The Review of Financial Studies

Risk Price Variation: The Missing Half of Empirical Asset Pricing

Andrew J. Patton
Duke University, USA

Brian M. Weller
Amazon.com, USA

Equal compensation across assets for the same risk exposures is a bedrock of asset pricing theory and empirics. Yet real-world frictions can violate this equality and create apparently high Sharpe ratio opportunities. We develop new methods for asset pricing with cross-sectional heterogeneity in compensation for risk. We extend k-means clustering to group assets by risk prices and introduce a formal test for whether differences in risk premiums across market segments are too large to occur by chance. We find significant evidence of cross-sectional variation in risk prices for almost all combinations of test assets, factor models, and time periods considered. (*JEL* G12, G14, C21, C55)

Received May 11, 2020; editorial decision November 30, 2021 by Editor Stijn Van Nieuwerburgh.

Academic models of asset prices often assume away complications like trading frictions and segmented markets, but in practice these simplifying assumptions rarely hold. As the global financial crisis made painfully apparent, arbitrage frictions matter and, even in ordinary times, institutional and informational frictions impede investor participation across markets and make for good deals in the markets' dusty corners (Fama and French 2010; Hou, Xue, and Zhang's 2020). Frictions such as these can generate heterogeneous prices of risk across market segments and the appearance of very profitable trading strategies. Consequently, cross-sectional differences in average returns may derive from differences in both risk exposures and *risk premiums.*

For helpful comments we thank the editor (Stijn Van Nieuwerburgh) and two reviewers; Tim Bollerslev, Svetlana Bryzgalova, Anna Cieslak, Stefano Giglio, Cam Harvey, Jia Li, Zhengzi (Sophia) Li, Dong Lou, Zhongjin (Gene) Lu, Shrihari Santosh, Michael Weber, and Dacheng Xiu; and seminar participants at the AQR Insight Award Conference, the ASU Sonoran Winter Finance Conference, Duke Fuqua, the INSEAD Finance Symposium, the International Symposium on Financial Engineering and Risk Management, the RAPS Conference at Baha Mar, the SFS Cavalcade, the TCU Finance Conference, Tim Bollerslev's 60th Birthday Conference, and the University of Chicago Stevanovich Center. We thank Qian Zhu for research assistance, and Duke Research Computing and, especially, Tom Milledge for large-scale computing support. Send correspondence to Andrew Patton, andrew.patton@duke.edu.

The Review of Financial Studies 35 (2022) 5127–5184

https://doi.org/10.1093/rfs/hhac012 Advance Access publication February 22, 2022

In an idealized financial market, cross-sectional differences in risk prices are tamed by two forces. First, risk sharing among market participants typically makes the consumption of a "representative" household the key determinant of risk prices for *all* assets. Second, sophisticated arbitrageurs eliminate differences in compensation for risk that arise on account of short-lived demand pressures. The global financial crisis broke both intuitions. In response, recent models, such as Brunnermeier and Pedersen (2009) and He and Krishnamurthy (2012), have placed intermediaries at the heart of the pricing kernel—intermediaries' rather than households' constraints and marginal value of wealth determine equilibrium prices of risk. Likewise, Gârleanu and Pedersen (2011) and Gromb and Vayanos (2018), among others, develop theoretical cross-sectional asset pricing implications of arbitrageur borrowing frictions in exogenously segmented markets.[1] We find empirically that differences in risk prices are so pervasive that such limits to arbitrage must take a central role in asset pricing within and across markets, in crises and in normal times.

Existing empirical work on asset pricing with heterogeneous risk premiums has drawn on a priori knowledge to conjecture groups of similarly priced assets (e.g., Fama and French 1993; Foerster and Karolyi 1999; Griffin 2002, among others). However, only in certain cases do we know how to group assets ex ante, and moreover any conjectured market segments may be incorrect or less informative than other dimensions of variation in risk prices. Further, given the sensitivity of estimated risk prices to how assets are grouped together, a skeptical empiricist should impose the same data-snooping hurdle on market segments as on factors in expected returns.

We propose a new approach to asset pricing with multiple prices of risk. Our methodology extends existing clustering algorithms to "let the data speak" in identifying groups of assets with similar risk prices. In so doing we address the empirical challenge of identifying segmented sets of assets in a wide range of economic settings. We then build on recent work on panel data models, for example, Lin and Ng (2012), Bonhomme and Manresa (2015), and Su, Shi, and Phillips (2016), to propose methods for formally testing whether multiple risk prices are needed in a given data set, or whether frictionless frameworks suffice to explain the cross-section of expected returns.

The core of our estimation technique consists of estimating group assignments and cross-sectional slopes via the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977). This algorithm cycles between (1) estimating cross-sectional slopes given conjectured group assignments and (2) reallocating portfolios to groups given estimated cross-sectional slopes (similar to Lloyd's algorithm in k-means clustering (MacQueen 1967). The first step generalizes standard cross-sectional techniques, such as the

[1] Such markets are necessarily "incomplete," that is, certain risks cannot be traded, and there does not exist a unique stochastic discount factor.

Fama and MacBeth (1973) regression, to settings with multiple risk prices, whereby we estimate cross-sectional slopes period-by-period and group-by-group. The second step then reallocates each portfolio to the group whose cross-sectional slopes best describe its return dynamics. Iterating these steps identifies progressively more important dimensions of heterogeneity in risk prices, and at convergence, the algorithm delivers an optimal set of group assignments and cross-sectional slopes.[2]

The structure of our estimation problem prevents the use of popular, off-the-shelf clustering technologies like k-means. In typical clustering applications, the algorithm groups assets using characteristics that are not a function of other assets, or of the group assignment, and hence can be taken as fixed. In our setting, the grouping attributes are cross-sectional slopes, that is, our estimates of risk premiums for each factor and date. It is the fact that these grouping attributes are not fixed for each asset and instead depend on the risk and return characteristics of other cluster members that makes standard clustering approaches inappropriate in our setting. Our new approach can be interpreted as a generalization of k-means that accommodates this dependence.

To test whether a set of portfolios has segmented risk prices, we need to evaluate whether the incremental explanatory power of multiple clusters warrants the additional parameters we estimate in a multiple-cluster model. A naive test of whether risk prices differ across clusters fails because cluster assignments are *estimated* in our approach, and standard tests for the significance of risk-price differences across clusters severely over-reject in finite samples.[3] We propose a simple alternative to overcome this problem based on subsamples of the data: we estimate the cluster assignments using one subsample, and we test for differences of estimated risk prices in a second subsample using the group assignments from the first subsample. We then construct modified F-statistics for equal risk prices across groups.[4]

The economics of our test differ in important ways from standard tests of asset pricing models, such as the GRS test of Gibbons, Ross, and Shanken (1989). Starting from a baseline factor model, the GRS test evaluates whether

---

[2] We pair this iterative approach with multistart and genetic algorithm global search methods to locate the global-best group assignments and corresponding cross-sectional slopes.

[3] By contrast, no such problems arise when cluster assignments are known ex ante, as in Foerster and Karolyi (1999), since in that case the problem of overfitting (stemming from the estimation of the group assignments) does not arise. In the leading example of testing a null of $G=1$ cluster (i.e, no heterogeneity in risk prices) against $G=2$ clusters with estimated assignments, we must consider the behavior of a model that allows for two clusters when the true number of clusters is one. Existing work on inference for k-means clustering and related methods that takes into account estimated assignments (e.g., Bonhomme and Manresa 2015) has assumed that clusters are "well separated," making such theory inapplicable to this hypothesis test.

[4] We find that standard asymptotic F-tests have poor size control in finite samples, so we instead obtain critical values using a permutation method in which we shuffle group assignments to obtain the distribution of the test statistic under the null (see, e.g., Lehmann and Romano, 2005). We show that this approach has good size control in realistic simulation designs.

adding test assets (or factors) improves Sharpe ratios enough to justify the additional flexibility granted in forming the in-sample mean-variance efficient portfolio. Our test evaluates whether adding *clusters* improves model fit enough to justify the additional flexibility allowed by group-specific slopes to explain the cross-section of returns (we use subsamples with fixed groups to shut down flexibility in group assignments). Rejections in each test signifies that a candidate factor model is incomplete—in not spanning priced risks in the GRS test and in not capturing cross-sectional variation in risk prices in our test.

We apply these tools to analyze risk price heterogeneity in a variety of economic settings, including domestic stock portfolios, international stocks portfolios, and multi-asset-class portfolios.[5] To these portfolio sets we apply leading factor models including the capital asset pricing model (CAPM), the Fama and French (1992) three-factor model; the Carhart (1997) four-factor model, the Fama and French (2015) five-factor model, the He, Kelly, and Manela (2017) intermediary-capital factor model, and the Hou, Xue, and Zhang (2015) q-factor model. Finally we split our data samples into subintervals to evaluate segmentation over time within each market-model pair.

Our analysis delivers two new empirical facts on cross-sectional heterogeneity in risk prices. First, we find segmented risk prices in almost every setting examined. We reject unified risk pricing for all but one of 173 combinations of test assets, benchmark factor models, and time periods, with a modal p-value of zero (to three decimal places). By contrast, we find no evidence of segmentation in a set of placebo portfolios, which are homogeneous by construction, beyond what is expected by chance. Second, we find variation in risk prices contributes significantly to explaining the cross-section of expected returns. Among U.S. stock portfolios, where segmentation is least pronounced, adding multiple clusters increases cross-sectional dispersion in expected returns from 3% to 161%, with 25th and 75th percentiles increases of 7% and 61%, respectively. This increase in cross-sectional explanatory power is comparable to replacing the Fama and French three-factor model with their more recent five-factor model. Ex post Sharpe ratios of global factor models also increase markedly when including factor variants local to each market segment; for domestic equity portfolios, the 25th and 75th percentiles of changes in maximal annual Sharpe ratios are 0.32 and 0.98. Potential gains from acknowledging multiple market segments are comparably large in our analyses of international equity portfolios and portfolios of assets across multiple asset classes.[6] We conclude that risk-price heterogeneity is a pervasive and economically important feature of real-world financial markets.

---

[5] We consider a wide range of portfolio sets in the spirit of Lewellen, Nagel, and Shanken (2010), who emphasize the importance of considering a diverse set of test assets when evaluating asset pricing models.

[6] These improvements in Sharpe ratios can also be interpreted as a measure of the size of the investment frictions among market segments, as arbitrageurs could equalize Sharpe ratios absent such frictions.

While we successfully explain substantial additional variation in the cross-section of average returns, our paper fundamentally differs from empirical asset pricing papers that introduce new candidate factors to accomplish the same objective. There, an empiricist selects factors from a large and unobserved model space, and the many degrees of freedom in data mining the factor (and its empirical implementation) carry no penalty.[7] Our approach penalizes such fishing expeditions: once we have selected a conventional factor model, the freedom to choose group assignments and fit additional cross-sectional slopes is explicitly accounted for in our statistical tests of unified versus segmented pricing. Moreover, our market segments *use the same risk factors* as the posited factor model, thereby imposing significant economic discipline relative to multifactor models in which new global factors are mined from the data.

Like most unsupervised learning techniques, including principal components analysis and clustering methods, our approach does not assign labels to the estimated clusters. Nevertheless, labels can often be intuited from cluster assignments, as we find in the settings we consider. Specifically, in our analysis of domestic equity portfolios we find they cluster (mostly) into small-cap and large-cap stock portfolios, thus contributing to the longstanding debate on whether factors earn differential compensation across stocks of different sizes—for example, Hong, Lim, and Stein (2000), Grinblatt and Moskowitz (2004), and Israel and Moskowitz (2013)—by identifying market capitalization as the most important determinant of differences in factor premiums among common U.S. stock portfolios. Our analysis of international stock portfolios finds them to segregate perfectly by geographic region, confirming international boundaries as a well-known source of market segmentation (see, e.g., Griffin 2002; Hou, Karolyi, and Kho 2011; Fama and French 2012). Finally, our analysis of multiasset class analysis of stock, commodity, bond, options, and currency portfolios finds that they cleave into (at least) five clusters, each corresponding to one or two asset classes with similar risk prices. Here, we contrast with He, Kelly, and Manela (2017) by rejecting the null of equal pricing across asset classes in a model with market and intermediary capital risk factors.

Another challenge we face is distinguishing between omitted clusters and omitted factors. In much of our analysis, we take the factor models as given and complete, and we conduct inference on whether the data support one or multiple prices of risk. However, as we show in Section 5, omitted factors can manifest as clusters, and vice versa. Omitting factors creates an omitted variable bias, and groups of assets with greater exposures on omitted factors may appear to have different risk prices from assets with smaller exposures. Similarly, omitting clusters generates a proliferation of apparent factors that capture the interaction of included factors and group membership dummies; the absence of momentum compensation in Japan can masquerade as a new,

[7] Harvey (2017) caricatures this search among candidate factors in the motivating example of the 2017 AFA Presidential Address.

Japan-specific factor on which only Japanese stocks load, for example. We use the model comparison test of Rivers and Vuong (2002) to distinguish between omitted factors and heterogeneity in risk prices, and we find that accounting for potential omitted factors leaves our main result intact; we find strong evidence of risk price heterogeneity in almost every setting considered. New dimensions of return variation call for new *clusters* rather than new *factors*.

In this vein, our work suggests a potential reason for the continual discovery of new factors in expected returns. The "factor zoo" described by Cochrane (2011) is too crowded to be plausible, standing at several hundred inhabitants according to Harvey, Liu, and Zhu (2016) and Hou, Xue, and Zhang's (2020) recent counts. Empirically, we find clusters even in the presence of potential omitted factors, but the reverse may not be true. In general, any factor betas or characteristics that are cross-sectionally correlated with omitted group membership dummies may appear to be new priced factors. Moreover, given the discreteness of clusters in segmented markets, continuous characteristics are unlikely to span group indicators, and several new "factors" may spuriously appear from just a few dimensions of latent market segmentation.

This paper is related to a large literature on market segmentation, the two main drivers of which are geographic boundaries (segmenting international markets) or the combination of transaction costs and heterogeneous market clientele (segmenting even the domestic stock market). In both cases, the a key implication of market integration is the equality of risk prices across markets. In the international context, Fama and French (1998) argue for a parsimonious, global market and value-factor model to explain value in international markets, while Griffin (2002) and Hou, Karolyi, and Kho (2011) favor a segmented-markets worldview. Similarly, the momentum factor performs well in most countries with the notable exception of Japan (Rouwenhorst 1998; Griffin, Ji, and Martin 2003; Fama and French 2012), contradicting a unified pricing of this factor.[8] Other papers that consider heterogeneous prices of risk include Errunza and Losq (1985), who propose a model of "mildly" segmented markets in which asymmetric barriers to trade break the equality of risk prices across countries, and Bekaert and Harvey (1995), who estimate a regime-switching model to consider time variation in the degree of cross-market integration.

While no study, to the best of our knowledge, systematically analyzes dimensions of segmentation within U.S. stocks., existing work on segmentation of U.S. stocks proposes several different dimensions of segmentation among U.S.-traded securities. In a seminal paper, Merton (1987) considers informational frictions in which only some traders are aware of an investment opportunity, effectively segmenting markets. Kadlec and McConnell (1994) and Foerster and Karolyi (1999) confirm Merton's "investor recognition hypothesis" in illiquid U.S. stocks and in international stocks listed on

[8] Related, Asness, Moskowitz, and Pedersen (2013) also find a particularly large Japanese value premium.

U.S. exchanges, respectively. Investor awareness, institutional ownership, and trading frictions vary with market capitalization, and several studies consider whether stocks earn the same risk premiums in large- and small-cap segments. Hong, Lim, and Stein (2000) and Grinblatt and Moskowitz (2004) estimate a negative interaction between momentum returns and market capitalization, while Israel and Moskowitz (2013) find this conclusion to be an artifact of the particular sample period. Fama and French (1993) consider common factors in stocks and bonds, and they find mixed support for integration of these markets: stock returns load on term-structure factors when other stock factors are included, but bond returns (mostly) do not load on stock factors when other bond factors are included. By contrast, He, Kelly, and Manela (2017) find that an intermediary-based asset pricing model explains the returns on stocks, bonds, options, currencies, commodities, and other asset classes, and with similar prices of risk. We note that with rare exceptions, conclusions on the degree of segmentation within and across markets depend critically on the choice of asset pricing model.

Motivated by the global financial crisis, a number of recent studies consider equilibrium asset prices in which frictions play a central role. Gârleanu and Pedersen (2011) show that margin constraints combined with limited capital prevent potential arbitrageurs from eliminating violations of the Law of One Price, in the spirit of Shleifer and Vishny's (1997) limits to arbitrage. Gromb and Vayanos (2018) extend Gromb and Vayanos (2002) to consider arbitrage dynamics across segmented markets when asset payoffs are not identical, and they obtain a similar result of arbitrageurs investing to maximize alpha per unit of collateral. Much like our paper, the authors focus on cross-sectional implications of constrained arbitrage, and like us, they assume that market segmentation occurs for exogenous reasons, including "regulation, agency problems, or a lack of specialized knowledge." New models focusing on inequality and intermediation provide additional rationales for differences in risk prices. Greenwald, Lettau, and Ludvigson (2016) and Lettau, Ludvigson, and Ma (2019) build on limited participation models (e.g., Mankiw and Zeldes 1991; Vissing-Jørgensen 2002) in which workers and shareholders have different ratios of capital income to labor income and do not share risks. In these setups, the representative agent's consumption does not price assets; rather, capital share-adjusted income enters the stochastic discount factor. He and Krishnamurthy (2012, 2013), Adrian, Erkko, and Muir (2014), and He, Kelly, and Manela (2017) replace the representative household's stochastic discount factor with that of marginal financial intermediaries to derive an intermediary-capital factor. Both classes of models generate imperfect correlation between the marginal value of wealth among households. Recent work on demand system asset pricing (see Koijen and Yogo 2019; Koijen, Richmond, and Yogo 2020) finds that preferences for asset characteristics such as size, value, and beta, and their demand elasticities for these characteristics, vary greatly across different classes of investors (households, mutual funds, investment advisors).

Such differences in preferences can manifest as heterogeneous prices of risk in the presence of limits to arbitrage (Pontiff 1996, 2006; Shleifer and Vishny 1997; Gromb and Vayanos 2018, among others) or sizeable trading frictions (Gabaix and Koijen 2020).

1. Detecting Heterogeneity in Risk Prices

1.1 Estimation of a factor model with multiple clusters

1.1.1 Economic setting. The economy consists of N assets and K asset pricing factors over T dates. Let f_t be a $K \times 1$ vector of asset pricing factors and r_t be an $N \times 1$ vector of asset excess returns at time t. The factors f_t are known and observable (we relax the former assumption in Section 5). As usual, we obtain factor exposures, β_i, by regressing the returns for asset i on the factors. The standard approach for estimating the risk premium (λ_t) associated with factor exposure is to then run cross-sectional regressions of returns on factor exposures:

$$r_{it} = \alpha_t + \lambda_t' \beta_i + \epsilon_{it}. \tag{1}$$

If factors are not traded, or there are other market frictions, then the realized risk premiums obtained from (1) may differ from the realized returns on the factor, that is:

$$\lambda_t \equiv \frac{cov(r_{it}, \beta_i)}{V(\beta_i)} \neq f_t, \tag{2}$$

If the factors are traded and markets are integrated risk prices will equal the factor realizations, so $\lambda_t = f_t$. Moreover, in such a case this equality holds not only for the full set of N assets, but also for all subsets of assets.

We consider the case with *clusters* of assets having the same realized risk premiums, but these premiums can differ across clusters. For example, assume that for one subset of assets (denoted g, say) it is possible to capture all, or nearly all, of the risk premium associated with f_t, leading to $\lambda_t^g \approx f_t$, while for another subset (g') this may be difficult, leading to $\lambda_t^{g'} \neq f_t$, implying that $\lambda_t^g \neq \lambda_t^{g'}$. If market frictions are absent, or they are present and they affect all assets in the *same* way, for example, through inducing a level shift in average returns, then we would find $\lambda_t^g = \lambda_t^{g'}$ for all pairs (g, g'), and the multicluster model reduces to the usual single-cluster model.[9]

Motivated by the work on segmented markets discussed in the introduction, the goal of this paper is to determine whether the model in (1) can be significantly improved by allowing for clusters of assets that exhibit

[9] Of course, the multicluster model will have lower finite-sample MSE than the single-cluster model, even when only a single cluster is needed, because of overfitting. Our procedure for testing for the presence of multiple clusters, decribed in Section 1.1.1.2, explicitly addresses this problem.

heterogeneity in the prices of risk. That is, whether there are clusters (g, g') such that $\lambda_t^g \neq \lambda_t^{g'}$. Importantly, we allow the data to estimate the optimal cluster assignments, using an estimation procedure described below, rather than forming clusters based on exogeneous characteristics.

1.1.2 Mapping to models of market segmentation. Several classes of economic models give rise to cross-sectional variation in risk prices; stated generally, the necessary ingredients are a source of market segmentation or incompleteness and heterogeneous agents across market segments. In this section, we map two specific models to the economic setting described above.

First, from the international asset pricing literature, Errunza and Losq (1985) consider markets separated by "mild" segmentation in which all investors may trade in "eligible" securities (one segment), and a set of investors cannot trade in "ineligible" securities (another segment), for example, because of capital controls. The authors derive a two-factor model consisting of a global market factor and a hedge-portfolio factor (comprising the ineligible market portfolio return minus the best-approximating portfolio of only eligible securities). As discussed above, segments earn different compensation per unit of risk exposure: ineligible securities earn a larger premium on the global-market factor, and *only* ineligible securities earn a risk premium for exposure to the hedge-portfolio factor. Moreover, only the returns on ineligible securities covary with hedge-portfolio factor realizations, so the dynamics of realized factor premiums are useful for identifying market segments (a feature we exploit below).

More recently, Gromb and Vayanos (2018) introduce a model of financially constrained arbitrage. The authors consider two assets traded in different segmented markets. With the exception of a capital-limited arbitrageur, investors are restricted to trade in their segment. For our purposes, we can view these assets as factor-mimicking portfolios comprising securities in each segment. Even in the case of identical asset fundamentals, constrained arbitrageurs do not eliminate differences in factor-mimicking portfolio returns induced by differences in hedging demands among segments; consequently, factor-mimicking portfolio returns can differ from one another, and their unconditional risk premiums may differ, too. The authors also propose a model of risky arbitrage in which segmented assets have both common factor exposures and idiosyncratic risk. All else equal, assets with greater common factor exposures (betas) feature larger spreads in returns between segments, which in turn implies different slopes with respect to factor exposures between segments.

1.1.3 Generalized cross-sectional estimation. Our analysis proceeds by assuming that each asset i is a member of one of $G \geq 1$ groups, where G is fixed for now. We use $\gamma_i \in \{1, ..., G\}$ to denote the group membership of asset i. Deviations of the risk premiums for each group from the factors returns are

defined as $\phi_t^g = \lambda_t^g - f_t$, and we combine these group-specific deviations into a $K \times G$ matrix Φ_t across the G groups. In the spirit of Fama and MacBeth (1973) two-stage regressions, the empiricist first estimates time-series regressions for each asset i,

$$r_{it} = a_i + \beta_i f_t + \eta_{it}. \tag{3}$$

to obtain estimates of risk exposures and idiosyncratic volatility for each portfolio. We make two assumptions to interpret the parameters of this regression. First, we assume that the group-specific deviations Φ_t are orthogonal to f_t in the time series; in economic terms, the group-specific distortions are uncorrelated with the factors themselves. This assumption is innocuous in that if the group-specific components are not orthogonal to f_t, β_i recovers the sensitivity of returns to variation in the common stochastic component. Second, we assume that the time-series dimension of our panel is large relative to the cross-sectional dimension (formally, $N/T \to 0$ as $N, T \to \infty$), which implies that the measurement error in estimated βs is negligible with respect to the measurement error in estimated risk prices. This assumption is particularly benign in our empirical estimation given that we use static betas for well-diversified portfolios,[10] and have much larger time series than cross-sections.

After computing factor loadings from the time series, the empiricist estimates factor realizations for all factors and groups via cross-sectional regressions. The estimated slopes, $\lambda_t^{(g)}$, are combined into a $K \times G$ matrix Λ_t across the G clusters, and combined across time to form a $K \times G \times T$ array Λ. Estimating Λ is difficult because we do not know group assignments ex ante; the empiricist must estimate the N group assignments, γ, in addition to all of the groupwise factor realizations. We use expectation maximization to address this challenge.

To see why our setting requires a new solution method beyond traditional ordinary least squares (OLS) or maximum-likelihood estimation, consider the error term for asset i in cluster g, under (1) using the time-series betas from (3),

$$\hat{\epsilon}_{it} = r_{it} - \alpha_t^{(g)} - \beta_i \lambda_t^{(g)}, \text{ for } \gamma_i = g. \tag{4}$$

Up to a constant, the associated log likelihood is

$$\log L(\alpha, \Lambda, I) = -\frac{1}{2} \sum_g \sum_{\gamma_i = g} \sum_t \frac{1}{\sigma_i^2} \left(r_{it} - \alpha_t^{(g)} - \beta_i \lambda_t^{(g)} \right)^2, \tag{5}$$

where $\sigma_i^2 = V(\eta_{it})$. Up to constants, this is a Gaussian likelihood and estimation can be interpreted as quasi-maximum likelihood. Maximizing (5)

[10] We use static betas to avoid complicating our approach with estimation error in the generated regressors. An alternative would be to estimate conditional betas and carry the time-series estimation errors into the cross-sectional regressions. However this approach adds considerable computational expense because it requires replacing each cross-sectional OLS regression—of which there may be millions in a typical run—with nonlinear regressions that account for this measurement error. Such an approach is computationally infeasible even using modern cluster computing resources.

entails optimizing over $T \cdot G$ αs, $K \cdot T \cdot G$ λs, and N group assignments. Even holding fixed group assignments, attacking this maximization problem directly is difficult because of the number of parameters involved. Direct maximization of (5) becomes computationally impossible when we optimize over the latent group assignment parameters γ. The objective function in (5) lacks differentiability with respect to discrete group assignments, which complicates or invalidates standard solution techniques for smooth problems.[11] An exhaustive search of group assignments for a "small" problem with 75 portfolios and two groups entails searching over 10^{18} possible groupings just to solve the integer component of the mixed-integer programming problem.

Fortunately iterative conditional approaches, such as expectation maximization, excel in situations in which conditional maximization problems are straightforward but full-problem optimizations—often involving latent parameters—are difficult. In our application the EM algorithm consists of two steps to recover model parameters (α and Λ) and group assignments (γ). First, given the group assignments, we estimate the model parameters ("maximization"). Second, given the model parameters, we reestimate the group assignments ("expectation"). We then iterate between these steps until convergence.[12, 13] Both steps are straightforward maximization problems:

1. **Estimate α and Λ given γ.** The first-order conditions of (5) with respect to $\alpha_t^{(g)}$ and $\lambda_t^{(g)}$ are

$$0 = \sum_{\gamma_i = g} \frac{1}{\sigma_i^2} \begin{bmatrix} 1 \\ \beta_i' \end{bmatrix} \left(r_{it} - \alpha_t^{(g)} - \beta_i \lambda_t^{(g)} \right). \tag{6}$$

   Crucially, conditioning on γ delivers separability across time and across clusters. Equation (6) is simply the moment conditions of TG cross-sectional regressions with precision weights $1/\sigma_i^2$,

$$r_{it} = \alpha_t^{(g)} + \sum_k \beta_{ik} \lambda_{kt}^{(g)} + \epsilon_{it}, \ \forall i \text{ s.t. } \gamma_i = g, \text{ for } g = 1, \ldots, G \text{ and } t = 1, \ldots, T. \tag{7}$$

---

[11] Replacing discrete group assignments with continuous group assignments is an alternative, potentially more tractable, modeling choice. However, we are not aware of asset pricing models that deliver partial group memberships, nor is it clear how to incorporate partial memberships into cross-sectional regressions without applying *ad hoc* observation weights.

[12] Dempster, Laird, and Rubin (1977) and Wu (1983) prove that this iteration achieves a local solution to the full maximization problem under general conditions. In our setting, the proof of convergence in a finite number of steps is almost immediate. Each step weakly improves the likelihood function, and a (local) maximum exists because the number of possible group assignments is finite (albeit large). Hence the sequence of likelihoods converges to a maximum by the monotone convergence theorem. The algorithm does not "cycle" because each step must weakly improve upon previous steps. Putting these components together, the number of steps is bounded above by G^N. In our applications, the EM algorithm tends to converge far more rapidly, taking between 5 and 15 iterations for the typical starting value and economic setting.

[13] Appendix B details multistart and genetic algorithm techniques we use to obtain global optima. In practice we find these techniques to be important. For the three leading examples of Section 3.3.3, only in one case would a local optimizer be likely to arrive at the global best group assignments, and in another case, most locally optimal group assignments differ substantially from the global best.

In short, this step reduces to separate Fama-MacBeth regressions for each group g.

2. **Estimate γ given α and Λ.** Fixing the parameters from Step 1, we can focus on a single asset at a time. The maximization problem reduces to finding the group with the smallest sum of squared errors,

$$\hat{\gamma}_i = \arg\min_{g\in\{1,\ldots,G\}} \sum_t \left(r_{it} - \alpha_t^{(g)} - \beta_i \lambda_t^{(g)}\right)^2. \tag{8}$$

This group assignment step is essentially immediate for each security, as it requires only the comparison of a set of G easily computed mean-squared errors. Note that group memberships are determined by the entire $T \times K$ matrix $\lambda^{(g)}$ rather than by its time-series average, $\bar{\lambda}^{(g)}$. Factor realizations at every date are helpful for identifying assets in a common cluster, and groups may be meaningfully distinct in their cross-sectional dynamics regardless of whether their unconditional average risk prices are similar.

This procedure has attractive economic properties. The first step generalizes Fama-MacBeth estimation of risk premiums to multiple market segments. The cross-sectional slopes $\lambda_t^{(g)}$ are the same as those in standard, single-cluster Fama-MacBeth regression if risk premiums are the same across assets. More generally, they are the factor-mimicking portfolio returns obtained using the assets within each cluster, for example, the best approximation of the global momentum factor using North American stocks for one cluster and Japanese stocks for another.[14] This step accommodates error structures with more complex cross-sectional or time-series dependence with suitable updates of the moment conditions for α and Λ.

The second step assigns assets to the clusters that minimize their pricing errors. Doing so resembles selecting the cluster for each security based on similarity in risk prices, but it is more economically robust because it weights cross-group differences in λs by asset betas; factors that are unimportant for explaining variation in portfolio returns do not affect group membership. This feature guards against grouping assets based on "junk" factors that explain little variation in the panel of realized returns.

Clustering by average lambdas is an alternative to (8). Indeed, doing so is preferable if risk premiums are constant and factor realizations are uninformative about cluster membership, because noisy factor realizations then

[14] This feature assumes that the given factor model is complete; otherwise, omitted factors may contaminate estimated risk prices (see, e.g., Giglio and Xiu 2021 for discussion). To address this concern, in our empirical analysis we consider a factor model augmented by the first three principal components of the portfolio return residuals ("Carhart+3"), as well as one based on prinicpal components extracted directly from the returns ("PC5"), much as those authors extract factors from test assets to clean their factor set. We also explicitly analyze omitted factors as an explanation for our results in Section 5.

obfuscate the key risk price heterogeneity among market segments. Instead—motivated in part by Errunza and Losq (1985) and Gromb and Vayanos (2018)—our clustering method allows for time variation in risk prices and factor-induced comovements to help identify cluster membership. Both models of market segmentation suggest that both mean and covariance information should be useful. Notwithstanding that we use this information in clustering, one of our tests (described below) uses only information on average risk prices to evaluate the hypothesis that portfolios have common risk premiums.

From step 1, we see that idiosyncratic volatility serves as observation weights in cross-sectional regressions. Any well-behaved set of weights delivers consistent estimates of α and Λ as $N \to \infty$, so potential errors in σ_i are immaterial in large samples for this step. Further, idiosyncratic errors do not enter the second step. Hence the choice of σ_i has no effect asymptotically for parameter consistency or group assignment. However, the way we estimate σ_i matters in practice in two ways. First, in finite samples, σ_i estimated from the time-series residuals maintains efficiency under the null of a single group, for which this choice of weights is optimal. In the worst case for finite N, the estimates in (6) are inefficient under the alternative model of multiple groups, and this inefficiency increases noise in λ estimates, decreases our ability to achieve dispersion among groups, and decreases the probability of rejecting the null. Second, observation weights affect the estimated maximized likelihood and the information-criterion selected number of groups. For this reason, our tests do not rely on a specific choice of G (we consider a range of values), and we use information criteria only when examining the detailed results from a specific multicluster model.

1.2 Testing for multiple clusters

The question at the heart of our paper is whether there exist latent market segments with different risk prices. In this section we develop a formal methodology to test the null of equal prices of risk in the cross-section (i.e., a single cluster) against the alternative of varying risk prices (i.e., multiple clusters).

We use subsamples of the data to implement our tests for multiple clusters. We partition our data into $\mathcal{R}$ and $\mathcal{P}$ disjoint subsamples, for example, the first and second halves of the sample, or odd- and even-dated observations. Let R and P denote the number of observations in each of these samples. For a fixed number of groups G, we estimate our cluster model on the $\mathcal{R}$ sample. This estimation yields parameters for each group $\hat{\alpha}_R^{(g)}$ and $\hat{\Lambda}_R^{(g)}$ as well as group assignments $\hat{\gamma}_R$. In estimating these parameters we make standard large-panel assumptions on each group: to estimate cross-sectional slopes consistently we need $R, N \to \infty$, and we also require the number of assets in each cluster, denoted N_g, $g = 1, 2, ..., G$, to be such that $\min_g N_g \to \infty$, that is, there are no small clusters.

Using the group assignments obtained from the $\mathcal{R}$ sample, $\hat{\gamma}_R$, we estimate Fama-MacBeth cross-sectional regressions on the $\mathcal{P}$ sample, separately for each group, obtaining the parameters $\hat{\alpha}_P^{(g)}$ and $\hat{\Lambda}_P^{(g)}$, which we use for testing. This step does not require the EM algorithm described in the previous subsection, as the group assignments (from the $\mathcal{R}$ sample) are taken as given in the $\mathcal{P}$ sample.

Our tests for multiple clusters take two forms. Our first test evaluates equality of average risk prices across groups for a total of $(G-1)K$ restrictions:

$$H_0:\ \bar{\lambda}_k^{(1)}=\bar{\lambda}_k^{(2)}=\ldots=\bar{\lambda}_k^{(G)}\ \forall k \tag{9}$$

$$\text{versus}\ \ H_1:\ \bar{\lambda}_k^{(g)}\neq\bar{\lambda}_k^{(g')}\ \text{for some}\ k,g,g'.$$

where $\bar{\lambda}_k^{(g)}\equiv E[\bar{\lambda}_{kt}^{(g)}]$. Our second test evaluates equality of cross-sectional slopes across groups at each point in time for a total of $(G-1)K\,P$ restrictions:

$$H_0:\ \lambda_{kt}^{(1)}=\lambda_{kt}^{(2)}=\ldots=\lambda_{kt}^{(G)}\ \forall k,t \tag{10}$$

$$\text{versus}\ \ H_1:\ \lambda_{kt}^{(g)}\neq\lambda_{kt}^{(g')}\ \text{for some}\ k,g,g',t.$$

The first test generalizes Fama-MacBeth style t tests to speak to differences in expected returns across market segments. The second test enriches the first by adding the information embedded in the dynamics of cross-sectional slopes to distinguish among groups of assets.[15] Indeed, groups may have different factor dynamics but identical unconditional risk prices. Intuitively, both tests assess whether adding clusters beyond the first generates differences in risk prices beyond what we would expect by chance. Note that neither test examines the equality of intercepts ($\bar{\alpha}^{(g)}$ or $\alpha_t^{(g)}$) because our focus is on risk price heterogeneity rather than on differences in zero-beta rates; however, both tests can easily be extended to include tests of equality of intercepts as well.[16]

We define test statistics for (9) and (10) analogously to F statistics for tests of equality of average slopes ("avg") and slope dynamics ("dyn"). Doing so requires some auxiliary quantities. First, let the estimated difference in cross-sectional slopes at date t for clusters g and g' be $\Delta\lambda_t^{(g,g')}$ and the time-series average of this quantity be $\Delta\bar{\lambda}^{(g,g')}$. Second, define $\hat{\Sigma}_{\lambda^{(g)}}$ as the cross-sectional covariance matrix of the parameter estimates at date t for cluster g, and define $\hat{\Sigma}_{\bar{\lambda}^{(g)}}$ as the time-series average of this value. Because cross-sectional slopes are estimated separately group-by-group and date-by-date, the covariance matrices

[15] A stronger form of both tests applies if factors are tradeable, namely, under the null hypothesis of no segmentation, the cross-sectional slopes for all segments should equal each other *and* the returns to the factor. We drop this latter condition to accommodate nontradeable factors, such as intermediary capital ratio innovations.

[16] That is, if all clusters have equal risk prices (λ) but they differ in their intercepts (α), then our estimation algorithm will, up to estimation error, detect the clusters. Our tests, however, will not reject the null as in such a case the clusters differ in their degree of mispricing, not in their prices of risk. The latter are the focus of this paper, and thus are the subjects of the restrictions under the null hypotheses.

of $\Delta\lambda_t^{(g,g')}$ and $\Delta\bar{\lambda}^{(g,g')}$ are simply $\hat{\Sigma}_{\lambda^{(g)}} + \hat{\Sigma}_{\lambda^{(g')}}$ and $\hat{\Sigma}_{\bar{\lambda}^{(g)}} + \hat{\Sigma}_{\bar{\lambda}^{(g')}}$, respectively. Combining these quantities into test statistics for differences in averages and differences in dynamics between all groups and factors, we obtain

$$F^{Avg} = \frac{1}{(G-1)K} \sum_{g=1}^{G-1} \Delta\bar{\lambda}^{(g,g+1)\prime} \left(\hat{\Sigma}_{\bar{\lambda}^{(g)}} + \hat{\Sigma}_{\bar{\lambda}^{(g+1)}} \right)^{-1} \Delta\bar{\lambda}^{(g,g+1)}, \tag{11}$$

$$F^{Dyn} = \frac{1}{(G-1)KP} \sum_{g=1}^{G-1} \sum_{t \in \mathcal{P}} \Delta\lambda_t^{(g,g+1)\prime} \left(\hat{\Sigma}_{\lambda_t^{(g)}} + \hat{\Sigma}_{\lambda_t^{(g+1)}} \right)^{-1} \Delta\lambda_t^{(g,g+1)}. \tag{12}$$

Incidentally, both tests downweight between-group differences in factor premiums on "junk" factors for which the dispersion in betas is low because the test statistics normalize differences in λs by the precision of λ estimates—which themselves are proportional to the cross-sectional variation in βs—on average or date by date.

The test statistics use estimated parameters from the $\mathcal{P}$ sample taking group memberships estimated on the $\mathcal{R}$ sample as given. If the dependence between observations in the $\mathcal{R}$ and $\mathcal{P}$ samples is limited, then the overfitting problem that arises when the same sample is used for both group membership estimation and testing is eliminated.[17] In principle, we can then use standard hypothesis testing methods to implement a test for multiple clusters: for example, the test of equal average risk prices is a simple F-test. However in simulation studies with realistic data generating processes, we found poor size control using this approach. We instead adopt a permutation testing approach (see, e.g., Lehmann and Romano 2005) to obtain critical values for the tests of hypotheses (9) and (10). The permutation tests randomly shuffle group assignments to obtain a distribution for the test statistic under the null of a single cluster, and are thus particularly well-suited for evaluating whether groups differ in some coefficient(s) of interest, for example, the set of cross-sectional slopes. We will show in the next section that this approach largely eliminates the tendency to over-reject.

We implement our permutation tests as follows. Given a set of group assignments $\hat{\gamma}_R$ obtained on the $\mathcal{R}$ sample, we draw random group assignments using as probability weights the group proportions implied by $\hat{\gamma}_R$. Next we calculate the statistics in Equations (11) and (12) on the $\mathcal{P}$ sample using the random group assignments. We repeat this procedure M=5,000 times to obtain a distribution of test statistics. Each permutation acts similar to a bootstrap draw, and generating many permutations fills out the distribution of the test

[17] To ensure that the dependence between our subsamples is negligible, we leave a gap of one year between the $\mathcal{R}$ and $\mathcal{P}$ samples. If the number of years in our sample is odd, we use the middle year as the "gap"; otherwise we split the sample evenly and then drop the last year of the $\mathcal{R}$ sample.

statistic under the null hypothesis.[18] p-values are then simply computed as the proportion of permutation statistics larger than the test statistics. The advantage of this approach is that it adjusts for possible issues arising from departures from our assumptions, such as finite group sizes and sample lengths, while requiring less structure than a traditional bootstrap design.

The test for multiple clusters requires the researcher to specify the number of clusters under the alternative, G. Rather than imposing a specific value, we instead implement the test for a range of values $G = 2, 3, 4, 5$.[19] We set the largest value under the alternative to five to balance the requirement that each cluster be "large" against the possibility of many market segments in the data. We account for the fact that this method obtains a joint test based on four individual tests by using a simple Bonferroni correction (see, e.g., Lehmann and Romano 2005). In practice, this means we calculate the overall p-value as the minimum of the individual p-values multiplied by four, and we then compare the overall p-value to desired size of the test (e.g., 5%).

1.3 Finite sample properties of the test for multiple clusters

We analyze the finite-sample properties of the above tests for multiple clusters via an extensive simulation study. Ensuring that the proposed tests have satisfactory finite-sample properties is a necessary condition for interpreting the test results that we present in the next section. We consider a range of values of *N, T, K,* and G, to match the various specifications that we consider in our empirical analysis in Section 3.

The DGP for the simulation study is obtained as follows, and it uses data described in detail in the next section. We first estimate factor means, μ_f, and covariance matrices, Σ_f, for two representative asset pricing models, the CAPM ($K = 1$) and Carhart ($K = 4$) factor models, at daily and monthly frequencies. Next we estimate time-series betas and idiosyncratic volatilities for each of 234 domestic equity portfolios (the portfolio set "P3" described in Section 2) for each factor model at daily and monthly frequencies.

For each simulation s, we draw with replacement N=75 stocks or N=225 stocks, where each draw consists of $(\{\beta_{i1}, \ldots, \beta_{iK}\}, \sigma_{i\epsilon})$ pairs. Asset returns are obtained as $r_{it} = \beta_i f_t + \epsilon_{it}$, with simulated factor returns $f \sim N\left(\mu_f, \Sigma_f\right)$ and idiosyncratic returns $\epsilon_i \sim N\left(0, \sigma^2_{i\epsilon}\right)$. We simulate factor returns with two time-series dependency structures: factor returns are either i.i.d. or exhibit time-series dependence as in a GARCH(1,1) process. We include simulations with volatility

[18] It is possible to consider all possible permutations of the assets when N and G are very small; however, for the values of N in our applications this approach is infeasible. As noted in Lehmann and Romano (2005), using a randomly drawn subset of all possible permutations, as we do, also controls the level of the test.

[19] One might instead consider selecting the number of groups to use under the alternative via an information criterion, such as the AIC or BIC. Unfortunately, doing so complicates the distribution of the test statistic, as in that case the alternative is data driven, and furthermore, neither of these information criteria is known to yield consistent estimates of the number of groups, even in simpler settings than considered here. For both of these reasons, we instead consider a range of values of G and adopt a multiple testing approach.

dependence to confirm that our procedure is robust to (1) heteroskedasticity and (2) long-range volatility dependence that may introduce dependence between the $\mathcal{R}$ and the $\mathcal{P}$ samples. Our GARCH processes use Zivot's (2009) parameters of $a=0.18$ and $b=0.78$ for monthly simulations and $a=0.09$ and $b=0.89$ for daily simulations estimated from S&P 500 data for 1986–2003, and we impose constant conditional correlation among the factors (Bollerslev 1990), with correlations estimated for the Carhart model on the entire 1963–2016 sample.

For the "monthly" design, we set T=300, and for the "daily" design we set T=10,000. We estimate the multicluster factor models using the methods described in Section 1.1.1, using data from our $\mathcal{R}$ sample (the first half minus a year at the end), and we implement tests on the $\mathcal{P}$ sample (the second half). We report rejection frequencies for nominal 0.05 level tests,[20] for a single alternative $G=2,3,4,5$, and for a multiple comparison $G\in\{2,3,4,5\}$ using a Bonferroni correction. We simulate under each design S=500 times. Because of computation constraints, we reduce the number of permutations within each simulation from M=5,000 to M=500. A test of correct size should reject 5% of the time, up to simulation variability.

Table 1 shows that the proposed testing procedure generally has rejection rates that are comparable to the nominal test sizes. For the "monthly" simulation design, in the left panel, rejection frequencies are generally close to the nominal 0.05 level in the i.i.d. case, though they are sometimes larger when N=225. In this case our assumption of T/N being "large" is less plausible. For the "daily" simulation design, in the right panel, almost all rejection frequencies are between 0.03 and 0.08. Rejection rates do not appear sensitive to the choice of factor model (CAPM or Carhart) with the exception of the large N and small T design.[21] Overall, we conclude that the test has satisfactory size control in finite samples, though the simulations suggest caution in interpreting borderline results in the monthly samples with larger factor models.

2. Data

Our data consist of risk factors and well-diversified portfolios common throughout the empirical asset pricing literature. To ensure that our conclusions on market segmentation are robust to choices of a particular factor model, we include several leading factor models in our analysis. These models include the CAPM; the Fama and French (1992) three-factor model ("FF3F"); the Carhart (1997) four-factor model ("Carhart"); the Fama and French (2015) five-factor model ("FF5F"); the He, Kelly, and Manela (2017) intermediary-capital

[20] The size control for significance levels of 0.01 and 0.10 are very similar to the 0.05 results presented here.

[21] Note that this testing approach is robust to the inclusion of factors with zero risk prices, that is, useless factors. In this case the price of risk for that factor is also homogeneous for all assets (it is zero) and thus satisfies the null hypothesis, and so the size of the test is unaffected. In finite samples including useless factors will affect power: if some factors have risk prices that differ across clusters, the inclusion of useless factors (which have the same price, zero, across clusters) will make the clusters appear closer together, that is, closer to the null hypothesis.

Table 1
Finite sample rejection rates

Design			$T=300$ months					$T=10,000$ days				
N	K	GARCH	G = 2	= 3	= 4	= 5	∈2–5	G = 2	= 3	= 4	= 5	∈2–5
A. Tests for equality of $\bar{\lambda}_k$ across groups												
75	**1**	**No**	0.04	0.04	0.07	0.07	0.07	0.05	0.07	0.07	0.05	0.06
75	**4**	**No**	0.07	0.06	0.06	0.05	0.07	0.06	0.06	0.04	0.05	0.07
225	**1**	**No**	0.05	0.05	0.05	0.06	0.06	0.04	0.06	0.06	0.07	0.07
225	**4**	**No**	0.08	0.08	0.11	0.08	0.10	0.04	0.03	0.07	0.07	0.06
75	**1**	**Yes**	0.04	0.05	0.07	0.07	0.07	0.06	0.08	0.07	0.05	0.07
75	**4**	**Yes**	0.06	0.06	0.06	0.04	0.07	0.05	0.05	0.04	0.05	0.07
225	**1**	**Yes**	0.04	0.05	0.05	0.03	0.05	0.04	0.06	0.07	0.06	0.05
225	**4**	**Yes**	0.07	0.09	0.09	0.08	0.11	0.06	0.06	0.06	0.06	0.08
B. Tests for equality of λ_{kt} across groups												
75	**1**	**No**	0.03	0.03	0.03	0.10	0.06	0.08	0.04	0.03	0.08	0.08
75	**4**	**No**	0.06	0.04	0.03	0.05	0.05	0.04	0.01	0.04	0.08	0.07
225	**1**	**No**	0.04	0.05	0.07	0.06	0.06	0.05	0.11	0.09	0.09	0.08
225	**4**	**No**	0.10	0.12	0.09	0.06	0.12	0.08	0.07	0.06	0.03	0.07
75	**1**	**Yes**	0.05	0.05	0.05	0.13	0.08	0.09	0.04	0.01	0.08	0.07
75	**4**	**Yes**	0.06	0.03	0.02	0.05	0.04	0.04	0.01	0.03	0.07	0.04
225	**1**	**Yes**	0.06	0.06	0.05	0.06	0.06	0.07	0.08	0.09	0.08	0.08
225	**4**	**Yes**	0.13	0.13	0.12	0.07	0.17	0.07	0.08	0.07	0.04	0.06

This table reports the proportion of simulations in which we reject a single cluster in favor of multiple clusters using the two F-tests described in Section 1.1.2 when the data comes from a single cluster model. We consider 16 simulation designs: $N=75$ and $N=225$ simulated portfolios; CAPM ($K=1$) and Carhart ($K=4$) factor models; i.i.d. and GARCH(1,1) factor realizations; and $T=300$ months (left columns) and $T=10,000$ days (right columns). We use $S=500$ simulations of each design. GARCH processes for the factors use parameters $a=0.18$ and $b=0.78$ for monthly simulations and $a=0.09$ and $b=0.89$ for daily simulations (both from Zivot 2009), and factors have constant conditional correlation (Bollerslev 1990). All tests are at the 0.05 nominal level. We report rejection frequencies for a single alternative $G=2,3,4$, or 5, and for a multiple comparison $G\in\{2,3,4,5\}$ using a Bonferroni correction.

factor model ("HKM");[22] the Hou, Xue, and Zhang (2015) q-factor model ("HXZQ"); the Carhart (1997) four-factor model augmented with the first three principal components of return residuals ("Carhart+3"); and a model based on the first five principal components of the test asset returns ("PC5").[23,24] Kenneth French's website provides the Fama and French factors and the momentum

[22] The intermediary capital factor is available monthly from January 1970 and daily from January 2000. For sample periods in which daily data are available for portfolio returns, but not for the intermediary capital factor, we use monthly intermediary capital factors and monthly portfolio returns to estimate betas and idiosyncratic volatilities, and we convert monthly volatility estimates to daily estimates by dividing by $\sqrt{21}$. In periods for which both data frequencies are available, the cross-sectional correlation in betas and idiosyncratic volatilities estimated using daily and monthly intermediary capital factors is about 90%, with slight variation depending on the portfolio set considered.

[23] Very similar results are obtained when using two, three, or four principal components. The results when using just a single principal component are essentially identical to those for the CAPM.

[24] Daniel and Titman (1997) and Daniel et al. (2020) show that portfolios may have different average returns but the same factor loadings if characteristics determine expected returns. Factors can appear to have different risk prices across portfolios if the model for expected returns is misspecified. The PC5 model avoids the need to take a stand on a specific factor model, and instead selects the "factors" that best describe return covariances in the data. In principle, including enough principal components will absorb the unpriced confounding factors described in Daniel et al. (2020). We choose five principal components to balance including "many" factors and ensuring that we can estimate risk prices well in clusters with few portfolios ($N_g \gg K$). Appendix A provides further discussion of the issue of mismeasured factors.

factor both in domestic and in global versions (as in Fama and French 2012). Asaf Manela's website provides the intermediary capital factor. Lu Zhang shares his investment and return-on-equity factors for the q-factor model. We adjust the time-series length and sampling frequency of the factors to match the corresponding portfolio sets described below.

Throughout our analysis we use value-weighted portfolios as test assets. While in principle our approach can be applied to individual stocks, at extreme computational cost, we use portfolios rather than individual securities for the usual reasons: (1) to increase the stability of security characteristics over time, including their risk exposures and their cluster assignments; (2) to decrease the measurement error in betas through diversification of idiosyncratic risk;[25] and (3) to reduce the sparsity of the matrix of realized returns. A focus on portfolios rather than individual assets also follows from Merton's (1973) intertemporal CAPM, in which all multifactor-minimum variance efficient investments are spanned by $K+1$ factor-mimicking portfolios. Importantly, risk premiums estimates obtained using portfolio returns do not generally apply to portfolio constituents because comovements between securities strongly influence portfolio dynamics.

We obtain characteristic-sorted equity portfolio data from Kenneth French's website. Our domestic equity data are daily for 1963 to 2016. The domestic equity portfolios are formed using standard sorting characteristics, including market capitalization, book-to-market ratios, prior returns, investment rates, operating profitability, market beta, and industry group. We also use double-sorted portfolios based on market capitalization and market beta, book-to-market ratio, prior return, investment rate, and operating profitability. In addition to these standard equity portfolios, we construct a set of 100 placebo equity portfolios to evaluate whether we detect segmentation where none exists. Our base assets are U.S. common stocks (sourced from CRSP) from 1963 to 2016 to parallel our domestic equity portfolios. Rather than sorting by characteristics, we instead assign each stock a portfolio number in 1 to 100 with equal probability and value-weight returns to obtain a portfolio return. Because these portfolios are large and formed at random, we expect their characteristics and risk prices to be the same up to sampling variation.[26]

The international stock portfolio data are monthly and run from 1991 to 2016.[27] They consist of double-sorted size-book-to-market and size-prior return

[25] A related benefit of using portfolios rather than individual stocks is that idiosyncratic risk cannot drive differences in apparent risk prices. Well-diversified portfolios eliminate such risk, so differences in risk prices across these test assets are a violation of the (approximate) arbitrage pricing theory (e.g., Chamberlain and Rothschild 1983) up to transaction costs. By contrast, small and finite cross-sectional differences in risk prices are allowable under the approximate APT, and idiosyncratic risk may prevent such differences from disappearing (Pontiff 2006 surveys this literature in discussing "Myth 5" of rational arbitrage).

[26] To ensure that results are robust to a particular set of random draws, we repeat this procedure to produce five sets of placebo portfolios. Results are similar across placebo portfolios throughout, so we report for only the first set.

[27] We use monthly data rather than daily data to avoid confusing asynchronous returns with segmented pricing. In a previous version of this paper we used daily data and obtained nearly identical results.

portfolios for developed markets in North America, Europe, Japan, and Asia Pacific (excluding Japan) regions. See Fama and French (2012) for details on the underlying data and security set. In our analysis, we pair these portfolios with global versions of the Carhart factors, and we substitute the global market factor for the U.S. market factor in the intermediary-capital factor model. Because return on equity is constructed using accounting data that varies across countries and is not readily accessible, we drop the q-factor model in studying risk price heterogeneity for these portfolios.

The cross-asset class sample is mainly courtesy of Asaf Manela, and our data is monthly for 1986 to 2010. In addition to domestic equities, these data include commodities, U.S. Treasuries and corporate bonds, sovereign bonds, options, and currencies.[28] Unlike the domestic and international stock portfolios, the diverse He, Kelly, and Manela (2017) portfolios necessarily start and end at different dates as new asset classes come into being and database availability changes. While Manela's data include some asset classes from 1970, to maintain a near-balanced panel, we initialize the sample at the start of the year that commodity and options data begin (1986) and end the sample at the end of the year that foreign exchange data end (2010). He, Kelly, and Manela (2017) describes portfolio construction and primary sources for these data in greater detail.

Table 2 summarizes the seven collections of standard portfolios. The first component of each subtable marks the constituents of each portfolio set. Our smallest and largest portfolio sets consist of 75 (P1) and 234 (P3) domestic equity portfolios, respectively. All portfolio sets sort on at least three variables (e.g., region, size, and book-to-market ratio) to ensure coverage of several potential dimensions of market segmentation. The second component of each subtable reports summary statistics for each set of portfolios or asset class. We report the first and second moments of average returns and return volatility for each group. Dispersion in average returns and volatility within each portfolio set is on the order of several percent per year, indicating considerable variation in factor exposures or risk prices within each group. These quantities also vary across sorts, indicating that different sorting variables capture different dimensions of heterogeneity. We omit summary statistics for the placebo portfolios because by construction they have no natural grouping dimensions.

For each factor model-portfolio set combination, we repeat our analysis using both the full time-series and shorter subsamples. Just as the risk characteristics of portfolios may change over time, so too may the market frictions that separate portfolios into segments with different risk prices. We split the data into two halves for international equity and cross-asset class analyses, and we increase the number of splits to three for domestic equity and placebo portfolio analyses because domestic equity data are available for a longer time period.

[28] We exclude credit default swaps and sovereign bonds from our analysis because data for these asset classes are available only in the second half of our sample period (2001–2012 and 1995–2011, respectively).

Table 2
Summary statistics for collections of test portfolios

A. Domestic equity portfolios, daily data, 1963–2016

#	Size-β_{mkt}	Size-B/M	Size-mom	β_{mkt}	B/M	Size	Mom	Ind	Inv	Prof	Size-Inv	Size-Prof
P1	✓	✓	✓									
P2	✓	✓	✓	✓	✓	✓	✓					
P3	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Ave $\bar{r}$	12.71	13.10	12.91	11.13	12.45	12.53	10.90	11.88	11.71	10.84	13.04	12.65
Ave σ_r	17.36	17.18	17.78	17.77	16.71	16.47	18.06	21.57	16.69	16.93	16.98	17.08
$\sigma(\bar{r})$	1.76	2.63	4.01	0.67	1.84	0.90	3.73	1.90	1.60	1.21	2.26	1.95
$\sigma(\sigma_r)$	4.58	1.72	2.94	4.93	1.39	0.98	3.32	4.88	1.57	1.45	1.71	1.56
N	25	25	25	10	10	10	10	49	10	10	25	25
Source	KF	KF	KF	KF	KF	KF	KF	KF	KF	KF	KF	KF

B. International equity portfolios, daily data, 1991–2016

	Size-value 25				Size-momentum 25			
#	North America	Asia-Pacific	Europe	Japan	North America	Asia-Pacific	Europe	Japan
P5	✓	✓	✓	✓				
P6	✓	✓	✓	✓	✓	✓	✓	✓
Ave $\bar{r}$	11.67	9.63	8.39	4.48	11.70	8.09	6.61	4.72
Ave σ_r	19.32	17.69	16.71	21.74	19.60	18.36	17.41	22.21
$\sigma(\bar{r})$	2.30	4.00	2.16	2.18	4.43	6.38	5.14	2.29
$\sigma(\sigma_r)$	2.13	1.90	2.85	2.06	2.84	2.94	3.51	2.34
N	25	25	25	25	25	25	25	25
Source	KF	KF	KF	KF	KF	KF	KF	KF

C. Cross-asset class portfolios, monthly data, 1986–2010

#	Size-B/M	Size-β_{mkt}	Size-mom	Commod	U.S. bonds	Options	FX
P7	✓			✓	✓	✓	✓
P8	✓	✓	✓	✓	✓	✓	✓
Ave $\bar{r}$	12.61	12.92	12.40	5.61	7.11	8.23	3.20
Ave σ_r	19.83	19.26	20.88	25.87	3.78	16.07	8.21
$\sigma(\bar{r})$	2.64	1.49	3.59	6.22	1.58	5.79	3.41
$\sigma(\sigma_r)$	3.13	4.93	4.68	6.24	1.70	2.03	1.41
N	25	25	25	23	20	18	12
Source	KF	KF	KF	HKM/CRB	HKM/CRSP/N	HKM/CJS	HKM/LMM/MSSS
HKM start year	1970	N/A	N/A	1986	1974	1986	1976
HKM end year	2012	N/A	N/A	2012	2012	2012	2010

This table reports the composition and summary statistics for eight main collections of test portfolios. The first panel consists of domestic test portfolios. The top component indicates whether a set of single- or double-sorted portfolios is included in portfolio collections P1–P3. "Mom," "Inv," and "Prof" abbreviate prior return, investment, and operating profitability sorted portfolios, respectively, and "Ind" is the Fama-French 49 industry portfolios. We download daily value-weighted portfolios from the data library on Kenneth French's website ("KF"). The middle component reports the average of the mean daily returns (ave $\bar{r}$) and standard deviation of daily returns (ave σ_r) within a portfolio set, as well as the standard deviation of these quantities $\sigma(\bar{r})$ and $\sigma(\sigma_r)$, all in annualized percentage terms. The second panel consists of international test portfolios P5–P6 (skipping P4 as our placebo portfolio set). These portfolios parallel the value-weighted daily size-value and size-momentum portfolios for our four global regions. The third panel consists of diversified domestic and international equity and nonequity assets P7–P8 used in He, Kelly, and Manela (2017) ("HKM"). We download the corresponding monthly portfolio returns for nonequity assets from Asaf Manela's website. Where data are not compiled by the authors, we report primary sources including the Commodities Research Bureau ("CRB"), the CRSP Fama Bond Portfolios ("CRSP"), Nozawa (2017) ("N"), Constantinides, Jackwerth, and Savov (2013) ("CJS"), Lettau, Maggiori, and Weber (2014) ("LMM"), and Menkhoff et al. (2012) ("MSSS"). We add start and end years for data availability in this panel because they differ across asset classes.

Splitting the sample allows our methodology to accommodate time-varying risk characteristics and segmentation without incurring the extreme computational costs of rolling estimation of group assignments. At the same time it allows us to evaluate stability in these characteristics, as we do in Appendix C.

3. Segmentation Everywhere

3.1 Testing for market segmentation

Table 3 presents our main empirical finding of strong and pervasive evidence of segmentation across choices of test assets, benchmark factor models, and time periods. Focusing first on the left columns of Table 3, we find evidence of differing *average* prices of risk across clusters for most domestic equity portfolio sets and factor models, and for almost all international equity and multiasset class portfolio sets and factor models. The main exceptions to this strong evidence of multiple clusters arise when using portfolio set P1, which is comprised of the most common double-sorted domestic equity portfolios, or the CAPM factor model. The weaker evidence against the null of equal average risk prices for portfolio set P1 is an indication that the assets in P1 are homogeneous or only weakly heterogeneous. As P1 contains three sets of 5×5 portfolio sorts on size-beta, size-value, and size-momentum, failing to find strong evidence of heterogeneity here is consistent with intuition about these portfolios as relatively easy to trade. The failure to reject the null of equal average risk prices using the CAPM is consistent with the expectation that most U.S. investors can trade the market factor at low cost, particularly in the most recent period.[29] We find that differences in unconditional risk premiums are important for almost all other environments with richer cross-sections of assets or better factor models for returns.

Tests of equal risk *dynamics* in the right columns make much stronger statements about segmentation. We find that *every test, except one* (of 173 in total), rejects the null hypothesis of a single cluster at the 5% significance level, and all p-values are less than 0.1% for portfolio sets P2–P8 (excluding the placebo set, P4).[30] In addition, of the 62 segmentation tests applied to the placebo portfolio set, reported in panel B, we reject unified pricing at the 5% level in just three of them, or 4.8% of our tests. This rate is quite similar to what we observe in the simulation study of Section 1.1.3. We conclude that where segmentation does not exist, we do not find evidence against unified pricing different from what we would expect by chance.

From these strong rejections we conclude that cross-sectional variation in risk prices is ubiquitous. In addition to the international and cross-asset class

[29] The failure to reject the null using the CAPM is also consistent with this model having only weak explanatory power for the cross-section of expected returns, leading to low power of the test. When using more informative models, we find strong evidence against the null.

[30] Our simulation study suggests that these tests somewhat over-reject for monthly data with larger factor models, but the p-values reported here are far from borderline cases that warrant statistical caution.

Table 3
Bonferroni-adjusted *p*-values for tests of multiple clusters against a single cluster

A. Domestic equity portfolios

		Equal average risk prices $\bar{\lambda}_k$				Equal risk prices λ_{kt}			
Portfolios	Model	1963–2016	1963–1980	1981–1998	1999–2016	1963–2016	1963–1980	1981–1998	1999–2016
P1	CAPM	0.312	0.003	0.025	0.561	0.026	0.026	0.012	0.006
	FF3F	0.057	0.112	0.166	0.011	0.000	0.000	0.000	0.000
	Carhart	0.050	0.279	0.287	0.102	0.000	0.000	0.000	0.000
	FF5F	0.078	0.054	0.009	0.375	0.000	0.074	0.000	0.000
	HKM	0.236		0.134	0.057	0.037		0.010	0.001
	HXZQ	0.000	0.896	0.000	0.671	0.000	0.000	0.000	0.000
	Carhart+3	0.005	0.017	0.046	1.000	0.000	0.000	0.000	0.000
	PC5	0.005	0.120	0.139	1.000	0.000	0.000	0.000	0.000
P2	CAPM	0.001	0.050	0.014	0.219	0.000	0.000	0.000	0.000
	FF3F	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Carhart	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000
	FF5F	0.001	0.002	0.000	0.100	0.000	0.000	0.000	0.000
	HKM	0.345		0.004	0.038	0.000		0.000	0.000
	HXZQ	0.000	0.009	0.000	0.000	0.000	0.000	0.000	0.000
	Carhart+3	0.168	0.110	0.002	1.000	0.000	0.000	0.000	0.000
	PC5	0.000	0.123	0.000	0.323	0.000	0.000	0.000	0.000
P3	CAPM	0.052	0.001	0.000	0.006	0.000	0.000	0.000	0.000
	FF3F	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000
	Carhart	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000
	FF5F	0.000	0.018	0.002	0.000	0.000	0.000	0.000	0.000
	HKM	0.543		0.000	0.007	0.000		0.000	0.000
	HXZQ	0.000	0.007	0.000	0.000	0.000	0.000	0.000	0.000
	Carhart+3	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.000
	PC5	0.004	0.000	0.000	0.003	0.000	0.000	0.000	0.000

(*Continued*)

Table 3
(Continued)

B. Domestic equity placebo portfolios

Portfolios	Model	Equal average risk prices $\bar{\lambda}_k$				Equal risk prices λ_{kt}			
		1963–2016	1963–1980	1981–1998	1999–2016	1963–2016	1963–1980	1981–1998	1999–2016
P4 (placebo)	CAPM	0.082	0.092	1.000	0.618	0.086	0.070	0.650	1.000
	FF3F	0.579	1.000	0.186	0.101	1.000	1.000	1.000	1.000
	Carhart	0.594	0.242	0.155	0.822	0.602	1.000	0.691	0.262
	FF5F	0.153	1.000	0.021	0.883	1.000	1.000	1.000	1.000
	HKM	1.000		0.066	0.711	0.015		0.073	0.466
	HXZQ	0.246	0.243	1.000	0.197	1.000	1.000	1.000	0.198
	Carhart+3	0.599	0.491	1.000	1.000	0.118	0.346	0.382	0.041
	PC5	1.000	0.424	1.000	0.561	1.000	1.000	1.000	1.000

C. International equity portfolios

Portfolios	Model	Equal average risk prices $\bar{\lambda}_k$			Equal risk prices λ_{kt}		
		1991–2016	1991–2003	2004–2016	1991–2016	1991–2003	2004–2016
P5	CAPM	0.000	0.036	0.000	0.000	0.000	0.000
	FF3F	0.000	0.000	0.000	0.000	0.000	0.000
	Carhart	0.000	0.000	0.000	0.000	0.000	0.000
	FF5F	0.000	0.000	0.000	0.000	0.000	0.000
	HKM	0.000	0.007	0.000	0.000	0.000	0.000
	HXZQ						
	Carhart+3	0.002	0.000	0.001	0.000	0.000	0.000
	PC5	0.000	0.000	0.073	0.000	0.000	0.000
P6	CAPM	0.000	0.001	0.000	0.000	0.000	0.000
	FF3F	0.000	0.000	0.000	0.000	0.000	0.000
	Carhart	0.000	0.000	0.000	0.000	0.000	0.000
	FF5F	0.000	0.000	0.000	0.000	0.000	0.000
	HKM	0.000	0.188	0.000	0.000	0.000	0.000
	HXZQ						
	Carhart+3	0.000	0.000	0.000	0.000	0.000	0.000
	PC5	0.011	0.016	0.000	0.000	0.000	0.000

(*Continued*)

Table 3
(Continued)

D. Cross-asset class portfolios

Portfolios	Model	Equal average risk prices $\bar{\lambda}_k$			Equal risk prices λ_{kt}		
		1986–2010	1986–1997	1998–2010	1986–2010	1986–1997	1998–2010
P7	CAPM	0.000	0.000	0.000	0.000	0.000	0.000
	FF3F	0.000	0.000	0.000	0.000	0.000	0.000
	Carhart	0.000	0.000	0.000	0.000	0.000	0.000
	FF5F	0.000	0.010	0.000	0.000	0.000	0.000
	HKM	0.000	0.000	0.000	0.000	0.000	0.000
	HXZQ	0.000	0.000	0.000	0.000	0.000	0.000
	Carhart+3	0.000	0.000	0.000	0.000	0.000	0.000
	PC5	0.000	0.000	0.000	0.000	0.000	0.000
P8	CAPM	0.000	0.000	0.000	0.000	0.000	0.000
	FF3F	0.000	0.016	0.000	0.000	0.000	0.000
	Carhart	0.000	0.000	0.000	0.000	0.000	0.000
	FF5F	0.000	0.053	0.000	0.000	0.000	0.000
	HKM	0.000	0.000	0.000	0.000	0.000	0.000
	HXZQ	0.000	0.000	0.000	0.000	0.000	0.000
	Carhart+3	0.000	0.000	0.000	0.000	0.000	0.000
	PC5	0.000	0.000	0.000	0.000	0.000	0.000

This table reports p-values from F-tests described in Section 1.1.2 for comparing a multiple-clusters model to a single-cluster model. We perform pairwise comparisons of G=2,3,4,5 clusters against one cluster and obtain a Bonferroni-adjusted p-value by taking the minimum individual test p-value and multiplying it by four. The left columns report tests of equality of average risk prices across clusters for all factors ($\bar{\lambda}_k$), and the right columns report tests of equality of cross-sectional slopes across clusters for all factors and dates (λ_{kt}). Portfolio sets and factor models are described in Section 2. For the He, Kelly, and Manela (2017) factor model, we use daily data for the most recent time period and monthly data for earlier time periods. We do not have sufficient coverage for their intermediary capital factor for 1963–1980 to include it in the domestic equity portfolio analysis. The q-factor (HXZQ) model is excluded from the international portfolio analysis because we do not have global return-on-equity factor data.

contexts, where we may have anticipated segmentation a priori, even one of the world's most-developed and liquid markets, such as U.S. stocks, exhibits significant variation in compensation to factor exposure.[31] Given the large size of our typical panels, these statistical rejections of the null of unified risk prices may not correspond with *economically* meaningful differences in cross-sectional dispersion of realized returns and expected returns. The next subsection confirms that they do.

Despite the strong rejections of integrated markets, our results are conservative in several respects. First, our ability to identify market segments depends on whether the selected factors have heterogeneous risk prices. In this respect our paper joins most others in the segmentation literature in depending on the choice of factor model. We address this issue in part by evaluating market segmentation with a battery of leading factor models and diverse portfolio sets. Second, the Bonferroni adjustment for penalizing multiple tests is too severe if tests are correlated, as our tests likely are. Third, to the extent that group assignments are time varying, our subsample approach for testing for multiple clusters will have lower power, as using the "wrong" clusters for the $\mathcal{P}$ sample decreases the improvement in model fit with multiple clusters relative to a single cluster. Evidence of segmented markets comes through strongly in Table 3 despite these features.

3.2 The economic importance of market segmentation

To assess the economic contribution of heterogeneous risk-price models of expected returns, we first define two measures of the dispersion in expected returns explained by a particular model. Perhaps of greatest academic interest is the estimated explanatory power for the cross-section of expected returns. Our first measure of economic importance is the variation in returns explained by our model, measured as the model-implied variance of unconditional average returns across test assets:

$$\sigma^2_{G^*}(\bar{r}) \equiv var_i\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_t^{(g)} + \beta_i \lambda_t^{(g)}\right)\right), \tag{13}$$

where we select the number of clusters $G^* \in \{2,3,4,5\}$ as the value that minimizes the AIC.[32] We calculate the variance in expected returns analogously

[31] Of course, others (e.g., Hong, Lim, and Stein 2000; Grinblatt and Moskowitz 2004) have noted variation in risk prices by market capitalization, for example, small-cap versus large-cap momentum premiums. Our interpretation of the results for portfolio sets P1-P3 is that investors are impeded from pursuing a long-short strategy that nets out factor exposure by (1) frictions that are larger than we think (see, e.g., Lesmond, Schill, and Zhou 2004; Patton and Weller 2020) and/or (2) risks to such a strategy (e.g., crash risks) that are larger than we see (e.g., Shleifer and Vishny 1997; Brunnermeier, Pedersen, and Nagel 2008; Daniel and Moskowitz 2016). Determining the source(s) of segmentation is an interesting avenue for future research; we take one step in that direction in Section 4.

[32] Note that we use the AIC here only as a rule of thumb for choosing one reasonable value of G; we do not use this in our testing approach.

for the one-cluster model using standard, full-sample Fama-MacBeth estimates for α_t and λ_t, and we report the ratio of $\sigma^2_{G^*}(\bar{r})/\sigma^2_1(\bar{r})$.

Our second measure of economic importance quantifies the Sharpe ratio improvements from using factors constructed from multiple clusters rather than from one. This measure is the difference in maximal in-sample Sharpe ratios achievable using factor-mimicking portfolios from all clusters and from a single combined cluster,

$$\Delta SR_{G^*} \equiv \sqrt{\mu'_\Lambda \Sigma_\Lambda^{-1} \mu_\Lambda} - \sqrt{\mu'_\lambda \Sigma_\lambda^{-1} \mu_\lambda}. \tag{14}$$

If the factors are tradeable, $\sqrt{\mu'_\lambda \Sigma_\lambda^{-1} \mu_\lambda}$ is simply the maximal in-sample Sharpe ratio achievable using the factors alone, $\sqrt{\mu'_f \Sigma_f^{-1} \mu_f}$.

The second measure takes the perspective of an arbitrageur able to frictionlessly invest in all portfolio sets. To the extent that risk prices vary in the cross-section, investing across clusters improves the maximal in-sample Sharpe ratio by enabling (1) tilts toward "local" factors with particularly high compensation per unit risk and (2) diversification of risk across imperfectly correlated mimicking portfolios. ΔSR_{G^*} answers how much better our hypothetical unconstrained investor can do in mean-variance terms by recognizing risk price heterogeneity. Equivalently, from a no-arbitrage perspective, it represents the magnitude of segmentation frictions in the size of gains remaining *despite* the activities of sophisticated potential arbitrageurs. In this respect, ΔSR_{G^*} measures cross-sectional limits of arbitrage described by Gromb and Vayanos (2018).

Table 4 reports these measures for all combinations of test assets, benchmark factor models, and time periods. Despite not being the optimization objective, dispersion in average returns increases almost everywhere, and often considerably. Starting with the domestic portfolios and setting aside outlier CAPM and He, Kelly, and Manela (2017) (HKM) rows, increases in cross-sectional dispersion in average returns range from 3% to 161%, with with 25th and 75th percentiles increases of 7% and 61%, respectively. These improvements are on par with adding additional factors to standard asset pricing models; for example, augmenting the Fama-French three-factor model with investment and profitability factors, as in the Fama-French five-factor model, increases cross-sectional dispersion in average returns by 49% for P1 portfolios, 16% for P2 portfolios, and 28% for P3 portfolios. Performance gains are much larger for the CAPM and HKM models because (1) both models perform quite badly in explaining cross-sectional variation in average returns, and (2) additional clusters likely pick up omitted factors, a possibility we address in Section 5. Conversely, while our statistical tests reject the null of a single cluster for all portfolio sets, the improvement in likelihood from adding additional clusters is sometimes small. The AIC selects a single cluster in several instances for portfolio set P1. This portfolio set is sufficiently homogeneous that the

Table 4
Contribution of clusters to expected return variation and Sharpe ratio improvements

A. Domestic equity portfolios

		1963–2016		1963–1980		1981–1998		1999–2016	
Portfolios	Model	Var($\bar{r}$)	ΔSR	Var($\bar{r}$)	ΔSR	Var($\bar{r}$)	ΔSR	Var($\bar{r}$)	ΔSR
	CAPM	3.77	0.26	55.78	1.04	2.31	0.77	161.88	0.15
	FF3F	1.81	0.74	1.19	0.26	1.44	0.70	1.80	−0.09
	Carhart	1.03	0.10	*	*	1.03	0.13	1.38	0.16
P1	FF5F	*	*	*	*	1.95	1.41	1.10	0.15
	HKM	8.25	0.33			34.85	1.18	5.30	0.38
	HXZQ	1.16	0.67	*	*	*	*	2.67	0.29
	Carhart+3	*	*	*	*	*	*	*	*
	PC5	1.01	0.69	*	*	1.23	1.37	1.18	0.87
	Median	1.49	0.50	28.49	0.65	1.70	0.98	1.80	0.16
	CAPM	6.48	0.55	80.02	1.61	1.02	0.20	52.70	0.32
	FF3F	2.15	0.69	1.40	0.40	2.67	1.93	2.48	0.74
	Carhart	1.05	0.07	1.17	0.29	1.17	0.95	1.30	0.61
P2	FF5F	1.61	0.53	1.15	0.27	1.71	0.79	1.05	0.24
	HKM	2.60	0.17			7.72	0.68	4.16	0.35
	HXZQ	1.23	0.47	0.97	0.25	1.76	2.11	2.31	0.68
	Carhart+3	*	*	1.01	0.15	1.12	1.41	1.07	0.27
	PC5	1.02	0.32	1.06	0.20	1.05	0.44	1.13	0.32
	Median	1.61	0.47	1.15	0.25	1.44	0.87	1.81	0.34
	CAPM	2.75	0.17	22.66	0.84	1.22	0.46	6.16	0.48
	FF3F	1.67	0.82	1.41	0.96	1.67	1.57	1.75	0.47
	Carhart	1.41	0.86	1.16	0.35	1.07	0.49	1.55	0.85
	FF5F	1.49	0.54	1.29	0.84	2.57	1.79	1.22	0.14
P3	HKM	6.25	0.08			55.58	1.49	10.42	0.72
	HXZQ	1.51	0.69	1.03	0.11	1.20	1.90	2.39	0.69
	Carhart+3	1.20	0.51	1.15	0.70	1.39	1.63	1.23	0.32
	PC5	1.11	0.44	1.07	0.41	1.04	0.78	1.03	0.78
	Median	1.50	0.53	1.16	0.70	1.31	1.53	1.65	0.59

(*Continued*)

four- and five-factor models we consider capture most of the economically interesting cross-sectional variation in returns, and additional clusters do not add much in terms of explained expected-return variation and increased ex post Sharpe ratios. For the placebo portfolio set, the AIC selects more than one cluster in only 2 of the 31 settings considered (results not tabulated). Even for those two cases,[33] we find improvements to maximal Sharpe ratios of only 0.03 and 0.08, respectively. Our large Sharpe ratio improvements for our nonplacebo portfolio sets do not arise mechanically.

Turning attention to the second and third panels, we see larger improvements in cross-sectional dispersion in model-implied average returns. The smallest increase in the international setting is 2%, and the largest increases are several hundred percent, again excluding the global CAPM and HKM models from consideration because of clearly omitted factors. As we will discuss in a detailed example in the next section, international markets have region-specific risk prices that make global factor models problematic. Intriguingly, by

[33] Namely, the CAPM applied to 1999–2016 data and the HKM model applied to 1981–1998 data.

Table 4
(Continued)

B. International equity portfolios

		1991–2016		1991–2003		2004–2016	
Portfolios	Model	Var($\bar{r}$)	ΔSR	Var($\bar{r}$)	ΔSR	Var($\bar{r}$)	ΔSR
	CAPM	7.20	0.55	11.10	0.23	1.34	0.06
	FF3F	5.00	0.51	2.91	0.60	1.25	0.13
	Carhart	5.61	1.31	3.17	0.74	1.07	0.25
	FF5F	1.33	0.54	1.41	1.00	1.11	0.92
P5	HKM	4.15	0.40	7.84	0.30	1.30	0.48
	HXZQ						
	Carhart+3	2.22	0.64	1.29	1.11	1.12	0.77
	PC5	1.37	0.89	1.02	0.35	3.12	0.90
	Median	4.15	0.55	2.91	0.60	1.25	0.48
	CAPM	3.98	0.89	368.25	0.66	1.21	0.71
	FF3F	3.06	1.13	1.75	0.44	1.46	0.93
	Carhart	4.07	1.62	1.14	1.05	1.37	0.75
	FF5F	2.10	1.27	2.43	1.45	1.07	0.70
P6	HKM	3.62	1.16	17.59	0.62	1.23	0.79
	HXZQ						
	Carhart+3	2.34	1.61	1.20	1.49	1.08	0.99
	PC5	1.10	0.99	1.10	1.38	1.27	0.69
	Median	3.06	1.16	1.75	1.05	1.23	0.75

C. Cross-asset class portfolios

		1986–2010		1986–1997		1998–2010	
Portfolios	Model	Var($\bar{r}$)	ΔSR	Var($\bar{r}$)	ΔSR	Var($\bar{r}$)	ΔSR
	CAPM	11.97	1.03	1.12	0.48	69.22	1.70
	FF3F	1.33	0.55	1.05	0.19	1.84	0.87
	Carhart	0.81	0.61	0.97	0.11	1.13	0.68
	FF5F	1.15	0.42	1.22	0.24	2.15	0.93
P7	HKM	7.48	0.79	1.03	0.33	19.69	1.40
	HXZQ	1.05	0.44	1.33	0.25	2.69	1.06
	Carhart+3	0.90	0.56	1.87	0.72	0.97	1.12
	PC5	1.54	1.04	1.34	0.72	1.74	1.16
	Median	1.24	0.59	1.17	0.29	2.00	1.09
	CAPM	4.93	1.11	1.23	1.00	5.95	0.68
	FF3F	1.27	0.80	1.03	0.48	1.34	1.26
	Carhart	1.08	0.87	1.10	0.46	2.48	1.80
	FF5F	1.23	0.95	1.18	0.39	2.80	1.93
P8	HKM	4.47	0.87	0.95	0.50	2.40	1.02
	HXZQ	1.21	0.90	1.28	0.47	1.56	1.41
	Carhart+3	1.18	1.47	1.84	1.40	1.08	1.37
	PC5	0.99	0.89	1.47	1.43	1.47	1.60
	Median	1.22	0.90	1.21	0.49	1.98	1.39

This table reports the ratio of cross-section variance in expected returns explained by multiple-cluster models to single-cluster models and the difference in maximal in-sample Sharpe ratios between multiple-cluster models and single-cluster models. We construct these measures as follows. First we estimate group assignments, risk prices, and likelihoods for models with $G = \{1, 2, 3, 4, 5\}$ groups using the entirety of each sample period. Next, we select the number of groups using the AIC. '*' indicate instances in which the AIC selects a single cluster. For the one-cluster and G^* cluster models, we then calculate the cross-sectional variance in average returns (var($\bar{r}$)), as well as the maximal in-sample Sharpe ratio attainable using the mimicking portfolios ($\sqrt{\mu_\Lambda' \Sigma_\Lambda^{-1} \mu_\Lambda}$). We tabulate the ratio of cross-sectional variances $\text{var}(\bar{r})^{(G^*)}/\text{var}(\bar{r})^{(1)}$ as "Var($\bar{r}$)" and increases in annualized Sharpe ratios $\sqrt{\mu_\Lambda' \Sigma_\Lambda^{-1} \mu_\Lambda} - \sqrt{\mu_\lambda' \Sigma_\lambda^{-1} \mu_\lambda}$ as "ΔSR." We repeat this procedure for all combinations of portfolio sets, risk models, and sample periods. Portfolios and models are described in the text. We omit results for the placebo portfolio set (P4) because multiple clusters are selected only twice of 27 models for these portfolios. For the He, Kelly, and Manela (2017) factor model, we use daily data for the most recent time period and monthly data for earlier time periods. We do not have sufficient coverage for their intermediary capital factor for 1963–1980 to include it in the domestic equity portfolio analysis. The q-factor (HXZQ) model is excluded from the international portfolio analysis because we do not have global return-on-equity factor data.

both metrics, we observe declines in inter-regional segmentation, suggesting that barriers to international arbitrage have decreased over time. Finally, in the cross-asset class context, we see comparable gains in performance for explaining the cross-section of average returns, even though with the exception of HKM we limit ourselves to models designed with only domestic equities in mind. However, in this panel we also observe occasional reductions in cross-sectional explanatory power when adding clusters to the Carhart and augmented Carhart models. This situation can occur if portfolios with relatively extreme betas belong to groups with low compensation for risk, as in Frazzini and Pedersen (2014), who find that high-market beta stocks have particularly low compensation for risk.

Perhaps of greater practical interest are increases in Sharpe ratios that can be attained by unconstrained arbitrageurs who use "local" versions of the factors. As with the dispersion in expected returns metric, using multiple clusters is as important for expanding the mean-variance frontier as using cutting-edge factor models. Returning to the domestic portfolios of the first panel of Table 4, typical improvements in annual Sharpe ratios are comparable to the market's Sharpe ratio of approximately 0.4, and some are as large as the maximal Carhart model Sharpe ratio of 1.3. Moving to the second and third panels, Sharpe ratio improvements of the multiple-cluster models are comparable with those of the first panel. Large improvements to attainable Sharpe ratios for international portfolios reinforce Asness, Moskowitz, and Pedersen's (2013) point that cross-region, multifactor strategies have highly desirable risk-return characteristics, even when compared with similar strategies within a single region. However, our implications contrast with theirs in that rather than investing in factor strategies everywhere, it may be even more profitable to trade the basis between factor strategies in different segments, for example, to go long North American momentum and short Japanese momentum.

Taken together, Tables 3 and 4 reveal segmentation everywhere. Cross-sectional differences in factor compensation represent an important new dimension of heterogeneity that is absent from standard asset pricing models. This dimension is as important as differences between risk models, even in the low-friction setting of U.S. equity markets, and especially so in more challenging international and cross-asset class settings. Of course, evidence for segmentation everywhere begs the question of from whence it comes.

3.3 Risk price heterogeneity: Detailed examples

In this section we investigate the dimensions of cross-sectional heterogeneity in risk prices in three representative settings: domestic portfolios with the Carhart four factors; international equity portfolios with the global Carhart factors; and cross-asset class portfolios with the market and He, Kelly, and Manela (2017) intermediary capital factors. The portfolios correspond to P3, P6, and P8, respectively, in Table 2.

Up to this point, our statistical tests and economic interpretation have focused on whether one or multiple sets of risk prices obtain in the data. Here, we are more interested in interpreting levels and cross-group differences in average cross-sectional slopes, and a different underlying theory is needed. Specifically, to interpret results from Fama-MacBeth tests for differences in average slopes from zero and from each other, we must account for the impact of estimated group memberships on the cross-sectional slopes.

Bonhomme and Manresa (2015) provide general conditions for problems related to ours under which the parameter estimates based on estimated group memberships have the same limiting distribution as the (infeasible) parameter estimates based on *true* group memberships. A critical condition for the asymptotic negligibility of the error in estimated group memberships is that the clusters are "well separated," that is, that there are indeed multiple clusters in the data. Table 3's strong rejection of a single cluster for these examples indicates that this separation condition is met. In our application, this equivalence of the limiting distributions implies that we can estimate the group memberships (γ) and Fama-MacBeth parameters (α and Λ) using the EM algorithm described in Section 1.1.1 and then conduct inference on the Fama-MacBeth parameters using standard methods (e.g., t- and F-tests on the time series of the estimated coefficients).

3.3.1 Domestic equity portfolios. Table 5 reports Fama-MacBeth regression estimates from one- and two-cluster models. The top panel reports p-values associated with tests of G clusters against one cluster (the null hypothesis). Tests of equality of average risk prices rejects the null against alternatives of $G=2$, $G=5$, and, more generally, $G>1$ clusters with Bonferroni-adjusted p-values.[34] Tests of equality of all cross-sectional slopes strongly reject the null for all $G \in \{2,3,4,5\}$. The top panel also reports the log likelihoods and AICs associated with G-cluster models in the full sample. We select two clusters using the AIC on this full sample.

The bottom panel presents standard Fama-MacBeth regression estimates in the leftmost column ("All"), as well as group-specific Fama-MacBeth estimates in columns for each group ("Group 1" and "Group 2"). ρ rows report the time-series correlation of factor-mimicking portfolio returns (cross-sectional slopes) and factor realizations, and the last column reports standard F tests in the style of Fama-MacBeth for equality of average risk prices across groups.[35]

---

[34] Intriguingly our test does not reject equality of means for $G=3$ and $G=4$ clusters. In these cases the optimal group assignments consist of clusters with smaller dispersion in average factor premiums, at least for the typical cluster pair. Maximizing (5) in the three- and four-cluster cases generates groups that instead differ more in their factor dynamics than in their average risk premiums.

[35] Namely, the F statistic is constructed as the sum of squared between-cluster differences in average λs normalized by the inverse covariance matrix of the time series of λs. This F statistic equals the square of the usual Fama-MacBeth t-statistic in the case of a single factor and pair of clusters.

Table 5
Domestic equity portfolios (P3) example: Domestic Carhart, 1963–2016

Sample	# clusters	1	2	3	4	5
$\mathcal{P}$	$p_F(\bar{\lambda}_k)$: 1 vs. G	–	**0.00**	0.49	0.49	0.00
	$p_F(\lambda_{kt})$: 1 vs. G	–	**0.00**	0.00	0.00	0.00
Full	LL ($\times 10^{-6}$)	6.44	**6.51**	6.53	6.54	6.55
	AIC ($\times 10^{-6}$)	−12.81	**−12.89**	−12.86	−12.82	−12.79

	One-cluster model	Two-cluster model		
	All	Group 1	Group 2	$p_F(\bar{\lambda}=)$
$\bar{\lambda}_{MKT}$	−1.13	−0.52	2.33	0.20
t-stat	(−0.44)	(−0.16)	(1.02)	
$\rho_{f,\lambda(g)}$	[0.83]	[0.75]	[0.77]	
$\bar{\lambda}_{HML}$	3.79	2.12	8.12	0.00
t-stat	(2.26)	(1.35)	(3.66)	
$\rho_{f,\lambda(g)}$	[0.95]	[0.92]	[0.82]	
$\bar{\lambda}_{SMB}$	1.60	2.21	−2.54	0.03
t-stat	(0.98)	(1.21)	(−0.86)	
$\rho_{f,\lambda(g)}$	[0.98]	[0.78]	[0.59]	
$\bar{\lambda}_{UMD}$	7.11	5.57	10.43	0.00
t-stat	(3.46)	(2.91)	(4.01)	
$\rho_{f,\lambda(g)}$	[0.99]	[0.96]	[0.92]	
ME 1-3	81	0	81	
ME 4-5	54	54	0	
Industry	49	44	5	
Other	50	50	0	
N_G	234	148	86	
T	13,469	13,469	13,469	

The top table reports p-values from F-tests described in Section 1.1.2 for comparing a multiple-clusters model to a single-cluster model. The underlying factor model is the Carhart four-factor model, and our sample period is 1963–2016. The bottom table reports full-sample Fama-MacBeth estimates of average cross-sectional slopes and associated t-statistics for each group in a model with one cluster ("All") and in a model with $G^* = 2$ clusters selected by the full-sample AIC. Standard errors are Newey-West with 252 daily lags. $\rho_{f,\lambda(g)}$ are the correlations of the factor return and the factor-mimicking portfolio return for cluster g. $p_F(\bar{\lambda}=)$ is the p-value associated with equality of factor means for the particular factor assuming fixed group memberships. Average cross-sectional R^2s are reported both within each cluster (R^2_G) and across clusters ($R^2_{Combined}$).

We also report average within- and across-cluster R^2s for the one- and G^*-cluster models.

The clusters have unequal sizes of 148 portfolios in group 1 and 86 portfolios in group 2. Comparing the columns for groups 1 and 2, group 2 has larger factor premiums for the market, value, and momentum factors, and a smaller factor premium for the size factor. Between-group differences in risk premiums for nonmarket factors are economically large, at about 5%–6%, and these differences are highly significant statistically as judged by F tests for equality of average premiums. Both clusters approximate the time-series factors for the market, value, and momentum reasonably well, with factor correlations ranging from 75% to 96% for these factors. Group 2's SMB-mimicking portfolio only achieves a correlation of 59% with the SMB factor; by comparison, Group 1's SMB-mimicking portfolio achieves a correlation of 78%.

From the portfolio counts for the two groups reported at the bottom of Table 5, we see that group 2 contains all 81 single- and double-sorted portfolios in which market equity is below the 60th percentile of NYSE stocks, as well as five industry portfolios. Group 1 contains all double-sorted portfolios in the largest and second-largest market-capitalization groups, as well as single-sorted market-capitalization decile portfolios 7–10 and 44 of the 49 industry portfolios. Because our portfolios are value-weighted, the single characteristic-sorted portfolios behave similarly to the highest market capitalization stocks, and they generate similar premiums to the high market capitalization portfolios of group 1. In short, the most important dimension of heterogeneity within our collection of U.S. stock portfolios is that of market capitalization, whereby small stocks earn greater risk premiums than large stocks on all but the size factor.

Our findings agree with Hong, Lim, and Stein (2000), Grinblatt and Moskowitz (2004), and Israel and Moskowitz (2013) in identifying market capitalization as an important determinant of cross-sectional differences in risk prices. Unlike these studies, we let the data inform us that size matters for anomaly compensation, and the split between the small and large size quintiles we identify is exactly that determined by previous studies. Notably, the existence of heterogeneity in risk prices is not sample dependent, as Israel and Moskowitz (2013) caution, although other time periods may see more or less heterogeneity in risk prices along the size dimension: we reject equal risk prices across all sets of domestic equity portfolios and across all time periods. Moreover, with the possible exception of the earliest sample period, the P3/Carhart group assignments are stable over time, as discussed in Appendix C; 86% of the group assignments agree between the first and second subsamples, and more than two-thirds of assignments in the first and third and second and third subsamples agree with one another.

While group 1 and group 2 portfolios differ in their risk premiums for all nonmarket factors, the momentum factor provides the strongest evidence of between-group segmentation. Both groups' mimicking portfolios nearly span the dynamics of the momentum factor, but the average compensation differential between portfolios is nearly 5% per year. Other factor-mimicking portfolios covary less across groups and have smaller compensation differentials per unit β, and hence going long one group and short the other has a worse risk-return trade-off.

The significant gap between large-cap and small-cap momentum compensation signals either high-Sharpe ratio opportunities ("good deals") or arbitrage frictions (a duality that permeates our paper). If arbitrageurs were aware of the differential return to momentum across market segments *and* could frictionlessly trade both portfolio sets or a portfolio set and the factor itself, these differences would be driven to zero. Given the considerable prior attention paid to cross-sectional differences in momentum premiums, a friction-based rationale seems more likely: either real-world market participants suffer large

implementation costs in replicating the momentum factor, as Novy-Marx and Velikov (2016) and Patton and Weller (2020) argue, or trading long-short momentum factor portfolios is especially costly in certain parts of the domestic equity universe, as Lesmond, Schill, and Zhou (2004) and others suggest.

In discussing these results, we should mention a caveat of our approach: more than half of our portfolios are sorted on the dimension of market capitalization, so our methodology has high resolution to detect differences in compensation among portfolios with different average market capitalizations. While we use an expansive portfolio set, these sets only include portfolios that have previously appeared in the literature, and so we are equipped only to detect heterogeneity in pricing along dimensions analyzed by other researchers. Hence, we can only conclude that market capitalization is the most important determinant of cross-sectional variation in risk prices among our stock portfolios. The study of other domestic equity test assets may find dimensions along which risk premiums vary even more strongly; we investigate one such dimension (liquidity) in Section 4.

3.3.2 International equity portfolios. Table 6 presents Fama-MacBeth regression estimates for the global Carhart model and 200 size-value and size-momentum-sorted portfolios. The top panel of Table 6 indicates that the AIC selects four clusters of risk prices as striking the best balance between model fit and parsimony. From the rightmost column, the greatest differences in risk prices are found for momentum. In particular, the fourth cluster has a small and statistically unreliable momentum premium, whereas the second and third clusters earn more than 12%/year per unit of momentum exposure. By contrast with the previous example, between-group factor dynamics are quite different among portfolios: for example, group 1's mimicking portfolio returns are between 62% and 94% correlated with the global factors, whereas group 2 and 4's mimicking portfolio returns are only 45%–63% correlated with these factors. Allowing for heterogeneous factor dynamics also contributes to a 17% improvement in average cross-sectional R^2s, suggesting that each cluster has strong within-cluster or local factor structures.

The estimated assignments tabulated in Table 6 clusters assets perfectly by geographic region: North America (group 1), Asia-Pacific excluding Japan (group 2), Europe (group 3), and Japan (group 4). None of the clusters include out-of-region portfolios. In the bottom-right plot, we see the well-documented failure of momentum in Japan (e.g., Rouwenhorst 1998; Griffin, Ji, and Martin 2003) set against the larger momentum premiums of the other three regions. Likewise, we see a greater value premium in Japan than in North America or Europe, consistent with Asness, Moskowitz and Pedersen's (2013) argument that a combined value-momentum strategy performs well across all major international regions.

This analysis inverts the standard asset pricing paradigm of selecting regions and comparing risk premiums or mean-variance efficient portfolios, as in

Table 6
International equity portfolios (P6) example: Global Carhart, 1991–2016

Sample	# clusters	1	2	3	4	5
$\mathcal{P}$	$p_F(\bar{\lambda}_k)$: 1 vs. G	–	0.00	0.00	**0.00**	0.00
	$p_F(\lambda_{kt})$: 1 vs. G	–	0.00	0.00	**0.00**	0.00
Full	LL ($\times 10^{-4}$)	5.50	6.00	6.20	**6.32**	6.33
	AIC ($\times 10^{-4}$)	−10.85	−11.72	−11.96	**−12.06**	−11.94

	One-cluster model	Four-cluster model				
	All	Group 1	Group 2	Group 3	Group 4	$p_F(\bar{\lambda}=)$
$\bar{\lambda}_{MKT}$	4.24	−1.57	−12.06	−0.55	2.03	0.12
t-stat	(0.73)	(−0.33)	(−2.38)	(−0.14)	(0.26)	
$\rho_{f,\lambda}(g)$	[0.47]	[0.62]	[0.53]	[0.66]	[0.46]	
$\bar{\lambda}_{HML}$	1.07	2.86	9.09	4.38	6.48	0.11
t-stat	(0.42)	(1.22)	(2.36)	(1.35)	(2.22)	
$\rho_{f,\lambda}(g)$	[0.76]	[0.89]	[0.55]	[0.85]	[0.60]	
$\bar{\lambda}_{SMB}$	0.05	2.82	−0.55	0.73	2.93	0.61
t-stat	(0.02)	(1.64)	(−0.16)	(0.37)	(1.02)	
$\rho_{f,\lambda}(g)$	[0.83]	[0.78]	[0.44]	[0.73]	[0.51]	
$\bar{\lambda}_{UMD}$	8.07	5.79	18.69	12.16	4.00	0.00
t-stat	(2.79)	(1.96)	(3.59)	(3.70)	(0.93)	
$\rho_{f,\lambda}(g)$	[0.98]	[0.94]	[0.60]	[0.87]	[0.63]	
NA	50	50	0	0	0	
AP	50	0	50	0	0	
EU	50	0	0	50	0	
JP	50	0	0	0	50	
N_G	200	50	50	50	50	
T	312	312	312	312	312	

The top table reports p-values from F-tests described in Section 1.1.2 for comparing a multiple-clusters model to a single-cluster model. The underlying factor model is the Carhart four-factor model with global factors described in Fama and French (2012), and our sample period is 1991–2016. The bottom table reports full-sample Fama-MacBeth estimates of average cross-sectional slopes and associated t-statistics for each group in a model with one cluster ("All") and in a model with $G^*=4$ clusters selected by the full-sample AIC. Standard errors are Newey-West with 12 monthly lags. $\rho_{f,\lambda}(g)$ are the correlations of the factor return and the factor-mimicking portfolio return for cluster g. $p_F(\bar{\lambda}=)$ is the p-value associated with equality of factor means for the particular factor assuming fixed group memberships. Average cross-sectional R^2s are reported both within each cluster (R^2_G) and across clusters ($R^2_{Combined}$).

Griffin (2002), Hou, Karolyi, and Kho (2011), and Fama and French (2012). Instead, we confirm that the cross-section of portfolio returns itself suffices to identify the region to which assets belong. That our estimated group boundaries coincide with geographic ones, taken by others previously as given because of institutional barriers to arbitrage, serves as reassurance that our methodology can detect important sources of segmentation, and encourages us to apply it in settings for which the critical dimensions of segmentation are not known ex ante.

3.3.3 Cross-asset class portfolios. Table 7 reports results from our clustering methodology applied to the He, Kelly, and Manela (2017) market-intermediary capital factor model with 148 cross-asset class portfolios (P8). In conducting this exercise we use the authors' monthly data and a similar set of test assets.

Table 7
Cross-asset class portfolios (P8) example: HKM factors, 1986–2010

Sample	# clusters	1	2	3	4	5
$\mathcal{P}$	$p_F(\bar{\lambda}_k)$: 1 vs. G	–	0.00	0.00	0.00	**0.00**
	$p_F(\lambda_{kt})$: 1 vs. G	–	0.00	0.00	0.00	**0.000**
Full	LL ($\times 10^{-4}$)	5.19	5.46	5.55	5.61	**5.66**
	AIC ($\times 10^{-4}$)	−10.30	−10.75	−10.83	−10.88	**−10.89**

	One-cluster model	Five-cluster model					
	All	Grp 1	Grp 2	Grp 3	Grp 4	Grp 5	$p_F(\bar{\lambda}=)$
$\bar{\lambda}_{MKT}$	7.14	45.85	10.18	10.57	−2.39	7.33	0.00
t-stat	(2.22)	(4.77)	(1.30)	(2.53)	(−0.48)	(2.10)	
$\rho_{f,\lambda(g)}$	[0.98]	[0.33]	[0.55]	[0.75]	[0.66]	[0.98]	
$\bar{\lambda}_{HKM}$	9.30	−48.38	22.84	14.43	−8.47	9.91	0.06
t-stat	(1.18)	(−1.34)	(1.75)	(1.15)	(−0.87)	(1.14)	
$\rho_{f,\lambda(g)}$	[0.62]	[0.16]	[0.36]	[0.45]	[0.51]	[0.56]	
Options	18	18	0	0	0	0	
Commod.	23	0	14	5	0	4	
U.S. Bonds	20	0	16	0	0	4	
FX	12	0	0	11	0	1	
Stocks	75	2	0	16	46	11	
N_G	148	20	30	32	46	20	
T	300	300	300	300	300	300	

The top table reports p-values from F-tests described in Section 1.1.2 for comparing a multiple-clusters model to a single-cluster model. The underlying factor model is the He-Kelly-Manela two-factor model with the intermediary capital factor from He, Kelly, and Manela (2017), (HKM), and our sample period is 1986–2010. The bottom table reports full-sample Fama-MacBeth estimates of average cross-sectional slopes and associated t-statistics for each group in a model with one cluster ("All") and in a model with $G^*=5$ clusters selected by the full-sample AIC. Standard errors are Newey-West with 12 monthly lags. $\rho_{f,\lambda(g)}$ are the correlations of the factor return and the factor-mimicking portfolio return for cluster g. $p_F(\bar{\lambda}=)$ is the p-value associated with equality of factor means for the particular factor assuming fixed group memberships. Average cross-sectional R^2s are reported both within each cluster (R^2_G) and across clusters ($R^2_{Combined}$).

However, rather than finding support for unified factor pricing, we instead find (at least) five clusters in the data, as selected by the AIC. We stop at five clusters to reduce the chance of an estimated cluster having a small number of members and poorly estimated risk prices.

Estimates of the equity premium range from −2.4% per year in cluster group 4 to 45.9% per year in cluster group 1, with a full-sample average equity premium of 7.14% per year. The intermediary capital factor price varies from −48.4% per year in group 1 to 22.8% per year in group 2. This large variation is not due to the clusters being too small to estimate λs well—our smallest group has 20 portfolios, and our most extreme risk prices come from clusters of 20 and 30 portfolios (by comparison five of the eight asset classes analyzed individually in He, Kelly, and Manela (2017) have 20 or fewer portfolios). An test for equality of average risk premiums across clusters rejects equality for both factors with p-values of 0.00 and 0.06.

Table 7 also reveals important differences in the dynamics of risk premiums across clusters missed by the test of differences in average risk prices. While the market factor is highly correlated with market factor-mimicking portfolios

for each cluster other than the first (similar to the international equities portfolio example), the intermediary capital factor looks very different from its mimicking portfolios. Three of the five clusters' local variants are less than 50% correlated with the global factor, and the average correlation is only 41%. This finding reinforces the main point of Haddad and Muir (2021), who discover important cross-asset class differences in the time variation in risk premiums as a function of barriers to direct participation by households (equivalently in their model, the degree of intermediation).

The estimated cluster assignments, tabulated in Table 7, reveal that group 1 consists of all U.S. options along with two stock portfolios; group 2 is evenly split between commodities and U.S. bonds; group 3 includes all but one foreign exchange portfolio, along with 16 small-cap stock portfolios, and five commodities; group 4 exclusively comprises U.S. stock portfolios; and group 5 has a mix of large-cap stock portfolios, commodity, U.S. bond, and foreign exchange portfolios. Some of these splits might be readily conjectured ex ante, for example, options and stocks look different from bonds and commodities, but some of these splits are not, for example, commodities and U.S. bonds are priced similarly. Our approach is uniquely positioned to find such unconventional partitions of the data by risk prices. That pricing approximately segments by asset class suggests that either the HKM measure does not capture intermediaries' pricing kernel, or that intermediary asset pricing fails to unify risk prices.

4. Man versus Machine: Ex Ante versus Estimated Clusters

This section presents an investigation of the performance of our data-driven method for finding clusters in comparison with clusters that are well-motived ex ante based on economic arguments. We focus on the U.S. equity returns discussed in Section 3.3.3, but unlike the previous analyses we construct new portfolios of individual stocks using firm characteristics generally thought to be related to market segmentation, rather than taking existing sort portfolios "off the shelf."[36] The first two characteristics relate to the institutional ownership, as used in Gompers and Metrick (2001), Chen, Hong, and Stein (2002), and Lewellen (2011), among others. We consider the institutional ownership ratio (IOR), which is the proportion of a firm's outstanding equity held by institutional investors, and the institutional ownership concentration (IOC), which uses the Herfindahl-Hirschman index to measure how dispersely (or not) the firm's equity is held. We obtain these measures from Thomson-Reuters 13F data available on WRDS.[37] We also consider two measures of liquidity based on

[36] We thank a referee for suggesting this analysis.

[37] Specifically, we use standardized WRDS code for institutional ownership concentration and breadth ratios provided at https://wrds-www.wharton.upenn.edu/pages/support/applications/institutional-ownership-research/institutional-ownership-concentration-and-breadth-ratios/, updating the end date on line 17 through 2016.

high-frequency data: the average effective quoted spread and price impact (both in percent); see, for example, Holden and Jacobsen (2014). We obtain these data from the Daily Intraday Indicators file on WRDS (`ESpreadDollar_VWi` and `TSignSqrtDVol1`, respectively). For this analysis, our sample period is 1993–2016 because intraday liquidity measures draw on the Trades and Quotes database, which starts in 1993.

Using the IOR and quoted spreads measures, we simply form decile portfolios, sorting on the median nonmissing value for each PERMNO for the year preceding the end of June (following Fama and French's convention). For the IOC measure we create a "portfolio 11" containing all firms with IOC exactly equal to one, and we form decile portfolios from the remaining firms. For the price impact measure we create a "portfolio 0" containing all firms with apparently negative price impact, which can happen due to sampling variability, and we form decile portfolios on the remaining firms. We reverse the order of the IOR portfolios so that all four measures go from high liquidity to lower liquidity. In total we have 42 one-way sort portfolios. Next we construct two-way sort portfolios, where we first sort stocks by market capitalization quintile, and then form quintile portfolios by liquidity measure (with the additional IOC-11 and PriceImpact-0 portfolios), leading to a total of 110 two-way sort portfolios. Thus in total we have 152 portfolios with an ex ante clear ordering: portfolios with higher illiquidity are harder to short sell and may exhibit risk prices that differ from assets that are easy to short sell.

Analogous to Table 5 of the previous section, we consider two clusters and allocate all portfolios with liquidity in the upper 60% of the distribution to the "high liquidity" group and the remainder to the "low liquidity" group. We use the Carhart four-factor model and report parameter estimates in Table 8. Consistent with work in the extant literature, we find evidence of variation

Table 8
Domestic equity liquidity sort portfolios: Ex ante sorts, 1993–2016

	One-cluster model	Two-cluster model		
	All	High Liq	Low Liq	$p_F(\bar{\lambda}=)$
$\bar{\lambda}_{MKT}$	13.53	11.16	22.88	0.00
t-stat	(4.12)	(2.76)	(2.93)	
$\bar{\lambda}_{HML}$	2.33	0.46	4.31	0.00
t-stat	(1.19)	(0.21)	(1.44)	
$\bar{\lambda}_{SMB}$	1.24	2.40	−1.18	0.20
t-stat	(0.55)	(1.16)	(−0.41)	
$\bar{\lambda}_{UMD}$	12.02	10.99	26.59	0.00
t-stat	(3.45)	(1.85)	(3.60)	
N_G	152	91	61	
T	4,657	4,657	4,657	

The underlying factor model is the Carhart four-factor model, and our sample period is 1993–2016. This table reports full-sample Fama-MacBeth estimates of average cross-sectional slopes and associated t-statistics for each group in a model with one cluster ("All") and in a model with two clusters based on whether the portfolio is in the upper 60% of the distribution ("high liquidity") or the lower 40% ("low liquidity"). Standard errors are Newey-West with 252 daily lags. $p_F(\bar{\lambda}=)$ is the p-value associated with a standard F-test of the equality of factor means for the particular factor.

Table 9
Domestic equity liquidity sort portfolios: Testing for segmentation, 1993–2016

		G				
N		2	3	4	5	Bonf.
152	$p_F(\bar{\lambda}_k)$: 1 vs. G	0.00	0.00	0.03	0.06	0.00
	$p_F(\lambda_{kt})$: 1 vs. G	0.00	0.00	0.00	0.00	0.00
42	$p_F(\bar{\lambda}_k)$: 1 vs. G	0.37	0.13	0.57	0.80	0.54
	$p_F(\lambda_{kt})$: 1 vs. G	0.00	0.29	0.33	0.21	0.01

This table reports *p*-values from *F*-tests described in Section 1.1.2 for comparing a multiple-clusters model to a single-cluster model. The underlying factor model is the Carhart four-factor model with global factors described in Fama and French (2012), and our sample period is 1993–2016. The top two rows report test results for 42 one-way liquidity-sported portfolios and 90 two-way sort portfolios, where stocks are first sorted on market capitalization and then on liquidity. The bottom two rows report test results for just the 42 one-way sort portfolios. The columns 2, 3, 4, and 5 report the results of a test of a single cluster against that specific alternative; the last column reports the Bonferroni-adjusted *p*-value for the joint test.

in risk prices across the two groups: risk premiums on the market, value and momentum are all significantly larger for the "low liquidity" group of assets, with *p*-values on the differences from the "high liquidity" group all less than 0.005. A joint test that each of four factors has equal risk premiums across the two groups is rejected with a *p*-value of 0.003.

Next we apply our data-driven method for detecting clusters to this set of test assets. We firstly conduct a test for multiple clusters, allowing the number of clusters under the alternative, G, to range from 2 to 5. The *p*-values from each test, as well as the Bonferroni-adjusted *p*-value for the joint test, are presented in the top two rows of Table 9. Consistent with the evidence from the ex ante grouping of assets into high and low liquidity groups, we reject the null of a single cluster for all values of G (using the lower-powered mean test, the *p*-value is .06 when $G=5$; it is .03 or less for all other values of G). The joint test strongly rejects the null of a single cluster.

To compare the optimal estimated clusters with the ex ante clusters, we focus on the $G=2$ case and examine the estimated cluster assignments. These assignments cleave quite cleanly: group 1 contains all of the one-way sorts, except IOC-11, and all of the two-ways sorts that involve the largest quintile of market capitalization, a total of 62 test assets. Group 2 contains all of the two-way sort portfolios involving market capitalizations in the bottom 80% of the distribution, and IOC-11, a total of 90 test assets. Recalling that the one-way sort portfolios are value-weighted, and so each of them alone is a "large cap"-like portfolio, the two clusters can be reasonably labeled "Large cap" and "Small cap."

Taken together, the findings above reveal that U.S. equity returns are segmented. If one forms groups using ex ante information on asset liquidity, strong evidence is found for market segmentation, consistent with past literature. If one estimates the optimal groups from the data, segmentation is also strongly significant, and the leading source of segmentation appears to be market capitalization. This finding, on a completely different set of test assets

(although constructed, naturally, from the same universe of U.S. equities), is consistent with the results reported in Section 3.3.3.

Note that any data-driven method for finding segmentation is affected by the set of test assets under consideration, and the set considered here included two-way portfolios sorted on market capitalization, possibly increasing the chances of finding market cap as a source of segmentation. To explore this idea further, we consider a test of market segmentation using a subset of the test assets above, namely the 42 one-way sort portfolios. The lower two rows of Table 9 presents the p-values from this test. The lower-power mean test finds no evidence of segmentation, while the dynamic test finds strong evidence when $G=2$ (p-value of less than 0.01), and the Bonferroni-adjusted joint test also has a p-value of less than 0.01, strongly rejecting the null of no segmentation. Thus, even this small set of portfolios exhibits significant differences in risk prices.

The estimated group assignments for this smaller set of test assets reveal a result that is reminiscent of the ex ante groups: group 1 contains 23 test assets, and nearly 75% of them are portfolios in the upper 60% of liquidity; group 2 contains the remaining 19 test assets, and is dominated by the low liquidity portfolios. Thus the two estimated clusters can reasonably be labeled "high liquidity" and "low liquidity," revealing that when we eliminate market capitalization as a possible source of heterogeneity, liquidity emerges as a driver of risk price heterogeneity.

5. Omitted Factors or Fundamental Heterogeneity?

5.1 Clusters as factors and factors as clusters

In this section we consider the possibility of omitted factors manifesting as differences in risk prices and vice versa.[38] To start, consider two simple models:

$$\text{Model 1: } r_{it}=\alpha_t^{(1)}\mathbf{1}(\gamma_i=1)+\alpha_t^{(2)}\mathbf{1}(\gamma_i=2)+\epsilon_{it}, \tag{15}$$

$$\text{Model 2: } r_{it}=\tilde{\alpha}_t+\beta_i\eta_t+\tilde{\epsilon}_{it}. \tag{16}$$

The first model consists of two clusters, with cluster memberships determined by $\gamma_i\in\{1,2\}$, each with time-varying average returns, $\alpha_t^{(g)}$, within the cluster. The second model consists of a single cluster with time-varying average returns and factor realizations as well as heterogeneous risk exposures β_i. Note that β_i can be a time series beta or a characteristic. To complete our notation, let N_g be the number of assets in cluster g, and define $\Delta\alpha_t\equiv\alpha_t^{(1)}-\alpha_t^{(2)}$.

Firstly, suppose that the true data-generating process (DGP) is (15), but we instead estimate (16). To simplify this case, assume $\epsilon_{it}\perp\beta_i$ at each date. Within

[38] An in-depth analysis of clusters as a source of the proliferation of factors in the cross-section of expected returns is the subject of ongoing research.

each cross-section, the OLS estimate for η_t is

$$\hat{\eta}_t = \frac{cov(r_{it},\beta_i)}{var(\beta_i)} = \frac{\left(\alpha_t^{(1)} - \alpha_t^{(2)}\right) cov(\mathbf{1}(\gamma_i = 1),\beta_i) + cov(\epsilon_{it},\beta_i)}{var(\beta_i)},$$

$$= \Delta\alpha_t \frac{var(\mathbf{1}(\gamma_i = 1))}{var(\beta_i)} \left(E\left[\beta_i \,|\, \gamma_i = 1\right] - E\left[\beta_i \,|\, \gamma_i = 2\right]\right). \tag{17}$$

The time series average of $\hat{\eta}_t$ replaces $\Delta\alpha_t$ with $\Delta\bar{\alpha}$. If average returns in each cluster are different, that is, $\bar{\alpha}^{(1)} \neq \bar{\alpha}^{(2)}$, then $\bar{\eta} \neq 0$, and *any* characteristic that varies on average across groups will appear to be priced. This expression is readily extended to a multivariate context, and it extends to more than two clusters by replacing (17) with a $G-1$ set of indicators corresponding to membership in each cluster other than the last one.

Next, suppose that the true DGP is (16), but we instead estimate (15). We assume $E[\tilde{\epsilon}_{it} | i \in G_1] = 0$ to keep the exposition as simple as possible. Within each cross-section, the OLS estimate for $\Delta\alpha_t$ is

$$\widehat{\Delta\alpha_t} = \frac{cov(r_{it},\mathbf{1}(\gamma_i = 1))}{var(\mathbf{1}(\gamma_i = 1))} = \frac{cov(\beta_i\eta_t,\mathbf{1}(\gamma_i = 1)) + cov(\tilde{\epsilon}_{it},\mathbf{1}(\gamma_i = 1))}{var(\mathbf{1}(\gamma_i = 1))}$$

$$= \eta_t (E[\beta_i | i \in G_1] - E[\beta_i | i \in G_2]). \tag{18}$$

Hence, so long as (a) the factor exposure or characteristic β_i has different conditional means across clusters, and (b) the factor realization is not precisely zero, the difference in cross-sectional means $\widehat{\Delta\alpha_t}$ is also nonzero. Moreover, if the factor is priced with $\bar{\eta} \neq 0$, then the difference in average returns across clusters, $\overline{\Delta\alpha}$, will appear to be nonzero, and we will spuriously identify a second cluster.[39] By parallel with the "clusters as factors" case, this expression can also be extended to a multifactor model with one or more omitted factors. We find segmentation in every portfolio set regardless of the choice of factor model considered in Section 3; however, we cannot rule out that these models miss important variation in expected returns that is driven by undiscovered or rarely included factors.

5.2 Economic restrictions of a segmented-markets model

To link the stylized example in the previous subsection to our empirical analysis below, we now show that an asset pricing model with K factors and G clusters

[39] Note that our permutation-based tests are designed to correct for exactly this feature, answering the question "what is the probability of finding differences in risk prices at least this large by chance?" For the CAPM and HKM models, the answer is sometimes – random reshufflings of group assignments sometimes generate apparent differences in risk prices, and this means that the difference implied by the optimal clustering model must be even larger in order to reject the null of integrated markets. Consistent with this, for the U.S. equity test assets (portfolio sets P1-P3), evidence of segmentation is generally the weakest when using the CAPM and HKM models. For the collections of international equity returns (P5-P6) and multiasset class returns (P7-P8) we find evidence of segmentation even using these two models.

is a special case of a larger, single-cluster, asset pricing model with GK factors. To show this equivalence, we rewrite (1) as

$$r_{it} = \alpha_t^{(1)} + \sum_k \beta_{ik}\left(f_{kt} + \phi_{kt}^{(1)}\right) + \epsilon_{it} + \sum_{g=2}^{G}\left[\left(\alpha_t^{(g)} - \alpha_t^{(1)}\right) + \sum_k \beta_{ik}\left(\phi_{kt}^{(g)} - \phi_{kt}^{(1)}\right)\right] 1(\gamma_i = g), \tag{19}$$

The first line of (19) is a standard realized-return model for assets in an integrated market (the presence of $\phi_{kt}^{(1)}$ allows for the realized risk premiums to differ from the observable factor), with the first cluster serving as the reference cluster. When the assets are homogeneous and all clusters are identical, the second line of (19) vanishes. Market segmentation adds group-specific zero-beta rates $\alpha_t^{(g)}$ as well as group-specific factor disturbances $\phi_t^{(g)}$.[40] The coefficients on the group-specific factors take one of two values. For assets in segment g, the loadings on group-specific factors, $\phi_{kt}^{(g)}$, are the same as those on the corresponding global factors (β_{ik}). For assets in other segments, the loadings on $\phi_{kt}^{(g)}$ are zero. Thus, $\phi_{kt}^{(g)}$ can be interpreted as a "local" factor with a specific pattern in its loadings.

The factor-mimicking portfolio interpretation of Fama-MacBeth cross-sectional slopes helps to clarify differences between segmented-market and extended-factor models. Equation (6) delivers factor-mimicking portfolio returns group by group, that is, the cross-sectional slopes are each segment's approximation of the global factor return given the assets only in that segment. In the international example of Section 3.3.3, these factor-mimicking portfolios are the approximations to global value, momentum, etc., using each region's size-value and size-momentum portfolios. Imposing the redundant-or-zero structure on betas in (19) maintains this feature. By contrast, in global models with unrestricted betas on all factors, the cross-sectional slopes take on a different interpretation. The second-stage slopes on β_{ik} represent the mimicking portfolio return using all assets, regardless of market segment, and zeroing out the local components.

5.3 Empirically distinguishing between factors and clusters

As the preceding examples make clear, it is challenging to distinguish between omitted factors and multiple clusters without imposing structure on what omitted factors might look like and how numerous they might be. Further, as shown in the previous section, factor models of arbitrary length nest cluster-based models as a special case. For this reason, we compare omitted-factor and cluster-based models of comparable size (defined in a variety of ways).

[40] The focus of our analysis is on factor premiums, so we assume for discussion that zero-beta rates are the same across groups.

Rivers and Vuong (2002) provide our framework for model comparison. Specifically, for two non-nested models that minimize in-sample (weighted) squared errors (e.g., estimated by linear regression), under weak conditions the models can be easily compared using their (weighted) cross-sectional mean squared errors (MSE_t), date by date. Under the null that the models M_1 and M_2 are equally accurate, we have

$$\frac{1}{\sqrt{T}}\sum_t \left(MSE_t^{(M_1)} - MSE_t^{(M_2)}\right) \sim N(0, V), \tag{20}$$

where V is the asymptotic variance of the difference in MSEs, which is computed using a HAC estimator, for example, Newey and West (1987). Hence, given a particular choice of cluster and factor model, we can use a simple t test on the difference in (weighted) MSEs to evaluate the null of equal fit against the alternative of unequal fit. Importantly, this test does not require either model to be correct. It may be that neither a multiple-cluster model nor an extended-factor model captures the richness of cross-sectional variation in returns, but the test reveals which approximation better describes the data. As in our estimation, we use idiosyncratic variance as weights in the MSE.

The challenge is how to choose the cluster and factor models to compare. For this purpose, we retain the $\mathcal{R}$ and $\mathcal{P}$ partitions described in Section 1.1.2. We first use the in-sample AIC on the $\mathcal{R}$ subsample to select the "best" cluster model and fix group assignments. We then estimate this model on the $\mathcal{P}$ subsample, using the group assignments from the $\mathcal{R}$ sample. Next, we retain the time-series regression estimates of factor betas and residual variation, and we extract additional factors from these residuals on $\mathcal{R}$ using principal components analysis. These additional factors maximize explanatory power of the variation in returns not spanned by the included factors, and they serve as our candidate omitted factors; in this respect our methodology mimics Giglio and Xiu's (2021) approach for extracting omitted factors that may bias second-stage estimates of risk premiums.[41] We retain factor loadings (βs) estimated on the $\mathcal{R}$ subsample to fix factor identities and employ them in the cross-sectional regressions on the $\mathcal{P}$ subsample.

Because we wish to speak to a large range of possible omitted-factor models, we consider three choices for the "best" omitted-factor model rather than a single one. We enumerate these models on the basis on the number of additional included principal components, K_1^*, K_2^*, or K_3^*. In the K_1^* models, we consider factor models augmented with the first three principal components of the panel of residual returns on the $\mathcal{R}$ partition. Our choice of three PCs is *ad hoc*

[41] As noted previously, the interpretation of Fama-MacBeth slopes as factor-mimicking portfolio returns requires that a given factor model be complete. Throughout our main analysis, we assume that posited factor models are correct up to the number of market segments. Here, we explicitly relax this assumption and extract additional factors from the residuals. In so doing, estimated cross-sectional slopes are cleaned of omitted-variable bias, and the mimicking-portfolio interpretation is restored.

Table 10
Comparison of multiple-cluster and omitted-factor models

A. Domestic equity portfolios

		1963–2016			1963–1980			1981–1998			1999–2016		
Portfolios	Model	K_1^*	K_2^*	K_3^*	K_1^*	K_2^*	K_3^*	K_1^*	K_2^*	K_3^*	K_1^*	K_2^*	K_3^*
	CAPM	−−	0	+++	−−−	−−−	−−−	−	+++	+++	−−	0	−−
	FF3F	−−−	0	0	0	+++	+++	0	+++	+++	−−	0	−−
	Carhart	+++	+++	+++	*	*	*	+++	+++	+++	++	+++	+++
P1	FF5F	*	*	*	*	*	*	+++	+++	+++	−	+++	0
	HKM	−−−	0	0				−−−	0	0	++	+++	+++
	HXZQ	+++	+++	+++	*	*	*	*	*	*	*	*	*
	Carhart+3	*	*	*	*	*	*	*	*	*	*	*	*
	CAPM	−−−	−−−	0	−−−	−−−	−	−−−	−−−	−−−	−−	−−	−−−
	FF3F	−−−	0	−−−	0	+++	+++	0	+++	+++	0	+++	−−
	Carhart	++	+++	+++	+++	+++	+++	0	+++	0	−−−	+++	−−−
P2	FF5F	−	+++	+++	+++	+++	+++	0	+++	0	−−	+++	0
	HKM	−−−	0	0				0	+++	+++	0	++	−−−
	HXZQ	0	+++	+++	+++	+++	+++	+++	+++	+++	0	+++	0
	Carhart+3	++	+++	+++	+++	+++	+++	0	+++	0	+++	+++	+++
	CAPM	+++	+++	+++	+++	+++	−−−	+++	+++	+++	0	−−	0
	FF3F	+++	+++	+++	+++	+++	+++	+++	+++	+++	0	+++	−−−
	Carhart	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	−−
P3	FF5F	+++	+++	0	+++	+++	+++	+++	+++	+++	0	+++	−−
	HKM	0	+++	+++				+++	+++	+++	+++	+++	−−−
	HXZQ	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	0
	Carhart+3	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++

(*Continued*)

and intended only to provide a uniform omitted-factor benchmark. In the K_2^* models, we consider G^*-1 additional factors, where G^* is the number of clusters selected by the AIC on $\mathcal{R}$. We make this choice of the number of additional factors to give both cluster and factor models similar flexibility in fitting the data. A model with G^* clusters adds G^*-1 partitions of the data on which to fit K-factor models, and a model with a single cluster and $K+K_2^*$ factors also has an G^*-1 additional degrees of freedom to fit each cross-section. The K_3^* models give maximum flexibility to extended-factor models to fit the data; just as we choose the number of clusters using the in-sample AIC, so too do we let the AIC on the $\mathcal{R}$ partition dictate the number of factors. We impose one restriction to make the selection process comparable, that is, much as we do not allow the AIC for the cluster models to select $G^*>5$, we do not let the number of additional factors selected exceed $(G^*-1)(K+1)-1$. This upper bound is one less than the number of additional variables that allow a factor model to perfectly replicate a cluster model. In settings with large N and large T, factor models with so much additional flexibility cannot "lose" to cluster models, making the model comparison uninformative and violating the non-nestedness condition of Rivers and Vuong (2002).

Equipped with models chosen on the $\mathcal{R}$ partition, Table 10 reports discretized t statistics for the hypothesis of equal mean squared errors on $\mathcal{P}$ for factor models augmented with additional clusters and our three choices of additional factors. We code p-values below .1, .05, and .01 with one, two, or three – or +,

Table 10
(Continued)

B. International equity portfolios

		1991–2016			1991–2003			2004–2016		
Portfolios	Model	K_1^*	K_2^*	K_3^*	K_1^*	K_2^*	K_3^*	K_1^*	K_2^*	K_3^*
	CAPM	+++	+++	−−	0	+++	0	+++	+++	0
	FF3F	+++	+++	+++	0	+++	0	0	+++	0
	Carhart	+++	+++	+++	0	+++	0	0	+++	0
P5	FF5F	−−−	+++	−−−	+++	+++	+++	0	+++	0
	HKM	+++	+++	+++	++	+++	−−	+++	+++	−
	HXZQ									
	Carhart+3	+++	+++	+++	+++	+++	+++	+++	+++	+++
	CAPM	+++	+	−−−	+++	++	0	+++	+++	0
	FF3F	+++	+++	+++	0	+++	0	+++	+++	0
	Carhart	+++	+++	+++	+	+++	−−	+++	+++	+++
P6	FF5F	+++	+++	+++	+++	+++	0	+++	+++	−−
	HKM	+++	+++	−−−	+++	+++	0	+++	+++	0
	HXZQ									
	Carhart+3	+++	+++	+++	+++	+++	0	+++	+++	0

C. Cross-asset class portfolios

		1986–2010			1986–1997			1998–2010		
Portfolios	Model	K_1^*	K_2^*	K_3^*	K_1^*	K_2^*	K_3^*	K_1^*	K_2^*	K_3^*
	CAPM	+++	+++	+++	++	++	++	+++	+++	+++
	FF3F	+++	+++	+++	++	++	++	+++	+++	+++
	Carhart	+++	+++	+++	+	+	+	+++	+++	+++
P7	FF5F	+++	+++	+++	+	++	0	+++	+++	++
	HKM	+++	+++	+++	++	++	++	+++	+++	+++
	HXZQ	+++	+++	+++	+	++	+	+++	+++	+++
	Carhart+3	+++	+++	+++	++	++	++	+++	+++	+++
	CAPM	+++	+++	++	++	+	0	+++	+++	0
	FF3F	+++	+++	0	+	++	0	+++	+++	+++
	Carhart	+++	+++	+++	++	++	0	+++	+++	+++
P8	FF5F	+++	+++	+++	++	++	+	+++	+++	0
	HKM	+++	+++	+++	++	++	+	+++	+++	+++
	HXZQ	+++	+++	+++	++	++	+	+++	+++	+++
	Carhart+3	+++	+++	+++	+++	+++	+++	+++	+++	+++

Table reports discretized t-statistics for the average time-series differences in mean-squared errors between omitted-factor and multiple-cluster models. We code p-values below .1, .05, and .01 with one, two, or three – or +, respectively. Positive (negative) values indicate support for the multiple-cluster (omitted-factor) models. To estimate these quantities, we first estimate group assignments, risk prices, and likelihoods for models with $G=1,\ldots,5$ groups on the $\mathcal{R}$ partition (the first half of the sample). We then select the number of groups G^* using the AIC on the $\mathcal{R}$ sample. '*' indicate instances in which the AIC selects a single cluster. We compare this multiple-cluster model with extended-factor models with additional factors extracted using principal components on the $N \times R$ panel of time-series residuals from the conjectured factor model on the $\mathcal{R}$ partition. These models augment the conjectured model with $K_1^*=3$, $K_2^*=G^*-1$, and $K_3^*=\min\left(\text{argmin}_K\, AIC_K, \left(G^*-1\right)(K+1)-1\right)$ principal components. By parallel with the selection of the number of clusters, we fix the extracted principal component loadings and select models using the AIC on the $\mathcal{R}$ partition. t-statistics compare models across dates in the $\mathcal{P}$ partition, and standard errors are Newey-West with 252 daily or 12 monthly lags. We then repeat this procedure for all combinations of portfolio sets, risk models, and sample periods. Portfolios and models are described in the text. We omit results for the placebo portfolio set (P4) because multiple clusters are selected only twice of 27 models for these portfolios (as in Table 4). For the He, Kelly, and Manela (2017) factor model, we use daily data for the most recent time period and monthly data for earlier time periods. We do not have sufficient coverage for their intermediary capital factor for 1963–1980 to include it in the domestic equity portfolio analysis. The q-factor (HXZQ) model is excluded from the international portfolio analysis because we do not have global return-on-equity factor data.

respectively. Positive entries signify that the cluster model performs better, and negative entries signify that the extended-factor model performs better. We use Newey-West standard errors with 252 daily or 12 monthly lags to allow for serial dependence in squared errors.

Focusing first on the domestic equity portfolios, we obtain mixed results on the importance of clusters versus factors depending on the breadth of the portfolio set considered. For the narrowest portfolio set, P1, model comparisons are roughly split between favoring multiple-cluster models and multiple-factor models. This result parallels Table 4 in that we again find that this portfolio set is too limited in variety for risk price heterogeneity to be the dominant consideration for explaining variation in returns, and this feature is reinforced by the fact that the $\mathcal{R}$-partition AIC in some cases only chooses a single cluster. Progressing to P2 and P3, for which domestic equity portfolios become more diverse, so too does the importance of clusters relative to omitted factors increase: both portfolio sets point strongly to clusters as more important than missing factors for explaining variation in returns. We highlight two exceptions to this finding. First, for the CAPM rows, we see in both cases that omitted factors are still more important than risk price heterogeneity for P2. This result is unsurprising given the CAPM's limited ability to explain cross-sectional variation in expected or realized returns—clearly other factors are missing, and risk price heterogeneity on the market is less vital than for other factors. Second, our tests favor omitted-factor models relative to multiple-cluster models with K_3^* additional factors in the most recent time period. We interpret this finding as evidence of increased capital-market integration across U.S. stocks, perhaps associated with the sharp decline in transactions costs during this time. Nevertheless, risk price heterogeneity still dominates omitted-factor explanations with comparably sized models (K_2^*) and for models with only a small number of omitted factors (K_1^*).

Like the diverse domestic equity portfolios, the international equity and cross-asset class settings unambiguously favor heterogeneous risk prices over richer factor models. This feature echoes the strong rejections of equal risk prices for P5–P8 documented in Table 3. We conclude that any analysis of the cross-section of expected returns in such settings requires consideration of segmentation across asset classes, notwithstanding recent evidence of integrated intermediation across markets or reduced barriers to trade across global regions.

One notable feature shared with the domestic equity portfolios is that more diverse portfolios more strongly favor risk price heterogeneity over missing factors. Nothing in our analysis mechanically generates this result, and broader portfolio sets could well have required additional factors to explain other cuts of the investible universe (consider, e.g., the additional factors and sorted portfolios of Fama and French 2015). Rather, more exotic corners of the market or nonstandard portfolio formations illuminate pricing discrepancies even among the factors included in relatively parsimonious

models. As empirical asset pricing continues to examine increasingly diverse portfolio sets in response to greater data availability and data-mining concerns, we anticipate heterogeneous risk pricing models to become commensurately more important.

6. Conclusion

We present new methods for detecting and estimating heterogeneous risk prices in a cross-section of assets. Our approach marries traditional asset pricing methods for risk price estimation and machine learning methods for clustering data. Using this methodology, we find statistically significant and economically important evidence of market segmentation across all portfolios, factor pricing models, and time periods. Arbitrage frictions matter universally, not just in highly specialized assets or during crisis periods: even within low-friction, high-participation markets, we find that compensation per unit risk varies significantly across assets.

Segmented risk prices challenge leading models of risk and return. At best, segmentation implies that these factor models are incomplete and miss important cross-sectional variation in expected returns. However, our contribution is not simply another attack on commonly used factor models in finance. Rather, our findings give fresh motivation to consider limits to arbitrage in security markets. We offer a structured alternative for how, not just whether, frictionless factor models might be improved upon in empirical applications. In doing so we suggest a promising new direction for the study of the cross-section of expected returns.

Our findings have important practical implications. Given the ability to invest across groups of assets, sophisticated investors should direct their capital to markets with the highest compensation per unit risk. We offer concrete guidance for identifying these groups of assets. Likewise, potential arbitrageurs across segments can earn the difference between risk prices net of implementation costs. While such long-short strategies are not true arbitrage opportunities—there are too few market segments to be well-diversified, and local factors are imperfectly correlated across segments—they nonetheless represent "good deals" in the sense of Cochrane and Saá-Requejo (2000) and contribute to substantial improvements in ex post Sharpe ratios.

Risk price heterogeneity also provides a novel explanation for the "factor zoo" of Cochrane (2011) and Harvey, Liu, and Zhu (2016). The examples of Section 5.5.1 indicate that any factor whose loadings align with different segments may appear to be priced. Our analysis suggests that missing clusters are more important than omitted factors for most combinations of factor models and portfolio sets, with the possible exception of U.S. equities in the most recent period. The natural follow-up question, and the subject of ongoing work, is the extent to which so-called expected return factors in U.S. stocks are instead proxies for membership in market segments with different risk prices.

Appendix A. Mismeasured Factors and Mismeasured Betas

A.1 Mismeasured Factors

Our ability to detect segmented risk prices depends on the choice of factor model. We consider here the robustness of our analysis to mismeasured factors, for example, if HML does not perfectly track the underlying economic state variable responsible for the value premium. Specifically, suppose that our factor set f is a noisy rotation of the true factors $\tilde{f}$ (as in Giglio and Xiu 2021), $f = \tilde{f}H + \eta$. Mismeasurement of this form rotates and attenuates factor betas, which in turn imparts a factor structure to the residuals $\epsilon \equiv r - f\hat{\beta}'$. This case thus reduces to one in which factors are omitted from the conjectured factor model.

We address this possibility by including in our battery of models a version of the Carhart (1997) four-factor model augmented with the first three principal components of the residual covariance matrix ("Carhart+3"), as well as a model based on the first five principal components of the return covariance matrix ("PC5"). These principal components include the most important noisy rotation residuals or truly missing factors. In the "Carhart+3" model, the principal components capture information not spanned by the included factors, while the "PC5" model avoids taking a stand on any specific factor model altogether. Table 3 reports strong evidence for multiple clusters with both of these models. In addition, the test of 5.5.3 does not distinguish between sources of omitted factors—extracted principal components include the most important noisy rotation residuals or truly missing factors. There, too, we find that "completing" factor models explains the data less well than differences in risk prices among clusters.

A.2 Mismeasured Betas

Throughout our analysis, we use unconditional betas rather than conditional betas for two reasons: (1) our theory assumes betas are precisely measured, and this assumption only applies as $T \to \infty$ for each beta estimate; and (2) our study entails running tens of billions of regressions.[42] Using static betas facilitates significant computational gains relative to time-varying betas in which the right-hand side of the cross-sectional regressions varies each date. However, if betas change in tandem across securities, then clustering assets by estimated risk prices may instead group on common variation in betas.

We take three approaches to address this concern. First, we use stock portfolios rather than individual stocks. Relative to individual stocks, portfolios diversify away mean-reverting measurement error in security-level betas and rebalance to keep characteristics (and hence factor exposures) stable over time. Hence portfolios are less subject to common variation in risk exposures that might contaminate our risk price estimates. Second, we conduct subsample analyses in which we split the sample into thirds or halves. Betas are more likely to be stable over these shorter horizons. We find no evidence that segmentation is weaker over short horizons than at long horizons, as we would expect if time-varying risk exposures explained our findings. Notably, even our full-sample results reestimate betas between $\mathcal{R}$ and $\mathcal{P}$ partitions as part of our split-sample permutation-based testing procedure. Third, depending on the form of comovement in betas, omitted time variation in betas can manifest as omitted factors and, as discussed in the previous subsection, we address such potential omissions by including principal-component enriched factor models.

[42] Another motivation to use static betas is that conditional betas can worsen pricing errors when factor models are misspecified (as Ghysels 1998 argues). This feature is particularly problematic in our application because (1) the null model is misspecified when portfolios are segmented and (2) group assignments are a function of common variation in pricing errors.

Appendix B. Finding Global Optima

B.1 Numerical Issues and Equivalence with Generalized k-means

We face three related optimization challenges in implementing the EM algorithm: selecting starting values; achieving global rather than local optima; and avoiding empty clusters. The choice of starting values for γ matters because expectation maximization finds local solutions, and final group assignments may depend heavily on initial group assignments. Likewise, we need a procedure to escape local basins of attraction on the likelihood surface in order to achieve the global maximum likelihood. Finally, clusters may depopulate to fewer than K elements as portfolios are reshuffled after λs are set. However, such collections of group assignments cannot be global optima because repopulating these clusters with (at least) K elements increases the likelihood function unless all portfolios are perfectly fit.

While we apply EM as the solution method to our maximization problem (5), our procedure closely resembles Lloyd's algorithm in k-means clustering. Like our algorithm, Lloyd's algorithm consists of update and assignment steps. Typically the update step consists of generating "centroids" by averaging characteristics within a group. This step is equivalent to minimizing squared errors within each group using a model with only a constant term. Linear regression also minimizes squared errors within each group, but it accommodates multifactor models. Hence, the first step of our methodology is a straightforward extension of k-means in which group "characteristics" are slopes for each date t and factor k. The assignment step of our methodology—choosing the cluster that minimizes errors given slopes—is exactly the same as in standard k-means in the sense of selecting the group that minimizes squared errors. However, in contrast with standard k-means, the importance of each characteristic for group assignment varies across observations in proportion with betas; the larger β_{ik} in absolute value, the more important λ_ks are for determining asset i's cluster.

Because of the similarity of our approach to k-means, we borrow and extend a common initialization method, known as "k-means++" (Arthur and Vassilvitskii 2007). k-means++ is an algorithm designed to choose cluster centers such that Lloyd's algorithm achieves a clustering solution that is competitive with the global optimum. We discuss our extension of k-means++ in Appendix B.3. While proving the desirable properties of this algorithm is beyond the scope of our paper, we do find significant reductions in squared errors at the solution relative to initialization by choosing cluster memberships at random, suggesting that our variant inherits some of k-means++'s desirable properties.

We take two approaches to find global solutions. First, we initialize our version of k-means++ at $2N$ starting group assignments. We then run the EM algorithm from each assignment to find local optima. If our initializations cover the most promising basins of attraction, this step alone will suffice to locate the global maximum likelihood and corresponding group assignments. However, searching over γ is a high-dimensional problem requiring at least N group assignments to G groups, and $2N$ starting points may be insufficient to find global solutions. For this reason, we couple our multistart approach with an explicit global optimizer able to accommodate integer problems.

In particular, we use MATLAB's mixed-integer programming implementation of the genetic algorithm based on Deb (2000) and Deep et al. (2009). We initialize the population of the genetic algorithm with the $2N$ solutions of the EM algorithm as well as with N nonoptimized initializations of our variant of k-means++. In so doing we cover a large number of local optima, while allowing the algorithm to search new combinations and mutations toward a global solution. Note that we only need to search over γs, because the group assignments imply αs and Λs and likelihoods, and at the global best choice of groups, no groups need to be reassigned once αs and Λs are estimated. Once we have the final population from the genetic algorithm, we take the $2N$ highest likelihood-values from the local and global procedures and apply the EM algorithm to each to ensure that near-optima from the genetic algorithm are (at least) local optima. We select the highest-likelihood solution from this procedure and retain the corresponding group assignments and cross-sectional slopes.

In addition to using global optimization techniques directly, we also make a minor modification to the standard EM algorithm to avoid suboptimal assignments in which at least one cross-section is too small to obtain slopes. In particular, if after reassignment a cluster would have fewer than $K+1$

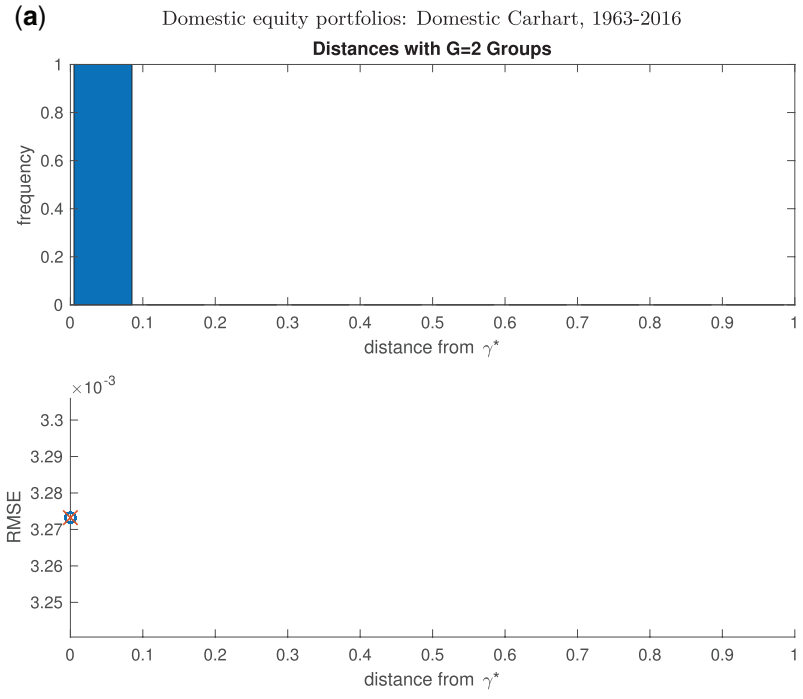

Figure B.I
Comparison of local and global solutions
Figure plots the distribution of optimizer solutions at the conclusion of the EM–genetic algorithm–EM procedure described in Section 1.1.1. We analyze the detailed examples described in 3.3.3: domestic equity portfolios (P3) with a Carhart four-factor model (top left); international equity portfolios (P6) with a global Carhart four-factor model (top right); and cross-asset class portfolios (P8) with two-factor intermediary capital factor model (bottom). In each case we use the AIC to select the number of groups, $G^* = \text{argmin}_G\, AIC_G$. Each subfigure displays two plots. The top plots are histograms of distances of the final $2N$ points from the likelihood-maximizing group assignment γ^*. Given cluster assignments γ^* and γ', we iterate over all possible permutations of the group labels for γ' and retain the permutation with the maximum number of groups in common. The histogram reports the distribution of this one minus proportion to obtain a "distance." The bottom of each subfigure plots z-scores of the $2N$ log-likelihoods against these distances. We mark the best optimization—on the y-axis with a γ^* distance of 0—with an x.

elements, we introduce a likelihood hurdle for moving portfolios. Elements can only be reassigned if the improvement in likelihood is greater than c, and we choose the smallest c such that no cluster after reassignment would have fewer than $K+1$ elements. Only then do we reassign portfolios to groups. Such a c always exists because in the worst case we can set c equal to the maximal change in likelihoods across observations to ensure no changes in group assignments occur. Once a group has $K+1$ elements it will not depopulate in the subsequent iteration because all portfolios are perfectly fit, and c may return to zero. This step avoids convergence to dominated local optima with vanishing groups.

B.2 Comparisons of Local and Global Solutions

Figure B.I compares local solutions to the global best solution to (5) using the EM–genetic algorithm–EM procedure described in the preceding section. To illustrate possible optimization outcomes, we analyze the detailed examples described in Section 3.3.3: domestic equity portfolios

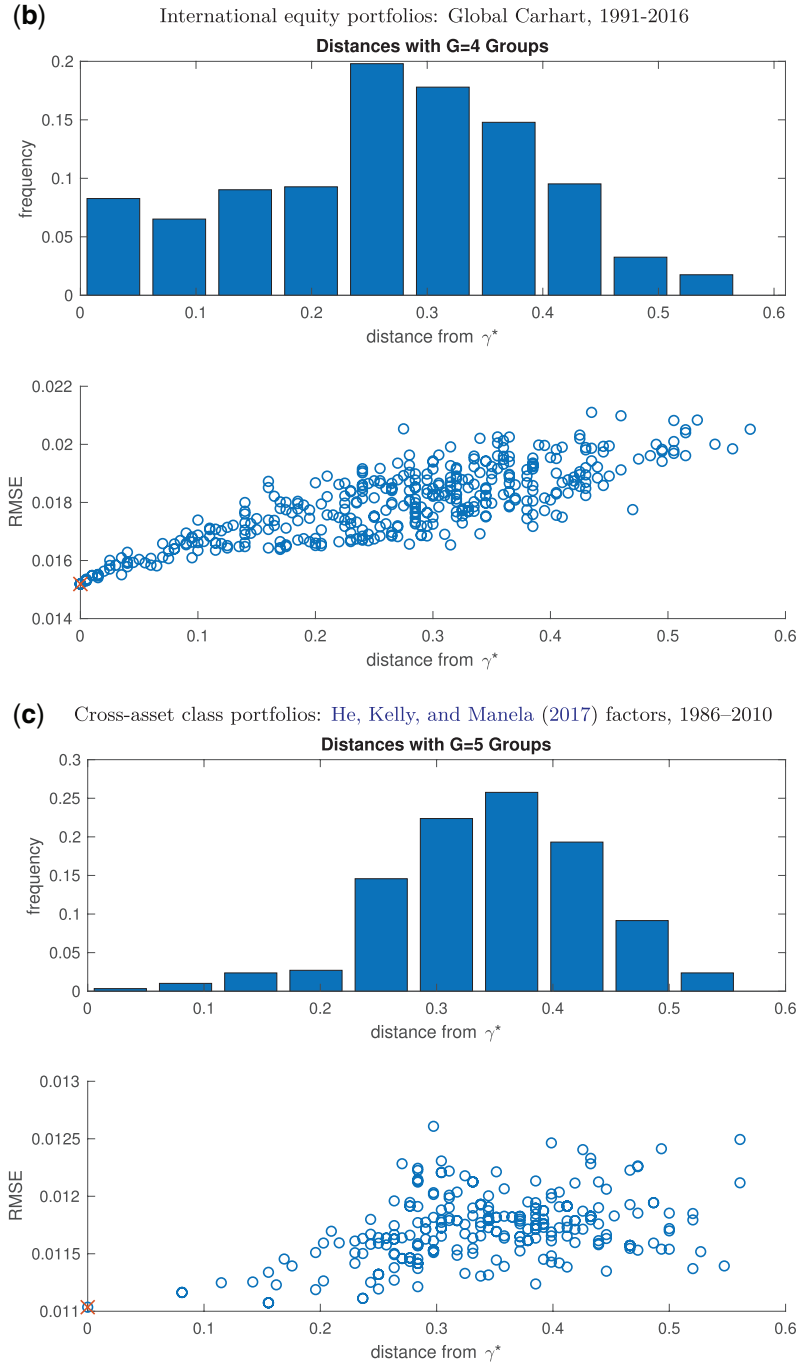

Figure B.I
(Continued)

with a Carhart four-factor model; international equity portfolios with a global Carhart four-factor model; and cross-asset class portfolios with a two-factor intermediary capital factor model.

As a preliminary step, we define two measures for summarizing the distance between two optimizations. The first distance uses group assignments. Let the group assignment of the global best solution be γ^*, and the group assignment of a candidate alternative solution be γ'. Because groups are only known up to a permutation of their labels—there's no difference between 1-2 and 2-1 in a two-asset economy, for example—we need our distance measure to be robust to relabelings. To enforce this robustness, given cluster assignments γ^* and γ', we iterate over all possible permutations of the group labels for γ' and retain the permutation with the maximum number of groups in common. We define our first distance as one minus this proportion in common. The second distance uses the root-mean-squared-error (RMSE), a standard measure of model quality. This metric is especially appropriate in our setting because the MSE is a linear function of (hard-to-interpret) log likelihoods, the optimizer maximand. We compare RMSEs to assess whether the optimizer reaches solutions of similar quality, even though the group assignments associated with those solutions may be quite different.

The first figure depicts a case in which all initial values converge to the same solution. Distances are zero for both metrics for all of the $2N$ final values of our procedure. This outcome suggests that our first setting is so well-behaved that EM converges to the global best solution from a wide range of starting values. The second figure indicates a large number of local maxima. Each point away from the global best represents group assignments different from γ^* and a higher associated RMSE. In this case the local solutions appear to populate a continuum in which distances gradually increase in group space and RMSE space. Here, a local optimizer has a low chance of finding the global best solution, but many local solutions are "close" to the global solution. The third figure similarly suggests many local maxima. However, in this case, we see no clustering around the best solution. From an optimization standpoint, most runs fall into other basins of attraction with disparate group assignments, and a local optimizer would be highly unlikely to stumble upon group assignments near the global best. This said, despite the moderate-to-large differences in assignment distances, other partitions capture risk price heterogeneity almost as well as the segments described in 3.3.3 (as judged by the RMSE). Multiple dimensions of risk price heterogeneity are important in the cross-asset class setting, and the null hypothesis of unified risk pricing is likely to be rejected along more than one of them.

B.3 Extension of k-means++ to Cross-Sectional Slopes

The k-means++ algorithm of Arthur and Vassilvitskii (2007) proceeds as follows:

1. Choose the first cluster center at random among the existing data points.
2. For the c centers that have already been chosen, calculate distances of all data points to all cluster centers. Define $D(x)$ as the distance of data point x to the nearest center.
3. Choose a new center from the data points with probabilities proportional to squared distance.
4. Repeat steps (2) and (3) until the desired number of clusters have been chosen.

This initialization ensures that all clusters are well spaced, and this spacing alone dramatically reduces squared errors (and often, the run time) of the k-means algorithm.

Adapting k-means++ requires only a change in how we define our data points. Because our characteristics are cross-sectional slopes, we require at least $K+1$ assets to define each "data point." We add two preliminary steps A1–A2 to the algorithm to accommodate this difference:

A1. Draw $H = NM$ groups of size $M = \lfloor 1.5(K+1) \rfloor$ from the N assets.

A2. For each group $h = 1, \ldots, H$, estimate $\alpha_t^{(h)}$ and $\lambda_{kt}^{(h)}$ for all k and t.

Note that if the only factor is a constant ($K = 0$) we are back to the k-means++ case up to using random draws of portfolios rather than the set of portfolios itself.

The first step (A1) lets the number of potential cluster locations grow with the number of portfolios and factors. To precisely parallel k-means++ would require picking all $\binom{N}{M}$ combinations of groups for candidate cluster centers, but this number of groups grows exponentially with K and is too large to be implementable. The second step (A2) obtains the characteristics (α and Λ) that determine cluster centers and cluster distances. Armed with these quantities we can proceed with k-means++ as before using α and Λ for each group in place of the underlying return data. In a zero-factor model, α *is* the underlying return data, and our algorithm again reduces to k-means++.

We also add a final step (B1) to k-means++ to prevent situations in which we obtain outlier (very small) or empty initial clusters:

B1. Drop the first selected cluster, and assign portfolios to the remaining clusters to obtain initial γs. Go back to step A1 if the size of the smallest cluster is too small relative to the size of the largest cluster, $N_{max}/N_{min} > 6$.

This modification ensures that we satisfy the theoretical requirement that the smallest cluster be large enough to use large-N asymptotics for the cross-sectional step. The first part makes a dominant first cluster less likely. The second part restarts the modified k-means++ algorithm from scratch when some clusters are too small.

Appendix C. Stability of Group Assignments

Our testing procedure assumes that group assignments are fixed over our sample period. However, just as the risk characteristics of portfolios may change over time, so too may the market frictions that separate portfolios into segments with different risk prices. We evaluate the stability of group assignments by comparing assignments across different subwindows for the same factor models and portfolio sets. Our stability measure is the proportion of groups assignments in common, where we take the highest proportion of common assignments over all permutations of group labels (as in our distance measure of Appendix B.2). We use the number of clusters with the smallest AIC from the full sample.

Table C.I reports the proportions in common for all pairs of subwindows for all sets of factor models and portfolio sets for which multiple subwindows are available. Generally stability is high throughout: the average and median stability value is 0.7, and 25% of values exceed 0.8. In some cases stability is low, for example, for the domestic equity portfolios, such as for the Fama-French five-factor model applied to the widest domestic equity set, or when using the principal component-based model (PC5), where the identities of the factors can change across subperiods. This limited stability suggests that segmentation indeed changes over time, at least for some risk factors and economic settings. Overall, we interpret Table C.I as indicating that full-sample analysis should be complemented by subwindow analyses to capture the dimensions of risk-price heterogeneity most relevant during any particular time period.

Table C.I
Stability of group assignments over time

A. Domestic equity portfolios

Model		P1	P2	P3		P1	P2	P3		P1	P2	P3
CAPM		0.83	0.70	0.56		0.83	0.67	0.55		0.89	0.64	0.61
FF3F	**Period 1** 1963–1980	0.72	0.74	0.44	**Period 1** 1963–1980	0.76	0.57	0.47	**Period 2** 1981–1998	0.96	0.51	0.63
Carhart		0.76	0.56	**0.86**		0.68	0.63	**0.70**		0.92	0.82	**0.68**
FF5F		0.55	0.70	0.71		0.56	0.67	0.69		0.85	0.96	0.90
HKM	**Period 2** 1981–1998				**Period 3** 1999–2016				**Period 3** 1999–2016	0.81	0.77	0.69
HXZQ		0.76	0.67	0.65		0.75	0.86	0.59		0.85	0.60	0.51
Carhart+3		1.00	0.63	0.75		1.00	0.62	0.51		1.00	0.82	0.58
PC5		0.72	0.86	0.51		0.45	0.77	0.50		0.48	0.45	0.85

B. International equity portfolios

Model		P5	P6
CAPM		0.87	0.68
FF3F	**Period 1** 1991–2003	0.61	0.84
Carhart		0.74	**0.78**
FF5F		0.99	0.70
HKM	**Period 2** 2004–2016	0.55	0.91
HXZQ			
Carhart+3		0.59	0.65
PC5		0.74	0.69

C. Cross-asset class portfolios

Model		P7	P8
CAPM		0.70	0.70
FF3F	**Period 1** 1986–1997	0.56	0.57
Carhart		0.57	0.55
FF5F		0.73	0.76
HKM	**Period 2** 1998–2010	0.62	**0.55**
HXZQ		0.56	0.57
Carhart+3		0.88	0.53
PC5		0.63	0.73

Table reports the proportion of stable group assignments across time periods. We first use the AIC to select the number of groups indicated in the full sample, $G^* = \text{argmin}_G AIC_G$. Then, for each date set, we estimate group assignments with G^* clusters. Given cluster assignments for two date sets, we report the proportion of group assignments that agree, taking into account that group labels are arbitrary. We repeat this procedure for all combinations of portfolio sets, risk models, and sample periods. Portfolios and models are described in the text. We omit the placebo portfolio set (P4) because the AIC almost always selects one cluster for these portfolios. Bolded values indicate proportions for the three examples of Section 3.3.3. For the He, Kelly, and Manela (2017) factor model, we use daily data for the most recent time period and monthly data for 1963–2016 and 1981–1998; we do not have sufficient coverage for their intermediary capital factor for 1963–1980 to include it. The q-factor (HXZQ) model is excluded from the international portfolio analysis because we do not have a global return-on-equity factor.

References

Adrian, T., E. Erkko, and T. Muir. 2014. Financial intermediaries and the cross-section of asset returns. *Journal of Finance* 69:2557–96.

Arthur, D., and S. Vassilvitskii. 2007. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 1027–35. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Asness, C. S., T. J. Moskowitz, and L. H. Pedersen. 2013. Value and momentum everywhere. *Journal of Finance* 68:929–85.

Bekaert, G., and C. R. Harvey. 1995. Time-varying world market integration. *Journal of Finance* 50:403–44.

Bollerslev, T. 1990. Modelling the coherence in short-run nominal exchange rates: A multivariate generalized arch model. *Review of Economics and Statistics* 72:498–505.

Bonhomme, S., and E. Manresa. 2015. Grouped patterns of heterogeneity in panel data. *Econometrica* 83:1147–84.

Brunnermeier, M. K., and L. H. Pedersen. 2009. Market liquidity and funding liquidity. *Review of Financial Studies* 22:2201–38.

Brunnermeier, M. K., L. H. Pedersen, and S. Nagel. 2008. Carry trades and currency crashes. *NBER Macroeconomics Annual* 23:313–47.

Carhart, M. M. 1997. On persistence in mutual fund performance. *Journal of Finance* 52:57–82.

Chamberlain, G., and M. Rothschild. 1983. Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica* 51:1281–304.

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

Cochrane, J. H. 2011. Presidential address: Discount rates. *Journal of Finance* 66:1047–108.

Cochrane, J. H., and J. Saá-Requejo. 2000. Beyond arbitrage: Good-deal asset price bounds in incomplete markets. *Journal of Political Economy* 108:79–119.

Constantinides, G. M., J. C. Jackwerth, and A. Savov. 2013. The puzzle of index option returns. *Review of Asset Pricing Studies* 3:229–57.

Daniel, K., and T. J. Moskowitz. 2016. Momentum crashes. *Journal of Financial Economics* 122:221–47.

Daniel, K., L. Mota, S. Rottke, and T. Santos. 2020. The cross-section of risk and returns. *Review of Financial Studies* 33:1927–79.

Daniel, K., and S. Titman. 1997. Evidence on the characteristics of cross sectional variation in stock returns. *Journal of Finance* 52:1–33.

Deb, K. 2000. An efficient constraint handling method for genetic algorithms. *Computer Methods in Applied Mechanics and Engineering* 186:311–38.

Deep, K., K. P. Singh, M. Kansal, and C. Mohan. 2009. A real coded genetic algorithm for solving integer and mixed integer optimization problems. *Applied Mathematics and Computation* 212:505–18.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B* 39:1–38.

Errunza, V., and E. Losq. 1985. International asset pricing under mild segmentation: Theory and test. *Journal of Finance* 40:105–24.

Fama, E. F., and K. R. French. 1992. The cross-section of expected stock returns. *Journal of Finance* 47:427–65.

———. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33:3–56.

———. 1998. Value versus growth: The international evidence. *Journal of Finance* 53:1975–99.

———. 2010. Luck versus skill in the cross-section of mutual fund returns. *Journal of Finance* 65:1915–47.

———. 2012. Size, value, and momentum in international stock returns. *Journal of Financial Economics* 105:457–72.

———. 2015. A five-factor asset pricing model. *Journal of Financial Economics* 116:1–22.

Fama, E. F., and J. D. MacBeth. 1973. Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy* 81:607–36.

Foerster, S. R., and G. A. Karolyi. 1999. The effects of market segmentation and investor recognition on asset prices: Evidence from foreign stocks listing in the united states. *Journal of Finance* 54:981–1013.

Frazzini, A., and L. H. Pedersen. 2014. Betting against beta. *Journal of Financial Economics* 111:1–25.

Gabaix, X., and R. S. Koijen. 2020. In search of the origins of financial fluctuations: The inelastic markets hypothesis. Working Paper, Harvard University.

Gârleanu, N., and L. H. Pedersen. 2011. Margin-based asset pricing and deviations from the law of one price. *Review of Financial Studies* 24:1980–2022.

Ghysels, E. 1998. On stable factor structures in the pricing of risk: Do time-varying betas help or hurt? *Journal of Finance* 53:549–73.

Gibbons, M. R., S. A. Ross, and J. Shanken. 1989. A test of the efficiency of a given portfolio. *Econometrica* 57:1121–52.

Giglio, S., and D. Xiu. 2021. Asset pricing with omitted factors. *Journal of Political Economy* 129:1947–90.

Gompers, P. A., and A. Metrick. 2001. Institutional investors and equity prices. *Quarterly Journal of Economics* 116:229–59.

Greenwald, D. L., M. Lettau, and S. C. Ludvigson. 2016. Origins of stock market fluctuations. National Bureau of Economic Research Working Paper 19818.

Griffin, J. M. 2002. Are the Fama and French factors global or country specific? *Review of Financial Studies* 15:783–803.

Griffin, J. M., X. Ji, and J. S. Martin. 2003. Momentum investing and business cycle risk: Evidence from pole to pole. *Journal of Finance* 58:2515–47.

Grinblatt, M., and T. J. Moskowitz. 2004. Predicting stock price movements from past returns: The role of consistency and tax-loss selling. *Journal of Financial Economics* 71:541–79.

Gromb, D., and D. Vayanos. 2002. Equilibrium and welfare in markets with financially constrained arbitrageurs. *Journal of Financial Economics* 66:361–407.

———. 2018. The dynamics of financially constrained arbitrage. *Journal of Finance* 73:1713–50.

Haddad, V., and T. Muir. 2021. Do intermediaries matter for aggregate asset prices? *Journal of Finance* 76:2719–61.

Harvey, C. R. 2017. Presidential address: The scientific outlook in financial economics. *Journal of Finance* 72:1399–440.

Harvey, C. R., Y. Liu, and H. Zhu. 2016. ... and the cross-section of expected returns. *Review of Financial Studies* 29:5–68.

He, Z., B. Kelly, and A. Manela. 2017. Intermediary asset pricing: New evidence from many asset classes. *Journal of Financial Economics* 126:1–35.

He, Z., and A. Krishnamurthy. 2012. A model of capital and crises. *Review of Economic Studies* 79:735–77.

———. 2013. Intermediary asset pricing. *American Economic Review* 103:732–70.

Holden, C. W., and S. Jacobsen. 2014. Liquidity measurement problems in fast, competitive markets: Expensive and cheap solutions. *Journal of Finance* 69:1747–85. doi:10.1111/jofi.12127.

Hong, H., T. Lim, and J. C. Stein. 2000. Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies. *Journal of Finance* 55:265–95.

Hou, K., G. A. Karolyi, and B.-C. Kho. 2011. What factors drive global stock returns? *Review of Financial Studies* 24:2527–74.

Hou, K., C. Xue, and L. Zhang. 2015. Digesting anomalies: An investment approach. *Review of Financial Studies* 28:650–705.

———. 2020. Replicating anomalies. *Review of Financial Studies* 33:2019–133.

Israel, R., and T. J. Moskowitz. 2013. The role of shorting, firm size, and time on market anomalies. *Journal of Financial Economics* 108:275–301.

Kadlec, G. B., and J. J. McConnell. 1994. The effect of market segmentation and illiquidity on asset prices: Evidence from exchange listings. *Journal of Finance* 49:611–36.

Koijen, R. S., R. J. Richmond, and M. Yogo. 2020. Which investors matter for equity valuations and expected returns? Working Paper, University of Chicago.

Koijen, R. S., and M. Yogo. 2019. A demand system approach to asset pricing. *Journal of Political Economy* 127:1475–515.

Lehmann, E., and J. Romano. 2005. *Testing statistical hypotheses, third edition*. New York: Springer.

Lesmond, D. A., M. J. Schill, and C. Zhou. 2004. The illusory nature of momentum profits. *Journal of Financial Economics* 71:349–80.

Lettau, M., S. C. Ludvigson, and S. Ma. 2019. Capital share risk in u.s. asset pricing. *Journal of Financial Economics* 74:1753–92.

Lettau, M., M. Maggiori, and M. Weber. 2014. Conditional risk premia in currency markets and other asset classes. *Journal of Financial Economics* 114:197–225.

Lewellen, J. 2011. Institutional investors and the limits of arbitrage. *Journal of Financial Economics* 102:62–80.

Lewellen, J., S. Nagel, and J. Shanken. 2010. A skeptical appraisal of asset pricing tests. *Journal of Financial Economics* 96:175–94.

Lin, C.-C., and S. Ng. 2012. Estimation of panel data models with parameter heterogeneity when group membership is unknown. *Journal of Econometric Methods* 1:42–55.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1*, 281–97. Berkeley, CA: University of California Press.

Mankiw, N., and S. P. Zeldes. 1991. The consumption of stockholders and nonstockholders. *Journal of Financial Economics* 29:97–112.

Menkhoff, L., L. Sarno, M. Schmeling, and A. Schrimpf. 2012. Carry trades and global foreign exchange volatility. *Journal of Finance* 67:681–718.

Merton, R. C. 1973. An intertemporal capital asset pricing model. *Econometrica* 41:867–87.

———. 1987. A simple model of capital market equilibrium with incomplete information. *Journal of Finance* 42:483–510.

Newey, W. K., and K. D. West. 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55:703–8.

Novy-Marx, R., and M. Velikov. 2016. A taxonomy of anomalies and their trading costs. *Review of Financial Studies* 29:104–47.

Nozawa, Y. 2017. What drives the cross-section of credit spreads?: A variance decomposition approach. *Journal of Finance* 72:2045–72.

Patton, A. J., and B. M. Weller. 2020. What you see is not what you get: The costs of trading market anomalies. *Journal of Financial Economics* 137:515–49.

Pontiff, J. 1996. Costly arbitrage: Evidence from closed-end funds. *Quarterly Journal of Economics* 111:1135–51.

———. 2006. Costly arbitrage and the myth of idiosyncratic risk. *Journal of Accounting and Economics* 42:35–52.

Rivers, D., and Q. Vuong. 2002. Model selection tests for nonlinear dynamic models. *Econometrics Journal* 5:1–39.

Rouwenhorst, K. G. 1998. International momentum strategies. *Journal of Finance* 53:267–84.

Shleifer, A., and R. W. Vishny. 1997. The limits of arbitrage. *Journal of Finance* 52:35–55.

Su, L., Z. Shi, and P. C. Phillips. 2016. Identifying latent structures in panel data. *Econometrica* 6:2215–64.

Vissing-Jørgensen, A. 2002. Limited asset market participation and the elasticity of intertemporal substitution. *Journal of Political Economy* 110:825–53.

Wu, C. F. J. 1983. On the convergence properties of the EM algorithm. *Annals of Statistics* 11:95–103.

Zivot, E. 2009. Practical issues in the analysis of univariate GARCH models. In T. Mikosch, J.-P. Kreiß, R. A. Davis, and T. G. Andersen, eds., *Handbook of Financial Time Series*, 113–55. Berlin, Germany: Springer.