < d$, the object of interest is $\int_{\times_{\ell=1}^{\delta}} \mu(\mathbf{z}) w\big(z^{(1)}, z^{(2)}, \ldots, z^{(\delta)}\big) dz^{(1)} dz^{(2)} \cdots dz^{(\delta)}$, 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. 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 equation (3):

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

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 provide valid inference.

Remark 5 (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 equation 3: $\mathsf{H}_0 : \mu(\mathbf{z}_H) - \mu(\mathbf{z}_L) = 0$. 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 the monotonicity of $\mu(\cdot)$ rather than using $\mu(\mathbf{z}_H) - \mu(\mathbf{z}_L)$ as a proxy (see the discussion in section II). Building on Cattaneo, Farrell, and Feng (forthcoming) and Cattaneo, Crump, Farrell, and Feng (2019), it may be 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 nontrivial additional technical work but would allow us to test monotonicity, concavity, and many other hypotheses of interest, such as testing for a U-shaped relationship (Hong, Lim, & Stein, 2000), or for the existence of any profitable trading strategy via $\mathsf{H}_0 : |\max_z \mu(\mathbf{z}) - \min_z \mu(\mathbf{z})| = 0$.

Remark 6 (Analogy to Cross-Sectional Regressions). As we discussed in remark 1, cross-sectional regressions are the parametric alternative to portfolio sorting. In practice, however, the more natural parametric alternatives 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 (7) of section IIIA. 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 3 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.

V. Mean Square 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 the choice of the number of portfolios. With the estimator defined as in equation (5), 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 the 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 five or ten portfolios. Here we 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. To simplify the calculations, this section assumes 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, not the corresponding rates (see Calonico, Cattaneo, & Titiunik, 2015, for a related example and more discussion). 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 and that the marginal quantiles of* $\mathbf{z}$ *are known. Then,*

$$\begin{aligned}\mathbb{E}\Big[\big(\big[\hat{\mu}(\mathbf{z}_H) - \hat{\mu}(\mathbf{z}_L)\big] - \big[\mu(\mathbf{z}_H) - \mu(\mathbf{z}_L)\big]\big)^2 \Big| \, \mathfrak{Z}, \\ \mathfrak{X}, \mathcal{F}_1, \ldots, \mathcal{F}_T \Big] \\ = \mathcal{V}^{(1)} \frac{J^d}{nT} + \mathcal{V}^{(2)} \frac{J^{2d}}{n^2 T} + \mathcal{B}^2 \frac{1}{J^2} + \mathcal{C} \frac{J^{3d/2}}{n^{3/2} T^{3/2}} \\ + O_{\mathbb{P}}\left(\frac{1}{nT}\right) + o_{\mathbb{P}}\left(J^{-2} + \frac{J^{2d}}{n^2 T}\right),\end{aligned}$$

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* $\mathcal{B}_t(\mathbf{z})$, $\mathcal{V}_t^{(1)}(\mathbf{z})$, $\mathcal{V}_t^{(2)}(\mathbf{z})$ *and* $\mathcal{C}$ *are defined in the supplementary appendix. The term* $\mathcal{C}$ *is (conditionally) mean 0, and 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 appropriate regularity conditions on the time-series structure (e.g., mixing conditions), it can be shown that $\bar{\mathcal{B}} = \text{plim}_{n,T\to\infty} \mathcal{B}$, $\bar{\mathcal{V}}^{(1)} = \text{plim}_{n,T\to\infty} \mathcal{V}^{(1)}$, $\bar{\mathcal{V}}^{(2)} = \text{plim}_{n,T\to\infty} \mathcal{V}^{(2)}$, where $\bar{\mathcal{B}}$, $\bar{\mathcal{V}}^{(1)}$, and $\bar{\mathcal{V}}^{(2)}$ are nonrandom and nonzero 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 (see Cattaneo, Crump, & Jansson, 2010, for another application of this logic). This approach is optimal in an inference-targeted sense because it minimizes the two leading terms not accounted for by the approximation in 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

$$J_t^\star = \left\lfloor \left(\frac{\bar{\mathcal{B}}^2}{d\bar{\mathcal{V}}^{(2)}} \left(n_t^2 T\right) \right)^{\frac{1}{2d+2}} \right\rfloor, \tag{10}$$

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 3: the condition required remains that $Bd < 2$, for $T \asymp n^B$, which limits the number of sorting characteristics or the length of time series allowed (see the discussion of assumption 3). To gain intuition for $J_t^\star$, consider the simple case of a univariate, homoskedastic linear model: $\mu(\mathbf{z}) = bz$, $\sigma_{it}^2 = \sigma^2$. Then $\mathcal{B}^2 \propto |b|^2$ and $\mathcal{V} \propto \sigma^2$, and so a steeper line (larger $|b|$) calls for more portfolios, whereas more idiosyncratic noise (larger σ^2) calls for fewer.

To make this choice practicable, we can select J to minimize a sample version of the MSE expansion underlying equation (10),

$$\widehat{\mathbb{MSE}}\left(\hat{\mu}(\mathbf{z}_H) - \hat{\mu}(\mathbf{z}_L); J\right) = \hat{\mathcal{V}}^{(2)} \frac{J^{2d}}{n^2 T} + \hat{\mathcal{B}}^2 \frac{1}{J^2}, \tag{11}$$

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 (11) (see the supplementary appendix for further details). Alternatively, if we had pilot estimates of $\mathcal{V}^{(2)}$ and $\mathcal{B}$, we could directly utilize the formula in equation (10) to obtain a choice for each J_t.

Remark 7 (Undersmoothing). A common practice throughout semi- and nonparametric 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 (see Calonico, Cattaneo, & Farrell, 2018, 2019, for more discussion). In contrast, the choice of $J_t^\star$ of equation (10) 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 $J_t^\star$.

Remark 8 (Parametric Component). 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. It can be shown that for any real, nonzero vector $\mathbf{a} \in \mathbb{R}^{d_x}$,

$$\frac{\frac{1}{T}\sum_{t=1}^T \mathbf{a}'(\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t)}{\sqrt{\frac{1}{T^2}\sum_{t=1}^T (\mathbf{a}'(\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t))^2}} \to_{\mathrm{d}} \mathcal{N}(0,1). \tag{12}$$

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 9 (Constructing Factors). Theorem 3 can be also be used when the goal is point estimation rather than inference. Using the leading variance term and the bias, we obtain

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

which is different in the constants but more important, 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 equities where the cross-sectional sample size is much larger than the number of time periods, it will be the case that $J_t^{\star\star} = o(J_t^\star)$—that is, that the optimal number of portfolios is smaller when constructing factors than when conducting inference on whether expected returns vary significantly with characteristics. Informally, this has been recognized in the empirical literature as the number of portfolios used to construct factors has been relatively small (Fama & French, 1993). As discussed in the supplement, a feasible version of $J_t^{\star\star}$ can be constructed following the steps as in equation (11), replacing $\hat{\mathcal{V}}^{(2)} J^{2d}/(n^2 T)$ with $\hat{\mathcal{V}}^{(1)} J^d/(nT)$.

VI. 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 (Banz, 1981; Reinganum, 1981) and the momentum anomaly (Jegadeesh & Titman, 1993).

A. *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 data to firms listed on the New York Stock Exchange (NYSE), American Stock Exchange (AMEX), or Nasdaq and use only returns on common shares (CRSP share code 10 or 11). To deal with delisting returns, we follow the procedure described in Bali, Engle, and Murray (2016). 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 twelve months ago ($t - 12$) until one month prior to the current month ($t - 2$). The one-month gap is to avoid confounding the momentum anomaly variable with the short-term reversal anomaly (Jegadeesh, 1990; Lehmann, 1990). 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. To construct the industry momentum variable, we form a value-weighted average of each individual firm's momentum variable within the industry. We use thirteenth-month lagged market capitalization to form weights so they are unaffected by any subsequent changes in price.

We implement the estimator introduced in section III as follows. Since the underlying data are monthly, 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 (11), 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. 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. Moreover, there are sharp jumps in cross-sectional sample sizes that occur in 1962 and 1972 that reflect the addition of firms listed on the AMEX and Nasdaq to the sample (see figure A1 in the supplemental appendix). Even for the subset of firms listed on the NYSE, 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 currently slightly below 1,500 firms.

B. *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 size of the firm. Thus, following the notation of section III, we have

$$R_{it} = \mu(\mathrm{ME}_{i(t-1)}) + \varepsilon_{it}, \quad i = 1, \ldots, n_t, \quad t = 1, \ldots, T. \quad (13)$$

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

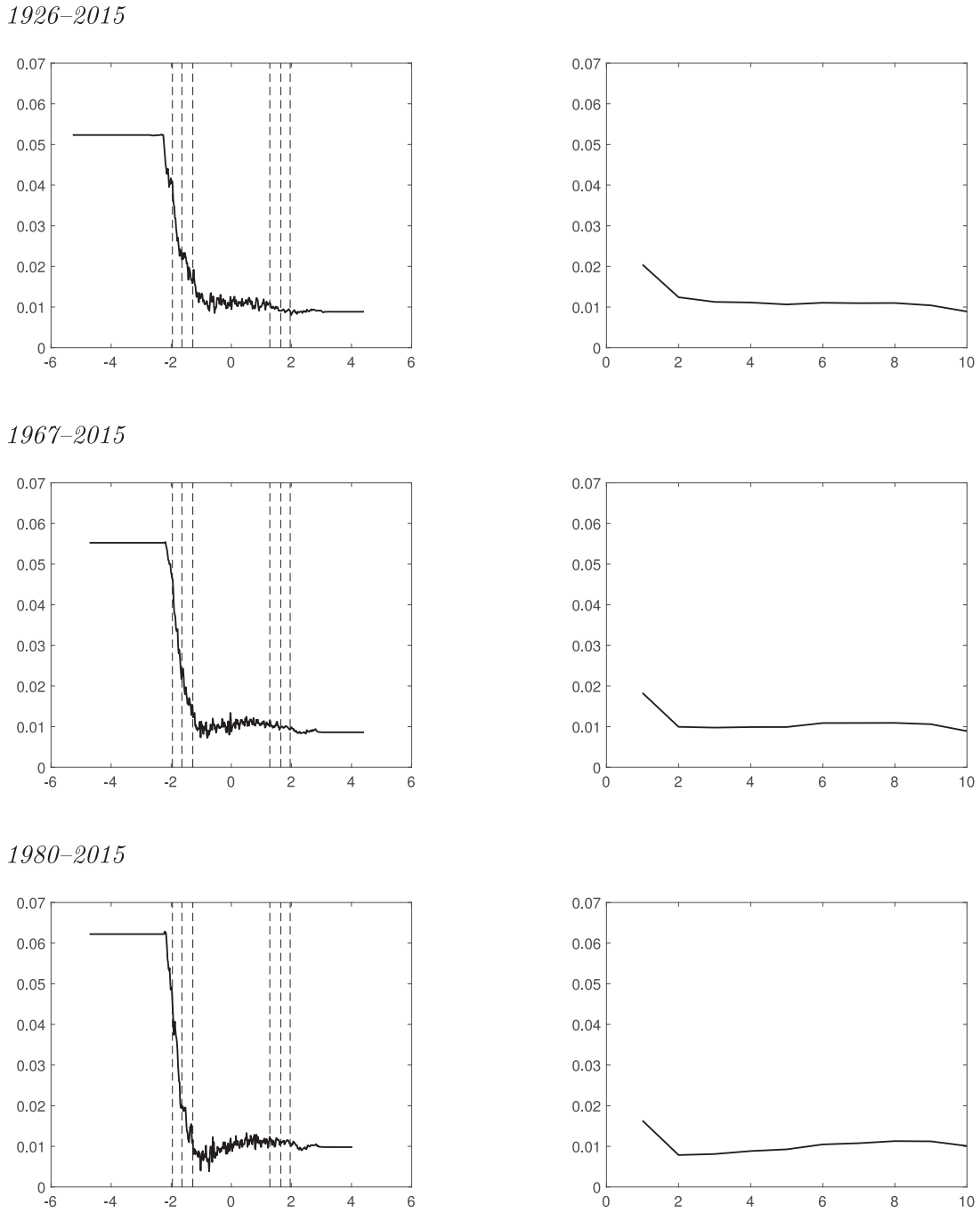
Figure 3 provides the estimates of the relationship between returns and firm size. The left column shows the estimate, $\{\hat{\mu}(z) : z \in \mathcal{Z}\}$, based on equation (5), whereas the right column plots the average return in each of ten portfolios formed based on the conventional approach currently used in the literature. The portfolio break points for the standard approach are commonly chosen using either deciles of the subsample of firms listed on the NYSE or deciles based on the entire sample. Here we choose deciles based on the latter as they ensure better comparability across estimators. 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 III and IV, 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.

The estimate, $\{\hat{\mu}(z) : z \in \mathcal{Z}\}$, is shown for three different subsamples in figure 3: 1926–2015, 1967–2015, and 1980–2015. The estimated shape between returns and size is generally very similar across the three subperiods 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 to 2015, and slightly above 6% over the sample 1980 to 2015.

Table 1 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 $(\Phi^{-1}(.9), \Phi^{-1}(.1))$, where $\Phi(\cdot)$ is the CDF of a standard normal random variable, shown as vertical lines in figure 3. The table also shows the point estimates and corresponding test statistics from the conventional approach using ten portfolios. Over all three subperiods, 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 to 2015, the t-statistic is -5.46. This is also the case when the evaluation points are shifted inward to $(\Phi^{-1}(.95), \Phi^{-1}(.05))$. 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 0 over the past 35 or so years. Instead, what has happened is that "larger" small firms are no longer producing higher returns in the last subsample. This pattern can be seen in the innermost set of evaluation points, $(z_H, z_L) = (\Phi^{-1}(.9), \Phi^{-1}(.1))$, where the size effect is estimated to be reversed, albeit statistically indistinguishable from 0.

To further investigate the results of table 1, we reconsider the estimates for the relationship between returns and firm

FIGURE 3.—SIZE ANOMALY: ALL STOCKS

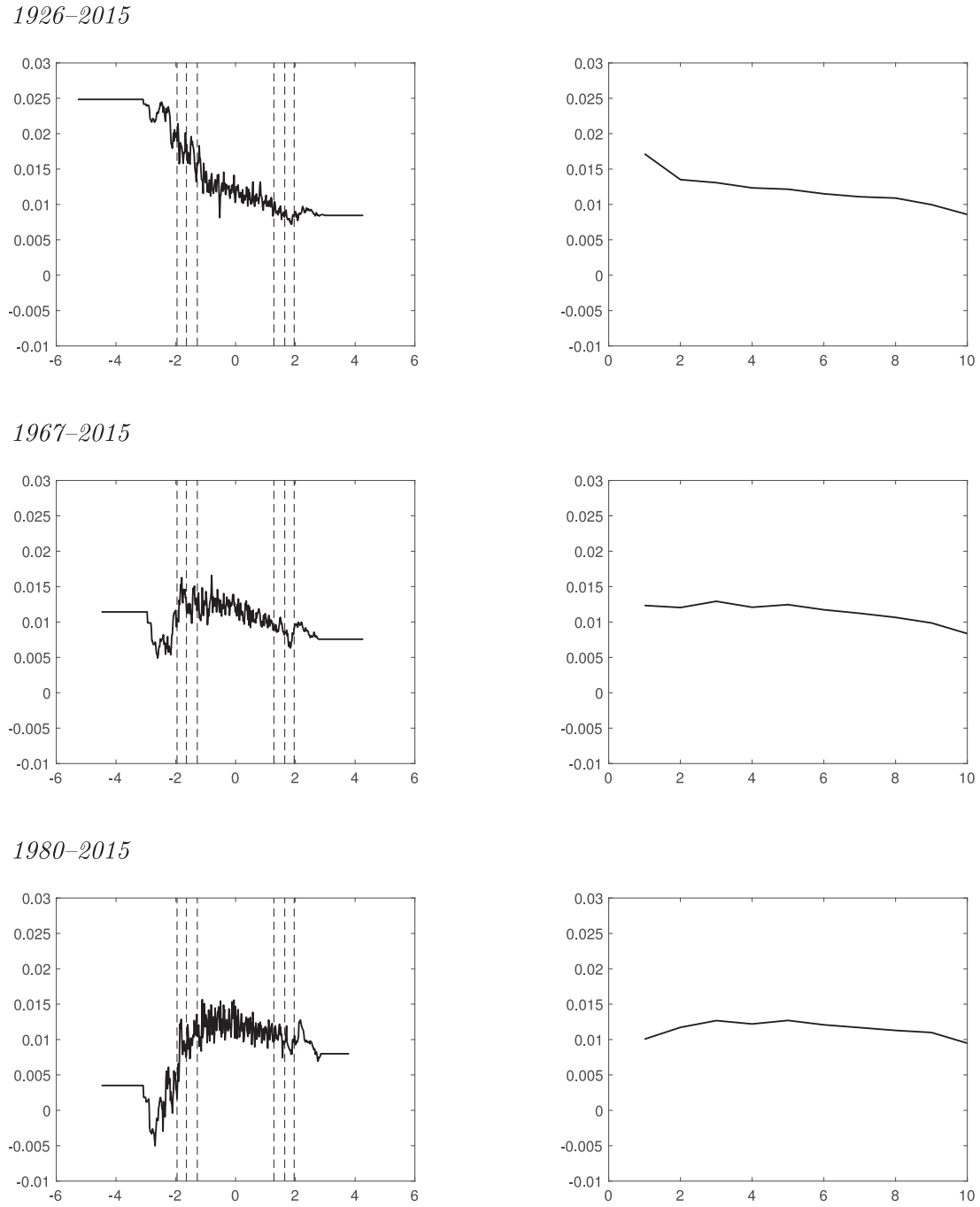


This figure shows the estimated relation between the cross section of equity returns and lagged market equity, equation (13). The left column displays $\hat{\mu}(\cdot)$ where J_t has been chosen based on equation (11), $z_H = \Phi^{-1}(.975)$, $z_L = \Phi^{-1}(.025)$. The right column displays the estimated relation using the standard portfolio-sorting implementation with $J = 10$. All returns are in monthly changes, and all portfolios are value-weighted based on lagged market equity.

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 to 2015 and 1980 to 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 1967–2015 and 1980–2015 samples.

The left panel of figure 5 shows time-series plots of the optimal number of portfolios in the sample for the size anomaly

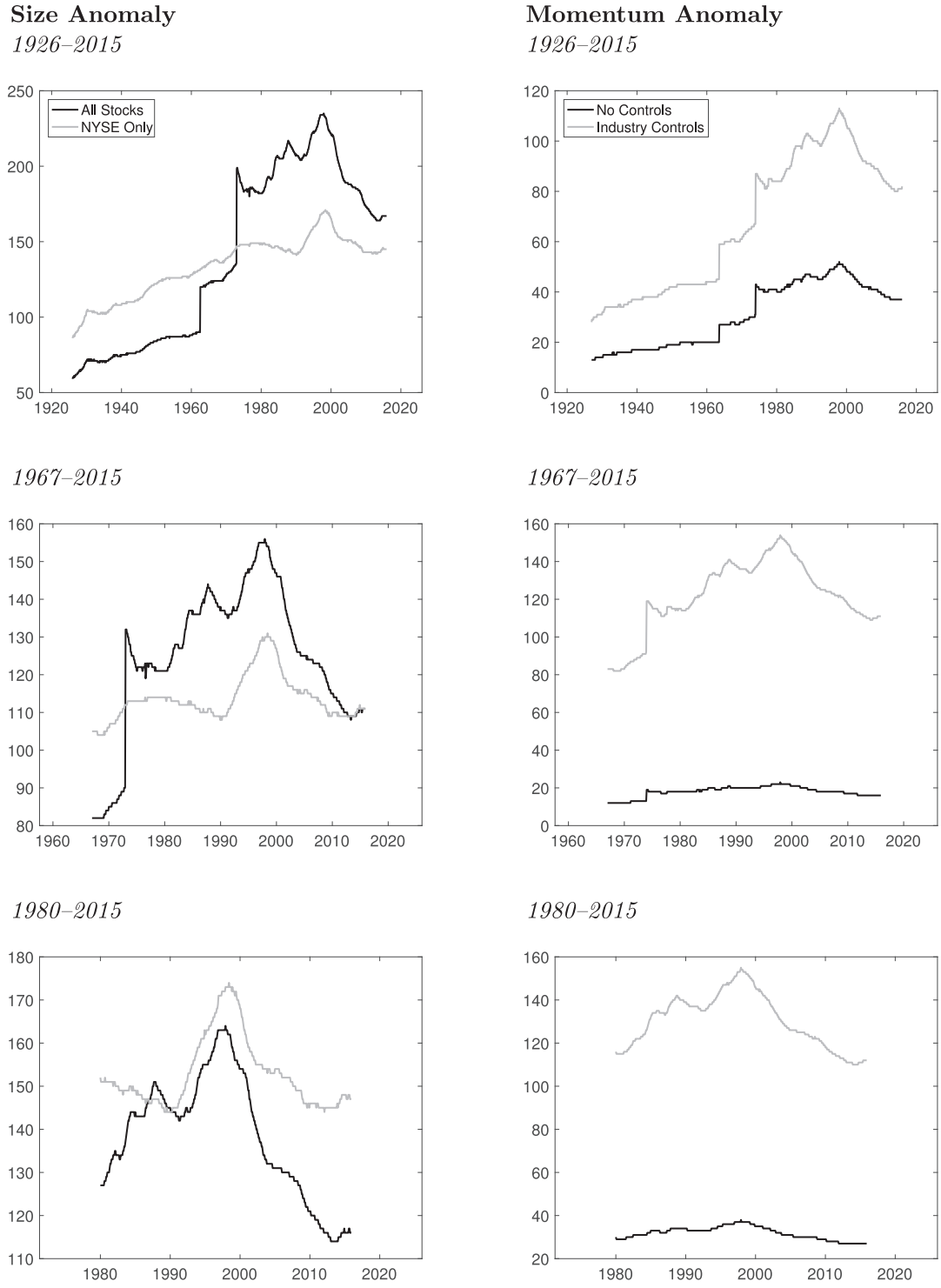
FIGURE 4.—SIZE ANOMALY: NYSE ONLY



This figure shows the estimated relationship between the cross section of equity returns and lagged market equity, equation (13), for NYSE firms. The left column displays $\hat{\mu}(\cdot)$, where J_t is based on equation (11), $z_H = \Phi^{-1}(.975)$, $z_L = \Phi^{-1}(.025)$. The right column displays the estimated relation using the standard portfolio sorting implementation with $J = 10$. All returns are in monthly changes, and all portfolios are value weighted based on lagged market equity.

chosen based on equation (11), using data for our three subperiods and based on $z_H = \Phi^{-1}(.975)$, $z_L = \Phi^{-1}(.025)$. 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 charts also show 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

FIGURE 5.—OPTIMAL PORTFOLIOS COUNTS



This figure shows the optimal number of portfolios for the estimated relationship between the cross section of equity returns and lagged market equity—equation (13), left column—and 12-2 momentum—equation (14), right column. J_t has been chosen based on equation (11), $z_H = \Phi^{-1}(.975)$, $z_L = \Phi^{-1}(.025)$.

not always the case that the 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.

C. *Momentum Anomaly*

We next consider the momentum anomaly—where firms that have had better relative returns in the nearby past also have higher relative returns on average. As discussed in section III, 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,

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

Here, MOM_{it}, represents the 12-2 momentum measure of firm i at time t transformed in the following way: at each cross section $t = 1, \ldots, T$, 12-2 momentum is demeaned and normalized by the inverse of the cross-sectional standard deviation (i.e., a z-score is applied). Unlike in the case of the size anomaly, no transformation is necessary to satisfy assumption 1c. 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.

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 appears concave with past "winners" (those with high 12-2 momentum values) earning about 2% in returns on average. The strategy of investing in past "losers" (those with low 12-2 momentum values) 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 subsample. This suggests that the short side of buying the spread portfolio appears to have become more profitable in recent years. This conclusion is robust to excluding the financial crisis and its aftermath. 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 1 shows the corresponding point estimates and test statistics for the momentum anomaly. The results strongly confirm that momentum is a robust 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 that have the largest magnitude in the 1980–2015 sample when evaluated at $(\Phi^{-1}(.975), \Phi^{-1}(.025))$ 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 the discussion in remarks 1 and 6). 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 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,

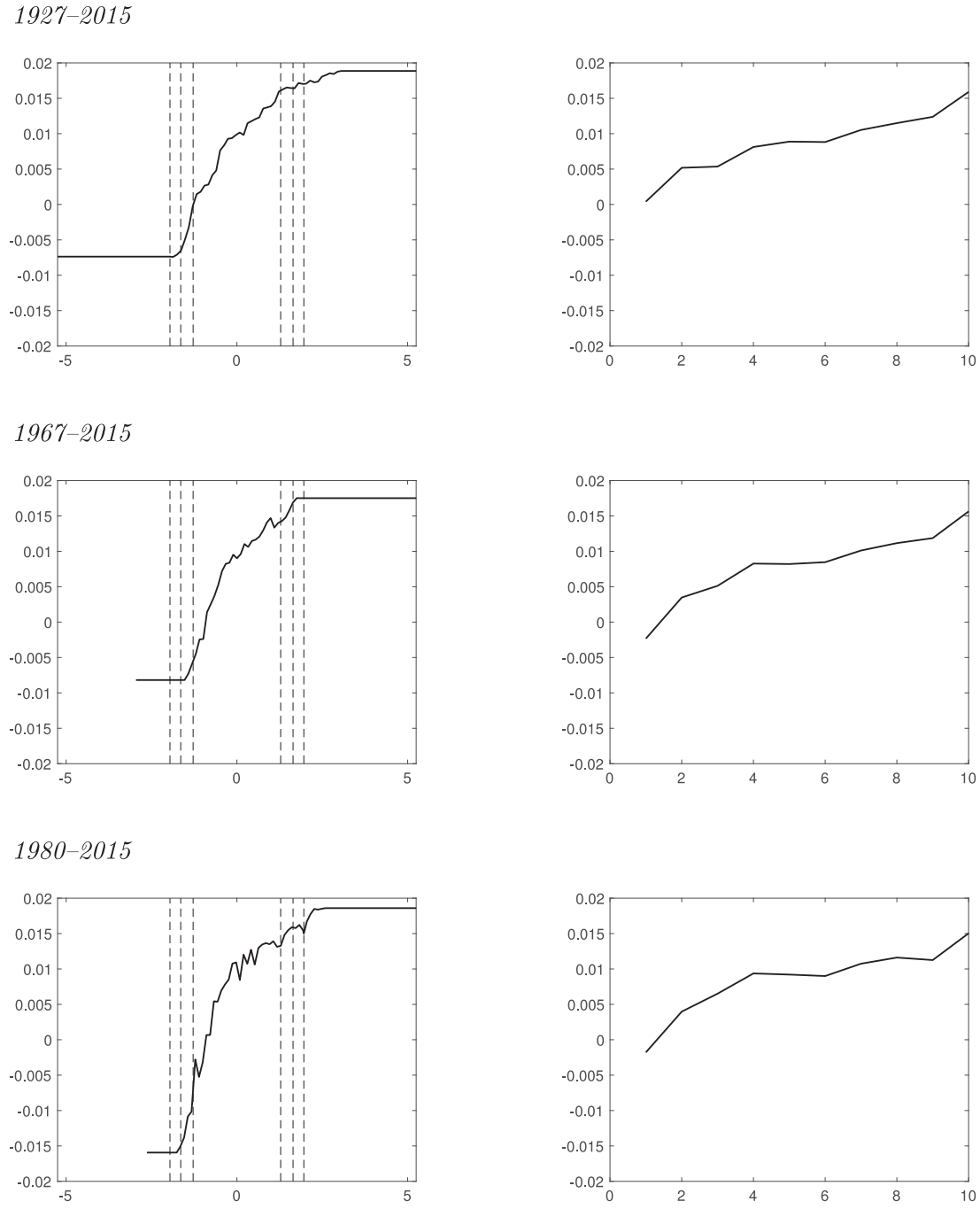
$$R_{it} = \mu(\text{MOM}_{it}) + \beta_1 \times \text{IMOM}_{it} + \beta_2 \times \text{IMOM}_{it}^2 + \beta_3 \times \text{IMOM}_{it}^3 + \varepsilon_{it}, \tag{15}$$

where 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 nonlinearities in this control.

Figure 7 shows the estimates of the relationship between returns and momentum controlling for industry momentum as in equation (15) (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—that is, based on equation (14)—for the same choice of the number of portfolios at each time t. 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 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 & Grinblatt, 1999; Grundy & Martin, 2001).

The bottom panel of table 1 provides point estimates and associated test statistics based on equation (15) in the rows labeled "w/controls." First, it is clear that the inclusion of industry momentum does have a noticeable effect on inference. In general, the magnitudes of the t-statistics for the high evaluation point, low evaluation point, and difference are shrunk toward 0. 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

FIGURE 6.—MOMENTUM ANOMALY

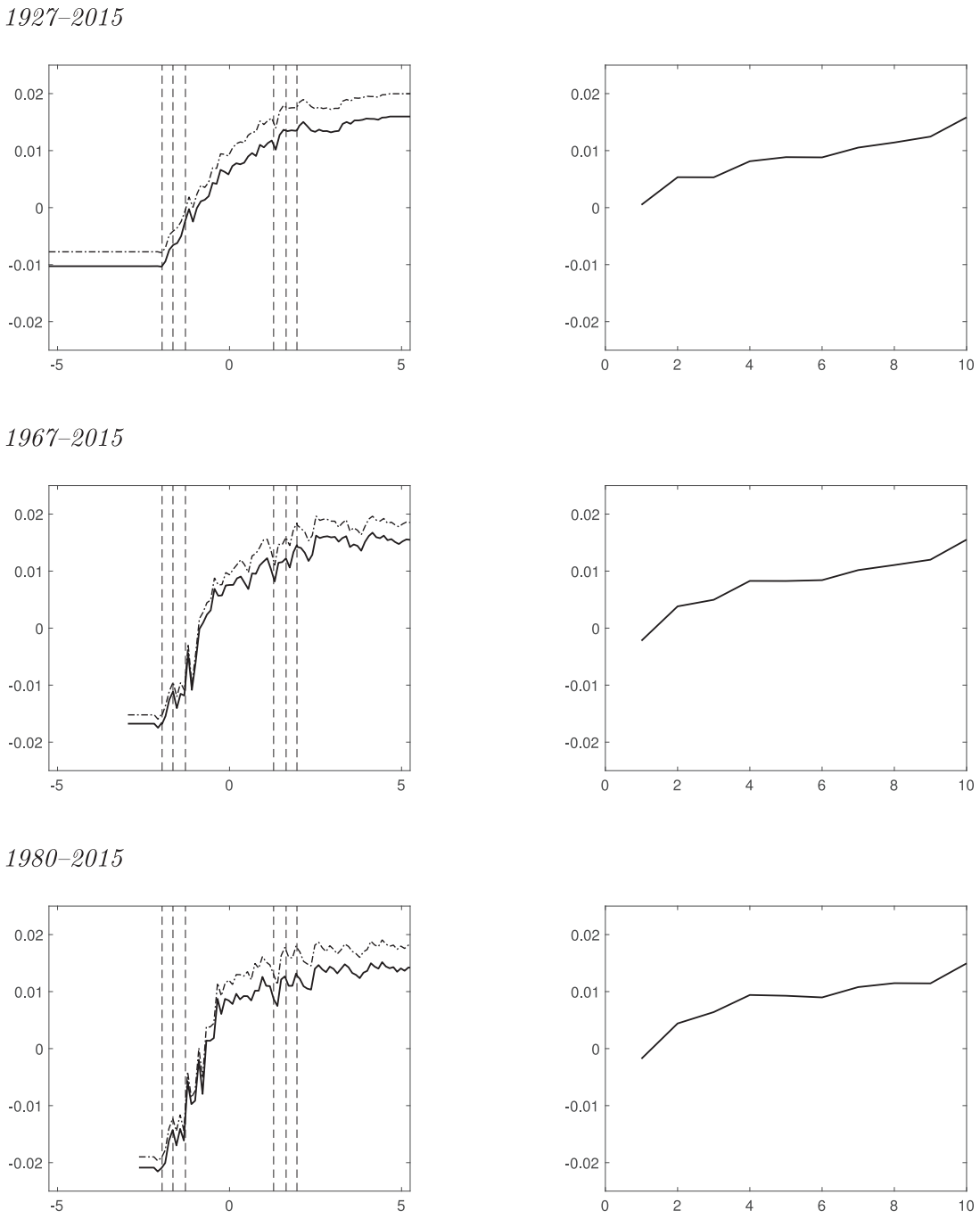


This figure shows the estimated relation between the cross section of equity returns and 12-2 momentum, equation (14). The left column displays $\hat{\mu}(\cdot)$, where J_t has been chosen based on equation (11), $z_H = \Phi^{-1}(.975)$, $z_L = \Phi^{-1}(.025)$. The right column displays the estimated relation using the standard portfolio-sorting implementation with $J = 10$. All returns are in monthly changes, and all portfolios are value weighted based on lagged market equity.

and low evaluation points results in a statistically significant return differential at the 5% level. This exercise illustrates the usefulness of our unified framework as it allows for the addition of control variables in a simple, straightforward manner.

Finally, the right panel of figure 5 shows time-series plots of the optimal number of portfolios in the sample for the momentum anomaly. 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. The charts also 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

FIGURE 7.—MOMENTUM ANOMALY: CONTROLLING FOR INDUSTRY MOMENTUM



This figure shows the estimated relation between the cross section of equity returns and 12-2 momentum. The left column displays $\hat{\mu}(\cdot)$ controlling for IMOM_{it}, IMOM_{it}^2, and IMOM_{it}^3 (solid line) as in equation (15) where J_t has been chosen based on equation (11), $z_H = \Phi^{-1}(.975)$, $z_L = \Phi^{-1}(.025)$. The dash-dotted line shows $\hat{\mu}(z)$ without control variables as in equation (14) for the same J_t. The right column displays the estimated relation using the standard portfolio-sorting implementation with $J = 10$ and no controls. All returns are in monthly changes, and all portfolios are value weighted based on lagged market equity.

explained only by 12-2 momentum. This lower variance results in a higher choice of J (see equation [10]). 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.

VII. 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 versus 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 provided practical guidance on how to implement this choice. In addition, we showed 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 multicharacteristic 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 bridged 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 found that the estimated relationship between returns and size appears to be monotonically decreasing and convex, 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 found that the estimated relationship between returns and past returns appears to be monotonically increasing and concave, with a significant and robust return differential. We found that the "short" side of the momentum spread trade has become more profitable in later subperiods. In both empirical applications, 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.

REFERENCES

Adrian, T., R. K. Crump, and E. Moench, "Regression-Based Estimation of Dynamic Asset Pricing Models," *Journal of Financial Economics* 118 (2015), 211–244.

Andrews, D. W. K., "Cross-section Regression with Common Shocks," *Econometrica* 73 (2005), 1551–1585.

Ang, A., J. Liu, and K. Schwarz, "Using Stocks or Portfolios in Tests of Factor Models," *Journal of Financial and Quantitative Analysis* (forthcoming), doi:10.1017/S0022109000255.

Bali, T. G., R. F. Engle, and S. Murray, *Empirical Asset Pricing: The Cross Section of Stock Returns* (Hoboken, NJ: Wiley, 2016).

Banz, R. W., "The Relationship between Return and Market Value of Common Stocks," *Journal of Financial Economics* 9 (1981), 3–18.

Basu, S., "Investment Performance of Common Stocks in Relation to Their Price-Earnings Ratios: A Test of the Efficient Market Hypothesis," *Journal of Finance* 32 (1977), 663–682.

Berk, J. B., "Sorting Out Sorts," *Journal of Finance* 55 (2000), 407–427.

Calonico, S., M. D. Cattaneo, and M. H. Farrell, "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference," *Journal of the American Statistical Association* 113 (2018), 767–779.

——— "Coverage Error Optimal Confidence Intervals for Local Polynomial Regression" (2019), arXiv:1808.01398.

Calonico, S., M. D. Cattaneo, and R. Titiunik, "Optimal Data-Driven Regression Discontinuity Plots," *Journal of the American Statistical Association* 110 (2015), 1753–1769.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng, "On Binscatter" (2019), arXiv:1902.09608.

Cattaneo, M. D., R. K. Crump, and M. Jansson, "Robust Data-Driven Inference for Density-Weighted Average Derivatives," *Journal of the American Statistical Association* 105 (2010), 1070–1083.

Cattaneo, M. D., and M. H. Farrell, "Optimal Convergence Rates, Bahadur Representation, and Asymptotic Normality of Partitioning Estimators," *Journal of Econometrics* 174 (2013), 127–143.

Cattaneo, M. D., M. H. Farrell, and Y. Feng, "Large Sample Properties of Partitioning-Based Series Estimators," *Annals of Statistics* (forthcoming).

Cochrane, J. H., "Discount Rates," *Journal of Finance* 66 (2011), 1047–1108.

Connor, G., M. Hagmann, and O. Linton, "Efficient Semiparametric Estimation of the Fama-French Model and Extensions," *Econometrica* 80 (2012), 713–754.

Conrad, J. S., M. J. Cooper, and G. Kaul, "Value versus Glamour," *Journal of Finance* 58 (2003), 1969–1996.

De Bondt, W. F. M., and R. Thaler, "Does the Stock Market Overreact?" *Journal of Finance* 40 (1985), 793–805.

Fama, E. F., and K. R. French, "The Cross-Section of Expected Stock Returns," *Journal of Finance* 47 (1992), 427–465.

——— "Common Risk Factors in the Returns on Stocks and Bonds," *Journal of Financial Economics* 33 (1993), 3–56.

——— "Dissecting Anomalies," *Journal of Finance* 63 (2008), 1653–1678.

Fama, E. F., and J. D. MacBeth, "Risk, Return, and Equilibrium: Empirical Tests," *Journal of Political Economy* 81 (1973), 607–636.

Gospodinov, N., R. Kan, and C. Robotti, "Spurious Inference in Reduced-Rank Asset-Pricing Models," *Econometrica* 85 (2017), 1613–1628.

Goyal, A., "Empirical Cross-Sectional Asset Pricing: A Survey," *Financial Markets and Portfolio Management* 26 (2012), 3–38.

Grundy, B. D., and J. S. Martin, "Understanding the Nature of the Risks and the Source of Rewards to Momentum Investing," *Review of Financial Studies* 14 (2001), 29–78.

Hong, H., T. Lim, and J. C. Stein, "Bad News Travels Slowly: Size, Analyst Coverage, and the Profitability of Momentum Strategies," *Journal of Finance* 55 (2000), 265–295.

Ibragimov, R., and U. K. Müller, "t-Statistic Based Correlation and Heterogeneity Robust Inference," *Journal of Business and Economic Statistics* 28 (2010), 453–468.

——— "Inference with Few Heterogenous Clusters," this REVIEW 98 (2016), 83–96.

Jegadeesh, N., "Evidence of Predictable Behavior of Security Returns," *Journal of Finance* 45 (1990), 881–898.

Jegadeesh, N., and S. Titman, "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency," *Journal of Finance* 48 (1993), 65–92.

——— "Profitability of Momentum Strategies: An Evaluation of Alternative Explanations," *Journal of Finance* 56 (2001), 699–720.

Kleibergen, F., and Z. Zhan, "Unexplained Factors and Their Effects on Second Pass *R*-Squared's," *Journal of Econometrics* 189 (2015), 101–116.

Lehmann, B. N., "Fads, Martingales, and Market Efficiency," *Quarterly Journal of Economics* 105 (1990), 1–28.

Lewellen, J., S. Nagel, and J. Shanken, "A Skeptical Appraisal of Asset Pricing Tests," *Journal of Financial Economics* 96 (2010), 175–194.

Lo, A. W., and A. C. MacKinlay, "Data-Snooping Biases in Tests of Financial Asset Pricing Models," *Review of Financial Studies* 3 (1990), 431–467.

Moskowitz, T. J., and M. Grinblatt, "Do Industries Explain Momentum?" *Journal of Finance* 54 (1999), 1249–1290.

Nagel, S., "Short Sales, Institutional Investors and the Cross-Section of Stock Returns," *Journal of Financial Economics* 78 (2005), 277–309.

Nagel, S., and K. J. Singleton, "Estimation and Evaluation of Conditional Asset Pricing Models," *Journal of Finance* 66 (2011), 873–909.

Patton, A. J., and A. Timmermann, "Monotonicity in Asset Returns: New Tests with Applications to the Term Structure, the CAPM, and Portfolio Sorts," *Journal of Financial Economics* 98 (2010), 605–625.

Reinganum, M. R., "Misspecification of Asset Pricing: Empirical Anomalies Based on Earnings' Yields and Market Values," *Journal of Financial Economics* 9 (1981), 19–46.

Romano, J. P., and M. Wolf, "Testing for Monotonicity in Expected Asset Returns," *Journal of Empirical Finance* 23 (2013), 93–116.

Shanken, J., and G. Zhou, "Estimating and Testing Beta Pricing Models: Alternative Methods and Their Performance in Simulations," *Journal of Financial Economics* 84 (2007), 40–86.

Stattman, D., "Book Values and Stock Returns," *Chicago MBA: A Journal of Selected Papers* 4 (1980), 25–45.