Quantity, Risk, and Return*

Yu An† Yinan Su‡ Chen Wang§

August 30, 2024

Abstract

We propose a new model of expected stock returns that incorporates quantity information from market trading activities into the factor pricing framework. We posit that the expected return of a stock is determined by not only its factor risk exposures (β) but also the factor's quantity fluctuations (q) induced by trading flows, and hence term the model *beta times quantity* (BTQ). The rationale is that sophisticated investors should require a greater factor premium when they are more exposed to that factor after noise traders sell lots of stocks with high exposures to that factor. The BTQ model provides a compelling risk-based explanation for stock returns, which is otherwise obscured without considering the quantity information. The cross-sectional risk-return association, which is nearly flat unconditionally, strongly depends on the quantity variable. The structured BTQ model reliably predicts monthly stock returns out of sample, and addresses the factor zoo problem by selecting a small number of factors.

Keywords: quantity, flow, noise trader, risk and return, cross section of return, return prediction, factor model, Lasso, PCA, BTQ

JEL Codes: G11, G12

*We thank Federico Bandi, Hank Bessembinder, Andrew Chen, Zhi Da, Darrell Duffie, Robin Greenwood, Zhiguo He, Ben Hébert, Shiyang Huang, Bryan Kelly, Serhiy Kozak, Jiacui Li, Dong Lou, Paolo Pasquariello, Nagpurnanand Prabhala, Seth Pruitt, Alessandro Rebucci, Paul Schultz, Dongho Song, Yang Song, Zhaogang Song, Semih Üslü, Wei Wu, Jun Yu; conference discussants, Aref Bolandnazar, Aditya Chaudhry, Thummim Cho, Fotis Grigoris, Badrinath Kottimukkalur, Xin Liu, Marcel Müller, Andrey Pankratov, Oleg Rytchkov, Andrea Vedolin; and conference and seminar participants at the Fed Board, Wolfe, MFA, Southern Methodist University, FMCG Conference, DC Junior, SoFiE Annual Conference, CICF, University of Macau, City University of Hong Kong, Chinese University of Hong Kong, SAFE Asset Pricing Workshop, UT Dallas Finance Conference, World Symposium on Investment Research, FMA Applied Finance Conference, and Michigan Mitsui Symposium for valuable comments and suggestions. A previous version of the paper was circulated under the title "A Factor Framework for Cross-Sectional Price Impacts."

†Carey Business School, Johns Hopkins University; yua@jhu.edu.

‡Carey Business School, Johns Hopkins University; ys@jhu.edu.

§Mendoza College of Business, University of Notre Dame; chen.wang@nd.edu.

1 Introduction

Explaining the expected returns of different stocks is a central question in asset pricing. The theoretical answer is clear—risk—investors are averse to risk and require compensation for bearing risk. Therefore, riskier investments should earn higher expected returns in equilibrium. However, the empirical answer has proven more complicated: evidence of the risk-return tradeoff, such as their positive association in the cross section, is elusive in data; and risk-based models hardly predict individual stock returns, in contrast to unstructured predictions with firm characteristics and machine learning models.[1] A revamped model is critically needed for the risk-based approach to expected returns.

This paper makes headway in this important area by incorporating a new aspect of risk's economic role in determining asset prices—the *quantity* variation in investors' risk holdings induced by trading flows. Many existing endeavors focus on the statistical aspects of risk, such as identifying the common factors and estimating factor premiums, and on the properties of the securities per se, such as risk exposures and firm characteristics.[2] We show the canonical risk framework equipped with the quantity variables, which are constructed from market trading activities and are about sophisticated investors' risk holding conditions, yields a compelling explanation for the cross-section of expected returns.

We integrate quantity into factor pricing by considering market trading activity's effect on sophisticated investors' risk holdings and, in turn, their required compensation for bearing risk. First, we acknowledge that the market is not populated by representative agents but is modeled with two groups of investors: noise investors (such as retail investors) and sophisticated investors (such as hedge funds and market makers). Noise investors generate large and correlated flows in individual stocks. Sophisticated investors take the other side of these trades, which causes fluctuations in the quantities of their holdings of the underlying

[1] See papers that report elusive risk-return association in Footnote 4 and those that predict stock returns in Footnote 5.

[2] These related topics constitute a large and growing body of literature. We contribute to three sub-areas with references listed in Footnotes 4, 5, and 6, respectively.

systematic risks. For example, if noise investors sell a large quantity of value stocks with high HML (high-minus-low) loadings, sophisticated investors' holdings of the HML risk will increase. The sophisticated investors are the marginal investors whose equilibrium conditions drive asset prices. We posit that they require greater compensation for a systematic risk factor when they hold more of it, i.e., they have less demand for that risk. This gives rise to our key innovation in model specification: a factor's premium varies with the factor's quantity fluctuations induced by trading flows. At the same time, sophisticated investors enforce no-arbitrage pricing across stocks, so the canonical factor pricing condition still holds. These two forces combined give rise to our main empirical model, in which the expected return of a stock is determined by the interactions of its factor risk exposures (β) and the factors' quantity fluctuations induced by trading flows (short for "quantity" or variable q throughout the paper), which we term the *beta times quantity* (BTQ) model.

This framework, though abstracted away from many details of the market microstructure, captures a significant economic force central to risk aversion that, nonetheless, has long been missing in empirical studies of risk and return. The new mechanism considered here is not new to the literature that studies the price impacts of noise trading flows for individual assets, factor portfolios, or asset classes.[3] Our contribution is integrating quantities into the factor pricing framework, which affords smooth upgrades of workhorse methods in cross-sectional asset pricing research. We demonstrate incorporating quantity information leads to important empirical discoveries in the following three aspects.

First, quantity information elicits risk-return tradeoff relationships that are otherwise obscured. Previous studies report a flat security market line (SML, which plots expected return $\mathbb{E}r$ against β in the cross section), inconsistent with the theoretical premise of high-risk-high-return.[4] However, a significant positive β-$\mathbb{E}r$ relation emerges *conditional on* high levels of accumulated flows on the market risk. That is, the risk-aversion implied high-

[3]See Gabaix and Koijen (2022) for a review. We discuss related papers in detail further below.

[4]Black (1972), Black, Jensen, and Scholes (1972), and Frazzini and Pedersen (2014) report a flat SML. Along this direction but with more involved investigations, Lopez-Lira and Roussanov (2020) question whether common factor exposure (β) really explains the cross-sectional variation in average returns.

risk-high-return relation holds when sophisticated investors have absorbed more market risk quantity. In this view, the previously reported flat SML is an unconditional average when the quantity information is ignored. Additional results that support this view are obtained with Fama-MacBeth regressions properly upgraded with quantity information and with tests that use factors other than the market.

Second, quantity information enables a risk-based model that predicts individual stock returns. A central goal of asset pricing is to explain (conditional) expected returns, and statistically predicting individual stock returns serves as a touchstone for proposed explanations. This task is empirically hard, and researchers have only recently made significant progress by resorting to unstructured machine learning models designed for forecasting and using a large number of firm characteristics, which inevitably sacrifice explainability. The state-of-the-art methods can reliably predict stock returns at the monthly horizon, even though the explained variation is small given the low signal-to-noise nature of market prices.[5] We build an economically grounded predictor that interacts stock-level factor exposures (β) with factor-level quantity fluctuations (q). Beta times quantity (BTQ) alone reliably predicts the panel of monthly individual stock returns with an OOS R^2 of around 1% in various robustness settings, a level comparable to high-dimensional machine learning models. Without quantity, the β-only model has almost no predictive power, consistent with previously reported null results (Lopez-Lira and Roussanov, 2020).

Third, quantity information offers a new and better perspective to address the factor zoo problem and provide new factor selection results. The proliferation of proposed factors challenges the asset pricing literature regarding which factors are important for expected returns and fundamental to investors' pricing decisions. The existing tests focus on factor premium: essentially, they ask whether high exposure to a factor is associated with high

[5]Studies on stock (and equity portfolio) return forecasting include Fama and French (2008), Welch and Goyal (2008), Koijen and Van Nieuwerburgh (2011), Rapach and Zhou (2013), and Lewellen (2014). More recent advances with machine learning include Gu, Kelly, and Xiu (2020), Feng, He, and Polson (2018), Freyberger, Neuhierl, and Weber (2020), Choi, Jiang, and Zhang (2023), and Kelly, Malamud, and Zhou (2024).

expected returns in the cross-section of stocks.[6] The new test asks an upgraded question: whether the expected return difference between high and low β stocks *widens* when the investors' factor risk quantity holding (q) is high (and vice versa). For one, using quantity as an instrument for factor premium provides more variation and, hence, greater identification power. More importantly, this upgrade is more informative of the economic mechanism in which risk aversion takes place and, therefore, should lead us closer to identifying the fundamental risks that investors are averse to. We find that the market is, first and foremost, the most prominent factor. A few other factors are also selected, including betting-against-beta, volatility, idiosyncratic risk, and value. However, size is dismissed in various settings, challenging its perceived importance as a fundamental risk factor. These results are obtained with a variable selection method (Lasso) that allows for the inclusion of a large number of candidate factors (including 153 factors from Jensen, Kelly, and Pedersen, 2023, henceforth JKP). Alternatively, pre-processing the candidate factors with principal component analysis (PCA) to "shrink the cross section" (Kozak et al., 2020) lead to a similar but even more parsimonious result in which only the first two principal components are selected as pricing factors.

In summary, these results suggest incorporating quantity into the factor pricing framework is crucial for empirically establishing a risk-based explanation of expected returns. These results are obtained with a unified model that upgrades a series of workhorse empirical asset pricing methods with quantity. We emphasize two key elements of the empirical method.

First, we construct factor-level time-series variables q that track sophisticated investors' factor risk exposure induced by retail trading flows. The process starts with mutual fund

[6]The proliferation of proposed factors to explain the cross-section of expected stock returns (a.k.a. the factor zoo problem) is noted by Cochrane (2011), Harvey, Liu, and Zhu (2016), McLean and Pontiff (2016), and Hou, Xue, and Zhang (2017). Existing studies address the problem by selecting or "shrinking" the factors (broadly speaking, estimating a low-dimensional factor space), including Feng, Giglio, and Xiu (2020), Lettau and Pelger (2020), Kozak, Nagel, and Santosh (2020), Giglio, Liao, and Xiu (2021), and Giglio and Xiu (2021). Essentially, they discipline a factor by its factor premium (i.e., cross-sectional risk-return association). In this sense, these are developments of the more traditional Fama and MacBeth (1973) method.

flows, which are used as a proxy for noise trading flows from retail investors. We use the flow-induced trading (FIT) metric (Coval and Stafford, 2007; Froot and Ramadorai, 2008; Lou, 2012) to construct flows into individual stocks. (A positive sign means retail investors sell and sophisticated investors buy, i.e., we view quantity changes from sophisticated investors' perspective.) Next, as an important step, individual stock flows are aggregated to the factor level. The aggregation accounts for each stock's factor exposure (β) in a way similar to "portfolio beta" used in risk management.[7] We focus on factor-level aggregated quantity rather than directly modeling the price impact of stock-level flows on stocks. This choice reflects the theoretical motivation that investors are averse to systematic risk, not to individual securities per se.

Second, we specify that each stock's expected return is a linear function of its factor exposure β times quantity q (BTQ model). This is a direct upgrade from the canonical factor pricing framework, in which the expected return is determined by a "β-only" model. The same BTQ model (including its non-parametric versions) is used consistently in the paper for the upgraded security market line (SML), Fama-MacBeth regressions, stock return prediction, and factor selection exercises. Variable q enters the model because it drives fluctuations in factor premium, whereas the traditional β-only model implies a constant factor premium. This is similar to the difference-in-differences (DID) analysis commonly used in applied microeconomics, since β captures the cross-sectional variation in expected returns while q provides the time-series variation. The "β-only" model that estimates constant factor premium only has the first difference in the DID. As discussed above, this upgrade brings greater identification power and economic relevance in selecting the factors that investors are averse to. Future empirical studies can easily subject a newly proposed factor to the BTQ factor pricing tests, given that the factor's BTQ term can be easily estimated. These properties speak to the advantages and general applicability of the new method.

[7] Section 3.3 provides details for constructing q time series. We use realized covariance rather than beta for technical reasons. We normalize factor-level flow by equity market capitalization for stationarity and use moving averages to account for the attenuating effects of older flows. Variable q is standardized to unit variance for ease of interpretation.

Two related frameworks in the literature differ from our research approach and objective. First, we do not treat flow or quantity fluctuations as a source of risk, and the constructed quantity time-series variables are not new risk factors.[8] Instead, we still use previously proposed factors, and the constructed factor-level quantity q complements factor risk exposure β in the specification of BTQ.

Second, in our model, trading flows or quantities do not directly or independently affect asset prices. The factor pricing condition is still front and center in our analysis. Individual stock flows are aggregated to the factor level and work through factor premiums to affect the cross-section of expected returns. However, this paper belongs to the general agenda that argues investor demand matters for asset prices, and that flow and quantity data can be adopted to improve asset pricing empirical research.[9] Our core economic mechanism recognizes that sophisticated investors have limited risk-bearing capacities, such that asset price (or, in our implementation, factor premium) responds to quantity inelastically (or, not perfectly elastically). We connect this idea more closely to the factor pricing framework because of its empirical relevance for studying the cross-section of expected returns. Kozak, Nagel, and Santosh (2018) argue that due to the cross-sectional arbitrage activities of sophisticated investors, asset returns should exhibit a factor structure even with the presence of noise traders. Combining the ideas in these papers, we find, indeed, that integrating the demand channel with factor pricing leads to exciting empirical discoveries.

Some existing papers have reported that flow is relevant for future returns in various settings. Examples include Teo and Woo (2004), Ben-David, Li, Rossi, and Song (2022a), Kang, Rouwenhorst, and Tang (2022), Li (2022), Li and Lin (2022), and Huang, Song, and Xiang (2024) in stock markets, Greenwood and Vayanos (2014) and Vayanos and Vila (2021) in bond markets, and Garleanu, Pedersen, and Poteshman (2008) in option markets. Unlike these papers, which focus on establishing the pricing effects for individual assets, factors, or

[8] This is the approach of De Long, Shleifer, Summers, and Waldmann (1990), Shleifer and Vishny (1997), Adrian, Etula, and Muir (2014), He, Kelly, and Manela (2017), and Dou, Kogan, and Wu (2022).

[9] Gabaix and Maggiori (2015); Koijen and Yogo (2019); Gabaix and Koijen (2022); Koijen, Richmond, and Yogo (2023).

markets, our primary emphasis is on integrating quantity into the factor pricing framework to investigate cross-sectional risk-return tradeoffs.[10]

In the remainder of the paper, Section 2 provides theoretical motivation, empirical model, and methods; Section 3 constructs the quantity q and other empirical measures; Section 4 presents empirical results for the BTQ model; Section 5 concludes.

2 Theoretical motivation, empirical model, and methods

2.1 Theoretical motivation

The theoretical reason why quantity information should be integrated into factor pricing is that market trading activities matter for sophisticated investors' risk holdings and in turn their required compensation for bearing risk. We focus on a prominent channel where a significant aspect of trading activities, the noise trading flows, matters for the central element of asset pricing, the factor premium, although there can be many other market microstructure mechanisms in which trading activities have price impacts. We outline this theoretical channel below.

Suppose the market is populated by two groups of investors: noise investors and sophisticated investors. Noise investors, such as retail traders, generate uninformed flows in and out of individual stocks over time. The noise flows are large and correlated across stocks, which can induce significant fluctuations when aggregated to the factor level.[11]

Sophisticated investors, such as hedge funds and market makers, take the other side of the retail trades by absorbing the noise flows and supplying liquidity. Therefore, noise flows induce fluctuations in the sophisticated investors' holding quantities of the underlying

[10] Additionally, Berk and Van Binsbergen (2016), Barber, Huang, and Odean (2016), and Ben-David, Li, Rossi, and Song (2022b) use a revealed preference approach to determine which factors investors care about. However, they do not focus on the asset pricing properties of the selected factors.

[11] Previous studies report (which we also confirm empirically) that the retail flows are not only significant in magnitude but also correlated across stocks due to the commonality in retail investors' trading behaviors. The correlation aligns with investment styles, such that, say in one period, they tend to sell growth stocks and in the next, they buy small (Li, 2022 and Huang, Song, and Xiang, 2024). This fact supports that retail flows can induce significant fluctuations in the quantity of risk when aggregated to the factor level.

systematic risks. For example, if retail investors sell lots of value stocks with high HML exposures, then sophisticated investors will accumulate more HML risk holdings. The aggregation from stock-level flows to factor-level risk quantities accounts for each stock's factor exposure (β) in the fashion of "portfolio beta" commonly used in risk management (see Section 3.3 for aggregation details). The sophisticated investors are the marginal investors whose risk-holding conditions drive asset prices. They require greater compensation for a systematic risk factor when they hold more of it. This gives rise to the key model specification that a factor's premium varies with the factor's quantity fluctuations induced by the trading flows, and we hypothesize that the relationship is positive. At the same time, sophisticated investors enforce no-arbitrage pricing across stocks, so the canonical factor pricing condition still holds.[12] These two forces combined imply the main empirical model specified below, in which both the stock's factor risk exposures (β) and factor risk quantities (q) determine its expected return.

2.2 Empirical model

The empirical model follows the canonical factor pricing framework, in which the cross-section of stock returns follows a factor structure

$$r_{i,t+1} = \sum_{k=1}^{K} \beta_{i,k,t} f_{k,t+1} + \epsilon_{i,t+1}, \qquad \forall i, t, \tag{1}$$

where $r_{i,t+1}$ is the return of stock i in month $t+1$, k indexes factors, f is factor return, and β is the stock's factor exposure, which we are going to estimate using realized daily returns up to month t. According to the APT (Ross, 1976), the cross-section of expected return follows the factor pricing,

$$\mathbb{E}_t[r_{i,t+1}] = \sum_{k=1}^{K} \beta_{i,k,t} \mu_{k,t}, \qquad \forall i, t, \tag{2}$$

[12] This is consistent with Kozak, Nagel, and Santosh (2018), who argue that cross-sectional no-arbitrage conditions are still valid in the presence of noise traders as long as there exist some sophisticated investors.

where $\mathbb{E}_t[r_{i,t+1}]$ is the (conditional) expected stock return, our research object, and $\mu_{k,t}$ is the factor premium conditional on time-t information.

The departure from the canonical framework is at the modeling of the factor premium. According to the theoretical motivation above, we specify that the factor premium is not a constant but varies with the factor's quantity fluctuations induced by trading flows.

$$\mu_{k,t} = \mu_k(q_{k,t}) = \mu_k + \lambda_k q_{k,t}, \qquad \forall k, t, \tag{3}$$

where the first equation is a general specification in which μ_k is an unspecified function of $q_{k,t}$. In most of the empirical settings, we implement a linear specification as in the second equation.[13] The first parameter μ_k corresponds to constant factor premium, which is the key interest of estimation in traditional factor pricing tests. The linear coefficient λ_k is the new central parameter of interest, which measures the sensitivity of the factor premium to the factor's quantity fluctuations.

Plugging the factor premium specification into the factor pricing condition (Eq. 3 into Eq. 2), we arrive at the main empirical model, the beta times quantity (BTQ) model of expected stock returns:

$$\mathbb{E}_t[r_{i,t+1}] = \left(\sum_{k=1}^{K} \mu_k \beta_{i,k,t}\right) + \sum_{k=1}^{K} \lambda_k \beta_{i,k,t} q_{k,t}, \qquad \forall i, t. \tag{4}$$

The first summation term is the traditional factor pricing model, what we refer to as the "β-only" model as the baseline in empirical comparisons. The second is the new beta times quantity (BTQ) term. In empirical implementation, we often find the β-only term is so close to zero (and so noisy for explaining expected returns), to the extent that having it in the BTQ model even hurts the empirical fit. Therefore, we typically omit the term in

[13] The linear specification can be microfounded using the standard theoretical framework with mean-variance utility and normally distributed payoffs (see Rostek and Yoon, 2023). In the upgraded Fama-MacBeth regressions of Section 4.2, we implement a non-parametric estimation of $\mu_k(\,\cdot\,)$. See Section 2.3 for an overview of the various parametric and non-parametric empirical methods.

parentheses and only include the BTQ term.

The key hypothesis implied by the theoretical motivation is that, for a "true" fundamental risk factor k, $\lambda_k > 0$. The hypothesis means that the cross-sectional return dispersion between high and low β stocks widens when the factor's quantity is high. This is similar to the difference-in-differences (DID) analysis as β captures the cross-sectional variation in expected returns while q provides the time-series variation. In other words, the observed factor risk aversion is stronger when q is high. This offers a new perspective compared to the traditional hypothesis $\mu_k > 0$, which asks whether higher exposure to that factor is associated with higher *average* returns, i.e., only the first "difference" in the DID analysis. The new test has more identification power provided by the time-series variation in q. More importantly, this test has more economic relevance since q variation tracks sophisticated investors' holding conditions so we are no longer inferring their risk pricing process from asset price information alone. Therefore, the new framework can lead us closer to identifying the fundamental risks that investors care about.

The model allows for multiple factors and allows each to have a different λ_k coefficient. This is useful for testing each factor's marginal importance in a joint setting controlling for other factors' contribution to expected returns.

An interesting property of the sign of λ_k is noted. Regardless of the sign of the factor (e.g., small-minus-big or big-minus-small), the sign of λ_k should, theoretically speaking, always be positive. This is because when factor return f flips its sign, both β and q flip their signs, and β times q remains unchanged. A positive λ_k estimate, nonetheless, is not empirically guaranteed. Thus, it provides another layer of testing for the risk-based theory, regardless of the specification of the factor's sign. A negative λ_k estimate would be an unambiguous rejection of the risk-based theory, and the empiricist could not blame the "wrong" sign of the factor as an excuse. Notice that μ_k in the β-only model does not have this property: big-minus-small would have a negative μ_k.

We focus on testing the hypothesis "$\lambda_k > 0$" in the cross-sectional setting of the BTQ

model (Eq. 4), not in the time series context of predicting factor returns $f_{k,t+1}$ with $q_{k,t}$. Although the BTQ model is theoretically motivated by the time-series specification of factor premium (Eq. 3), empirically, a positive time-series predictive coefficient between $q_{k,t}$ and $f_{k,t+1}$ is far from implying the cross-sectional hypothesis of $\lambda_k > 0$. The gap between the two is the cross-sectional variation of the risk exposures (β), which is not present in the time series setting. The gap is analogous to the traditional β-only setting: a long-short portfolio with a high average return does not guarantee that it is a priced factor in cross-sectional tests, such as the Fama-MacBeth regressions.

2.3 Empirical methods

We use a series of empirical methods to estimate and test the BTQ model from different perspectives. The methods are presented as upgrades of familiar procedures in asset pricing, such as the security market line, Fama-MacBeth factor premium estimates, and return prediction exercises, for ease of comparison and to demonstrate the value of incorporating quantity information into the factor model. We present an overview of the methods here, while the details are provided when presenting the empirical results in Section 4.

From the methodological perspective, the progression of the methods can be seen as gradually adding parameterization to the model of expected stock return. To start with, the familiar security market line (SML) can be seen as a simple non-parametric model, $\mathbb{E}_t[r_{i,t+1}] = Er(\beta_{i,k,t})$, where $Er(\,\cdot\,)$ is an unspecified function. (The SML is typically estimated with the market beta, i.e., $k = \text{MKT}$, but we implement it with other factors as well.) The conditional SML (Section 4.1) upgrades it to a bi-variate non-parametric model that includes q, $\mathbb{E}_t[r_{i,t+1}] = Er(\beta_{i,k,t}, q_{k,t})$. We estimate this non-parametric model with a simple kernel method by binning observations of β and q. This method is easy to interpret via the familiar SML plot, and clearly shows that q is a highly relevant variable in the expected return function (Er) with significant effects on the risk-return (β-$\mathbb{E}r$) relation.

The second method, the quantity upgraded Fama-MacBeth factor premium estimates, is

semi-parametric (Section 4.2). It imposes a linear relationship between risk (β) and return according to APT, but is still non-parametric about q's effect: $Er(\beta_{i,k,t}, q_{k,t}) = \beta_{i,k,t}\mu_k(q_{k,t})$, where the factor premium function $\mu_k(\ \cdot\ )$ is left unspecified. It is still estimated non-parametrically by binning q and averaging the returns of the Fama-MacBeth factor mimicking portfolio (FMP, which are coefficients of cross-sectional regression $r_{i,t+1}$ on $\beta_{i,k,t}$) within each bin.

Third, once the $\mu_k(\ \cdot\ )$ function is also specified as linear, we arrive at the parametric BTQ model $Er(\beta_{i,k,t}, q_{k,t}) = \lambda_k\beta_{i,k,t}q_{k,t}$. The parametric setting easily accommodates multiple factors, and is estimated with a linear predictive regression on the panel of stock returns $r_{i,t+1} = \sum_{k=1}^{K} \lambda_k\beta_{i,k,t}q_{k,t} + error_{i,t+1}$ (Section 4.3). Notice that each factor's beta times quantity (BTQ) together serve as a predictor, and the BTQ terms of different factors serve as multivariate predictors. Predicting stock returns has experienced significant progress with firm characteristics and machine learning models. We follow the literature's setup of the stock return panel and evaluate our model with the same metric of empirical success, the out-of-sample (OOS) prediction fit (R^2) besides the in-sample (IS) R^2.

Lastly, in response to the factor zoo problem, when the number of candidate factors (K) is large, the number of BTQ predictors grows accordingly to well more than 100. In such a setting, we use machine learning methods designed for high-dimensional prediction, such as Lasso, to select a small number of priced factors (Section 4.4). By inducing sparsity in the λ_k coefficients, Lasso allows us to select a small number of BTQ terms and reveal which factors are priced in a joint setting, controlling for other factors. Additionally, we follow Kozak et al. (2020) and pre-process the candidate factors with principal component analysis (PCA). Then we supply the principal component factors to the same BTQ construction and Lasso prediction exercise (Section 4.5). The potential benefit of this method is to "shrink the cross section" of factors and elicit latent factors that explain the most time-series return variation of the many candidate factors, which according to existing literature, can be more reliable candidate factors for explaining expected returns.

In summary, we put forward the message that integrating quantitative information into various empirical methods can lead to significant empirical discoveries. We implement the methods outlined above to support this message, although the methods here are far from exhaustive, given the vast asset pricing literature. We believe the quantity variables can similarly interact with many other existing methods and lead to a broad avenue of potential empirical discoveries.

3 Constructing quantity (q) and other variables

The data to run the BTQ predictive regression include the (unbalanced) panel of monthly stock returns $r_{i,t+1}$, and for each factor k, a panel of $\widehat{\beta}_{i,k,t}$ and a time series of $q_{k,t}$. Among them, $\widehat{\beta}_{i,k,t}$ is constructed from the time series of factor return $f_{k,t}$ as in the first stage of the Fama-MacBeth procedure. The construction of $q_{k,t}$ is new. It requires the stock-level retail flow in the same unbalanced panel structure as the returns, which is then aggregated to the factor level according to each stock's factor exposure measures. In summary, the source data are only the panel of returns and the panel of flows at the stock level, with which one can calculate both β and q of any factor from the time series of factor return $f_{k,t}$.

3.1 Constructing return and risk variables

We obtain return data from standard sources. Specifically, we use delisting-adjusted stock returns, denoted as $r_{i,t}$, from CRSP. Factor returns are obtained from two sources. First, Fama-French-Carhart (i.e., Fama and French, 1993, 2015; Carhart, 1997) factors are from Kenneth French's website. Additionally, we acquire the 153 factors constructed by Jensen, Kelly, and Pedersen (2023, JKP) from the authors' website. All returns are obtained in both daily and monthly frequencies. In total, we have around 1,644,000 stock-month observations in a full sample of 276 months from January 2000 to December 2022, or on average around 6,000 stock-month observations per month.

Each stock's exposure to a factor k in month t is

$$\widehat{\beta}_{i,k,t} := \frac{\widehat{\text{cov}}_t(r_{i,t}, f_{k,t})}{\widehat{\text{var}}_t(f_{k,t})}, \qquad \forall i, t, k, \tag{5}$$

where the realized $\widehat{\text{cov}}_t$ and $\widehat{\text{var}}_t$ are sample covariance and variance, respectively, estimated with daily returns in a 12-month rolling window up to month t. We estimate factor exposures to all 159 Fama-French-Carhart and JKP factors across all stocks in our sample. Notice $\widehat{\beta}_{i,k,t}$ corresponds to the regression coefficient of a *single*-factor model.[14]

3.2 Constructing flow variables

We construct the stock-level flow $flow_{i,t}$ panel using the mutual fund flow-induced trading (FIT) metric, proposed by Coval and Stafford (2007), Froot and Ramadorai (2008), and Lou (2012).

We use the standard mutual fund data source but carefully clean data errors by cross-validating several sources. In particular, we obtain monthly mutual fund returns and characteristics from the CRSP Survivorship-Bias-Free Mutual Fund database and quarterly holdings data from the Thomson/Refinitiv Mutual Fund Holdings Data (S12). Our sample period spans from January 2000 through December 2022.[15] The mutual fund sample comprises both active and passive mutual funds. To ensure accuracy in our flow measure, we cross-validate mutual funds' monthly returns and total net assets (TNA) obtained from the CRSP database with corresponding data from Morningstar and Thomson/Refinitiv. In the process, we manually correct several data input inaccuracies. Details regarding this process are in Appendix A.1.

[14] This is different from the original Fama-MacBeth procedure, where the first stage is a multi-factor regression. A single-factor beta is simply the realized covariance normalized by scalar variance and offers two advantages. First, multi-factor regressions can be unreliable when the number of factors is only moderately high. Second, a single-factor beta, and consequently each factor's BTQ term, can be constructed independently of other factors in the model, allowing for a more convenient empirical procedure.

[15] The mutual fund industry witnessed significant growth and sustained inflows throughout the 1990s (Lou, 2012; Ben-David, Li, Rossi, and Song, 2022a). In the post-2000 era, the monthly flows of mutual funds have generally maintained relative stability, prompting us to start our sample period from 2000, aligning with Gabaix and Koijen (2022).

The $flow_{i,t}$ panel construction procedure has three steps. First, we compute dollar mutual fund flows following the standard procedure,

$$flow_{m,t}^{\{\text{fund}\}} = \text{TNA}_{m,t} - \text{TNA}_{m,t-1}(1 + r_{m,t}^{\{\text{fund}\}}), \tag{6}$$

where $\text{TNA}_{m,t}$ is the total net assets of mutual fund m at the end of month t, and $r_{m,t}^{\{\text{fund}\}}$ is mutual fund m's net-of-fee return in month t.

Second, we allocate mutual fund flows to dollar stock-level flows, based on the established assumption in the literature that mutual funds buy or sell stocks in proportion to their prior holdings,

$$\widetilde{flow}_{i,t} = -\sum_{\text{fund } m} flow_{m,t}^{\{\text{fund}\}} weight_{i,m,\text{quarter}(t)-2}^{\{\text{fund}\}} \ . \tag{7}$$

The negative sign flips retail investors' outflows to sophisticated investors' inflows. In other words, a positive $\widetilde{flow}_{i,t}$ dollar number indicates that retail investors are selling stock i in month t, and sophisticated investors are buying. Moreover, we use the two-quarter-lagged mutual fund holding weight, denoted as $weight_{i,m,\text{quarter}(t)-2}^{\{\text{fund}\}}$. For instance, quarter(July) $-$ 2 = Q1.[16]

Lastly, we normalize stock-level dollar flows by the total US stock market capitalization from the preceding month,

$$flow_{i,t} = \frac{\widetilde{flow}_{i,t}}{\text{total stock market cap}_{t-1}}. \tag{8}$$

The normalization accounts for the significant upward trend in mutual fund dollar flows

[16]The use of a two-quarter lag deviates from the conventional one-quarter lag (Lou, 2012) to be more conservative and ensure that the constructed $flow_{i,t}$ is observable with information up to month t. In particular, mutual fund holding is reported with a maximum statutory delay of 45 days (Christoffersen, Danesh, and Musto, 2015), which means the end of Q2 holdings may not be observable in July. By using a two-quarter lag, July relies on the end of Q1 holdings, which are guaranteed to be available. Our results remain robust even when we apply the one-quarter lag commonly used in the literature. These results are available upon request.

by dividing by the total stock market capitalization. The normalization is the same across stocks i, so the cross-sectional variation is preserved. It reflects the view that the risk-bearing capacity of sophisticated investors is growing proportionally with the total stock market capitalization.

3.3 Constructing quantity (q) variables

The construction of $q_{k,t}$ has three steps. First, we aggregate stock-level flows into factor-level shocks, using the risk measures, $\widehat{\beta}_{i,k,t}, \widehat{\text{cov}}_t(r_{i,t}, f_{k,t}), \widehat{\text{var}}_t(f_{k,t})$, from Eq. 5,

$$\eta_{k,t} := \sum_i flow_{i,t} \widehat{\text{cov}}_t(r_{i,t}, f_{k,t}) = \sum_i flow_{i,t} \widehat{\beta}_{i,k,t} \widehat{\text{var}}_t(f_{k,t}), \qquad \forall k, t. \tag{9}$$

Second, these shocks are accumulated over time with a six-month moving average,

$$\widetilde{q}_{k,t} := \frac{1}{h} \sum_{h'=0}^{h-1} \eta_{k,t-h'}, \qquad \forall k, t, \qquad \text{with } h = 6. \tag{10}$$

Finally, the raw $\widetilde{q}_{k,t}$ is standardized as $q_{k,t} := \widetilde{q}_{k,t} / \sigma(\widetilde{q}_{k,t})$, where $\sigma(\widetilde{q}_{k,t})$ is the time-series standard deviation of $\widetilde{q}_{k,t}$ for each k evaluated over the full sample period. The standardization of $\widetilde{q}$ is for ease of interpreting its regression coefficients.

This construction procedure is intuitive but requires a few clarifications. First, stock-level flows are aggregated into the factor level by accounting for each stock's factor exposure, in a similar spirit to calculating the portfolio beta commonly used in risk management. The second expression in Eq. 9 is for explaining the intuition. We can interpret $flow_{i,t}$ as portfolio weights at month t. This is the marginal portfolio that the sophisticated investors add to their existing holdings in response to retail flows. This portfolio's risk characteristics are determined by its composition (portfolio weight $flow_{i,t}$), as well as each constituent stock's factor exposures ($\widehat{\beta}_{i,k,t}$) in aggregation. For example, if retail investors sell a large quantity of value stocks with high HML loadings, the sophisticated investors' HML risk quantity would experience a positive shock. In addition, $\widehat{\text{var}}_t(f_{k,t})$ modulates the portfolio's risk by the

Table 1: Summary statistics of flow-induced factor risk quantity $\widetilde{q}_{k,t}$ (unit: 10^{-5})

	Fama-French-Carhart factors				Across 153 JKP factors		
	MKT	SMB	HML	MOM	Q25	Median	Q75
Mean	0.33	0.05	0.15	-0.17	-0.06	-0.02	0.04
Std	2.15	0.33	0.74	0.94	0.26	0.45	0.88

Note: This table presents the mean and standard deviation of the flow-induced factor risk quantity $\widetilde{q}_{k,t}$ for the Fama-French-Carhart factors and JKP factors. The monthly observations span from January 2000 to December 2022.

time-series fluctuation in factor volatility. In this sense, we are indeed tracking the quantity of factor *risk*, not the physical quantity of securities or portfolios.

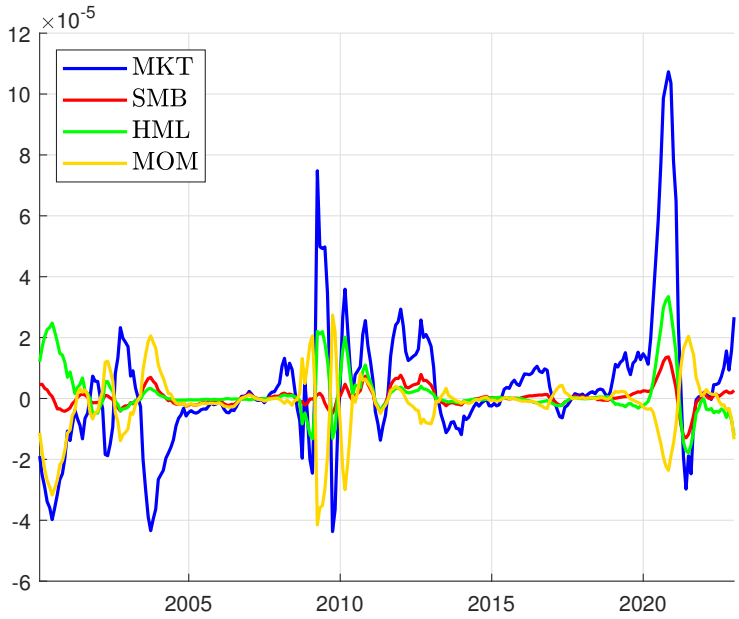
Second, the shocks $\eta_{k,t}$ are accumulated over time with a moving average to account for the persistent effects of older flows on future returns (Eq. 10). What matters for the expected return in month $t+1$ is the total risk quantity held at the end of month t, which is impacted by shocks in all previous periods, $\eta_{k,t}, \eta_{k,t-1}, \eta_{k,t-2}...$ The speed at which sophisticated investors can absorb the shocks and eliminate their effect on risk aversion is not our focus. We use the 6-month lagging moving average for simplicity and transparency to avoid a more involved parametric study of the speed. The empirical results are robust to alternative specifications.

Next, we present the summary statistics of the flow-induced factor risk quantity, $\widetilde{q}_{k,t}$, the central new variable introduced in this paper. Figure 1 shows the time-series plots of $\widetilde{q}_{k,t}$ for the Fama-French-Carhart (FF3C) factors. We plot the pre-standardized series $\widetilde{q}$ to show magnitudes.[17] Table 1 presents the full-sample statistics of FF3C $\widetilde{q}$ and summaries of these statistics across the 153 JKP factors.

Examining the basic time series properties of $\widetilde{q}_{k,t}$, we see the variation dominates its trend, so quantity fluctuation is the main feature compared to the secular trend of retail flows. The series have dynamic volatility clustering, similar to the more familiar factor

[17]The magnitudes of $\widetilde{q}$ are in the unit of 10^{-5}. The absolute level is irrelevant for empirical analysis, as the variables are standardized in regressions. To understand this magnitude, we know the monthly mutual fund flows are in the order of tens of billions of dollars, and the total market capitalization is in the order of tens of trillions of dollars (see Appendix Figure A.1). So the first term in (9) is in the order of 10^{-3} (given market β around 1). The last term, $\widehat{\text{var}}_t(f_{k,t})$ is in the order of 10^{-2}, so $\widetilde{q}$ is in the order of 10^{-5}.

Figure 1: Flow-induced factor risk quantity $\widetilde{q}_{k,t}$



Note: This figure plots the time series of the flow-induced factor risk quantity $\widetilde{q}_{k,t}$ for the Fama-French-Carhart factors. The monthly observations span from January 2000 to December 2022.

return time series.

Among the four factors plotted in Figure 1, MKT (in blue) has the most time-series variation in its quantity. The reason is that most stocks have positive market beta centered around one, so $\widetilde{q}_{\text{MKT},t}$ roughly aggregates the *overall* retail flows into (and out of) the entire mutual fund sector. In contrast, the three long-short factors have stock betas that are more evenly distributed around zero, so their $\widetilde{q}_{k,t}$ series reflect the *net* retail flows into (and out of) stocks of particular investment styles. For the same reason, these series are not mechanically correlated, even though they are all constructed from the same retail flow panel data. As a result, each long-short factor's $\widetilde{q}_{k,t}$ series, though less volatile, still offers valuable pricing information beyond that of the market.

Turning to the notable spikes in the plot, we note $\widetilde{q}_{\text{MKT},t}$ experiences significant increases during the Global Financial Crisis and the COVID-19 pandemic in the spring of 2020. These spikes are attributed to significant outflows from mutual fund investors during these periods. As a result, the sophisticated investors' risk holding quantity increases, making them

more "averse" to the market risk, which can be related to market crashes and subsequent rebounds. However, this is a highly simplified and anecdotal explanation of the main economic mechanism, as it does not consider cross-sectional variation in factor exposures, more nuanced fluctuations, or factors beyond MKT. Next, we turn to formal empirical analysis.

4 Empirical results

4.1 Security market line (SML) depends on quantity

The security market line (SML) is a simple and familiar tool to visualize the relationship between systematic risk exposure and expected return (β-$\mathbb{E}r$) in the cross-section of stocks without resorting to parametric modeling. We construct the empirical SML and the upgraded versions conditional on factor risk quantity q. We show the β-$\mathbb{E}r$ relationship is nearly flat unconditionally, which is consistent with the existing empirical results that factor exposure alone cannot adequately explain the cross-sectional variation in stock returns. However, once conditional on quantity information, the SML reveals interesting risk-return patterns that strongly support a risk-based explanation.

The unconditional SML is a non-parametric estimation of the β-$\mathbb{E}r$ relationship, which estimates the univariate regression function $\mathbb{E}_t[r_{i,t+1}] = Er(\beta_{i,k,t})$. We use a simple version of the kernel method by distributing the sample of stock-month observations into equally spaced bins according to $\widehat{\beta}_{i,k,t}$, and then calculate the average of $r_{i,t+1}$ within each bin. Notice, the return $r_{i,t+1}$ leads $\widehat{\beta}_{i,k,t}$ by one month, so it estimates conditional expected returns.

The upgraded SML conditional on quantity estimates the bi-variate function: $\mathbb{E}_t[r_{i,t+1}] = Er(\beta_{i,k,t}, q_{k,t})$, and the purpose is to show that q matters for the risk-return relationship. Again, we conduct a simple non-parametric estimation for transparency and intuitiveness. The estimation procedure is the same as the unconditional SML, but repeated in the subsamples consisting of months with high or low $q_{k,t}$, respectively, split at the time-series

median of $q_{k,t}$.[18]

Figure 2 presents single-factor models using the Fama-French-Carhart factors (MKT, SMB, HML, MOM), with black curves representing the unconditional SMLs, and red and blue for conditional on high and low $q_{k,t}$, respectively.

We find that the unconditional SML is nearly flat for the market factor, with a slight downward slope in the higher beta range. This implies that the market beta *alone* cannot explain the cross-sectional variation in expected returns, which is consistent with similar reports in the existing literature. Similar null results for unconditional SMLs are observed for SMB and MOM, while HML's SML is slightly upward sloping.

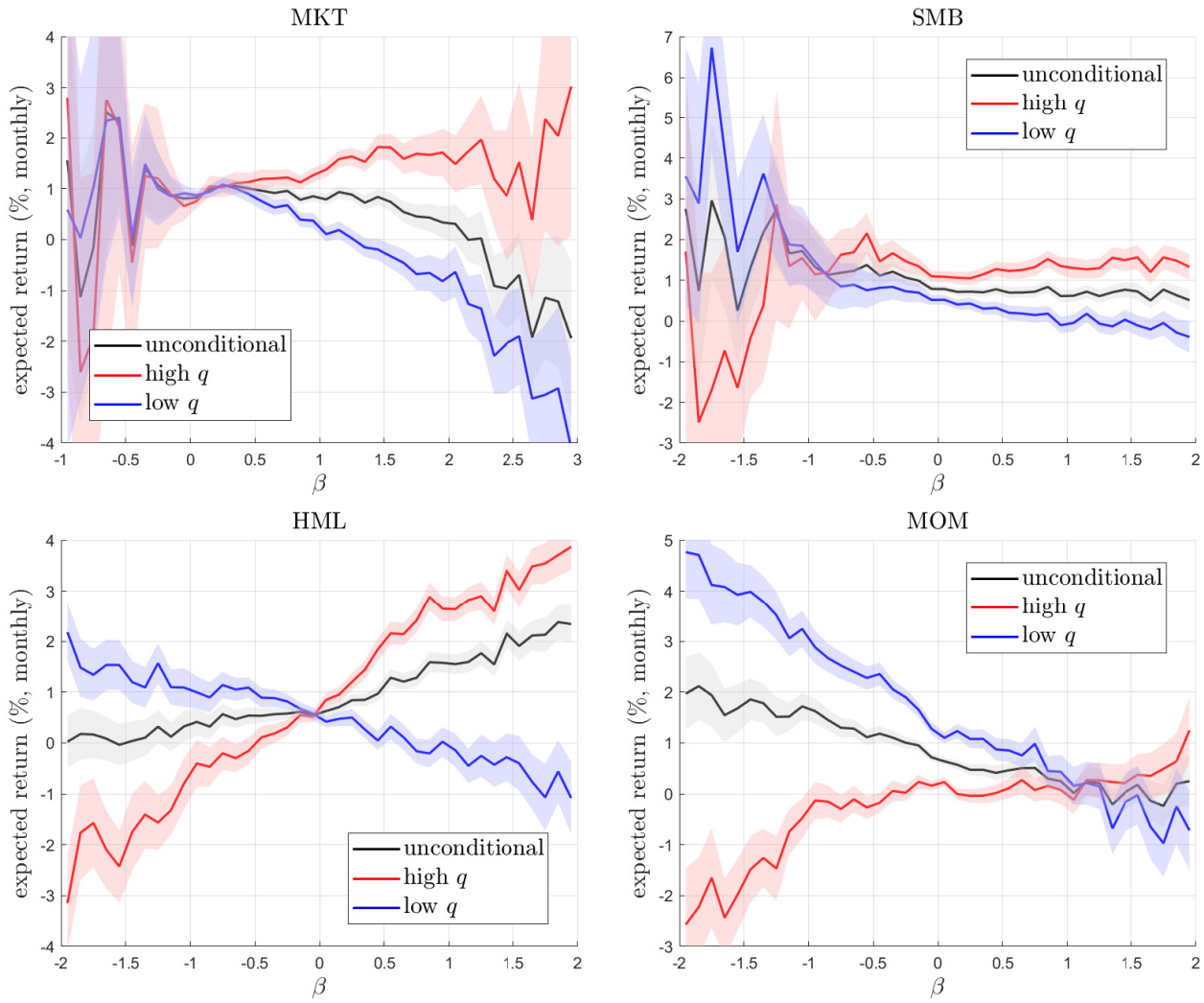
In contrast, the conditional SMLs show interesting risk-return patterns that are not observable without conditioning on q. The high-q SMLs exhibit a strong positive slope, while the low-q SMLs are downward sloping. The two conditional SMLs have distinct slopes, and the unconditional SML lies in between them as the mixed average. The positive high-q slope means the cross-sectional risk-return tradeoff is strong and positive, suggesting that sophisticated investors demand higher additional compensation for bearing high systematic risk investments in high-q environments. The negative low-q slope indicates a negative risk-return tradeoff, suggesting that investors are more willing to hold high-risk investments when they are required to sell lots of such stocks to retail traders in low-q months.[19] The salient gap in the slopes suggests that sophisticated investors' risk-holding conditions matter for their demand for risk, which has a significant impact on the pricing of risks in the cross-section.

The magnitude of q's effects is economically massive. For instance, a market beta-neutral stock has an expected return of around 0.9% per month regardless of q. In contrast, for a stock with a market beta of 1, the expected return can be as high as 1.4% in high-q months

[18]Formally, an unconditional bin is defined as $\{(n,t) \text{ s.t. } \widehat{\beta}_{n,k,t} \in [a,b)\}$; a "high q" bin is defined as $\{(n,t) \text{ s.t. } \widehat{\beta}_{n,k,t} \in [a,b) \text{ and } q_{k,t} \geq \text{median}(q_{k,t})\}$; and "low q" is the same "high q" but with "$\geq$" replaced by "$<$", where a and b are boundaries of the equally spaced bin and $\text{median}(q_{k,t})$ is the time-series median of $q_{k,t}$.

[19]The negative low-q slope is puzzling in the sense that it suggests a risk preference (rather than aversion) in low-q months. The frictions regarding sophisticated investors' risk management as described in Frazzini and Pedersen (2014) can be a potential explanation, which is an interesting direction for future research.

Figure 2: Security market line (SML) conditioning on quantity, $\mathbb{E}[r_{i,t+1}|\beta_{i,k,t}, q_{k,t}]$



Note: Security market line (SML) plots expected stock returns against β. The unconditional SML (black): bin the stock-month observations into equally-spaced intervals of $\beta_{i,k,t}$ (each with length 0.1) and plot the average return $r_{i,t+1}$ within each bin. The conditional SMLs (red for high q, blue for low q): the same process but in sub-samples split by the time-series median of $q_{k,t}$. The shaded bands are 95% confidence intervals.

or 0.3% in low-q months, with the average being still around 0.9%. The high v.s. low-q gap is around 1% per month or more than 10% annualized. Such a gap is even greater for higher market β stocks. For HML, the gap for a $\beta_{\text{HML}} = 1$ stock is around 30% annualized, while the HML-neutral stock does not depend on q. However, for the SMB factor, neither SMB's β nor its q significantly affects expected returns. This suggests that SMB is not a priced factor. We further support these findings and present more exact point estimates with parametric

estimations further below.[20]

All the SMLs are roughly close to straight lines regardless of their slopes, especially in the center range of β where the majority of stocks lie and sampling noise is less pronounced. This linearity in β is consistent with the cross-sectional law of one price (LOOP), although the slope (risk premium) can vary significantly with the q condition. Next, we specify the linearity in β, while still leaving the effect of q non-parametric in an upgraded Fama-MacBeth regression framework.

4.2 Fama-MacBeth factor premium increases with quantity

We specify a linear relationship between factor exposure (β) and expected return, where the linear coefficient (factor premium) is allowed to vary with quantity: $Er(\beta_{i,k,t}, q_{k,t}) = \beta_{i,k,t}\mu_k(q_{k,t})$.

To estimate this model, the first stage of Fama-MacBeth regressions provides factor risk exposures $\widehat{\beta}_{k,i,t}$ from time-series regression (already detailed in Section 3.1). The second stage of the Fama-MacBeth regression runs cross-sectional regression for each t:

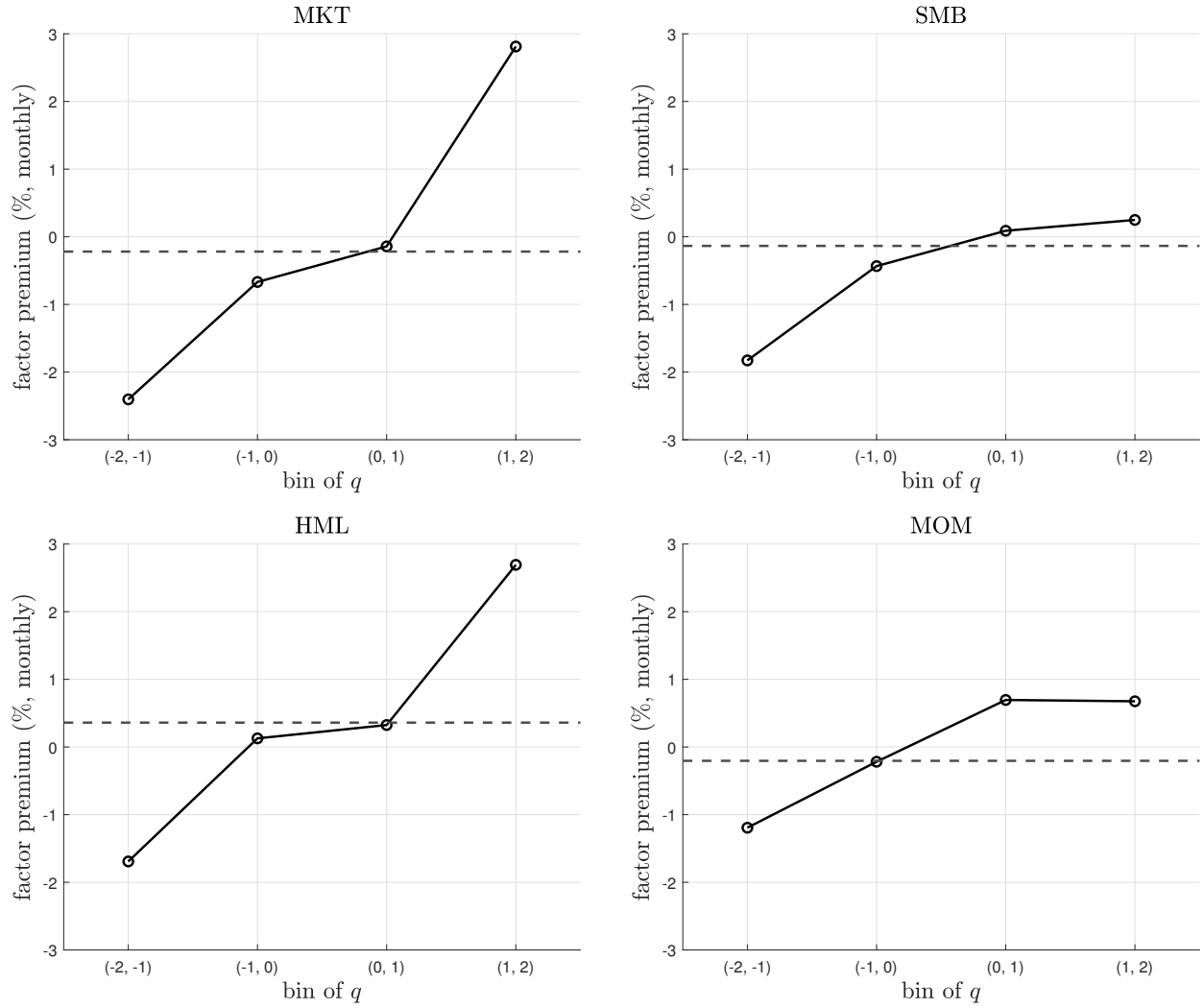
$$r_{i,t+1} = \gamma_{k,t+1}\widehat{\beta}_{i,k,t} + error_{i,t+1}, \qquad \forall i, \tag{11}$$

where $\gamma_{k,t+1}$ is the Fama-MacBeth factor mimicking portfolio (FMP) return. Canonically, the factor premium is estimated as the time-series average of $\gamma_{k,t+1}$. It measures the average cross-section association between factor loading and stock return. It is often cited as evidence against factor pricing because the unconditional Fama-MacBeth factor premium is close to zero (Lopez-Lira and Roussanov, 2020).

The innovation is that we estimate the mean of $\gamma_{k,t+1}$ conditional on $q_{k,t}$. To this end, we form four unit bins of $q_{k,t}$ (which is standardized) and calculate the average of $\gamma_{k,t+1}$ in

[20]It is also interesting to note that the crossings of the high-q and low-q SMLs are at around $\beta = 0$ for MKT, SMB, and HML, but not for MOM. Crossing at $\beta = 0$ is more consistent with the risk-based explanation because systematic risk-neutral stocks' expected returns should be irrelevant to fluctuations in factor quantity q. The fact that MOM's crossing occurs around $\beta_{\text{MOM}} = 1$ is an interesting discovery emerging from the new method and warrants further investigation.

Figure 3: Fama-MacBeth factor premium conditioning on quantity, $\mu_k(q_{k,t})$



Note: Fama-MacBeth factor mimicking portfolio returns (FMP, $\gamma_{k,t+1}$) averaged unconditionally (dashed line) and averaged within unit bins of $q_{k,t}$ (solid line).

each bin. Figure 3 presents the conditional (solid lines) and the unconditional (dashed lines) factor premiums for the four single-factor specifications.

The plot shows strong and consistent evidence that the Fama-MacBeth factor premium is not zero but increasing in factor quantity $q_{k,t}$. Specifically, the cross-section risk-return relationship is strong and positive when quantity $q_{k,t}$ is high. And the factor premium is negative when $q_{k,t}$ is low, suggesting that the risk-return tradeoff is reversed in low q environments. On average, the unconditional premium is close to zero, which only tells a small part of the interesting story that reveals itself when we condition on quantity.

The increasing relationship in $\mu_k(q_{k,t})$ is consistent across the four factors, although the market factor exhibits the most substantial variation. The market factor premium varies from less than -2% per month when market q is in the lowest $(-2,-1)$ standard deviation range to nearly $+3\%$ per month when market q is in the $(1,2)$ range. Consistent with the SML results, the magnitude of factor premium fluctuation driven by q can reach double-digit annualized percentage points.

4.3 Beta times quantity (BTQ) forecasts individual stock returns

The empirical results so far with non-parametric plots show factor risk quantity q significantly affects the cross-sectional risk-return relationship and each stock's expected return. Next, we turn to the parametric BTQ model, which allows us to include multiple factors, provide more formal point estimates, and conduct OOS model fit evaluation and factor selection tests. We show the BTQ model provides a compelling explanation for the expected return of individual stocks.

Once factor premium function $\mu_k(q_{k,t})$ is specified as the linear form $\mu_k(q_{k,t}) = \lambda_k q_{k,t}$, we arrive at the parametric BTQ model, which is estimated as the following panel-wise return predictive regression:

$$r_{i,t+1} = \sum_{k=1}^{K} \lambda_k q_{k,t} \widehat{\beta}_{i,k,t} + error_{i,t+1}, \qquad \forall i, t. \tag{12}$$

We compare it with the "β-only" model, which is implied by a constant factor premium μ_k:

$$r_{i,t+1} = \sum_{k=1}^{K} \mu_k \widehat{\beta}_{i,k,t} + error_{i,t+1}, \qquad \forall i, t. \tag{13}$$

We first present the results of the single-factor predictive regressions ($K = 1$) with $k =$ each of the four Fama-French-Carhart factors (MKT, SMB, HML, MOM) and the 153 JKP factors (Table 2).

The key finding is that the BTQ model provides a significantly better fit for stock returns

Table 2: Predicting stock returns with and without quantity, single factor

	Fama-French-Carhart factors				Across 153 JKP factors		
	MKT	SMB	HML	MOM	Q25	Median	Q75
Panel A: IS R^2 comparison, full sample 2000-2022 (%)							
BTQ	0.99	0.30	0.96	0.88	0.38	0.59	0.92
β-only	0.07	0.07	0.10	0.05	0.02	0.07	0.10
Panel B: OOS R^2 comparison, evaluation window 2010-2022 (%)							
BTQ	0.84	0.58	0.86	0.64	0.21	0.39	0.69
β-only	0.11	0.00	0.09	0.03	-0.03	0.04	0.13
Panel C: full-sample coefficient comparison: 2000-2022							
BTQ							
λ_k	1.78	0.72	1.45	1.74	0.61	0.96	1.45
t-stat	(4.14)	(2.75)	(3.44)	(3.30)	(2.16)	(2.92)	(3.65)
β-only							
μ_k	0.48	0.37	0.50	-0.46	-0.32	-0.12	0.24
t-stat	(1.34)	(1.48)	(1.54)	(-1.13)	(-1.53)	(-0.59)	(1.12)

Note: BTQ and β-only return predictions (Eq. 12 and 13), single-factor models ($K = 1$). The first four columns repeat the same prediction exercises with k = MKT, SMB, HML, and MOM, respectively. The last three columns report the summary statistics across the 153 JKP factors. The t-statistics (in parentheses) are calculated using standard errors clustered by month. Return prediction R^2 is calculated without demeaning ($R^2 := 1 - \sum_{i,t} (r_{i,t+1} - \widehat{r}_{i,t+1})^2 / \sum_{i,t} r_{i,t+1}^2$, where $\widehat{r}_{i,t+1}$ is predicted return) throughout the paper following Gu, Kelly, and Xiu (2020).

than the β-only model. This improvement in R^2 is substantial and remains consistent across different factor choices and in both in-sample and out-of-sample evaluations.[21] Even with only one factor, the BTQ model's OOS return predictive R^2 is around 0.8% for MKT and HML, which are among the ones with a high model fit in the 153 JKP factors. The median OOS R^2 across the 153 JKP factors is around 0.6%, and 140 out of the 153 factors have an OOS R^2 above 0. This level of predictability is economically significant and comparable to unstructured machine learning models that use a large number of firm characteristics

[21] For OOS evaluations, we estimate the model parameters (λ_k and μ_k) using the sample period from 2000 to 2009 and apply these estimates to calculate the OOS R^2 for the period from 2010 to 2022.

to predict stock returns. The state-of-the-art machine learning models typically achieve an OOS R^2 of around 1% to 2% in the literature. In contrast, the β-only models have a low R^2 around 0, and 50 out of the 153 JKP factors have a negative OOS R^2.

Turning to the coefficients estimates, the BTQ model's λ_k are significantly positive for all four Fama-French-Carhart factors and all of the 153 JKP factors. The economic magnitude of the λ_k estimates is substantial. For example, $\lambda_{\text{MKT}} = 1.78\%$, meaning for one standard deviation increase in market risk quantity, the expected return of a stock with a market beta of 1 increases by 1.78% per month, or $1.78\% \times 2 = 3.56\%$ per month for a stock with a market beta of 2, so on and so forth. In contrast, the β-only model's μ_k coefficients are mostly statistically insignificant from zero, and 88 out of the 153 JKP factors even have negative coefficient point estimates.

In summary, the single-factor results show the BTQ model already reliably predicts stock returns, the coefficients are consistent with the risk-based explanation, and the β-only model fails in both model fit and coefficient estimates.

In addition, Appendix Table A.1 presents an incidental empirical finding: each factor's return $f_{k,t+1}$ is predictable by its quantity $q_{k,t}$, with the predictive coefficients predominantly positive and statistically significant. As discussed in Section 2.2, while this time-series predictability is consistent with the BTQ model's cross-sectional return predictability, it is a much weaker result to argue for quantity's pricing power and peripheral to our research focus.

Moving onto multi-factor models, Table 3 presents the results for these models while maintaining a relatively low dimensionality with $K \leq 6$. This is achieved by using various combinations of the Fama-French-Five-Carhart factors. The BTQ model still significantly outperforms the β-only model in all multi-factor specifications. Allowing multiple factors further boosts BTQ's predictive accuracy with the best OOS R^2 values reaching above 1%. In contrast, the β-only model still barely predicts stock returns with low R^2 values even in sample.

Table 3: Predicting stock returns with and without quantity: multi-factor models

	CAPM	FF3	FF3C	FF5	FF5C
	$K = 1$	3	4	5	6
Panel A: IS R^2 comparison, full sample 2000-2022 (%)					
BTQ	0.99	1.13	1.15	1.14	1.17
β-only	0.07	0.18	0.21	0.18	0.23
Panel B: OOS R^2 comparison, evaluation window 2010-2022 (%)					
BTQ	0.84	1.07	1.10	0.50	0.71
β-only	0.11	0.19	0.25	-0.20	-0.02
Panel C: coefficients, full sample 2000-2022					
BTQ, λ_k (%) and t-statistics in parentheses					
MKT	1.78	1.29	1.17	1.31	1.19
	(4.14)	(2.10)	(1.99)	(2.03)	(2.01)
SMB		-0.23	-0.16	-0.20	-0.10
		(-0.75)	(-0.58)	(-0.68)	(-0.37)
HML		0.77	0.47	0.78	0.48
		(1.35)	(0.65)	(1.52)	(0.70)
MOM			0.52		0.76
			(0.69)		(0.94)
CMA				0.08	0.06
				(0.29)	(0.22)
RMW				-0.11	-0.27
				(-0.32)	(-0.72)
β-only					
— see Appendix Table A.2 —					

Note: BTQ and β-only return predictions (Eq. 12 and 13). Same as Table 2 but with multi-factor models ($K \geq 1$). The coefficients of the β-only model are relegated to Appendix Table A.2.

In terms of factor importance, controlling for other factors' contributions, MKT is the most prominent with the highest and most statistically significant coefficients across all multi-factor models, even though λ_{MKT} attenuates when more factors are included. HML and MOM also have positive coefficients but are statistically insignificant. When these factors are added to the model, both IS and OOS R^2 increase, indicating that their BTQ terms provide additional predictive power, and that they are priced factors. SMB, CMA, and RMW's coefficients are near zero or negative, indicating they are not priced factors according

to the BTQ model. This is also consistent with the fact that the OOS R^2 drops when adding these factors to the model. The β-only model's μ coefficients are all insignificant from zero or negative. (These numbers are relegated to Appendix Table A.2.)

Comparing BTQ's IS v.s. OOS model fits, we see slight reductions in R^2 when moving from IS to OOS for CAPM, FF3, FF3C. This indicates mild overfitting or parameter instability issues. It underscores the robustness of the BTQ model's predictive power, especially considering the inherent difficulty of forecasting monthly stock returns due to the low signal-to-noise ratio in stock prices. When moving to FF5 and FF5C, the IS R^2 keeps increasing slightly, while the OOS R^2 reverses to lower values of 0.5% and 0.7%. These levels of prediction accuracy are still economically significant, but the gap between IS and OOS R^2 indicates an overfitting issue. It suggests the ordinary least squares (OLS) estimation method has limitations for moderately higher-dimensional BTQ models. The additional factors might be noisy or redundant and introduce sample estimation errors. Next, we adopt a regularization method to select factors from a much greater number of candidates.

4.4 Taming the factor zoo with BTQ

The proliferation of proposed factors challenges the asset pricing literature, and the BTQ model offers a new method to select factors. This method has stronger identification power and economic relevance than traditional factor premium tests.

To implement it, we use the same return prediction framework (Eq. 12) but overload it with a large number of proposed factors ($K = 159$, including six from FF5C and 153 from JKP). It is well expected that many of these factors are noisy or redundant when controlling for other factors for pricing stock returns. Therefore, we use the Lasso method to induce sparsity in the predictive model and filter out the factors that are not priced according to the BTQ model.

Lasso is a regularization method that adds a penalty term to the OLS objective function to shrink and threshold the coefficients towards zero. Specifically, the parameter estimates

solve the following optimization problem:

$$\min_{\lambda_1 \ldots \lambda_K} \frac{1}{2|\text{IS}|} \sum_{i,t \in \text{IS}} \left(r_{i,t+1} - \sum_{k=1}^{K} \lambda_k \widehat{\beta}_{i,k,t} q_{k,t} \right)^2 + \omega \sum_{k=1}^{K} \frac{1}{\sigma(\widetilde{q}_{k,t})} |\lambda_k| \,, \tag{14}$$

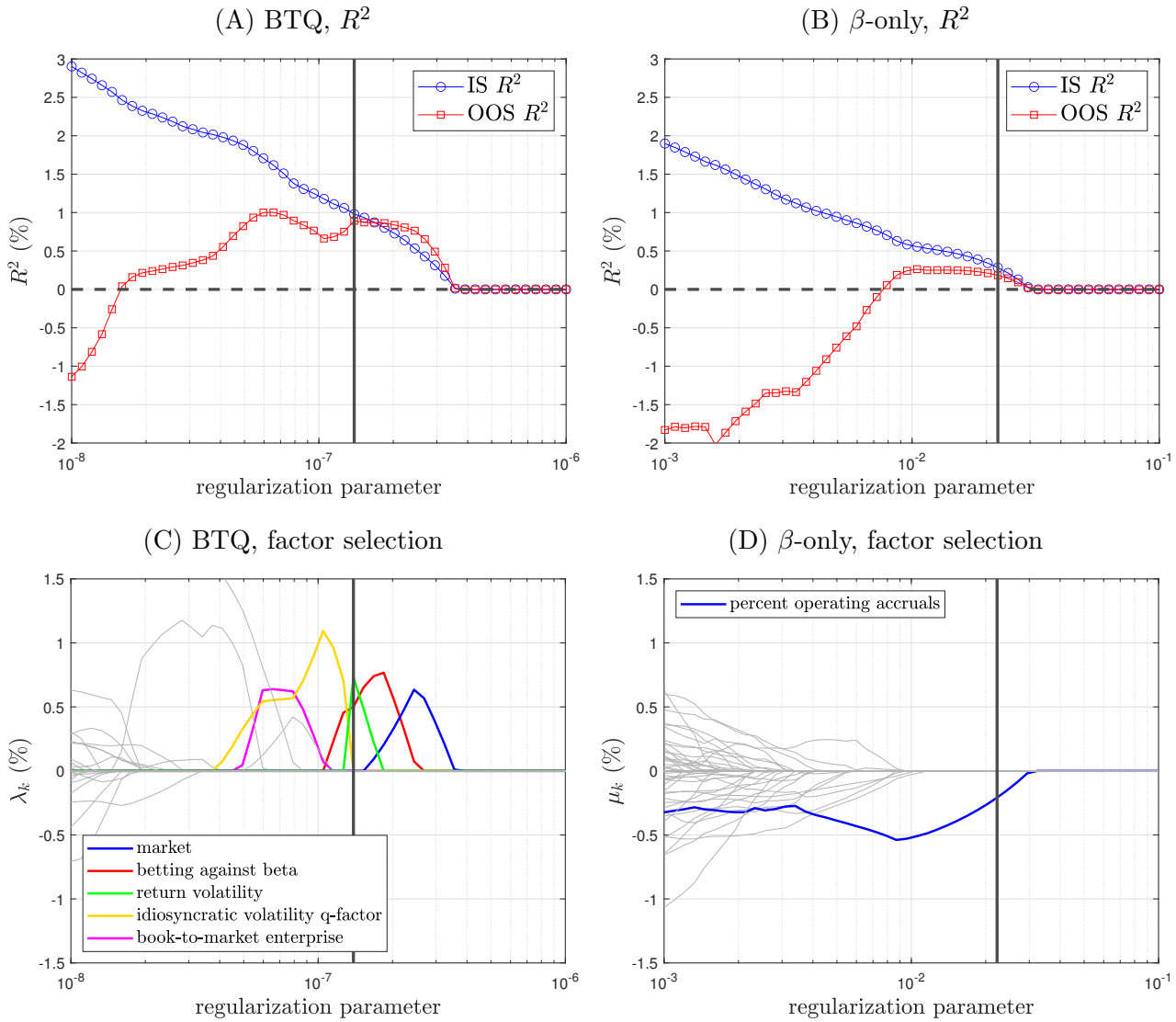
where |IS| is the number of stock-month observations in the training sample, and ω is the regularization parameter that controls the strength of the penalty term.[22]

Figure 4 plots the model fit and factor selection results for the BTQ and β-only models as the regularization parameter (ω) varies. (β-only model's Lasso implementation is similar, see technical details in Appendix A.2.) As ω increases, the fitted BTQ model becomes more parsimonious, as shown by the decreasing IS R^2 (Panel A blue curve) and the decreasing number of selected factors (those with non-zero λ_k in Panel C). This is the expected behavior of the Lasso method. More importantly, the OOS R^2 (Panel A red curve) displays a hump shape, with a broad and relatively stable peak that reaches around 1.0%. This suggests that the BTQ model's predictive power is strong and robust to choice of ω. In contrast, the β-only model's OOS R^2 never exceeds 0.3% and is only positive in a smaller range of ω values. This comparison once again highlights that quantity is essential for a risk-based explanation of expected stock returns.

The most important use of the BTQ + Lasso setup is a new way to investigate which factors are important for pricing stock returns. We find that only a few factors out of the factor zoo are sufficient for the models' high predictive power. The selected factors (those with non-zero λ_k when OOS R^2 peaks) are colored in Panel C. We find MKT is the first and most important factor, consistent with the observations in previous sections (4.1 to 4.3). The MKT factor is central to multi-factor pricing theory such as Merton's (1973) ICAPM model, and is historically the most important factor in workhorse empirical models such as the CAPM and Fama French. However, some research casts doubt on whether market

[22] The penalty on λ_k is normalized by the standard deviation of $\widetilde{q}_{k,t}$, so that the regularization is "fair" across different factors whose quantity $\widetilde{q}_{k,t}$ have different scales of fluctuations before standardization ($q_{k,t} = \widetilde{q}_{k,t}/\sigma(\widetilde{q}_{k,t})$). See technical details in Appendix A.2.

Figure 4: Return prediction with factor selection



Note: Model fit and parameter estimates as regularization parameter (ω, horizontal axis) varies. In Panels A, B, the IS R^2 is evaluated in the training window (2000-2009), and the OOS R^2 is the same models evaluated in the testing window (2010-2022). Panels C, D plots the parameter estimates from the training window, which are also brought out of the sample for evaluating the OOS R^2 in Panels A, B. The selected factors (colored curves) are, for BTQ: market (`mkt`), betting against beta (`betabab_126d`), return volatility (`rvol_21d`), idiosyncratic volatility from the q-factor model (`ivol_hxz4_21d`), and book-to-market enterprise value (`bev_mev`); for β-only, percent operating accruals (`oaccruals_ni`). The unselected factors are in gray. The vertical black line indicates the tuned ω based on ten-fold cross-validation (see Appendix A.2 for tuning detail).

beta is indeed related to expected returns (Black, 1972; Black, Jensen, and Scholes, 1972; Frazzini and Pedersen, 2014). Our results show that the market factor *equipped with quantity* variation is still very effective for explaining expected stock returns. However, this conclusion

cannot be achieved with β-only models.

The other selected factors include three based on technical information, betting against beta, return volatility, and idiosyncratic volatility from Hou, Xue, and Zhang's (2015) q-factor model, and one based on fundamental information, book-to-market enterprise value (which is a variant of the HML factor). These are among the usual suspects in the literature, while our results reinforce their importance when considering quantity. On the other hand, SMB and other factors related to size are dismissed by the Lasso selection.

Moreover, notice that the λ estimates of these selected factors from the BTQ model are all positive, which is consistent with the risk-based explanation as discussed in Section 2.2.

The β-only model only selects one factor, percent operating accruals (Panel D). It has a negative coefficient, which is inconsistent with the risk-based explanation. We believe this is not a reliable result in a mis-specified model, as indicated by β-only's low model fit.

Additionally, choosing ω based on the OOS R^2 peak is sufficient for the purpose of interpreting the BTQ model's factor selection. However, for the purpose of forecasting stock returns, it has the look-ahead bias. To overcome the problem, we provide the tuned ω using only IS information via ten-fold cross-validation, shown as the vertical black line (see technical details in Appendix A.2). The IS tuned ω is close to the OOS R^2 peak, suggesting the robustness of prediction and selection results.

4.5 BTQ with latent factors

Latent factors estimated using statistical methods to fit the realized time-series variation of returns have shown superior explanatory power for expected returns.[23] We demonstrate the BTQ framework can be applied to latent factors as well, and it leads to a strong two-factor structure with high predictive power for stock returns that is unattainable with the β-only counterpart.

We extract the principal components of the factor zoo portfolio returns, which are the

[23]See, for example, Kozak, Nagel, and Santosh (2020), Kelly, Pruitt, and Su (2019), and Lettau and Pelger (2020).

linear combinations of the factor returns that capture the most time-series variation.[24] Then, we construct $\widehat{\beta}$ and, in turn, quantity q with respect to each of these PC factors from scratch following the same procedure reported in Section 3. Based on these variables, we conduct the same BTQ predictive regression with Lasso as in the previous section. The new set of $\widehat{\beta}$ and q variables provides some external validation of our method's robustness and generalizability.

Figure 5 shows that the BTQ model with PC factors has a strong predictive power for stock returns, with the OOS R^2 peaking at around 1.0%, similar to the previous Figure 4 with the original factors. The high OOS R^2 is, once again, robust to the choice of ω, shown with the broad peak of the OOS R^2 hump-shaped curve. In contrast, the β-only model with PC factors hardly delivers any predictive power, with the OOS R^2 less than 0 for almost all ω values.

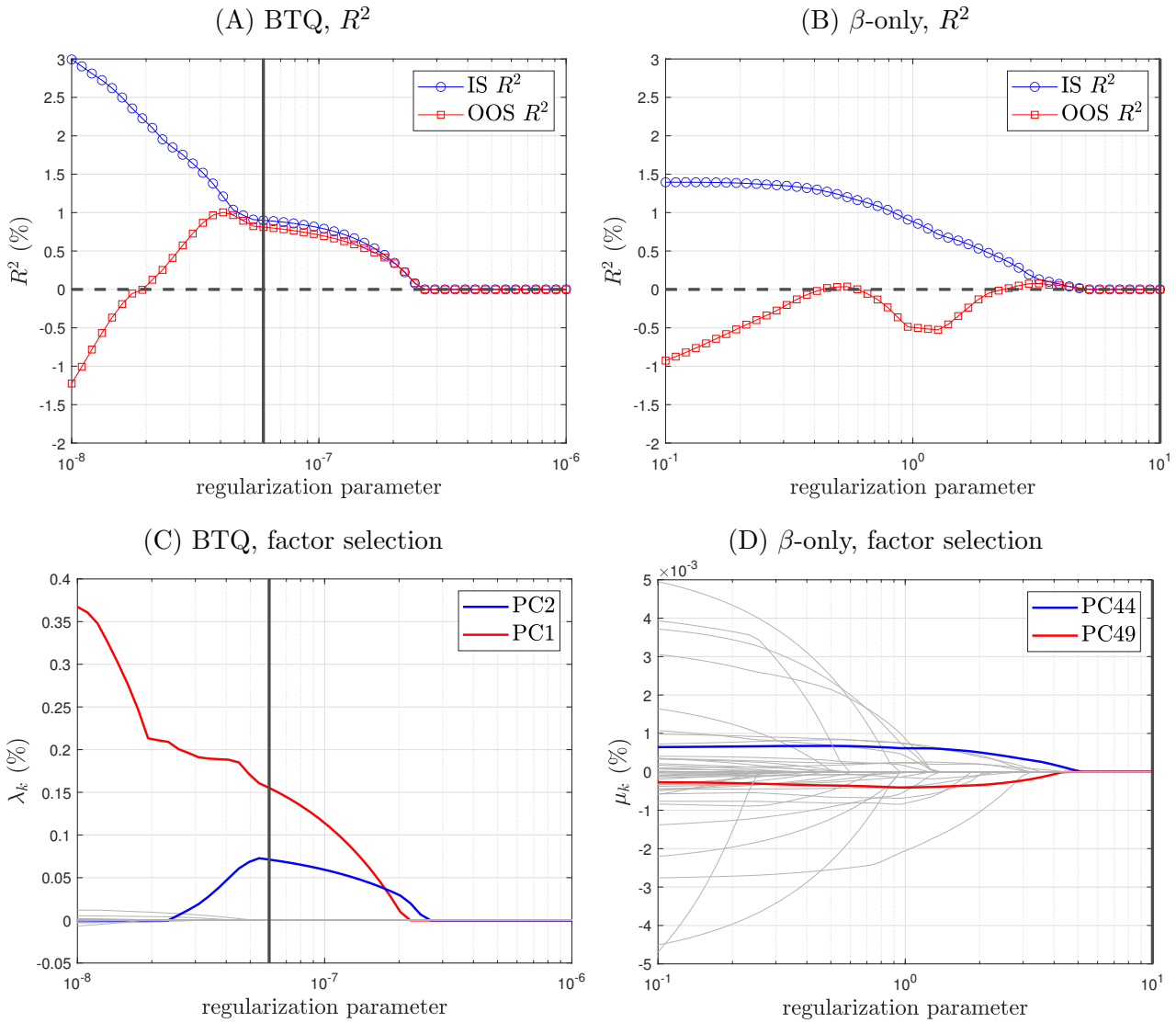
More importantly, Panel C shows a strong two-factor structure, with PC1 and PC2 selected as the most important factors for predicting stock returns. The magnitude of their λ estimates dominates the subsequent PC factors (gray curves). The parsimonious structure attained with the BTQ model with latent factors can explain expected stock returns well with high OOS R^2. This is consistent with the literature that suggests latent factors are helpful to "shrink the cross section" and reduce the dimensionality of the factor zoo.[25]

Notice, once again, the signs of the λ estimates for the selected factors, PC1 and PC2, are both positive. This is required by the risk-based theory no matter how the signs of the PCs are specified. In contrast, the β-only model's selection and parameter estimates do not show a discernible pattern, which we believe are mostly estimation noise under the mis-specified model.

[24] Specifically, we use the first 50 principal components estimated from the monthly returns of the FF5C and 153 JKP factors over the period from 1970 to 2009.

[25] See Kozak et al. (2020). Note that Kozak et al. (2020) conduct asset pricing tests on various portfolios as test assets, whereas our test assets are individual stocks, which arguably clears a higher bar.

Figure 5: Return prediction with factor selection: latent factors



Note: Model fit and parameter estimates as regularization parameter (ω, horizontal axis) varies. In Panels A, B, the IS R^2 is evaluated in the training window (2000-2009), and the OOS R^2 is the same models evaluated in the testing window (2010-2022). Panels C, D plots the parameter estimates from the training window, which are also brought out of the sample for evaluating the OOS R^2 in Panels A, B. We perform Lasso regression using the first 50 principal components derived from the monthly returns of the FF5C and JKP factors over the period from 1970 to 2009. The unselected factors are in gray. The vertical black line indicates the tuned ω based on ten-fold cross-validation (see Appendix A.2 for tuning detail).

5 Conclusion

This paper considers a new but important aspect of risk's economic role in determining asset prices—the *quantity* variation in investors' risk holdings induced by trading flows. The

economic rationale is simple: when sophisticated investors hold more of a systematic risk factor, they require greater compensation for bearing that risk, which in turn drives the expected return of every stock exposed to the factor. Yet the empirical model yields a compelling risk-based explanation for expected stock returns.

We show incorporating quantity into canonical factor pricing has important implications for asset pricing studies in three main aspects of new findings. First, quantity variation elicits risk-return tradeoff relationships, which have been hard to capture with β only and cast doubt on whether risk is the main driver of expected returns. We find the cross-sectional association between factor exposures and expected returns (β-$\mathbb{E}r$) strongly depends on factor quantity variation, and the previous null result is a mixed average unconditional on quantity. Second, quantity enables a risk-based predictive model (termed beta times quantity, BTQ) for monthly stock returns. The model delivers high prediction accuracy in this hard empirical task dominated by unstructured machine learning models and firm characteristics. Third, quantity provides a new way for factor selection and thereby new answers to the factor zoo problem. Instrumenting factor premium with quantity variation has not only greater identification power but also more economic relevance than traditional factor premium tests. We find that a few factors out of the factor zoo are selected for the model's high predictive power, and in a latent factor setting, the first two principal components overwhelmingly dominate the remaining components.

This paper fits into the general agenda of using market trading and investor holding quantity data to better model investors' *demand* for risk and the resulting implication for asset prices. We provide a simple and actionable way to incorporate quantity into workhorse asset models, ranging from the security market line, Fama-MacBeth regressions, to predicting stock returns. We are confident that future research can similarly incorporate factor quantity variation into many existing asset pricing methods to yield new insights for various specific research questions.

References

Adrian, Tobias, Erkko Etula, and Tyler Muir, 2014, Financial intermediaries and the cross-section of asset returns, *Journal of Finance* 69, 2557–2596.

Barber, Brad M, Xing Huang, and Terrance Odean, 2016, Which factors matter to investors? evidence from mutual fund flows, *The Review of Financial Studies* 29, 2600–2642.

Ben-David, Itzhak, Jiacui Li, Andrea Rossi, and Yang Song, 2022a, Ratings-driven demand and systematic price fluctuations, *Review of Financial Studies* 35, 2790–2838.

Ben-David, Itzhak, Jiacui Li, Andrea Rossi, and Yang Song, 2022b, What do mutual fund investors really care about? *Review of Financial Studies* 35, 1723–1774.

Berk, Jonathan B, and Jules H Van Binsbergen, 2016, Assessing asset pricing models using revealed preference, *Journal of Financial Economics* 119, 1–23.

Black, Fischer, 1972, Capital market equilibrium with restricted borrowing, *Journal of business* 45, 444–455.

Black, Fischer, Michael C Jensen, and Myron Scholes, 1972, The capital asset pricing model: Some empirical tests, *Unpublished working paper* .

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

Choi, Darwin, Wenxi Jiang, and Chao Zhang, 2023, Alpha go everywhere: Machine learning and international stock returns, *Available at SSRN 3489679* .

Christoffersen, Susan Kerr, Erfan Danesh, and David K Musto, 2015, Why do institutions delay reporting their shareholdings? Evidence from form 13F, Working paper, University of Toronto.

Cochrane, John H, 2011, Presidential address: Discount rates, *The Journal of finance* 66, 1047–1108.

Coval, Joshua, and Erik Stafford, 2007, Asset fire sales (and purchases) in equity markets, *Journal of Financial Economics* 86, 479–512.

De Long, J. Bradford, Andrei Shleifer, Lawrence H. Summers, and Robert J. Waldmann, 1990, Noise trader risk in financial markets, *Journal of Political Economy* 98, 703–738.

Dou, Winston, Leonid Kogan, and Wei Wu, 2022, Common fund flows: Flow hedging and factor pricing, *Journal of Finance* Forthcoming.

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

Fama, Eugene F, and Kenneth R French, 2008, Dissecting anomalies, *The journal of finance* 63, 1653–1678.

Fama, Eugene F, and Kenneth R French, 2015, A five-factor asset pricing model, *Journal of Financial Economics* 116, 1–22.

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

Feng, Guanhao, Stefano Giglio, and Dacheng Xiu, 2020, Taming the factor zoo: A test of new factors, *Journal of Finance* 75, 1327–1370.

Feng, Guanhao, Jingyu He, and Nicholas G Polson, 2018, Deep learning for predicting asset returns, *arXiv preprint arXiv:1804.09314* .

Frazzini, Andrea, and Lasse Heje Pedersen, 2014, Betting against beta, *Journal of Financial Economics* 111, 1–25.

Freyberger, Joachim, Andreas Neuhierl, and Michael Weber, 2020, Dissecting characteristics nonparametrically, *The Review of Financial Studies* 33, 2326–2377.

Froot, Kenneth A, and Tarun Ramadorai, 2008, Institutional portfolio flows and international investments, *Review of Financial Studies* 21, 937–971.

Gabaix, Xavier, and Ralph SJ Koijen, 2022, In search of the origins of financial fluctuations: The inelastic markets hypothesis, Working paper, Harvard University.

Gabaix, Xavier, and Matteo Maggiori, 2015, International liquidity and exchange rate dynamics, *The Quarterly Journal of Economics* 130, 1369–1420.

Garleanu, Nicolae, Lasse Heje Pedersen, and Allen M Poteshman, 2008, Demand-based option pricing, *The Review of Financial Studies* 22, 4259–4299.

Giglio, Stefano, Yuan Liao, and Dacheng Xiu, 2021, Thousands of alpha tests, *The Review of Financial Studies* 34, 3456–3496.

Giglio, Stefano, and Dacheng Xiu, 2021, Asset pricing with omitted factors, *Journal of Political Economy* 129, 1947–1990.

Greenwood, Robin, and Dimitri Vayanos, 2014, Bond supply and excess bond returns, *The Review of Financial Studies* 27, 663–713.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2020, Empirical asset pricing via machine learning, *The Review of Financial Studies* 33, 2223–2273.

Harvey, Campbell R, Yan Liu, and Heqing Zhu, 2016, ... and the cross-section of expected returns, *Review of Financial Studies* 29, 5–68.

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

Hou, Kewei, Chen Xue, and Lu Zhang, 2015, Digesting anomalies: An investment approach, *The Review of Financial Studies* 28, 650–705.

Hou, Kewei, Chen Xue, and Lu Zhang, 2017, A comparison of new factor models, *Fisher college of business working paper* 05.

Huang, Shiyang, Yang Song, and Hong Xiang, 2024, Noise trading and asset pricing factors, *Management Science* Forthcoming.

Jensen, Theis Ingerslev, Bryan Kelly, and Lasse Heje Pedersen, 2023, Is there a replication crisis in finance? *Journal of Finance* 78, 2465–2518.

Kang, Wenjin, K Geert Rouwenhorst, and Ke Tang, 2022, Crowding and factor returns, Working paper, Yale University.

Kelly, Bryan, Semyon Malamud, and Kangying Zhou, 2024, The virtue of complexity in return prediction, *The Journal of Finance* 79, 459–503.

Kelly, Bryan T, Seth Pruitt, and Yinan Su, 2019, Characteristics are covariances: A unified model of risk and return, *Journal of Financial Economics* 134, 501–524.

Koijen, Ralph SJ, Robert J Richmond, and Motohiro Yogo, 2023, Which investors matter for equity valuations and expected returns? *Review of Economic Studies* Forthcoming.

Koijen, Ralph SJ, and Stijn Van Nieuwerburgh, 2011, Predictability of returns and cash flows, *Annu. Rev. Financ. Econ.* 3, 467–491.

Koijen, Ralph SJ, and Motohiro Yogo, 2019, A demand system approach to asset pricing, *Journal of Political Economy* 127, 1475–1515.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, 2018, Interpreting factor models, *Journal of Finance* 73, 1183–1223.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, 2020, Shrinking the cross-section, *Journal of Financial Economics* 135, 271–292.

Lettau, Martin, and Markus Pelger, 2020, Factors that fit the time series and cross-section of stock returns, *The Review of Financial Studies* 33, 2274–2325.

Lewellen, Jonathan, 2014, The cross section of expected stock returns, *Forthcoming in Critical Finance Review, Tuck School of Business Working Paper* .

Li, Jiacui, 2022, What drives the size and value factors? *Review of Asset Pricing Studies* 12, 845–885.

Li, Jiacui, and Zihan Lin, 2022, Prices are less elastic at more aggregate levels, Working paper, University of Utah.

Lopez-Lira, Alejandro, and Nikolai L Roussanov, 2020, Do common factors really explain the cross-section of stock returns?, *Jacobs Levy Equity Management Center for Quantitative Financial Research Paper* .

Lou, Dong, 2012, A flow-based explanation for return predictability, *Review of Financial Studies* 25, 3457–3489.

McLean, R David, and Jeffrey Pontiff, 2016, Does academic research destroy stock return predictability? *The Journal of Finance* 71, 5–32.

Merton, Robert C, 1973, An intertemporal capital asset pricing model, *Econometrica* 867–887.

Rapach, David, and Guofu Zhou, 2013, Forecasting stock returns, in *Handbook of economic forecasting*, volume 2, 328–383 (Elsevier).

Ross, Stephen A, 1976, The arbitrage theory of capital asset pricing, *Journal of Economic Theory* 13, 341–60.

Rostek, Marzena J, and Ji Hee Yoon, 2023, Imperfect competition in financial markets: Recent developments, Working paper, University of Wisconsin - Madison.

Shleifer, Andrei, and Robert W. Vishny, 1997, The limits of arbitrage, *Journal of Finance* 52, 35–55.

Teo, Melvyn, and Sung-Jun Woo, 2004, Style effects in the cross-section of stock returns, *Journal of Financial Economics* 74, 367–398.

Vayanos, Dimitri, and Jean-Luc Vila, 2021, A preferred-habitat model of the term structure of interest rates, *Econometrica* 89, 77–112.

Warther, Vincent A., 1995, Aggregate mutual fund flows and security returns, *Journal of Financial Economics* 39, 209–235.

Welch, Ivo, and Amit Goyal, 2008, A comprehensive look at the empirical performance of equity premium prediction, *The Review of Financial Studies* 21, 1455–1508.

Appendix

A Additional technical details

A.1 Construction and cleaning of mutual fund flows

In this appendix, we present details involved in constructing and cleaning mutual fund flows.

Our primary data source is the CRSP Survivorship-Bias-Free Mutual Fund database. We start with all funds' return and total net assets (TNA) data at the share-class level. A mutual fund may include multiple share classes. We first drop observations with no valid CRSP identifier, `crsp_fundno`. A few fund-share classes report multiple TNAs in a given month. These are likely data duplicates, so we keep only the first observation of the month. In what follows, we discuss the cleaning steps for returns and TNA at the share-class level. After cleaning, we aggregate the share-class level data to the fund level.

A.1.1 Return cleaning

We first correct data errors in monthly net returns, `mret`.

First, we address extremely positive returns. We study the case in which a particular fund share has returns greater than 100% and has existed for more than one year.[26] We manually check the entire time series of each share class in this subsample. Most of these extreme returns reflect misplaced decimal points, which confound returns in decimal and percentage formats. For these cases, we divide the faulty returns by 100.

Second, we address extreme negative returns. Similarly, we study the case in which a particular fund share has existed for more than one year and has returns lower than -50%. With extremely negative returns, we need to distinguish data errors from significantly negative returns before a fund's closure. Thus, we manually check only the subsample of

[26] We use the one-year threshold because mutual fund return and TNA during the first year are sometimes inaccurate in the CRSP database. For example, return and TNA can be stale or reported using a placeholder number such as 0.1. We address these issues by cross-checking with the alternative database.

negative returns that occur at least one year prior to the last observation of a closed fund. We manually check whether these extreme returns reflect data-input errors for this subsample. For the cases with misplaced decimal points, we divide the faulty returns by 100.

A.1.2 TNA cleaning

Unlike many prior studies that construct percentage mutual fund flows, we study dollar-value flows to preserve the cross-sectional relative magnitudes. The mutual fund size distribution features a very long right tail. Winsorizing the extreme dollar-value TNA likely removes both valid large values and input errors, generating significant bias. We devise an algorithm to identify and correct erroneous observations of TNA:

1. Using the sample with corrected returns, we calculate dollar flows as in (6) at the share-class level.

2. We study the top and bottom 0.5% of all dollar flows.[27] We manually check this subsample's TNA time series of all share classes. We identify several common errors:

   - Misplaced decimal points (usually by hundredths or thousandths).
   - Stale TNA observations from CRSP, typically when a fund reorganizes its share class offering (e.g., adding a new share class and moving assets into a single share class).
   - CRSP sometimes sets TNA = 0.1 for the first few months of a new fund or a new share class.

   We correct the misplaced decimal issue. For funds suffering from the latter two problems, we obtain their TNA from Morningstar.[28] Morningstar's TNA data (`Net_Assets_ShareClass_Monthly`) suffer to a lesser extent from these issues than

[27] The choice of the top and bottom 0.5% is motivated by the distribution of dollar flows, where most extreme values tend to occur at these tails.

[28] We merge the CRSP and Morningstar databases using a fund's ticker.

CRSP's TNA data. We conclude by further cross-checking other third-party vendors (e.g., Yahoo Finance and Bloomberg Terminal). Hence, whenever a fund's CRSP TNA deviates more than 50% from its Morningstar TNA, we use the Morningstar TNA.

3. We repeat the previous steps one more time to ensure that we have accounted for most, if not all, extreme errors.

4. We compare our cleaned TNA with total assets (`assets`) from Thomson/Refinitiv Holdings data.[29] Following Coval and Stafford (2007) and Lou (2012), we drop observations whenever our cleaned TNA deviate more than 50% from `assets` from Thomson/Refinitiv.

Using cleaned return and TNA data, we calculate dollar flows at the share-class level using equation (6). We obtain a fund's flows by adding up the flows of all share classes in the same fund. The final sample contains 1,707,742 fund×month observations.

A.1.3 Cross-validating the data-cleaning procedure

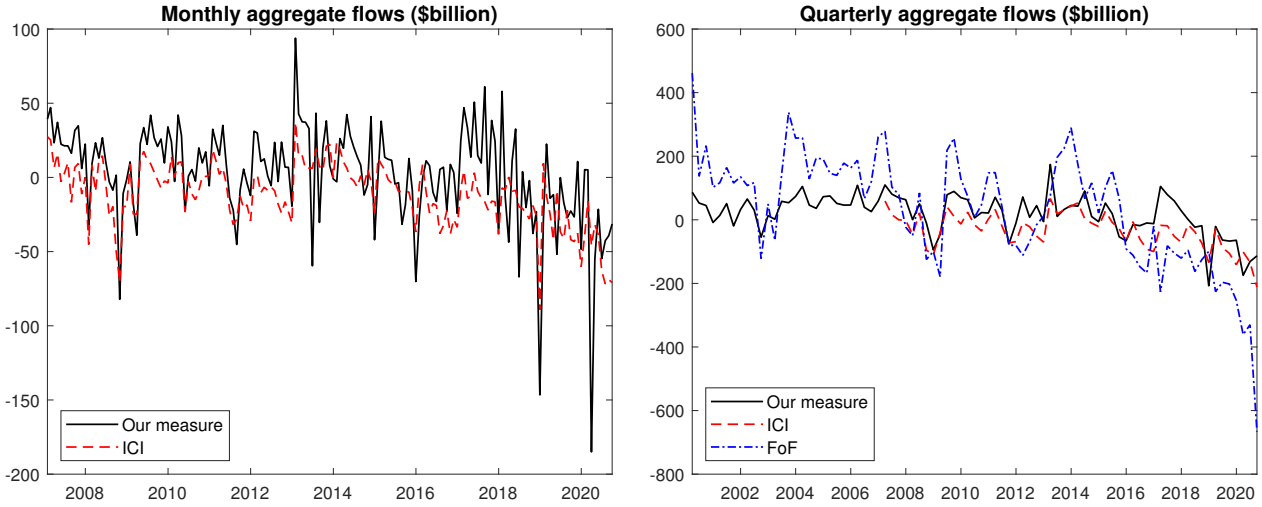
We cross-validate our data-cleaning procedure. We compute the aggregate mutual fund flows in dollar amounts each month. We compare our aggregate flow measures with alternative sources, including the Investment Company Institute (ICI) and the Flow of Funds (FoF).

The ICI provides aggregate monthly mutual fund flows. We obtain a version of ICI aggregate flows data from 2007 to 2020. We use the ICI's Total Equity mutual fund flows, which feature a close coverage scope to mutual funds in our sample. The FoF data (now known as "Financial Accounts of the United States - Z.1") are published quarterly by the Federal Reserve Board. We use mutual fund flow (Line 28) of Corporate Equities (Table 223) and unadjusted flows (FU). We use the December 2021 vintage of the data because the Federal Reserve revises historical FoF data every quarter.

Figure A.1 plots the time series of aggregate mutual fund flows from various sources.

[29] We merge the two databases via the linking table MFLINKS, which WRDS provides.

Figure A.1: **Time series of aggregate mutual fund flows from various sources**



Note: The left panel plots the monthly time series of our measure of aggregate mutual fund flows and Investment Company Institute (ICI) flows. The right panel plots the quarterly time series of our measure, ICI flows, and Flow of Funds (FoF) flows.

The left panel shows the monthly time series of our measure and ICI flow, and the right panel shows the quarterly time series of all three sources. Our measure of aggregate mutual fund flows is broadly consistent with the other two sources. The correlation between our aggregate flow measure and ICI flow is 0.68 at the monthly level and 0.80 at the quarterly level. The correlation between our measure and FoF flow is 0.55 at the quarterly level.

The differences in Figure A.1 between the three measures of aggregate flows likely reflect differences in mutual fund coverage. Specifically, the ICI flow tracks virtually all U.S. equity mutual funds that invest in both domestic and world equity markets.[30] The FoF flow, sourced from unpublished ICI data, aggregates unadjusted flows into and out of all U.S. mutual funds (including variable annuity long-term mutual funds). The FoF flow is calculated based on mutual fund assets in common stock, preferred stock, and rights and warrants.[31] In comparison, our mutual fund sample contains U.S. mutual funds that CRSP covers. CRSP collects historical data from various sources.[32] Due to the nature of the data collection

[30]The ICI is a trade association for the mutual fund industry, and virtually all U.S. mutual funds are ICI members (Warther, 1995).

[31]See https://www.federalreserve.gov/apps/fof/SeriesAnalyzer.aspx?s=FA653064100&t=F.223&suf=Q.

[32]The sources include the Fund Scope Monthly Investment Company Magazine, the Investment Dealers Digest Mutual Fund Guide, Investor's Mutual Fund Guide, the United and Babson Mutual Fund Selector,

process, CRSP's coverage is smaller than ICI's coverage.

A.2 Technical details of Lasso implementation

In optimization (14), adding "$1/\sigma(\widetilde{q}_{k,t})$" is technically necessary because we have already standardized $\widetilde{q}_{k,t}$ to $q_{k,t} = \widetilde{q}_{k,t}/\sigma(\widetilde{q}_{k,t})$ (see Section 3.3). Optimization (14) is equivalent to running the vanilla Lasso on the pre-standardized BTQ ($\widehat{\beta} \times \widetilde{q}$)

$$\min_{\widetilde{\lambda}_1 \ldots \widetilde{\lambda}_K} \frac{1}{2|\text{IS}|} \sum_{i,t \in \text{IS}} \left(r_{i,t+1} - \sum_{k=1}^{K} \widetilde{\lambda}_k \widehat{\beta}_{i,k,t} \widetilde{q}_{k,t} \right)^2 + \omega \sum_{k=1}^{K} \left| \widetilde{\lambda}_k \right|, \tag{A.1}$$

and then standardizing the coefficients for economic interpretation: $\lambda_k = \widetilde{\lambda}_k \sigma(\widetilde{q}_{k,t})$. Although we standardize $\widetilde{q}_{k,t}$ for interpretability, we do not want to lose the information contained in the original quantity $\widetilde{q}_{k,t}$ during the Lasso selection. A factor with greater variation in $\widetilde{q}_{k,t}$ will have an inflated λ_k after standardizing to $q_{k,t}$, but we do not want to penalize it more for that reason. Standard Lasso implementation where the economic interpretation is not a concern would recommend standardizing the predictor (BTQ together) across the i, t panel. We are effectively creating a customized standardization based on the required economic interpretation.

Similarly, for the β-only model, the Lasso implementation is

$$\min_{\mu_1 \ldots \mu_K} \frac{1}{2|\text{IS}|} \sum_{i,t \in \text{IS}} \left(r_{i,t+1} - \sum_{k=1}^{K} \mu_k \widehat{\beta}_{i,k,t} \right)^2 + \omega \sum_{k=1}^{K} |\mu_k| . \tag{A.2}$$

We perform ten-fold cross-validation to tune hyperparameter ω based on only in-sample information (from 2000 to 2009). For each fold, we exclude one year of observations and solve the Lasso problem (A.1) using the remaining nine years of in-sample data. The model is evaluated in the left-out year to form predicted returns $\widehat{r}_{i,t+1}^{[\text{cv}]}$. After enumerating all folds and forming predicted returns for all in-sample observations, we calculate the cross-validated

and the Wiesenberger Investment Companies Annual Volumes.

Table A.1: Predicting factor return $f_{k,t+1}$ using quantity $q_{k,t}$

	Fama-French-Carhart factors				Across 153 JKP factors		
	MKT	SMB	HML	MOM	Q25	Median	Q75
λ_k (%)	1.04	0.49	0.82	1.10	0.25	0.66	1.00
t-stat	(3.25)	(2.45)	(2.89)	(1.76)	(1.40)	(2.00)	(2.64)
μ_k (%)	0.38	0.19	0.08	0.36	-0.20	-0.01	0.27
t-stat	(1.39)	(1.16)	(0.38)	(1.41)	(-1.41)	(-0.10)	(1.63)
R^2 (%)	5.05	2.48	5.59	4.35	1.50	4.74	7.35

Note: This table reports the results of the time-series regression $f_{k,t+1} = \lambda_k q_{k,t} + \mu_k + \epsilon_{k,t+1}$ for the Fama-French-Carhart factors and JKP factors, using monthly observations from January 2000 to December 2022. The table presents the point estimates, t-statistics based on Newey-West standard errors, and the ordinary R^2. The coefficients are expressed as percentages. For example, the first cell indicates that a one standard deviation increase in $q_{k,t}$ predicts a 1.04% increase in market return in the following month.

(CV) in-sample mean squared errors (MSE) as $\sum_{i,t\in \text{IS}} \left(r_{i,t+1} - \widehat{r}^{[\text{cv}]}_{i,t+1}\right)^2$. Hyperparameter ω is tuned by choosing the one with the minimum CV MSE.

B Additional empirical results

Table A.1 presents the results of the time-series regression $f_{k,t+1} = \lambda_k q_{k,t} + \mu_k + \epsilon_{k,t+1}$ for various factors. The estimated λ_k is predominantly positive and statistically significant for all Fama-French-Carhart factors and most JKP factors. This indicates that each factor's return is predictably related to its quantity, with the correct sign.

Table A.2 extends Table 3 by providing the full-sample coefficient estimates for both the BTQ model (also shown in Table 3) and the β-only model. The μ_k coefficients in the β-only model are all either statistically insignificant or negative.

Table A.2: Table 3 continued, full-sample coefficient estimates

	CAPM	FF3	FF3C	FF5	FF5C
BTQ model: λ_k (%, monthly), t-statistics in parentheses					
MKT	1.78	1.29	1.17	1.31	1.19
	(4.14)	(2.10)	(1.99)	(2.03)	(2.01)
SMB		-0.23	-0.16	-0.20	-0.10
		(-0.75)	(-0.58)	(-0.68)	(-0.37)
HML		0.77	0.47	0.78	0.48
		(1.35)	(0.65)	(1.52)	(0.70)
MOM			0.52		0.76
			(0.69)		(0.94)
CMA				0.08	0.06
				(0.29)	(0.22)
RMW				-0.11	-0.27
				(-0.32)	(-0.72)
β-only model: μ_k (%, monthly), t-statistics in parentheses					
MKT	0.48	0.54	0.44	0.67	0.62
	(1.34)	(1.07)	(0.94)	(1.46)	(1.42)
SMB		-0.04	0.06	-0.02	0.10
		(-0.12)	(0.20)	(-0.07)	(0.30)
HML		0.53	0.46	0.50	0.39
		(1.58)	(1.44)	(1.34)	(1.08)
MOM			-0.39		-0.46
			(-1.02)		(-1.20)
CMA				0.04	0.11
				(0.20)	(0.52)
RMW				0.12	0.16
				(0.46)	(0.63)

Note: Full-sample coefficient estimates of BTQ and β-only return predictions (Eq. 12 and 13). Same as Table 2 but with multi-factor models ($K \geq 1$).