Who Owns What?
A Factor Model for Direct Stockholding[*]

Vimal Balasubramaniam[†] John Y. Campbell[§]
Tarun Ramadorai[‡] Benjamin Ranish[¶]

February 8, 2021

Abstract

We build a cross-sectional factor model for investors' direct stockholdings, by analogy with standard time-series factor models for stock returns. We estimate the model using data from almost 10 million retail accounts in the Indian stock market. We find that stock characteristics such as firm age and share price have strong investor clienteles associated with them. Similarly, account attributes such as account age, account size, and extreme underdiversification (holding a single stock) are associated with particular characteristic preferences. Coheld stocks tend to have higher return correlations, suggestive of the importance of clientele effects in the stock market.

[*] We thank seminar participants at Imperial College Business School, Harvard Business School, UC Boulder, and Dartmouth, and Xavier Gabaix and Robin Greenwood for comments.

[†] Balasubramaniam: Queen Mary University of London E1 4NS, UK, and CEPR. Email: v.balasubramaniam@qmul.ac.uk.

[‡] Ramadorai: Imperial College, London SW7 2AZ, UK, and CEPR. Email: t.ramadorai@imperial.ac.uk.

[§] Campbell: Department of Economics, Littauer Center, Harvard University, Cambridge MA 02138, USA, and NBER. Email: john_campbell@harvard.edu.

[¶] Ranish: Board of Governors of the Federal Reserve System. Email: ben.ranish@frb.gov.

1 Introduction

How should investors combine stocks into portfolios, and how do they actually do so?

The first question is central to modern financial economics, and has been answered under many different sets of assumptions. Since the original mean-variance analysis of Markowitz (1952), financial economists have considered the implications of capital market equilibrium (Sharpe 1964, Lintner 1965), exogenous income risk hedging (Mayers et al. 1972), intertemporal hedging (Merton 1973), multifactor structure in returns (Ross 1976), liquidity demand (Amihud and Mendelson 1986), and investor tastes for firm attributes such as ethical and environmental quality (Hong and Kacperczyk 2009, Pástor et al. 2020).

Much less progress has been made in answering the second question, a task that we take up in this paper. Both measurement and conceptual challenges have hampered descriptive research on the construction of portfolios from individual stocks. Measurement of household portfolios is challenging because surveys rarely ask about the individual stocks that investors hold, while administrative data from brokerage firms may not capture the complete portfolios of investors with multiple accounts. Administrative data from Scandinavian countries have been used in recent research such as Calvet et al. (2007), but the important role played by mutual funds in these countries makes it hard to interpret individual stockholdings without also looking through mutual fund holdings to the underlying stocks held by funds. In this paper we use Indian data on direct stockholdings, exploiting the very limited share of mutual funds in India emphasized by Campbell et al. (2014).

Conceptually, the challenge is to model a sparse holdings matrix of N stocks by H households, where both N and H are large (3103 and 9.7 million, in our dataset for August 2011). Our response in this paper is to specify a cross-sectional factor model for stockholdings across households that is analogous to the classic time-series factor model for stock returns over time. This allows us to exploit numerous insights and methods from the time-series factor literature.

We work with observable factors, as in the modern empirical literature following Fama and French (1993). However we also use methods from the unobservable factor literature (Chamberlain and Rothschild 1983, Connor and Korajczyk 1986, 2019, Ahn and Horenstein 2013) to characterize the potential importance of omitted unobservable factors.

The observable factors in our model come in two varieties. Some factors are attributes of stockholding accounts that do not depend on the particular stocks held by these accounts, such as account age, size, location, and the number of stocks held. These factors are analogous to macroeconomic factors in a time-series model. Other factors are based on characteristics of the stocks held in each account: these factors are analogous to return-based factors such as the Fama-French SMB and HML factors. We estimate the loadings of stocks on these factors using unrestricted cross-sectional regressions, and relate these estimated loadings to observable stock characteristics.

We begin by showing that households do not maximally diversify their portfolios, conditional on the number of stocks held. We find that while the stocks that are popular in single-stock portfolios do tend to be large stocks with relatively low idiosyncratic risk, household portfolios that contain relatively few stocks are far from optimally diversified. This result holds true both when the mean-variance optimal portfolio is described by the market portfolio (i.e., CAPM), as well as when the optimal portfolio is defined by exposure to the market and three additional factors.

As a first step towards a better description of household portfolio choice, we measure the importance of different stock characteristics by the strength of the investor clienteles associated with them. Our measure of clientele strength for each characteristic is the variance of the holdings-weighted characteristic across households. A characteristic that is strongly positive for some household portfolios and strongly negative for others is a characteristic that appears to matter in household portfolio formation: we say that such a characteristic has a strong clientele effect. This logic is analogous to using the return variance of a portfolio such as HML or SMB as a measure of the tendency for value stocks or growth stocks to move together.

Among the characteristics we consider, stock age has the strongest investor clientele but stock price, past returns, and turnover also have strong clienteles. While some of these characteristics are cross-sectionally correlated with one another, we find very similar results when we orthogonalize the characteristics across stocks. We find that the characteristics of stocks that are emphasized in time-series factor models such as the Fama-French (1993) model are relatively less important, a finding that also carries over to our more elaborate multifactor model-based analysis.

To develop our understanding of clienteles further, we adapt the same approach, and measure how particular investor attributes group together in stock clienteles. To measure this, we analogously measure the variance of the holdings-weighted attributes across stocks. A household attribute that is strongly positive in the investor base of some stocks and strongly negative in the investor base of other stocks is an attribute that appears to matter in the formation of investor clienteles: we say that such an attribute has a strong clientele effect. In the data, we find that both account attributes and portfolio attributes (i.e., the composition of the other stocks in households' portfolios) help to identify clienteles. The age of an account and the number of stocks that it trades are particularly useful account attributes; and the average share price, age and market capitalization of household stockholdings are particularly useful portfolio attributes in this regard.

We include both account and portfolio attributes in our empirical multifactor model of household portfolio choice. The model is well able to capture the empirically observed propensities for households to simultaneously hold particular stocks (which we call "coholdings"). The performance of the model also compares favorably with an unobserved factor model based on principal components analysis. The model reveals that investor clienteles in the Indian data form around clusters of stock characteristics: size and share price; turnover and beta; and book-market, volatility, and skewness. Moreover, these clienteles are also well characterized by account and portfolio attributes. For example, larger accounts appear to prefer large, established growth companies, while accounts with high turnover prefer smaller, cheaper, high-turnover stocks with "lottery-like" characteristics.

We conclude by exploring the relation between measured coholdings and return comovement across stocks. We show that there is a visibly strong positive correlation across stock pairs between coholdings and stock return correlations, which is another way to see that households' portfolios are not optimally diversified and which suggests that investor clienteles may be important drivers of stock return comovements.

Related literature

The literature on positive household portfolio choice is quite limited. Because of the difficulty in measuring the complete portfolios of individual investors, many papers focus on households' trading behavior and realized returns rather than their portfolio composition. Examples include Barber et al. (2009), Barber and Odean (2000, 2001), Grinblatt and Keloharju (2000), Kaniel et al. (2008), Odean (1998), Seru et al. (2010).

Among those papers that do study household portfolio choice, it is common to study choices of mutual funds rather than direct stockholdings. For example Grinblatt et al. (2016) highlight the impacts of IQ on mutual fund choice by Finnish investors using detailed data on mutual fund choices alongside less detailed information on direct equity investment. Betermeier et al. (2017) use Swedish data to estimate value and growth tilts in household portfolios, but they estimate these tilts directly for mutual funds and do not attempt to look through to the implied weights on individual stocks.

Within the smaller literature on households' direct stockholdings, one precursor to note is Dorn and Huberman (2010) which identifies idiosyncratic volatility as a relevant attribute of stocks that investors pay attention to in their stock selection. Massa and Simonov (2006) and Døskeland and Hvide (2011) ask whether Scandinavian households use undiversified equity holdings to hedge their specific labor income risks. They find that if anything households hold stocks that have a more positive correlation with their labor income than average, indicating a tendency to “anti-hedge”.

Some of our findings on household stockholding behavior have parallels in the literature on institutional stockholding. For example Coval and Moskowitz (1999) document local bias in the stocks held by US mutual fund managers, and we document a similar pattern among Indian households. Our work can be regarded as complementary to efforts such as Koijen and Yogo (2019) to empirically characterize the structure of institutional investors' portfolio demands.

Organization of the paper

The organization of our paper is as follows. Section 2 lays out the factor structure that we use to organize our empirical research. Section 3 describes our Indian dataset. Section 4 asks to what extent Indian investors appear to be optimally diversifying their portfolios, conditional on the number of stocks they hold. Section 5 measures the strength of investor clienteles over a range of stock characteristics and attributes of investors. Section 6 estimates multifactor models of stockholdings, not only our model with observable factors but also models with unobserved principal-components-based factors. Section 7 compares empirically observed coholdings with those predicted by the factor models, uses our observable factors to explain clienteles, and relates coholdings and clienteles with return covariances. Section 8 concludes. An online appendix, Balasubramaniam et al. (2020), provides additional details on the empirical analysis.

2 Factor Structure in Stockholding

In this section, we introduce some concepts that we use to structure our empirical investigation of cross-sectional patterns in stockholding. We first define the holdings matrix, which summarizes the holdings of N stocks by H households. From this we derive coholdings matrices for stocks and for investors, and show how they can be used to measure the strengths of different types of investor clienteles. Finally, we define a cross-sectional factor model for stockholdings, analogous to the familiar time-series factor models used to describe stock returns.

2.1 Holdings, Coholdings, and Clientele Strength

Traditional time-series factor models for stock returns work with stocks $i = 1, ..., N$ observed over time periods $t = 1, ..., T$. Our goal is to empirically describe the patterns in market participants' stockholding decisions. This means that we are interested in another important dimension, namely, $h = 1, ..., H$, which indexes households in our current application, but could also capture institutional investors or other types of market participants more generally. To reduce the dimension of the problem, we begin by collapsing the time dimension into a single period.[1] This eliminates the need for time subscripts in our notation.

The holdings matrix

We first define an N by H **holdings matrix** of households' stock holdings Q. The choice of this letter refers to "quantum" or "quantity". The elements Q_{ih} are positive whenever household h holds stock i, and zero otherwise. We denote the N-vector of household h's stock holdings (i.e., the h'th column vector of Q) by Q_h, and call it the household's **holdings vector**. We denote the H-vector of stock i's investor base (i.e. the i'th row vector of Q) by Q_i, and call it the stock's **investor vector**.

2.1.1 Characteristics and attributes

We are interested not so much in specific stocks or specific households as in characteristics of stocks, and attributes of households. For each stock characteristic of interest, we

[1] In our empirical application, we study a single month, August 2011, which is the last month in our sample period and therefore provides us the maximum past history for each investor. We have re-run our analysis across all of the months in the dataset to check the persistence of the relationships that we estimate. An alternative procedure would be to average all the periods observed in the raw data and conduct the analysis on empirical time-averages of holdings.

define c as a zero-mean N-vector of the rank of each stock's characteristic on the interval $[-0.5, 0.5]$. Similarly, for each household attribute of interest, we define a as a zero-mean H-vector of the rank of each household's attribute on the interval $[-0.5, 0.5]$.

We seek to define and study "clientele effects" for particular stock characteristics, and to study the attributes of the investors that group together in their holdings of particular stocks. We introduce and develop these concepts further below.

The stock coholdings matrix and characteristic clientele strength

Although each stock characteristic has an equal-weighted average of zero across all stocks, the holdings-weighted average characteristic need not be zero. Households on average may tilt their portfolios towards stocks with certain characteristics, but in general, these tilts will reflect the supply of those characteristics as well as household demand. Accordingly, we focus not on the average characteristic but on the variance of the holdings-weighted characteristic across households. A characteristic that is strongly positive for some household portfolios and strongly negative for others is a characteristic that appears to matter in household portfolio formation: we say that such a characteristic has a strong clientele effect.

To begin with, consider the demeaned holdings vector for household h:

$$\tilde{Q}_h = Q_h - H^{-1} \sum_{h'=1}^{H} Q_{h'}, \tag{1}$$

where demeaning takes place across all households. The empirical **stock coholdings matrix**, defined over N stocks, is the $N \times N$ matrix

$$\Omega_h = H^{-1} \sum_{h=1}^{H} \tilde{Q}_h \tilde{Q}'_h. \tag{2}$$

The diagonal elements of Ω_h capture the intensity with which particular stocks are held, and the off-diagonal elements capture the intensity with which particular pairs of stocks are held, averaging across all households. Equivalently, one might say that the diagonal elements measure the popularity of each stock among investors, and the off-diagonal elements measure the popularity of each pair of stocks.

To develop intuition about the stock coholdings matrix, we note that it is analogous to the familiar empirical covariance matrix of stock returns. To construct the stock return covariance matrix, we also begin with a single time period and calculate the outer product matrix of returns in that period (after time-series demeaning returns), and subsequently average these outer products over time. Thus, the empirical stock return covariance matrix uses time periods where the stock coholdings matrix uses households, but otherwise the two matrices have the same structure. The stock coholdings matrix must be positive semi-definite whenever $H > N$, just as the empirical covariance matrix of stock returns must be positive semi-definite whenever $T > N$.

To study the clientele for a particular characteristic we define **characteristic clientele strength** as the empirical variance of $c'Q_h$ across households:

$$\sigma^2(c'Q_h) = c'\Omega_h c. \tag{3}$$

In the time-series analysis of returns, the analogous approach is to argue that a pervasive characteristic represents a potentially important risk if a long-short portfolio formed by sorting stocks on this characteristic has a relatively high return variance (Kozak, Nagel, and Santosh 2018).

The investor coholdings matrix and attribute clientele strength

Similarly, although each household attribute has an equal-weighted average of zero across all households, the holdings-weighted average attribute need not be zero if attributes are correlated with holdings. Accordingly, we focus not on the average attribute but on the variance of the holdings-weighted attribute across stocks. A household attribute that is strongly positive in the investor base of some stocks and strongly negative in the investor base of other stocks is an attribute that appears to matter in the formation of investor clienteles: we say that such an attribute has a strong clientele effect.

More specifically, we define the demeaned investor vector for stock i as

$$\tilde{Q}_i = Q_i - N^{-1} \sum_{i'=1}^{N} Q_{i'}, \tag{4}$$

where demeaning now takes place across all stocks. The empirical **investor coholdings matrix**, defined over H households, is the $H \times H$ matrix

$$\Omega_i = N^{-1} \sum_{i=1}^{N} \tilde{Q}_i \tilde{Q}'_i. \tag{5}$$

The diagonal elements of Ω_i capture the stockholding intensity of particular households, and the off-diagonal elements capture the intensity of coholdings of particular pairs of households, averaging across all stocks.

The time-series analog of the investor coholdings matrix would be a matrix whose diagonal elements contain the cross-sectional variance of stock returns in each period, and whose off-diagonal elements contain the cross-sectional covariance between stock returns in each possible pair of periods. Such a matrix is a way to measure the similarity of stock return behavior across periods. Of course, when $T > N$ such a matrix is singular and similarly, in our context with $H > N$, the investor coholdings matrix is singular.

Finally, we define **attribute clientele strength** as the empirical variance of $a'Q_i$ across stocks:

$$\sigma^2(a'Q_i) = a'\Omega_i a. \tag{6}$$

This measure can be calculated without explicitly constructing the extremely large matrix Ω_i, and it is well behaved even though Ω_i is singular.

2.1.2 Measures of quantity

So far, we have not yet defined the elements of the holdings matrix Q. There are several ways to do this, and we make choices that have desirable properties in our empirical setting. In common with other studies of retail investors (see, e.g., Grinblatt and Keloharju (2001), and Liao et al. (2020)) our setting has extreme variability across households in the number of stocks held and the amount invested, and across stocks in the size of the investor base. Accordingly we define elements of Q to try to ensure that our conclusions about characteristic clientele strength reflect a representative stock investor, and that our conclusions about attribute clientele strength reflect a representative stock.

Specifically, we consider two approaches, Q^v and Q^s, which equalize the sum of the elements of Q across households and stocks respectively, i.e. $\iota' Q^v_h = 1$ for all h and $\iota' Q^s_i = 1$ for all i, where ι is a vector of ones.

In the matrix Q^v, the column vector for household h, Q^v_h, is the vector of portfolio shares for household h, which of course sums to one. Accordingly $c' Q^v_h$ is the value-weighted average characteristic c of household h's stockholdings, and characteristic clientele strength is the variance of this across households.[2]

Our choice of Q^v makes our characteristic clientele strength measure $c' \Omega^v_h c$ representative of a typical household's stock investment. This necessarily means that it primarily reflects investor preferences within the set of widely held stocks. In other words, elements of the stock coholdings matrix are small in magnitude for the numerous stocks which are rarely held by households. As a check that our conclusions about characteristic clientele strength are applicable to the broader universe of stocks, we alternately exclude the most widely held 10 or 50 stocks, and recompute c, Q^v, and $c' \Omega^v_h c$ using this reduced dataset.[3]

[2] As an alternative, we have also considered a measure Q^e that uses equal weights for each stock held by household h. This pure extensive margin measure is very similar, as Q^v_h and Q^e_h are always zero simultaneously. The difference between the two reflects both investor decisions about the scale of stock purchases and the history of stock returns during the investor's holding period.

[3] An alternative approach would be to assign weights to equalize variance for each stock, effectively computing the empirical correlation matrix of Q^v_h rather than its empirical covariance matrix. However this approach gives a great deal of weight to extremely rarely held stocks and so we do not pursue it here.

In the matrix Q^s, the row vector for stock i, Q_i^s, assigns equal weight of $1/s_i$ to each household invested in stock i, where s_i is the number of shareholders in the stock. Accordingly $a'Q_i^s$ is the average attribute of stock i's investor base.

Our choice of Q^s makes our attribute clientele strength measure $a'\Omega_i^s a$ representative of a typical stock, necessarily meaning that it will disproportionately reflect investors who hold more stocks, and particularly, investors who hold relatively obscure stocks with few shareholders. As a check that our conclusions about attribute clientele strength are applicable to the broader universe of investors, we recompute the variance measure after excluding investors who hold more than 10 stocks.

2.2 A Factor Model for Holdings

So far, we have discussed measures of clientele strength for stock characteristics and for investor attributes separately, but we have not linked the attributes of investors to the characteristics of the stocks they hold. A natural way to do that is to estimate a factor model, a cross-sectional analogue of the models commonly used to describe the variation in stock returns over time. Focusing on the portfolio share holdings matrix Q^v, for each stock i we can estimate a cross-sectional regression:

$$Q_{ih}^v = \alpha_i + \sum_{k=1}^{K} \beta_{ik} F_{kh} + \varepsilon_{ih}, \qquad h = 1, ..., H, \tag{7}$$

where β_{ik} is the loading of stock i on factor k, and F_{kh} is the factor realization for household k.

In equation (7), the factors could be attributes of the household, such as account size or account age, which are not affected by the composition of the household's portfolio. Pursuing the analogy with factor models of stock returns, these are like time-series factors that are estimated without reliance on the behavior of other stocks, such as shocks to inflation or industrial production. However, the factors could also be attributes of the household portfolio, like the average size or book-to-market ratio of the other stocks held by the household. This is analogous to using the contemporaneous returns on other stocks to create factors such as HML and SMB in the usual Fama-French time-series analysis.[4] This use of household portfolio characteristics as attributes further connects the analysis of stock characteristics and investor attributes developed in the previous subsection.

The β_{ik} coefficients inform us about the average attributes of the investor clientele for each stock i. In other words, they tell us which types of households (the "who" in the paper's title) tend to hold stock i ("what"). We estimate these coefficients freely, stock by stock, but we report weighted averages of the coefficients using important stock characteristics as weights. This enables us to measure the determinants of clienteles not only for individual stocks, but also for stock characteristics.[5]

The factor model (7) simplifies the structure of the stock coholdings matrix $\tilde{Q}_h$. Consider a situation where $\alpha_i = 0$, as will be the case if equation (7) is estimated using household-demeaned holdings $\tilde{Q}_h^v$ and zero-mean factors. Assume in addition that the factors are orthogonal to one another, and that enough factors are included to make the error terms ε_{ih} uncorrelated across households h for all stocks i. Under these conditions the diagonal elements of the stock coholdings matrix Ω_h^v take the form:

$$\Omega_{h,i,i}^v = \sum_{k=1}^{K} \beta_{ik}^2 \sigma_k^2 + \sigma_i^2, \tag{8}$$

where σ_k^2 is the cross-sectional variance of F_{kh} and σ_i^2 is the cross-sectional variance of ε_{ih}. Under the same assumptions, the off-diagonal elements of the stock coholdings matrix

[4] In time-series factor analysis, it is common practice to construct factors using all stock returns, so that an individual stock's betas are estimated from a regression in which that individual stock's return influences the explanatory variables as well as the dependent variable. This practice is generally harmless because factor portfolio returns are well diversified across stocks. In our context, however, many households have concentrated portfolios so we are careful to exclude own holdings when we construct stockholding characteristic factors.

[5] An alternative procedure would be to restrict the betas of individual stocks on account and stockholding characteristics to be linear functions of stocks' characteristics, and to estimate the restricted version of equation (7) as a panel regression. We do not pursue this alternative here.

take the form:

$$\Omega^{v}_{h,i,j} = \sum_{k=1}^{K} \beta_{ik}\beta_{jk}\sigma^2_k, \tag{9}$$

so the common factors determine the coholdings propensities for pairs of stocks i and j. Factors with large standard deviations or dispersed loadings are influential determinants of coholdings.

These properties of the model follow from the linearity of equation (7). A disadvantage of (7) is that it is a linear probability model whose fitted values may lie outside the theoretically appropriate range from zero to one. An alternative approach would be to estimate a nonlinear bounded model for holding probabilities such as a probit or logit model, but in this case the implied coholdings matrix would no longer have the simple structure of equations (8) and (9).

3 Indian Equity Market Data

3.1 Equity Ownership

Our data on Indian stockholdings, which are also used in Campbell et al. (2014), Anagol et al. (2018), Campbell et al. (2019), and Anagol et al. (2021), come from India's two share depositories with the approval of India's apex capital markets regulator, the Securities and Exchange Board of India (SEBI). We observe data from the beginning of February 2002, but because the cross-sectional relationships we study are fairly stable over time, we focus primarily on August 2011. This is the last month of data in our sample, and consequently, provides us the maximum past history for each account and correspondingly more precise estimates of the factors.

The older and larger of the two depositories, National Securities Depository Limited (NSDL), accounts for 64% of the roughly 9.7 million individual accounts we study in August 2011, with the remainder held at Central Depository Services Limited (CDSL).

These two depositories together record almost all trading in and holdings of Indian equity at the account-issue level at a monthly frequency.[6]

We do not observe data on holdings of equity derivatives or mutual funds. However, during our sample period derivatives and mutual funds are relatively unimportant for Indian individual equity investors. While single-stock futures markets are quite active in India (Martins et al. 2012, Vashishtha and Kumar 2010), a minority of accounts invest in equity derivatives over our sample period.[7] Moreover, while mutual funds have grown in popularity in India, the typical investor that holds individual equities in our sample has no bonds or mutual funds.[8] Additionally, we estimate that 89% of individuals' aggregate equity holdings in 2011 were direct, as opposed to holdings of equity mutual funds, unit trusts and unit-linked insurance plans.[9]

The sensitive nature of these data mean that there are limitations on the demographic information provided to us. The information we do have includes the state in which the investor is located, whether the investor is located in an urban, rural, or semi-urban part of the state, and the type of investor. We use investor type to identify individual investor accounts.[10] A given individual investor can hold multiple accounts, so we aggregate accounts that share the same Permanent Account Number (PAN)—a unique identifier issued to all taxpayers by the Income Tax Department of India. This aggregation may not always correspond to household aggregation if a household has several PAN numbers,

[6] The share depositories were established to promote dematerialization, i.e., the transition of equity ownership from physical stock certificates to electronic ownership records. While equity securities in India can be held in both dematerialized and physical form, settlement of all market trades in listed securities in dematerialized form is compulsory. To facilitate the transition from the physical holding of securities, the stock exchanges do provide an additional trading window, which gives a one time facility for small investors to sell up to 500 physical shares. However, the buyer of these shares has to dematerialize such shares before selling them again, thus ensuring their eventual dematerialization. Statistics from the Bombay Stock Exchange (BSE) and the National Stock Exchange (NSE) highlight that virtually all stock transactions take place in dematerialized form.

[7] A 2011 SEBI survey estimates that fewer than one million Indian households invest in derivatives. See: https://www.sebi.gov.in/sebi_data/attachdocs/1326345117894.pdf

[8] A 2009 SEBI survey found that about 65% of Indian households owning individual equities did not own any bonds or mutual funds. See: http://www.sebi.gov.in/mf/unithold.html

[9] See Table A1 of the internet appendix to Campbell et al. (2014).

[10] We exclude "individuals" that hold at least 5% of a stock with market capitalization above 500 million Rs (approximately $10 million), reclassifying these accounts as beneficial owners.

for example, if children or spouses have separate PANs. In addition, we are unable to link accounts by PAN between NSDL and CDSL. However, conversations with our data provider suggest that few retail investors have multiple depository relationships.

Given our interest in household portfolio construction, we restrict our current analysis to the portfolios of retail investors in the market, and do not at this stage consider the portfolios of institutions or government entities (which we also observe). We also exclude non-public equities, which the typical household may have difficulty acquiring. Furthermore, since there is no requirement in India that publicly listed equities should have a large investor base, we remove de-facto private equities. We define these as stocks in the bottom 25th percentile ranked by the number of shareholders invested at the end of the previous month. This cutoff corresponds to removing equities with fewer than 1,177 investors at the end of July 2011 from the August 2011 cross-section of stocks that we study. After applying these filters, our final sample comprises 3,103 Indian equities and the portfolios of 9.7 million individual accounts that hold at least one of these stocks at the end of August 2011.

3.2 Stock Characteristics

We match our data on Indian equity holdings to data on returns, dividends, market capitalization, share price, book value, turnover, and the age, industry, location, and business group affiliation of the firm. These data are primarily drawn from the CMIE Prowess database, with Datastream and Compustat Global used to supplement and validate these data.[11]

We handle missing stock characteristics as follows. For stocks missing an industry assignment, we assign values to their industry dummies equal to the fraction of stocks in

[11] Where data sources differ, we use the value from the data source(s) that are more consistently in close agreement. For stock returns, we also (1) manually validate the 25 largest and smallest percentage returns observed in the data and (2) manually collect and fill missing returns for the few instances in which a stock with a missing return comprises at least 1% of the average individual's stock portfolio.

the given industry. For other missing (continuous) characteristics, we use all available characteristics in a regression to impute values for any characteristics that are missing.[12] This has little impact on our results as our use of rank-normalized characteristics limits the influence of any measurement errors, and characteristics are missing for relatively few stock holdings.[13]

3.3 Summary Statistics

In the early 21st Century, equity market participation in India underwent dramatic expansion. The number of individual depository accounts increased roughly four-fold from 2.4 million in 2003 to 9.7 million at the end of our sample period in August 2011.[14] The period also saw a significant jump in the number of accounts in January 2008, when the extraordinarily large IPO of Reliance Power brought a large set of investors into the market.

Table 1 summarizes characteristics of the household accounts and the composition of their stock portfolios in the August 2011 cross-section that we study. The median account is slightly over four years old at this date (where age is measured from the first month in which the account holds any stock) and roughly 10% of accounts are ten or more years old. While some stockholders do exit the market, the large share of young accounts reflects the enormous growth in households holding equities during the years before 2011.

As documented in Campbell et al. (2019), the account size distribution is dispersed and right-skewed, with a median account size of US$ 780, and a mean account size of

[12] Prior to imputation, we apply a log-transformation to share price, market capitalization and popularity–as the distribution of these variables has a fat right-tail. We further winsorize values of book to market ratio, returns, volatility and skewness that are used for imputation purposes at the 5th and 95th percentile.

[13] Specifically, for August 2011, we impute stock age for 6.2% of stock holdings, book to market for 3.2% and lagged returns, volatility and skewness for about 0.24%. We impute industry for 2.7% of stock holdings. Other characteristics do not require imputation.

[14] We illustrate this fact in the top-left panel of online appendix Figure A.1. It does not reflect increases in dematerialization, as even at the beginning of our sample period, most Indian stocks were held in dematerialized form.

over US$ 11,000, close to the 90th percentile value of US$ 13,000. This distribution of account sizes is similar to the United States when accounting for the differences in per-capita GDP between the two countries, as we show in online appendix Figure A.2. Jayaraj and Subramanian (2008) show that the median (wealthiest) deciles of Indian households had average total asset values of about $3,000 ($35,000) in 2008, meaning that the stock portfolios we study represent a non-trivial share of wealth for many of the investors in the data.[15]

Our empirical work utilizes several other account characteristics, including the number of stocks held by the account (of the total set of 3,103 stocks that we consider), the number of stocks traded, and portfolio turnover. All these characteristics are dispersed and right-skewed. The median account in the data holds four stocks, and the mean number of stocks held is 8.45. Only the top decile of individual accounts holds 20 or more stocks. Relatedly, the median account makes trades in only one stock over the year prior to August 2011, while accounts at the 90th percentile trade 13 different stocks over the prior year. We also measure trading activity by account turnover, computed as the dollar value of shares traded between September 2010 and August 2011 divided by the current account value. We winsorize this ratio at the 99th percentile to remove the influence of outliers. This measure of trading activity is similarly dispersed and right-skewed.

The bottom half of Table 1 summarizes the characteristics of the stocks held in investor portfolios. We measure these stockholding characteristics as average ranks, ranking all the stocks in the available universe on each characteristic from -0.5 to 0.5, and then for each investor taking a portfolio-share-weighted average of the characteristic ranks of all stocks held. The stock characteristics we consider are share price, stock age (years since listing), realized volatility, market capitalization, realized returns, turnover, market beta,

[15] The accounts in our dataset are spread out across India. The wealthier west of India contributes 43% of all accounts, the east of India contributes roughly 11% of all accounts, and the remaining accounts are divided roughly equally between the north and south of India. Details are given in online appendix Figure A.1.

book-market ratio, and realized skewness.[16]

The table reveals that the median retail investor portfolio is at the 95th percentile of the firm size distribution. The other characteristics of the median stockholdings are in line with this tilt towards large stocks, since larger stocks in our sample tend to have higher share prices, lower book-market ratios, and lower past realized volatility and skewness. However stockholding characteristics vary significantly across accounts, with a standard deviation of close to 0.2 for most characteristics and as high as 0.27 for stock age. We explore the volatility of these factors in greater detail in section 5.

Details on other stockholding characteristics are reported in the online appendix. Figure A.1 in the appendix shows the average distribution of stockholdings across seven industries and business groups. Business groups—sets of independently listed companies with a large ownership stake and common control by a single underlying entity—are quite common in developing countries (e.g., Anagol and Pareek 2019), and in our data, 886 of the 3,103 stocks are affiliated with 266 business groups. In the average account, the top 10 business groups account for 31% of stockholdings, with remaining business groups accounting for a further 16% of stockholdings. We account for both industry group holdings and business group affiliations in our set of factors used to explain stockholdings and coholdings.

Figure A.3 in the online appendix shows correlations between account and stockholding characteristics. Within the set of account characteristics, the most correlated are account size, the number of stocks held, and the number of stocks traded, but all correlations are below 0.6. The largest correlations are within the set of stockholding characteristics. Stock size (market capitalization) is strongly positively correlated with share price (0.78) and negatively correlated with book to market (-0.57). Share price and market cap are both negatively correlated with realized volatility. Return-based characteristics, including

[16] Turnover and all the return-based stock characteristics are computed over the year from September 2010 through August 2011, using weekly data to compute return volatility and skewness. The book-market ratio is computed using the standard Fama-French methodology applied to Indian stocks.

realized returns, volatility, skewness, and beta, have strong correlations with one another that likely reflect the particular time period we consider. Correlations between account and stockholding characteristics are generally smaller. This suggests that an empirical model including both types of factors will be well-behaved.

Figure 1 plots the cross-sectional distribution of the number of investors holding each stock in August 2011. The most widely held stock is Reliance Power Limited, held by roughly 40% of all accounts, comprising roughly 4 million accounts. The top five stocks ranked by holdings are each held by over 10% of all individual accounts, and the top ten stocks are each held by over 7.5% of all accounts. At the other extreme, roughly 62% of all stocks in our sample are held by fewer than 0.1% of individual accounts.[17] The characteristics of stockholdings in the summary statistics, therefore, heavily reflect holdings of popular stocks. This distribution highlights the distinction between an analysis of the composition of a typical investor's portfolio and an analysis of the investor clientele for a typical stock.

Figure 2 similarly plots the cross-sectional distribution of the average portfolio share held in each stock, where the average is taken both over all individual investors and over those investors who hold the stock. (For example, roughly 80% of all stocks have a portfolio weight of 10% or less in the portfolios of investors who hold it, and roughly 80% of all stocks have a weight of approximately 0.02% on average across all investors' portfolios, including those who do not hold it). This figure further illustrates the extreme differences between a few stocks that are widely held with high portfolio weights, and many stocks that are rarely held and have low portfolio weights even when held.

[17] The left censoring of the distribution in Figure 1 results from the filter that we described in the data section, which results from dropping the bottom 25% of stocks based on the number of accounts holding the stock at the end of July 2011.

4 Diversification of Stockholdings

In this section we explore the ability of diversification motives to explain investors' stockholdings. The most widely held stocks tend to be large, with relatively low idiosyncratic volatility. This suggests that portfolio mean-variance optimization could play a role in stock selection. To evaluate this explanation, we compare observed portfolios to portfolios that are mean-variance optimized subject to a constraint on the number of stocks that can be held.

4.1 Constrained Mean-Variance Optimization under the CAPM

According to the CAPM, all investors should hold the market portfolio in order to maximize the Sharpe ratio of their portfolio returns. Most households in our Indian data hold a handful of individual stocks, consistent with results found in many other settings (Gomes et al. 2020). This means that it is straightforward to reject a strict interpretation of the CAPM's predictions for portfolio construction. Instead, we evaluate the hypothesis that household portfolio construction can be explained as a constrained optimization problem. That is, we check whether households h attempt to get as close to the market portfolio Sharpe ratio as possible, while operating under a constraint on the number of stocks N_h that they hold, as well as a constraint on short sales. Exogenous variation in N_h across households could arise from cognitive or real frictions associated with holding and trading multiple stocks, or simply from a lack of financial sophistication; we do not model these frictions in this paper.

To conduct this evaluation, we first assume that expected excess returns follow the CAPM, meaning that the market Sharpe ratio is ex-ante optimal. We then assume that households attempt to get as close to the market Sharpe ratio as possible subject to the constraint of holding N_h stocks, by building a portfolio that maximizes the fit to the returns on the market portfolio.

To generate an empirical benchmark for the constrained optimization problem faced by households, we implement a least absolute shrinkage and selection operator (LASSO) regression. We regress market portfolio returns on individual stock returns, using weekly total realized returns over the period September 2009 through August 2011, and for each value of N_h we adjust the LASSO regularization parameter to deliver a portfolio with exactly N_h stocks. That is, for lower (higher) N_h, the regularization parameter tightens (weakens) the constraint on the number of regressors included in the model. The estimated portfolios associated with each N_h trade off the regression fit against the number of regressors included, and are plausible solutions for the constrained optimization problem. For $N_h = 1$ we simply choose the stock which is maximally correlated with the market.

Panel A of Figure 3 plots the results from this exercise for N_h ranging from 1 to 50. The height of each grey bar in panel A indicates the maximum obtainable Sharpe ratio associated with each value of N_h on the horizontal axis using the LASSO implied portfolio of stocks.[18] This maximum Sharpe ratio roughly doubles as N_h increases from 1 to 5, increases more slowly as N_h increases further to 25, and has small gains beyond that point. For optimal portfolios with more than 25 stocks, the Sharpe ratios are very close to that of the market portfolio, which is shown as a black bar.

The blue triangles in panel A show the locations of the median estimated Sharpe ratios of investors' actual stock portfolios observed in the data over the same time period. Holding larger numbers of stocks is associated with a Sharpe ratio that is relatively larger compared to the constrained optimum. This finding could reflect the role of financial sophistication in jointly determining performance and N_h, or could simply reflect underlying heterogeneity in investors' preferences for taking idiosyncratic risk.

The dotted lines extending vertically above and below the triangles span the 10th to

[18] The Sharpe ratio on the market is estimated over a longer sample period from April 2003 through August 2011, since realized Sharpe ratios are noisy estimates of true Sharpe ratios over short sample periods.

90th percentiles of investors' estimated Sharpe ratios. Even at the 90th percentile, these values are below the empirical benchmark estimated using the LASSO approach for all values of N_h, with an especially large relative gap when N_h is low.

Of course, the CAPM may not be the best model for pricing Indian stocks. As an alternative, we next consider investors' performance under a popular four-factor model of returns.

4.2 Constrained Mean-Variance Optimization under a Four-Factor Model

We add three standard priced factors—size, value, and momentum—to the market return to create a four-factor model. The maximum Sharpe ratio is now achieved by the tangency portfolio of these four factors.[19] Once estimated, we compute the tangency portfolio's returns by applying its loadings to the factor returns. As before, we generate an empirical benchmark for the constrained optimization problem faced by households using LASSO regression that maximizes the fit of the returns to the tangency portfolio returns over September 2009 through August 2011, conditional on holding only N_h stocks with no short selling. To assess households' performance, we calculate their portfolio returns' fit to the tangency portfolio returns.

Panel B of Figure 3 plots the results from the four-factor exercise in a similar form to the CAPM exercise in Panel A. The unconstrained optimal Sharpe ratio is over twice as high as under the CAPM. However, the constrained optimal Sharpe ratio is only about 90% of the unconstrained optimal Sharpe ratio at $N_h = 50$, as the four-factor tangency portfolio applies negative weights to some stocks. In contrast to the optimal portfolios, households' Sharpe ratios are only modestly higher under the four-factor model and there-

[19] We use the dataset of Agarwalla, Jacob, and Varma (2013) available at `http://www.iimahd.ernet.in/~iffm/Indian-Fama-French-Momentum`. Following the procedure we used for the CAPM, we estimate the tangency portfolio's factor loadings and Sharpe ratio using weekly factor returns over the period April 2003 through August 2011.

fore are always even further below the constrained optimal level. This results because few portfolios lean heavily towards factors that have been well compensated historically, aside from the market factor—which accounts for less than half of the tangency portfolio.

As this exercise has shown little evidence of constrained mean-variance optimization, we now look for evidence of other preferences that may shape household portfolio choice.

5 Clientele Effects

In this section we discuss the strength of clientele effects in Indian stockholdings, focusing first on stock characteristic clienteles and then on investor attribute clienteles.

5.1 Stock Characteristic Clienteles

As discussed in section 2, the strength of the clientele effect for a stock characteristic can be measured by the empirical variance of the characteristic-weighted stockholding $c'Q_h^v$ across households, where c is a stock characteristic rank ranging from -0.5 to 0.5, and the i'th element of Q_h^v is the weight of stock i in investor h's portfolio.[20] The use of portfolio weights ensures that each investor has equal weight in the holdings matrix Q_h^v, although investors with concentrated portfolios will contribute more strongly to the stock coholdings matrix and to characteristic clientele strength.

In Figure 4 we plot this measure of clientele strength for the nine stock characteristics we described in Table 1. The figure has six panels organized in three rows and two columns. The top row uses all investors, the middle row excludes single-stock accounts, and the bottom row excludes all accounts with ten or fewer stocks. The left column uses raw characteristics, and the right column uses characteristics that have been orthogonalized to one another. Orthogonalization helps to ensure that the clienteles we identify for

[20] Stocks may have the same value of a given characteristic. For each unique value of the characteristic, we compute the average of the ranks spanned by stocks with this value, and assign this average to each of those stocks. Since a given value is never shared by many stocks, the distribution of ranks remains approximately uniform and the mean c remains zero.

particular characteristics are not merely the result of correlation in the cross-section of stocks between those characteristics and other characteristics which investors really care about. We proceed sequentially, first identifying the characteristic with the strongest clientele and orthogonalizing all other characteristics to it using kernel regression, then identifying the two strongest characteristics and orthogonalizing all other characteristics to them using multivariate kernel regression, and so forth.[21]

In each panel of the figure, the nine stock characteristics are presented in their order of clientele strength when all individual investors are included and when the characteristics are orthogonalized. That is, the characteristics are in descending order of clientele strength in the top right panel. The vertical axis is empirical variance normalized by a hypothetical high-variance benchmark in which investors choose stocks to maximize clientele strength, conditional on the number of stocks they hold.[22] Clientele effects in the data are indicated by red diamonds for each characteristic.

The strongest clientele effect in Figure 4 is associated with stock age. Some Indian individual investors strongly prefer to hold young companies (recent IPOs), while other investors strongly prefer established companies. The stock age effect is strongest regardless of whether we look at all investors or only those who hold more than 1 or more than 10 stocks.

The second strongest clientele effect is associated with share price. As noted ear-

[21] Specifically, in each iteration k, we identify the k out of C characteristics with the largest $c(k-1)'\Omega_h^v c(k-1)$, where $c(k-1)$ represents the set of characteristics resulting from the $k-1$st iteration and $c(0)=c$. Next we use this set of k vectors c to predict the remaining $C-k$ characteristics c using a multivariate kernel regression. This local linear regression uses a Euclidean distance measure over the k predictor characteristics, and applies a truncated gaussian kernel with bandwidth for each stock such that 10% of other stocks fall within twice the parameter, applying zero weight beyond. We define the rank of the residuals (over the interval -0.5 to 0.5) from this regression as $c(k)$, setting $c(k)=c(k-1)$ for the k strongest selected characteristics. After $C-1$ iterations, we have our sequentially orthogonalized set of characteristics $c^o = c(C-1)$.

[22] To maximize clientele strength given the empirical distribution of N_h, half of households with N_h stocks invest $1/N_h$ of their portfolio in each of the N_h stocks with the largest c (closest to 0.5), with the other half buying the N_h stocks with the smallest c (closest to -0.5). Since the vast majority of N_h are small relative to N, the resulting distribution of characteristic holdings closely approximates two equal point masses at 0.5 and -0.5—which is the maximum variance solution when the distribution of N_h is unconstrained.

lier, share price is strongly correlated (0.78) with market capitalization; but the clientele strength for share price is noticeably stronger than that for market cap when these characteristics are studied separately without orthogonalization (in the left column of Figure 4), and the share price characteristic drives the market cap characteristic to the bottom of the clientele strength ranking when we orthogonalize characteristics (in the right column of Figure 4). Thus we find evidence that some investors prefer to hold high-priced stocks, while others prefer low-priced stocks. The preference for market capitalization is more uniform (all Indian investors tilt towards large companies) which explains the weaker clientele effect for this characteristic.

Past realized returns also have a relatively strong clientele effect in our data. This may reflect the tendency for some investors to trade momentum while others trade reversal; we caution, however, that since our data are a snapshot at a point in time, we cannot distinguish momentum preference from a preference for some other unmeasured stock characteristic that happened to do well in the period September 2010–August 2011.

Turnover also has a moderately strong clientele effect when all investors are included (in the top row of Figure 4). In this case, turnover is the third strongest of the orthogonalized characteristics. It tends to weaken, however, as we consider more diversified portfolios lower down the figure. This tells us that among undiversified investors, some prefer high-turnover stocks while others prefer low-turnover stocks, but diversified investors have a more uniform preference for this characteristic. Turnover is positively correlated with volatility and market beta, and it drives these characteristics down the orthogonalized ranking.

Perhaps surprisingly, the clientele effects for Fama-French styles (market beta, book-market, and market capitalization) are weaker than those we have already discussed, indicating more limited investor heterogeneity in preferences for these style characteristics. However, even these clientele effects are quite strong in an absolute sense as we now discuss.

There are several ways to judge the absolute strength of a clientele effect. First, we can use the vertical scale of the figure which measures the strength of the clientele effect as a fraction of the theoretical maximum which would be obtained with the most extreme possible investor preferences for or against a given characteristic. When all investors are included, the age effect is 30% of the theoretical maximum and even the weakest clientele effect, for market capitalization, is 12% of the maximum without orthogonalization and 8% with orthogonalization.

Second, we can compare clientele strength to what we would observe under a series of simple alternative models. A natural alternative is random stockpicking conditional on the number of stocks held. If each stock is picked with equal probability, this delivers uniform clientele strength just below 15% of the theoretical maximum when all investors are included, around 7% when single-stock investors are excluded, and around 2% when we exclude investors holding 10 or fewer stocks. If stocks are picked with probability proportional to the free-float capitalization of the stock—which seems a more realistic description of random stockpicking—then we obtain different values for each characteristic which are plotted as green circles in Figure 4. Finally, if we assume that investors hold the optimally mean-variance diversified portfolios described in section 4, conditional on the number of stocks they hold, then we again obtain different values for each characteristic which are plotted in Figure 4 as squares for the CAPM case and triangles for the four-factor case.

Clientele effects in our Indian data are generally strong relative to all these alternatives. The age and stock price clientele effects in particular dominate all the alternatives; and in all specifications, clientele effects are much stronger than they would be with free-float-probability random stockpicking. When we look at all investors and do not orthogonalize, some of the weaker clientele effects could be explained by equal-probability random stockpicking or by mean-variance optimization; but if we orthogonalize or exclude the most concentrated portfolios, then all the clientele effects are too strong to be explained by these alternative models.

5.2 Investor Attribute Clienteles

As discussed in section 2, the strength of the clientele effect for an investor attribute can be measured by the empirical variance of the equal-weighted average investor attribute $a'Q_i^s$ across stocks, where a is an investor attribute rank ranging from -0.5 to 0.5, and the h'th element of Q_i^s is the reciprocal of the number of shareholders in stock i if investor h holds the stock, and zero otherwise (to deliver equal-weights across all investors holding the stock).[23]

Investor attributes include five account attributes that can be calculated without knowledge of the particular stocks held: account age and value, the number of stocks held and traded, and account turnover. We also calculate portfolio attributes that rely on knowledge of portfolio composition. These are the portfolio-weighted average ranks of the nine stock characteristics studied in Figure 4 (e.g., if a household held 50% of its portfolio in a stock at the extreme positive end of the distribution of a particular stock characteristic (rank 0.5), and 50% in a stock at the median (rank 0), this variable computed for that characteristic would take the value of 0.25 for this household). To avoid any mechanical relationship between a stock's characteristics and the attributes of portfolios that hold the stock, we compute the portfolio-weighted average ranks across each households' other stocks held (taking the same example as above, we would substitute values of 0 and 0.5 respectively depending on the stock in question).[24] This "leave-out" approach to factor construction is not typically applied in time-series analysis of factor

[23] Households may have the same value of a given attribute. For each unique value of the attribute, we compute the average of the ranks spanned by households with this value, and assign this average to each of those households. While certain household attributes (e.g. account turnover–often zero, number of stocks held) are shared by a significant fraction of all households, this assignment of ranks reduces the variance of a (across households) by less than 10 percent while preserving the mean of zero.

[24] For single-stock accounts, we set portfolio attributes equal to their average across households holding more than one stock. This ensures that single-stock accounts do not contribute to the clientele strength of portfolio attributes.

portfolios, but is important in our context because of the high concentration of many household portfolios.

In Figure 5 we plot clientele strength, again as a percentage of a hypothetical high-variance benchmark. Account attributes are plotted as solid red diamonds, while portfolio attributes are hollow red diamonds. Panel A works with raw attributes, while Panel B reports results for orthogonalized attributes.[25] As in Figure 4, we compare empirical clientele strength with what would be implied by alternatives. Equal-probability random stockpicking would imply clientele strength of almost exactly zero: since this model implies that every stock has the same large number of stockholders, if these stockholders are randomly selected then every stock has an almost identical shareholder base and there is almost no variance in investor attributes across stocks. Free-float-probability random stockpicking implies very slightly positive clientele strength, because some stocks have a small shareholder base and therefore some possibility of random variation in attributes. Empirical clientele strength exceeds both these benchmarks for all but the weakest orthogonalized attribute.

Among the account attributes, the strongest one is account age indicating a tendency for some stocks to be held by new investors and others to be held by experienced investors. Campbell, Ramadorai, and Ranish (2014) focuses more narrowly on this account age effect in the Indian data. The number of stocks traded and the size of the account also have fairly strong clientele effects when attributes are not orthogonalized, but the account value clientele effect is fairly weak after orthogonalization. Campbell, Ramadorai, and Ranish (2019) stresses that account size has an important impact on the degree of diversification and also influences style tilts; the results here indicate that some of these effects can be captured by portfolio attributes that are correlated with account size. The number of

[25] We orthogonalize household attributes in a similar manner to the orthogonalization of stock characteristics, but to reduce the computational burden of orthogonalization given the much larger size of a relative to c, we apply an OLS regression in each iteration instead of a multivariate kernel regression. Specifically, in iteration k, we regress the remaining attributes on a second degree polynomial of the k selected attributes. Both account and portfolio attributes are orthogonalized jointly, although orthogonalization has stronger effects within each category of attributes than across the two categories.

stocks held and account turnover are the account attributes with the weakest clientele effects.

One might expect that the portfolio attributes would have clientele effects that align with the characteristic clientele effects shown in Figure 4 and discussed in the previous subsection. Indeed share price, which has the second strongest characteristic clientele effect in Figure 4, has the strongest portfolio attribute effect in Figure 5. Some stocks have shareholders who tend to hold high-priced stocks while others have shareholders who tend to hold low-priced stocks, and this aligns with the previous finding that some household portfolios contain predominantly high-priced stocks while others contain predominantly low-priced stocks. However, market capitalization has a weak characteristic clientele effect in Figure 4 but a strong portfolio attribute clientele effect in Figure 5; and stock age has the strongest characteristic clientele effect in Figure 4 but only the third weakest portfolio attribute clientele effect in Figure 5.

These contrasting results reflect the different weighting scheme implicit in Figures 4 and 5. The measurement of characteristic clientele effects gives high weight to widely held stocks, while the measurement of attribute clientele effects weights all stocks equally. Hence, our results imply that there is little heterogeneity in household preference for large stocks among widely held stocks (which all tend to be large), but much more heterogeneity in this preference when we consider the full cross-section of stocks. Conversely, the heterogeneity in household preference for young or old stocks is concentrated in the widely held stocks, some of which are young (recent mega-IPOs) and others of which are long established companies.

Despite the existence of some contrasts in results, there is a clear message from both Figures 4 and 5. Investor clienteles exist in India and are stronger than can be explained by simple models of random or optimally diversifying stock selection. These clienteles are particularly associated with stock age and share price, both variables that are not normally the focus of research in asset pricing. Style characteristics such as the book-market ratio, the subject of an enormous academic literature on value vs. growth stocks, have comparatively weak investor clienteles. Given these results, we now turn to a factor model that we can use to identify and describe the types of households that make up specific characteristic clienteles.

6 Factor Models of Stockholdings

Our objective in this section is to link the two types of clientele effects we have discussed by estimating a multifactor model of the sort developed in Section 2. We begin by briefly describing the construction of our observed factors, then summarize stock-level coefficient estimates and the explanatory power of the different factors for the panel of all stocks. We conclude this section by contrasting the observed multifactor model results with those obtained from an unobserved (principal-component-based) factor analysis of stockholdings.

6.1 Observed Multifactor Model

6.1.1 Factor construction

We construct 14 account-attribute and portfolio-attribute factors from the account and portfolio attributes summarized in Table 1 and discussed in the previous section. To this we add several other sets of factors. First, we include 4 dummy variables to capture the broad geographical zones in which households are located. Second, we add industry factors which capture the share of the portfolio in each of 6 industry groups, namely, financial services; food agriculture and textiles; information technology; manufacturing; oil and gas; and other retail. Third, we add business group factors which capture the share of the portfolio in each of 11 business groups.[26] Finally, we add a dummy variable for

[26] To avoid collinear factors, we exclude the construction industry. We combine all business groups aside from the top ten into a single "other business group" for a total of 11 business group indicators.

single-stock accounts, for which portfolio attributes are unavailable given our "leave-out" construction, and we add a variable measuring the fraction of each investor's portfolio that pays dividends. The factors enter the model in raw form, without orthogonalization.

6.1.2 Estimation and results

We estimate stockholdings using all observed factors for each of our 3,103 stocks in our August 2011 sample. Each stock-specific cross-household regression is of the form shown in equation (7), and is run with 9.7 million household observations.

The factor loadings β_{ik} in these regressions are the product of unconstrained estimation, and have no mechanical correlation with the observable characteristics of any given stock. For example, it is entirely possible for a small stock to have a positive loading on the factor that measures the average size rank of households' stockholdings, if that small stock is typically co-held with large stocks. This allows our model to capture complex patterns of portfolio construction.

For ease of interpretation, we first divide each factor by its unconditional standard deviation in each stock-specific regression, and multiply it by 10^4 for readability. $\tilde{\beta}_{ik}$ is then the basis point increase in the portfolio weight of stock i for a one standard deviation increase in factor k.

Table 2 summarises the $\tilde{\beta}_{ik}$ estimated from the 3,103 stock-specific estimates of equation (7). The rows of the table correspond to the K factors, and the columns present various statistics of the cross-stock distribution of the betas estimated on these factors. The cross-stock mean $\tilde{\beta}_k$ measures the average loadings of a particular factor across 3,103 stocks. Our focus here is on the cross-sectional dispersion in these loadings as it is a necessary condition for a factor to be useful in predicting cross-sectional dispersion in household stockholdings. Given our focus, the first four columns of the table therefore summarize the cross-stock distribution of $\tilde{\beta}_{ik}$, presenting the cross-stock standard deviation, and the 10th, 50th, and the 90th percentiles of the cross-stock distribution of factor betas. The last two columns show the average $t-$statistic across all 3,103 regressions, and the percentage of estimated $\tilde{\beta}$â's that are statistically significantly different from zero at the 5% level.

Panel A of Table 2 shows the distribution of $\tilde{\beta}_{ik}$ for the account attribute factors, and Panel B summarizes the distribution of $\tilde{\beta}_{ik}$ for portfolio factors. The final two columns of both panels reveal that the majority of factors have high t-statistics on average, with a few exceptions such as realized skewness and some of the business group factors which are important for only a small number of stocks. In all cases, the fraction of coefficients that are statistically significant at the 5% level far exceeds the 5% that we would expect to see if our factors were noise uncorrelated with household portfolio decisions.

While the statistical significance of the factors is high on average, they exhibit very different levels of cross-stock variation. A necessary condition for a useful factor is that it helps to predict cross-sectional dispersion in household stockholdings. The equivalent in the standard returns setting is factors such as SMB and HML that exhibit a large cross-sectional spread in normalized factor loadings, and help to explain the time-variation in realized returns across stocks. We later discuss how specific stock characteristics are connected with account and portfolio attribute factors, but for now, we simply discuss the magnitude of the cross-stock spread in factor loadings seen in Table 2.

Account-attribute factors

The account-attribute factor with the highest cross-sectional standard deviation of factor loadings is the dummy for single-stock accounts. The cross-sectional distribution of loadings indicates that almost all stocks have a negative loading on this factor, but a few stocks—which we might call "entry-level" stocks—are particularly favored by single-stock investors and have a large positive loading.

The next most important account-attribute factor, again looking at the standard deviation of $\tilde{\beta}_{ik}$ across stocks, is account size, followed by turnover and account age. The numbers of stocks held and traded have smaller effects once we control for single-stock accounts using a dummy variable. The loadings on all these factors seem to be close to symmetric across stocks, as the median loading is close to zero.

There is some evidence of geography-based stock selection, which, as we show later, is mainly driven by local bias of the sort found by Coval and Moskowitz (1999) for US mutual funds. However, the geographical factors are only significant for 50-70% of stocks and are among the less important factors in the model. This may in part reflect the fairly coarse geographical information captured in our data.

Portfolio-attribute factors

Panel B of Table 2 turns to factors based on portfolio attributes. The table divides these factors into five categories, namely the Fama and French (1993) style factors capturing the size and value characteristics of household portfolios; return-based factors based on realized stock returns experienced in the portfolio; behavioral factors capturing revealed preferences through stockholdings for high or low share price, popular, old, high-turnover, or dividend-paying stocks; business group factors; and industry factors.

Once again using the factor loading cross-stock standard deviation as a guide, Panel B reveals strong variability for factors capturing industry holdings in information technology and manufacturing, as well as the market capitalization and stock age of other household portfolio holdings (recall that we use a leave-out construction). The importance of the market cap and stock age factors is consistent with their role in Figure 5, which like Table 2 gives equal weight to all stocks. However share price has a weaker effect than one might expect from Figure 5, while the book-market ratio has a stronger effect.

The loadings for most industry and business group factors are positively skewed, as we see from the negative median loadings. This reflects the fact that industry and business group factors strongly increase the probability of holding stocks in the same industry or business group, but weakly decrease the probability of holding all other stocks.

6.1.3 Explanatory power

Connor and Korajczyk (2019) introduce a way to assess the performance of specific groups of factors in multifactor models, classifying groups of factors as "natural rate," "semi-strong," and "weak." They define natural rate factors as those for which the sum of squared factor loadings increase proportionally to the number of assets in the data. In other words, natural rate factors explain the structure of additional data just as well as they explain the structure of data in the given sample. In contrast, semi-strong factors' sum of squared factor betas grow to infinity, but at a slower rate, and finally, weak factors are those with bounded eigenvalues. We later apply their asymptotic tests more rigorously to our data, but in a first step, we follow their methodology to conduct an informal analysis of the relative performance of the different groups of factors discussed above.

The approach that Connor and Korajczyk (2019) recommend is to first stack all $3,103$ stocks into a single pooled OLS regression. In our implementation, we regress the holdings of all stocks by all households onto stock dummies and stock dummies interacted with the set of observable factors F_k, effectively allowing stock-specific intercepts and factor loadings. In Table 3 we report the R^2 statistic from such a pooled regression. This measure of explanatory power captures the model's ability to explain which accounts hold the most widely held stocks, as these account for the bulk of the variance in the pooled stockholding data.

The first row of Table 3 shows that the R^2 of the full multifactor model is 7.61%. The remaining rows of the table show the contribution to explanatory power offered by each of the groups of factors included in the model. As suggested by Connor and Korajczyk (2019), we measure this contribution using the marginal R^2, which is the difference between the full-model R^2 and the R^2 of a model in which the set of factors under consideration is dropped. In each case, we express the contribution as a percentage of the full-model R^2. For example, the table shows that account-attribute factors contribute roughly 36% of the total explanatory power in the equally-weighted case, with portfolio-attribute factors accounting for roughly 67% of the total R^2. The two contributions do not add up to 100%, because the underlying factors are not orthogonal to one another.

Among the account-attribute factors, the single-stock dummy is most important and account size next most important. This is consistent with the patterns shown in Table 2 despite the fact that that table weights stocks equally rather than overweighting widely held stocks. This analysis helps to bring together disparate themes in prior literature on the influence of account characteristics on stockholding propensities into a common framework. For example, account size and wealth have been highlighted as important determinants of stockholdings behavior by Campbell et al. (2019) and Bach et al. (2020), and account age by Campbell et al. (2014) and Betermeier et al. (2017).

Among the portfolio-attribute factors, industry factors, Fama-French factors, and business group factors appear to contribute the most explanatory power, despite the typically modest level and cross-sectional spread seen in the loadings on these factors in Table 2.

6.2 Unobserved Multifactor Model

It is natural to ask how our observed multifactor model compares with an unobserved multifactor model based on principal components analysis (PCA). To construct such a model, we compute the principal components of the 3,103 by 3,103 covariance matrix of stockholdings derived from the 9.7 million accounts that we observe. The first principal component is the eigenvector of this covariance matrix which corresponds to the largest eigenvalue, and subsequent principal components are estimated as the eigenvectors associated with successively smaller eigenvalues of the covariance matrix. By construction, these principal components are orthogonal to one another, and are normalized linear combinations of household stockholdings that together summarize the total variance of stockholdings. They are ordered by the fraction of the total variance that they capture.

Following the statistical literature on factor models in stock returns, we briefly investigate the number of "natural rate" factors in the structure of coholdings using statistical tests suggested by Ahn and Horenstein (2013) and Connor and Korajczyk (2019). Natural rate factors are defined as those for which the sum of squared factor loadings increases proportionally with the number of assets in the data. In other words, natural rate factors explain the structure of additional data just as well as they explain the structure of data in the given sample.

Ahn and Horenstein (2013) suggest a procedure to detect the number of natural rate factors in the data. They show that if there are r natural rate factors, the ratio of successive eigenvalues should peak when comparing the eigenvalues of the rth and $r+$ 1th factors. Online appendix Figure A.2 shows this ratio of eigenvalues for the PCA unobserved factor model, using the Ahn and Horenstein (2013) recommended parameters applied to our empirical setting. This ratio peaks at two principal components, so the eigenvalue ratio test suggests that there are two unobserved factors.

In Table 4 we report the pooled R^2 statistic for a 10-factor PCA model and for the first, second, and third principal component. Since the principal components are orthogonal to one another by construction, their contributions can be added together to calculate the overall fit of single-factor, two-factor, and three-factor PCA models.

The left column of Table 4 works with the full sample, in which a single-factor PCA model has a considerably higher explanatory power (12.4%) than our observed multifactor model (7.6%). A ten-factor PCA model does even better with an explanatory power of 31.5%. However this explanatory power of PCA models is concentrated in a few widely held stocks. If we eliminate from the sample only the most widely held stock, Reliance Power, the explanatory power of the single-factor PCA model drops by almost ten percentage points to 22.0%, while that of the multifactor model declines more modestly to 6.1%. Once we eliminate the ten most widely held stocks from the sample, the observed multifactor model dominates even a ten-factor PCA model. This reflects the fact that PCA methodology applied to stockholdings concentrates on explaining patterns in the large number of holdings of a very few stocks.[27]

Figure 6 makes a similar point visually, presenting scatter diagrams that plot the explanatory power of our observed multifactor model against the explanatory power of a single-factor PCA model (panel A) or a ten-factor PCA model (panel B), separately for each stock in our sample. Points above the 45-degree line are stocks for which the observed multifactor model predicts holdings better than the PCA model. In panel A this is the case for all stocks except Reliance Power at the far right of the figure. In Panel B it is the case for almost all stocks, and the exceptions are mostly among the ten most widely held stocks (shown as red diamonds).

7 Multifactor Model Insights

In this section we put our multifactor models to work. We first ask how well they describe stock-level coholdings, comparing empirically observed coholdings across all pairs of stocks with those generated by our factor models. We then ask what our multifactor models tell us about the clientele for each of the stock characteristics that we discussed in section 5. Finally, we ask how the coholdings matrix relates to the covariance matrix of stock returns, and how our measures of clientele strength relate to the return variances of the corresponding factor portfolios.

7.1 Empirical and Model-Implied Coholdings

The elements of the model-predicted coholdings matrix are given by equation (9). To facilitate interpretation, given the substantial empirical variation in coholdings that is driven by variation in holdings intensity, we convert the predicted and actual coholdings matrices into coholdings correlation matrices by dividing the elements of the sample co-

[27] Figure A.7 in the online appendix further develops this theme by regressing the first ten PCA factors onto our observed factors. The largest coefficients are on business group factors, as the most widely held stocks tend to be in popular business groujps.

holdings matrix and the model predicted coholdings matrix by the geometric average of their corresponding diagonal elements. The diagonal elements of the (actual and predicted) holdings correlation matrices then equal 1, and the off-diagonal elements range between -1 and 1.

Most of these correlations are small, and a few are very large. In addition, some correlations are negative. To handle this, we rank the correlations and in Figure 7 we plot the correlation rank from the data against the correlation rank implied by the model. The observed factor model performs well as illustrated by the fact that a regression line has a slope of 0.627.

7.2 Stock Characteristics and Factor Loadings

In this section we investigate the extent to which the factor loadings of stocks are related to those stocks' own characteristics. By aggregating loadings using characteristic ranks, we can learn about the nature of the clientele for each stock characteristic. Taking the estimated loadings for each stock and each factor in our model, we construct weighted average loadings, using our various stock characteristics' demeaned ranks as weights. This tells us which types of accounts make up the clientele for each characteristic.

Table 5 shows the results of this exercise, with each column representing results for a different stock characteristic, orthogonalized in the manner described earlier. The top panel of the table reports results for account attributes listed in each of the row headers, and the bottom panel reports results for portfolio attributes, once again described in the row headers. Coefficients are colored to indicate the sign and strength of each relationship, with positive relationships in red, negative relationships in blue, and stronger relationships in darker colors.

Table 5 is initially most easily read by viewing each row in turn to see the types of stocks favored by particular types of accounts. For example, the first row of the top panel of the table shows that controlling for other account attributes, older accounts prefer older stocks with lower share prices, positive momentum (high past realized returns), lower beta, higher book-market ratios, higher volatility, and lower market capitalization. Many of these patterns are consistent with those documented by Campbell, Ramadorai, and Ranish (2014) in a study focusing exclusively on account age.

The second row of the top panel of Table 5 shows that larger accounts tend to hold older stocks with high share prices and high market capitalization, but low book-market, volatility, and realized skewness. In other words they prefer large, established growth companies. These style tilts are similar to those documented by Campbell, Ramadorai, and Ranish (2019) in a study focusing exclusively on account size.

The third row of the top panel of Table 5 shows that accounts with high turnover, by contrast, prefer smaller, cheaper stocks with high turnover, volatility, book-market, and realized skewness, in other words easily traded stocks with "lottery-like" characteristics. The pattern is similar and extremely strong for single-stock accounts. These accounts strongly favor low-priced, small stocks with high book-market and volatility, again lottery-like stocks, although they do not particularly tend to hold stocks with high turnover.

The bottom panel of Table 5 shows how investors' portfolio attributes are related to stock characteristics. The diagonal elements of the panel show the extent to which the tendency to hold a particular stock with a given characteristic can be predicted by holdings of other stocks with the same characteristic. The off-diagonal elements show the extent to which holdings of stocks with particular characteristics are predictive of investors' holdings of stocks with other characteristics. Interestingly, many of the off-diagonal elements are larger in absolute value than the diagonal elements, indicating that stock characteristics appear to cluster into groups, with similar investor clienteles holding constellations of these characteristics simultaneously. (This is not mechanical, as we have orthogonalized the characteristics against one another in the cross-section of stocks in this table.)

Reading down the columns of the bottom panel of Table 5, we see that holdings of stocks along the share price and market capitalization dimensions are predicted similarly by portfolio attributes, as are holdings of stocks along the dimensions of turnover and beta. Book-market, realized volatility, and realized skewness also seem to constitute such a group of characteristics. The columns of the top panel of Table 5 show a similar pattern. Thus, our multifactor model tells us that investor clienteles form around related stock characteristics: size and share price; turnover and beta; and book-market, volatility, and skewness.[28]

7.3 Coholdings and Return Correlations

Figure 8 plots the relationship between return correlations and measures of coholdings. The return correlation estimates are based on weekly Indian stock returns data for the year leading up to August 2011, when we estimate coholdings. The blue diamonds in the figure show the relationship between estimated return correlations on the vertical axis and empirical coholdings correlations on the horizontal axis, while the green circles replace empirical coholdings correlations with observed-factor model-implied coholdings correlations. The relationships are estimated by simply regressing return correlations on the vertical axis on one-percentile-point bin-dummies of the holdings correlation measures shown on the horizontal axis.

The figure shows a clear, statistically significant, and positive relationship between return correlations and coholdings correlations for stocks. The R^2 from a linear regression of return correlations on raw coholdings correlations is about 4% for the empirical coholdings, implying that the correlation between these two correlations is roughly 20%. This rises to $R^2 = 10\%$, or a correlation of roughly 31%, when return correlations are regressed on model-implied coholdings correlations. Moreover, the non-linearity visible with the raw coholdings correlation, especially at very low values, becomes less pronounced, and

[28] Again, it is worth noting that this clientele structure does not result mechanically from stock-level characteristics across correlations, because we have orthogonalized characteristics in the cross-section of Indian stocks.

the relationship more obviously monotonic when the model-implied coholdings correlation is employed on the x-axis of the figure.

While these are preliminary observations, the positive relationship between coholdings correlations and return correlations is intriguing. If Indian investors were attempting to diversify portfolios with a small number of stocks, they would tend to cohold stocks with relatively low return correlations. On the other hand, if investor clienteles buy and sell co-held stocks at the same time, that could lead to a positive relationship between coholdings and return correlations, and could increase the return volatility of characteristic portfolios that have strong clienteles. More generally, in equilibrium asset pricing models holdings and returns are jointly determined, and different models have different implications for the relationship between them. The results in this section warrant further investigation, as they are a first step to more deeply understanding the empirical relationships between holdings and returns.

8 Conclusion

In this paper we have suggested that a factor model for investors' stockholdings provides a natural way to understand household portfolio decisions and the structure of investor clienteles for different types of stocks. The model is a cross-sectional analogue to the time-series factor models that are commonly used to describe the variation in stock returns over time. We have applied the model to comprehensive administrative data from India, where direct stockholdings are the norm at the time of our analysis. While direct stockholdings have become less prevalent over the longer run in many advanced economies, we note that they are, at the time of writing this paper, experiencing an unusual resurgence around the world, accompanied by substantial increases in trading volume by retail investors.

Our main emphasis is on a model with multiple observable factors, some related to account characteristics such as the number of stocks held, and others related to the characteristics of accounts' stockholdings such as their average market capitalization. We find that this model exhibits good performance in comparison with an unobservable PCA-based factor model, and provides a good description of the empirical coholdings matrix.

Certain characteristics of stocks seem to have strong clientele effects associated with them, meaning that many investors' portfolios load either positively or negatively on these characteristics. The strongest characteristic clienteles are associated with firm age and share price, even though these are not characteristics that attract a great deal of attention in the asset pricing literature. Clientele effects are relatively weak for Fama-French style characteristics despite their importance in academic asset pricing research and in the organization of the US mutual fund industry.

We use our model to estimate which types of accounts hold which stocks and make up the clienteles for these characteristics. We find that single-stock accounts have strong preferences for particular types of stocks, as do older vs. younger accounts and larger vs. smaller accounts. By including all these account attributes in a single model, we are able to compare their importance rather than consider their effects on portfolio choice in isolation as most previous research has done.[29] Also, we find that characteristics form clusters with similar clienteles, even after we orthogonalize those characteristics in the cross-section of Indian stocks. Market capitalization and share price have similar clienteles, as do turnover and beta, and book-market, volatility, and skewness. In other words different types of investors show preferences for large, well known "quality" stocks, for highly traded stocks that move with the overall market, and for risky, lottery-like stocks.

Finally, we present a preliminary finding on the relation between coholdings and comovement of stock returns. Stocks that are more commonly coheld tend to correlate

[29] For example, account size and wealth have been highlighted as important determinants of stockholdings behavior by Campbell et al. (2019) and Bach et al. (2020), and account age by Campbell et al. (2014) and Betermeier et al. (2017).

more strongly with one another. This pattern runs counter to the view that investors optimally diversify their portfolios conditional on a constraint on the number of stocks held, but it reinforces the idea that clientele effects, captured by coholdings propensities, contribute to common variation in stock returns.

References

Agarwalla, S. K., J. Jacob, and J. R. Varma (2013). Four factor model in Indian equities market. Working Paper W.P. No. 2013-09-05, Indian Institute of Management, Ahmedabad.

Ahn, S. C. and A. R. Horenstein (2013). Eigenvalue ratio test for the number of factors. *Econometrica 81*(3), 1203–1227.

Amihud, Y. and H. Mendelson (1986). Asset pricing and the bid-ask spread. *Journal of Financial Economics 17*(2), 223–249.

Anagol, S., V. Balasubramaniam, and T. Ramadorai (2018). Endowment effects in the field: Evidence from India's IPO lotteries. *The Review of Economic Studies 85*(4), 1971–2004.

Anagol, S., V. Balasubramaniam, and T. Ramadorai (2021). The effects of experience on investor behavior: Evidence from India's IPO lotteries. *Journal of Financial Economics forthcoming.*

Anagol, S. and A. Pareek (2019). Should business groups be in finance? Evidence from Indian mutual funds. *Journal of Development Economics 139*, 229–248.

Bach, L., L. E. Calvet, and P. Sodini (2020). Rich pickings? risk, return, and skill in household wealth. *American Economic Review 110*(9), 2703–2747.

Balasubramaniam, V., J. Y. Campbell, T. Ramadorai, and B. Ranish (2020). Online appendix to who owns what? A factor model for direct stockholding.

Barber, B. M., Y.-T. Lee, Y.-J. Liu, and T. Odean (2009). Just how much do individual investors lose by trading? *Review of Financial Studies 22*(2), 609–632.

Barber, B. M. and T. Odean (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors. *Journal of Finance 55*(2), 773–806.

Barber, B. M. and T. Odean (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *Quarterly Journal of Economics 116*(1), 261–292.

Betermeier, S., L. E. Calvet, and P. Sodini (2017). Who are the value and growth investors? *Journal of Finance 72*(1), 5–46.

Calvet, L. E., J. Y. Campbell, and P. Sodini (2007). Down or out: Assessing the welfare costs of household investment mistakes. *Journal of Political Economy 115*(5), 707–747.

Campbell, J. Y., T. Ramadorai, and B. Ranish (2014). Getting better or feeling better? How equity investors respond to investment experience. Technical report, National Bureau of Economic Research Working Paper 20000, Available at SSRN: https://ssrn.com/abstract=2176222.

Campbell, J. Y., T. Ramadorai, and B. Ranish (2019). Do the rich get richer in the stock market? Evidence from India. *American Economic Review: Insights 1*(2), 225–40.

Chamberlain, G. and M. Rothschild (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica 51*(5), 1305–1324.

Connor, G. and R. A. Korajczyk (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of Financial Economics 15*(3), 373–394.

Connor, G. and R. A. Korajczyk (2019). Semi-strong factors in asset returns. *Available at SSRN 3419446*.

Coval, J. D. and T. J. Moskowitz (1999). Home bias at home: Local equity preference in domestic portfolios. *Journal of Finance 54*(6), 2045–2073.

Dorn, D. and G. Huberman (2010). Preferred risk habitat of individual investors. *Journal of Financial Economics 97*(1), 155–173.

Døskeland, T. M. and H. K. Hvide (2011). Do individual investors have asymmetric information based on work experience? *Journal of Finance 66*(3), 1011–1041.

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics 33*(1), 3–56.

Gomes, F., M. Haliassos, and T. Ramadorai (2020). Household finance. *Journal of Economic Literature, forthcoming*.

Grinblatt, M., S. Ikäheimo, M. Keloharju, and S. Knüpfer (2016). IQ and mutual fund choice. *Management Science 62*(4), 924–944.

Grinblatt, M. and M. Keloharju (2000). The investment behavior and performance of various investor types: A study of Finland's unique data set. *Journal of Financial Economics 55*(1), 43–67.

Grinblatt, M. and M. Keloharju (2001). How distance, language, and culture influence stockholdings and trades. *The Journal of Finance 56*(3), 1053–1073.

Hong, H. and M. Kacperczyk (2009). The price of sin: The effects of social norms on markets. *Journal of Financial Economics 93*(1), 15–36.

Jayaraj, D. and S. Subramanian (2008). Adjusting headcount deprivation for horizontal and spatial inequality: Some illustrative examples using census housing data. *Indian Journal of Human Development 2*(2), 425–434.

Kaniel, R., G. Saar, and S. Titman (2008). Individual investor trading and stock returns. *Journal of Finance 63*(1), 273–310.

Koijen, R. S. and M. Yogo (2019). A demand system approach to asset pricing. *Journal of Political Economy 127*(4), 1475–1515.

Kozak, S., S. Nagel, and S. Santosh (2018). Interpreting factor models. *The Journal of Finance 73*(3), 1183–1223.

Liao, J., C. Peng, and N. Zhu (2020). Price and volume dynamics in bubbles. *Available at SSRN 3188960*.

Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics 47*(1), 13–37.

Markowitz, H. M. (1952). Portfolio selection. *Journal of Finance 7*(1), 77–91.

Martins, R., H. Singh, and S. Bhattacharya (2012). What does volume reveal: A study of the Indian single stock futures market. *Indian Journal of Economics & Business 11*(2), 409–419.

Massa, M. and A. Simonov (2006). Hedging, familiarity and portfolio choice. *Review of Financial Studies 19*(2), 633–685.

Mayers, D. et al. (1972). Nonmarketable assets and capital market equilibrium under uncertainty. *Studies in the theory of capital markets 1*, 223–48.

Merton, R. C. (1973). An intertemporal capital asset pricing model. *Econometrica 41*, 867–887.

Odean, T. (1998). Are investors reluctant to realize their losses? *Journal of Finance 53*(5), 1775–1798.

Pástor, L., R. F. Stambaugh, and L. A. Taylor (2020). Fund tradeoffs. *Journal of Financial Economics*.

Ross, S. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory 13*(3), 341–360.

Seru, A., T. Shumway, and N. Stoffman (2010). Learning by trading. *Review of Financial Studies 23*(2), 705–739.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance 19*(3), 425–442.

Vashishtha, A. and S. Kumar (2010). Development of financial derivatives market in India: A case study. *International Research Journal of Finance and Economics 37*(37), 15–29.

Table 1
Summary Statistics

This table provides means, standard deviations and quantiles of the main variables of interest for the August 2011 cross-section of roughly 9.7 million individual investors in the 3,103 stocks in our sample. Age is the number of months since the investor opened their first depository account. Size is the investors' USD value of all holdings of stocks in our sample. Turnover is the investors' average monthly value of trades over the past year (Sep. 2010-Aug. 2011) divided by the lagged (August 2011) portfolio size. Turnover is winsorized at the 99th percentile. No. Stocks is the number of stocks in our sample held by the investor. No. Stocks Traded is the number of unique stocks traded by the investor over the past year. Stockholding characteristics represent investors' portfolio weighted average stock characteristic, where the stock characteristics are rank normalized on the interval [-0.5, 0.5]. Stock Age is the number of months since the stock began public trading. Book/Market is constructed using the latest book value as of December 2010. Turnover and Realized Volatility, Returns, Skewness, and Market Beta are measured over the previous year, using weekly data.

Variable Name	Mean	Std. Dev.	P10	P25	Median	P75	P90
Account Characteristics							
Age	61.30	36.89	16.00	39.00	52.00	84.00	124.00
Size ('000s USD)	11.54	533.43	0.04	0.14	0.78	3.54	13.01
Turnover	0.38	1.17	0.00	0.00	0.02	0.18	0.71
No. Stocks	8.45	16.48	1.00	1.00	4.00	9.00	20.00
No. Stocks Traded	4.74	11.24	0.00	0.00	1.00	5.00	13.00
Stockholding Characteristics							
Share Price	0.22	0.21	-0.06	0.13	0.26	0.38	0.44
Stock Age	-0.06	0.27	-0.43	-0.30	-0.08	0.15	0.34
Realized Volatility	-0.17	0.18	-0.35	-0.30	-0.21	-0.08	0.09
Market Capitalization	0.38	0.17	0.18	0.37	0.45	0.48	0.49
Realized Returns	-0.02	0.20	-0.28	-0.16	0.00	0.10	0.22
Turnover	0.08	0.19	-0.14	-0.02	0.07	0.22	0.33
Market Beta	0.11	0.18	-0.12	-0.02	0.12	0.23	0.34
Book/Market	-0.14	0.18	-0.33	-0.25	-0.19	-0.07	0.08
Realized Skewness	-0.15	0.19	-0.34	-0.30	-0.17	-0.05	0.12

Table 2
Multifactor Regression Estimates

For each stock i we run the regression specification in Equation (7) of the paper over the set of 9.7 million individual investors in August 2011. This table summarizes the coefficients across the 3,103 stocks, presented in terms of the basis point change in portfolio share per standard deviation change in the factor. Each row in Panel A corresponds to an account characteristic factor, and each row in Panel B corresponds to a stockholding characteristic factor. Columns show the standard deviation, 10^{th}, 50^{th}, 90^{th} percentiles of the cross-sectional distribution, respectively. The last two columns present the average of the absolute values of the $t-$statistic, and the percent of stocks that are statistically significant at the 5% level.

Panel A: Account Characteristics

	Std. Dev	10%	50%	90%	Avg. |t-stat|	Sig.(5% level)
Age	4.67	-0.93	-0.01	0.44	15.72	86.88
Size	14.37	-0.66	0.04	1.62	21.80	91.59
Turnover	5.98	-0.18	-0.01	0.26	6.54	65.03
No. Stocks	2.36	-0.22	-0.01	0.19	4.57	56.85
No. Stocks Traded	1.71	-0.16	0.01	0.29	4.89	55.40
Single Stock Dummy	40.54	-11.12	-1.28	-0.33	44.57	99.84
Geographic Region						
Southern	2.73	-0.22	0.01	0.38	6.43	57.94
Northern	3.71	-0.31	-0.01	0.15	4.68	52.59
Western	4.66	-0.46	0.02	0.27	6.44	69.93

Panel B: Stockholding Characteristics

	Std. Dev	10%	50%	90%	Avg. |t-stat|	Sig.(5% level)
Fama-French factors						
Book/Market	8.17	0.08	0.31	2.31	25.00	98.97
Market Capitalization	9.79	-2.87	-0.45	-0.13	32.58	99.90
Market Beta	4.90	-1.05	-0.15	-0.05	12.28	98.58
Return-based factors						
Realized Volatility	4.81	0.08	0.26	1.67	18.87	99.48
Realized Returns	2.58	0.03	0.11	0.84	9.55	92.72
Realized Skewness	2.20	0.03	0.12	0.83	8.88	95.71
Behavioral factors						
Share Price	1.64	-1.01	-0.20	-0.06	10.41	96.07
Stock Age	7.09	0.07	0.23	1.92	17.38	98.32
Turnover	4.92	-0.96	-0.15	-0.05	12.38	97.42
Dividend Paying	27.29	-6.87	-0.79	-0.20	32.38	99.52
Business Group Holdings						
Reliance (ADAG)	3.99	-1.26	-0.15	-0.04	11.91	98.26
Tata	1.01	-0.02	0.01	0.12	1.86	20.43
Reliance (DAG)	1.44	-0.29	-0.02	0.04	3.00	37.06
Birla Aditya	0.79	-0.13	-0.01	0.01	1.97	25.20
Jaypee	2.21	0.02	0.07	0.53	6.42	88.01
Jindal	0.85	-0.22	-0.02	0.00	2.72	40.77
Mahindra	1.90	-0.47	-0.04	0.00	4.71	61.23
Suzlon	0.99	-0.23	-0.02	0.02	2.89	43.38
Vedanta	0.53	0.00	0.01	0.13	1.80	23.65
Adani	0.73	-0.07	0.00	0.02	1.60	16.47
Others	9.58	-2.20	-0.29	-0.07	22.59	99.36
Industry Holdings						
Financial Services	7.11	-2.02	-0.19	-0.02	15.57	87.46
Food, Agri. and Textiles	4.26	-1.13	-0.13	-0.03	11.67	92.85
Information Technology	11.92	-3.03	-0.38	-0.10	29.28	99.87
Manufacturing	15.16	-3.95	-0.45	-0.12	28.10	99.77
Oil and Gas	4.11	-1.06	-0.12	-0.03	8.59	91.75
Other Retail	1.18	-0.01	0.02	0.22	2.73	32.77

Table 3

Explanatory Power

This table presents the relative contribution of different groups of factors to the explanatory power of the regressions summarized in Table 2. The first row in Panel A presents the full model R-squared from a pooled least squares model with stock-specific intercepts and loadings on F_k. In each row following the first, we re-estimate this model excluding factors corresponding to the characteristic(s) listed at left, and report the reduction in R-squared that results (i.e. the marginal R-squard) as a percentage of the full model R-squared.

	Unweighted
Full R-squared	7.61
	Percent of Full R-squared
Account Characteristics based Factors	36.43
One Stock Accounts	18.46
Size	7.34
Geographic factors	2.21
Turnover	2.33
Age	1.39
No. Stocks	0.22
Stockholding Characteristics based Factors	66.83
Industry factors	15.55
Business group factors	6.83
Fama-French factors	10.77
Return factors	1.95
Behavioral factors	3.46

Table 4

Contribution to Explanatory Power: Without Top N Stocks

The first two rows of this table compare the R-squareds of the observed factor model in Table 3 with the R-squared from the first ten principal components of the stockholding data. The left column presents an R-squared across all 3,103 stocks, with columns to the right presenting R-squareds excluding the top 1, 10 and 50 stocks from the calculation. The bottom three rows of the table present R-squareds associated with each of the first three principal components.

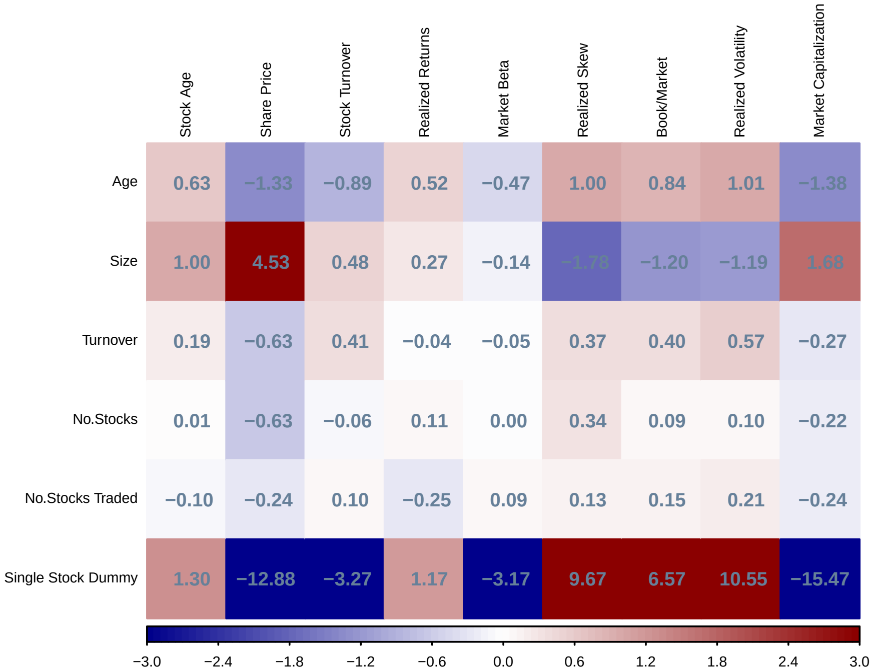
	Full sample	W/o Top 1	W/o Top 10	W/o Top 50
Observed Factor Model	7.61	6.05	2.00	0.76
PCA 1-10	31.51	21.99	1.36	0.04
PC1	12.37	0.33	0.04	0.01
PC2	6.77	7.58	0.03	0.01
PC3	2.00	2.28	0.02	0.00

Table 5

Factor Loadings and Stock Characteristic Clienteles: All factors

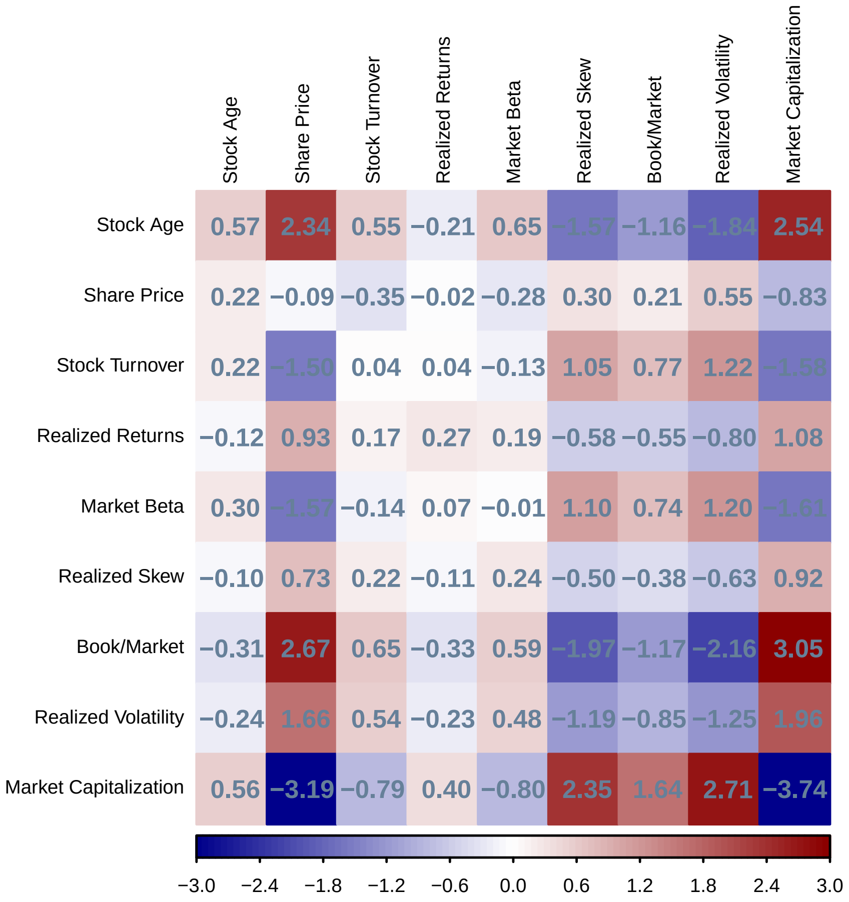
This table is structured as Table 5, but uses a regression of stock holdings Q_{ih}^{v} on both account and stockholding characteristics. Panel A presents stock characteristic weighted loadings on account characteristics, and Panel B presents stock characteristic weighted loadings on stockholding characteristics.

Panel A: Account Characteristics



	Stock Age	Share Price	Stock Turnover	Realized Returns	Market Beta	Realized Skew	Book/Market	Realized Volatility	Market Capitalization
Age	0.63	−1.33	−0.89	0.52	−0.47	1.00	0.84	1.01	−1.38
Size	1.00	4.53	0.48	0.27	−0.14	−1.78	−1.20	−1.19	1.68
Turnover	0.19	−0.63	0.41	−0.04	−0.05	0.37	0.40	0.57	−0.27
No.Stocks	0.01	−0.63	−0.06	0.11	0.00	0.34	0.09	0.10	−0.22
No.Stocks Traded	−0.10	−0.24	0.10	−0.25	0.09	0.13	0.15	0.21	−0.24
Single Stock Dummy	1.30	−12.88	−3.27	1.17	−3.17	9.67	6.57	10.55	−15.47

Panel B: Stockholding Characteristics



	Stock Age	Share Price	Stock Turnover	Realized Returns	Market Beta	Realized Skew	Book/Market	Realized Volatility	Market Capitalization
Stock Age	0.57	2.34	0.55	−0.21	0.65	−1.57	−1.16	−1.84	2.54
Share Price	0.22	−0.09	−0.35	−0.02	−0.28	0.30	0.21	0.55	−0.83
Stock Turnover	0.22	−1.50	0.04	0.04	−0.13	1.05	0.77	1.22	−1.58
Realized Returns	−0.12	0.93	0.17	0.27	0.19	−0.58	−0.55	−0.80	1.08
Market Beta	0.30	−1.57	−0.14	0.07	−0.01	1.10	0.74	1.20	−1.61
Realized Skew	−0.10	0.73	0.22	−0.11	0.24	−0.50	−0.38	−0.63	0.92
Book/Market	−0.31	2.67	0.65	−0.33	0.59	−1.97	−1.17	−2.16	3.05
Realized Volatility	−0.24	1.66	0.54	−0.23	0.48	−1.19	−0.85	−1.25	1.96
Market Capitalization	0.56	−3.19	−0.79	0.40	−0.80	2.35	1.64	2.71	−3.74

Figure 1

Number of Investors per Stock

This figure plots the cross-sectional distribution of the number of investors holding each stock in August 2011 sample. The $x-$axis plots the percentile cut-offs from 0 to 100, the left $y-$axis shows the number of investors (logarithmic scale), and the right $y-$axis shows the corresponding percent share of investors (%). The 10 most widely held stocks and the share of investors holding them are: Reliance Power limited (40%), Reliance Industries limited (26%), Reliance Communications limited (12%), National Hydro Power Corporation (12%), Power Grid Corporation of India (11%), Suzlon Energy limited (9.5%), National Thermal Power Corporation (8%), Tata Steel limited (8%), Larsen and Toubro limited (7.5%), Reliance Infrastructure limited (7.5%).

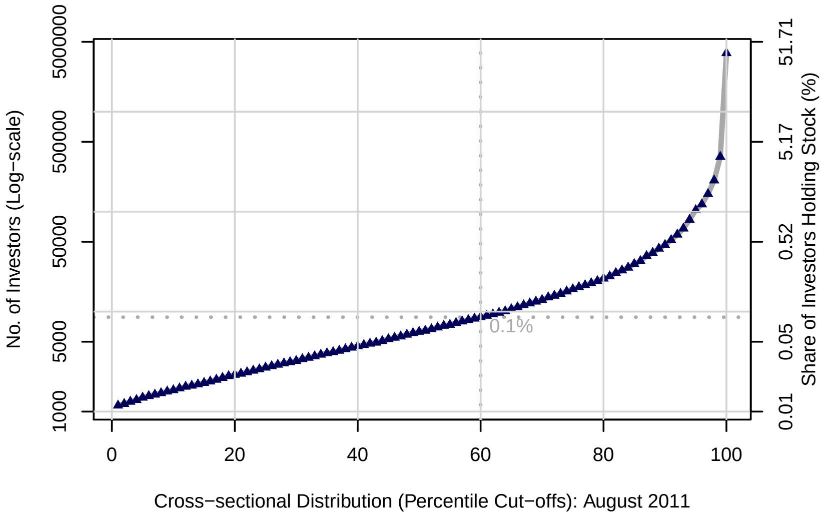


Figure 2
Average Portfolio Share

This figure plots the cross-sectional distribution of the average portfolio share of each stock in August 2011, both across all individual investors (blue curve, left hand side axis) and only those investors holding the stock (green curve, right hand side axis).

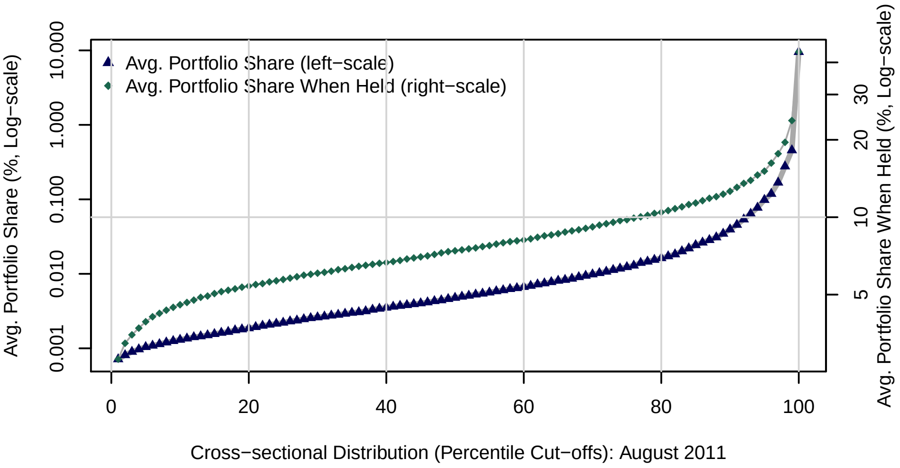
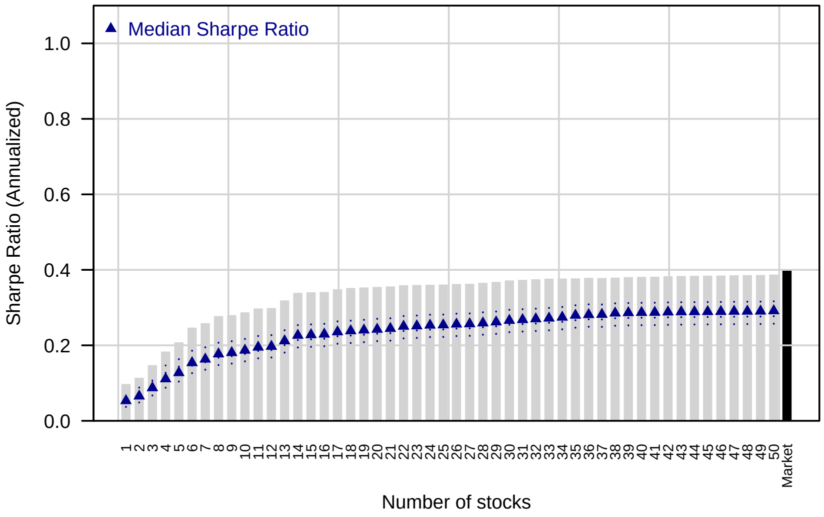


Figure 3

CAPM and Four Factor-Implied Sharpe Ratios

Panel A presents the annualized Sharpe ratio from the best N stock CAPM-implied portfolio. The $x-$axis represents the number of stocks in the portfolio, with the market portfolio as the last bar in the plot. The Sharpe ratio estimates are based on weekly returns data for the period March 2003 until August 2011. The triangle plots the median CAPM implied Sharpe ratio for accounts in our data, for the same time period, and the dotted lines represent the range from the 10th to the 90th percentile of the household Sharpe ratio distribution. Panel B presents the annualized Sharpe ratio from the best N stock Four Factor-implied portfolio for the same time-period, and the four factor-implied Sharpe ratio for households, similar to Panel A.

Panel A: CAPM-Implied Sharpe Ratio Estimates



Panel B: Four Factor-Implied Sharpe Ratio Estimates

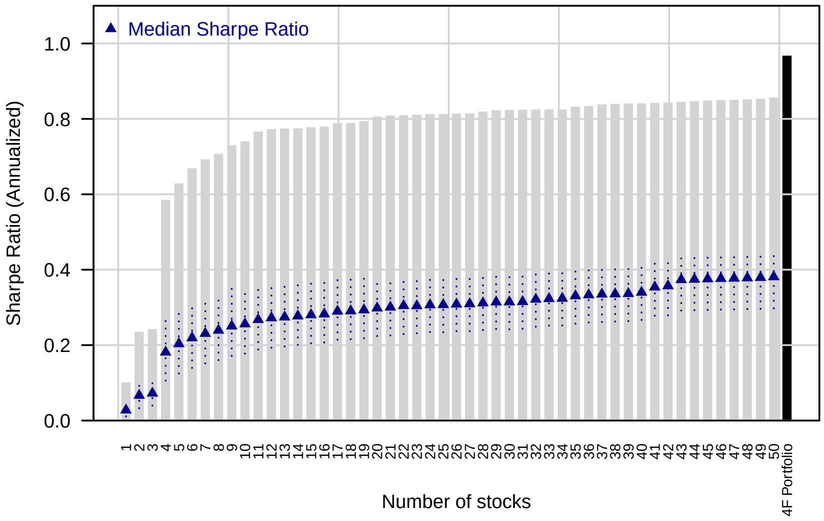
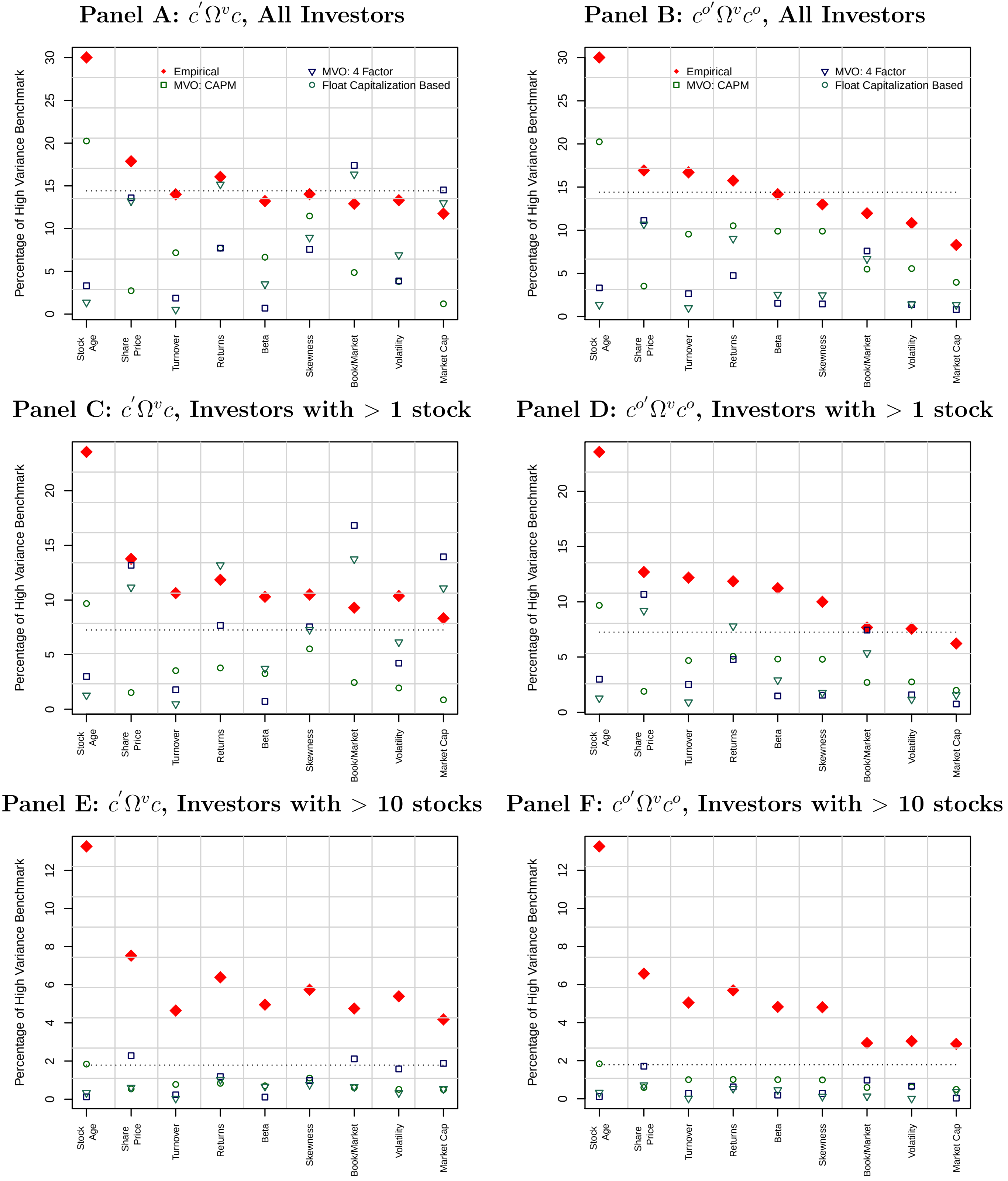


Figure 4

Stock Characteristic Clientele Strength

This figure compares the variance of stockholding characteristics across investors (clientele strength $c^{'}\Omega^{v}c$) in our data against the variances generated under alternatives described further in the text. The dashed lines represent variance under random stock selection, and all variances are expressed as a percentage of the maximum obtainable. Right hand side panels use stock characteristics that have been sequentially orthogonalized as described in the text. The top two panels include all investors, while the middle and lower panels exclude investors holding one stock and ten or fewer stocks respectively.



Panel A: $c^{'}\Omega^{v}c$, All Investors

Panel B: $c^{o'}\Omega^{v}c^{o}$, All Investors

Panel C: $c^{'}\Omega^{v}c$, Investors with $>$ 1 stock

Panel D: $c^{o'}\Omega^{v}c^{o}$, Investors with $>$ 1 stock

Panel E: $c^{'}\Omega^{v}c$, Investors with $>$ 10 stocks

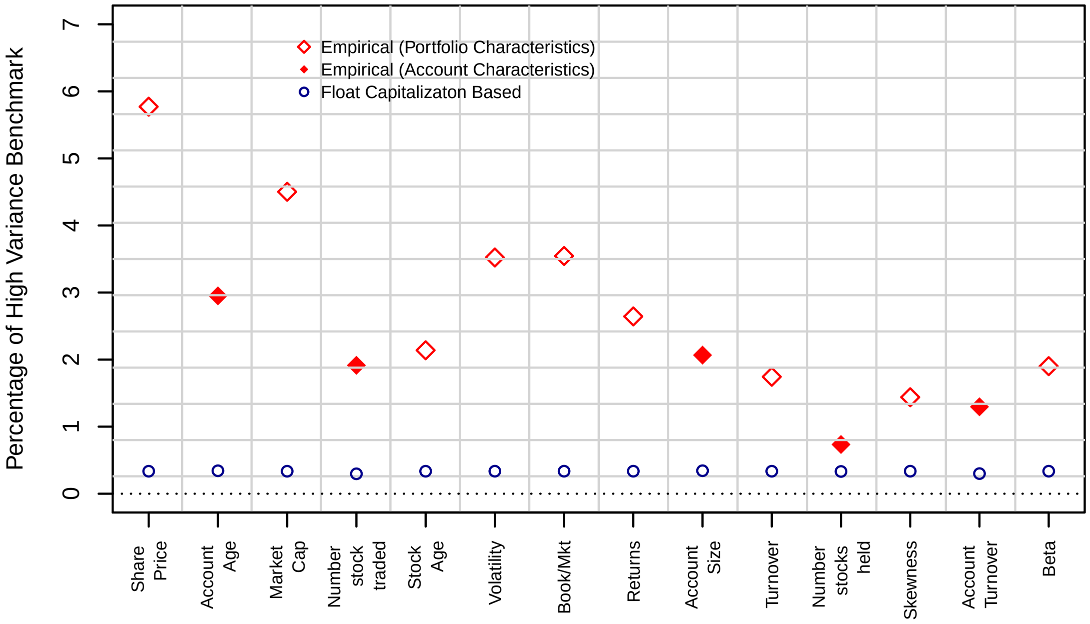
Panel F: $c^{o'}\Omega^{v}c^{o}$, Investors with $>$ 10 stocks

Figure 5

Investor Attribute Clientele Strength

This figure compares the variance of investor (account and stockholding) characteristics of investor bases across stocks in our data against the variances generated under alternatives described further in the text. The dashed lines represent variance under random stock selection, and all variances are expressed as a percentage of the maximum obtainable. The top panel uses ranked investor characteristics on the interval [-0.5, 0.5], and the bottom panel uses ranked sequentially orthogonalized characteristics as described in the text. Stockholding characteristics of investors in each stock are constructed based on investors' other stocks held.

Panel A: $a'\Omega^s a$



Panel B: $a^{o\prime}\Omega^s a^o$

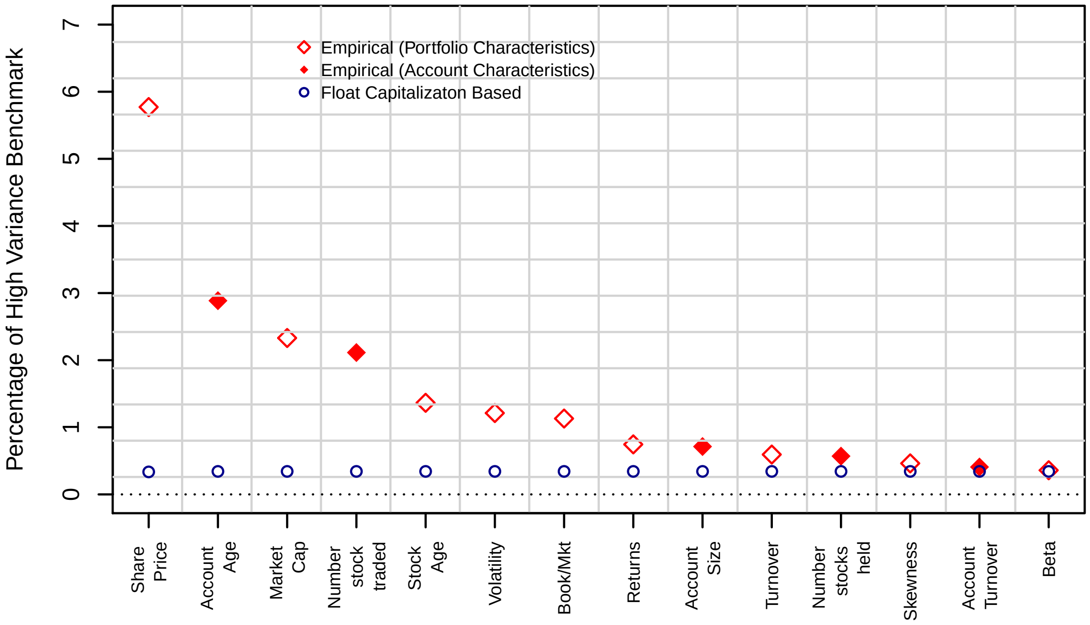
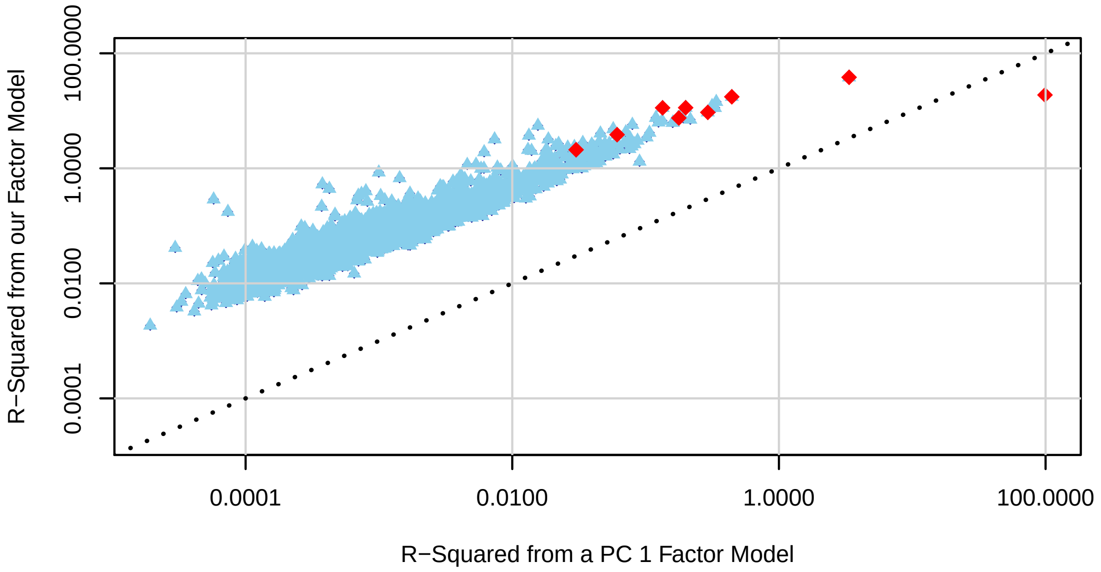


Figure 6

Comparison of Stock Level R-Squareds

This figure presents a stock-by-stock comparison of the R^2 estimates from the observed factor model ($y-$axis), and the unobserved factor model ($x-$axis), both on logarithmic scales. The dashed line marks the 45-degree line. The triangles (diamonds) are stocks in which the observed factor model does better (worse) than the unobserved PCA model. The red diamonds represent the top 10 stocks by the share of investors holding the stock. Panel A presents a comparsion to a 1-factor model, while panel B presents a comparison to a PC1-10 factor model.

Panel A: Observed Multifactor model vs. PC 1 Factor model



Panel B: Observed factor model vs. PC 1-10 Factor model

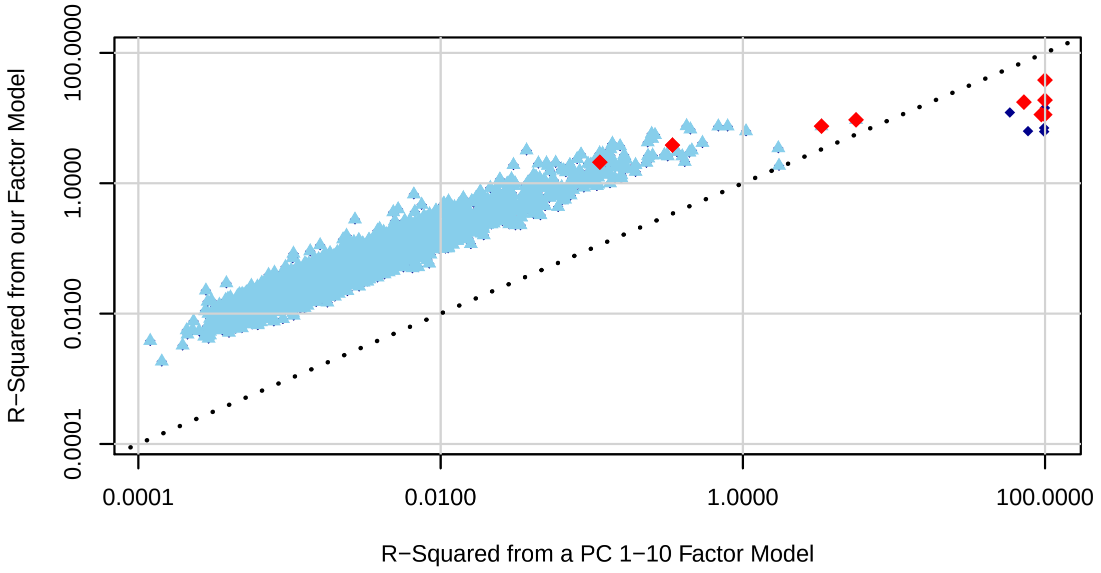


Figure 7
Empirical vs Model Coholdings

This figure plots the empirical coholding likelihood ($y-$axis), against the model-implied coholding likelihood measured as in Equation (9) ($x-$axis). Both variables have been rank normalized to be between zero and one. The grey points represent a 1% random sample of stock-pairs in the data, and the dotted lines the 45-degree line. The solid line represents the linear relationship between the variables, with a slope of 0.627 (s.e = 0.0004), and $R^2 = 39.3\%$.

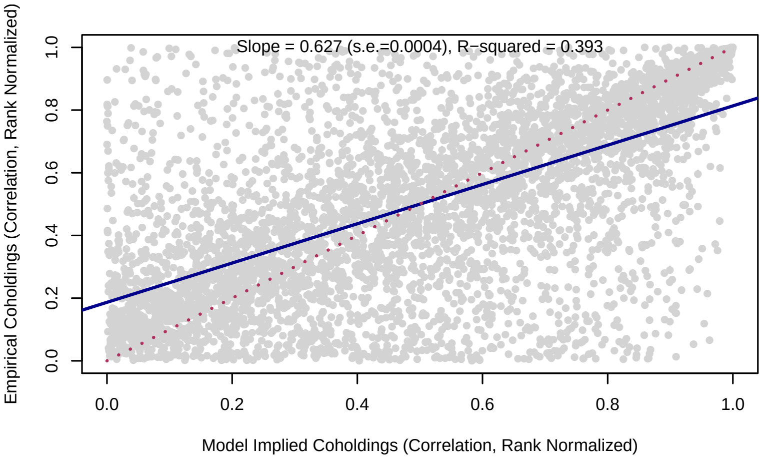
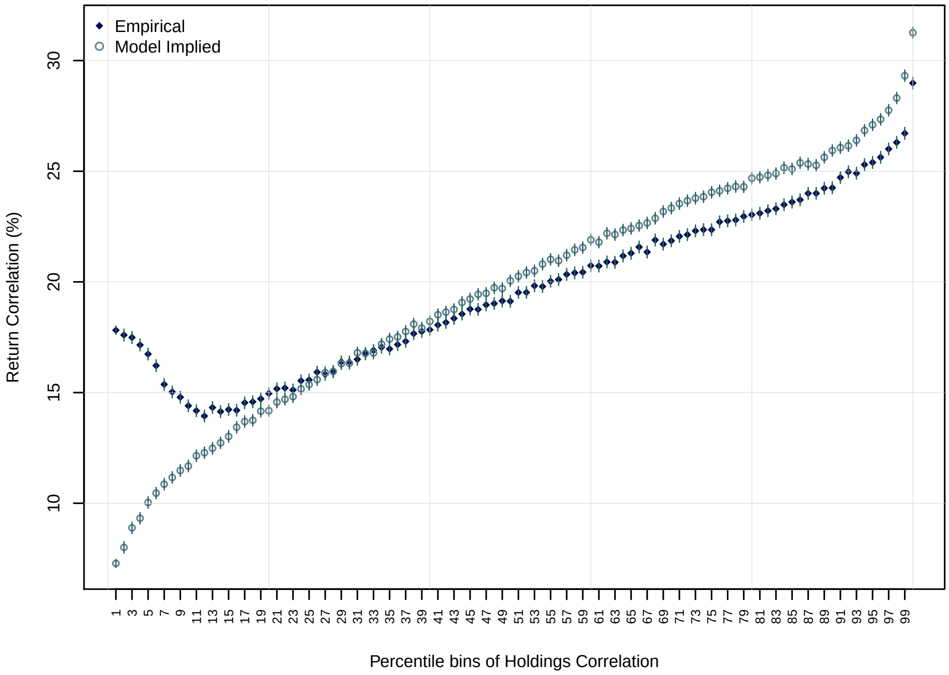


Figure 8
Return Correlation and Coholdings

This figure plots the relationship between cross-stock correlation ($y-$axis) and the coholding (correlation) measure ($x-$axis). Each point is the coefficient estimate from a regression with the cross-stock correlation as the dependent variable and percentile bins of the coholding (correlation) measure. Diamond points presents the relationship to empirical coholding estimates on the $x-$axis, while the hollow circles presents the observed factor model-implied coholding (correlation). The return correlation estimates are based on weekly returns data for a year leading up to August 2011. A simple linear relationship with empirical coholdings measure yields a slope of 0.202 (s.e = 0.0006) and $R^2 = 0.041$. With the model-implied coholdings measure, the slope is 0.309 (s.e=0.0005), and $R^2 = 0.096$.



Online Appendix

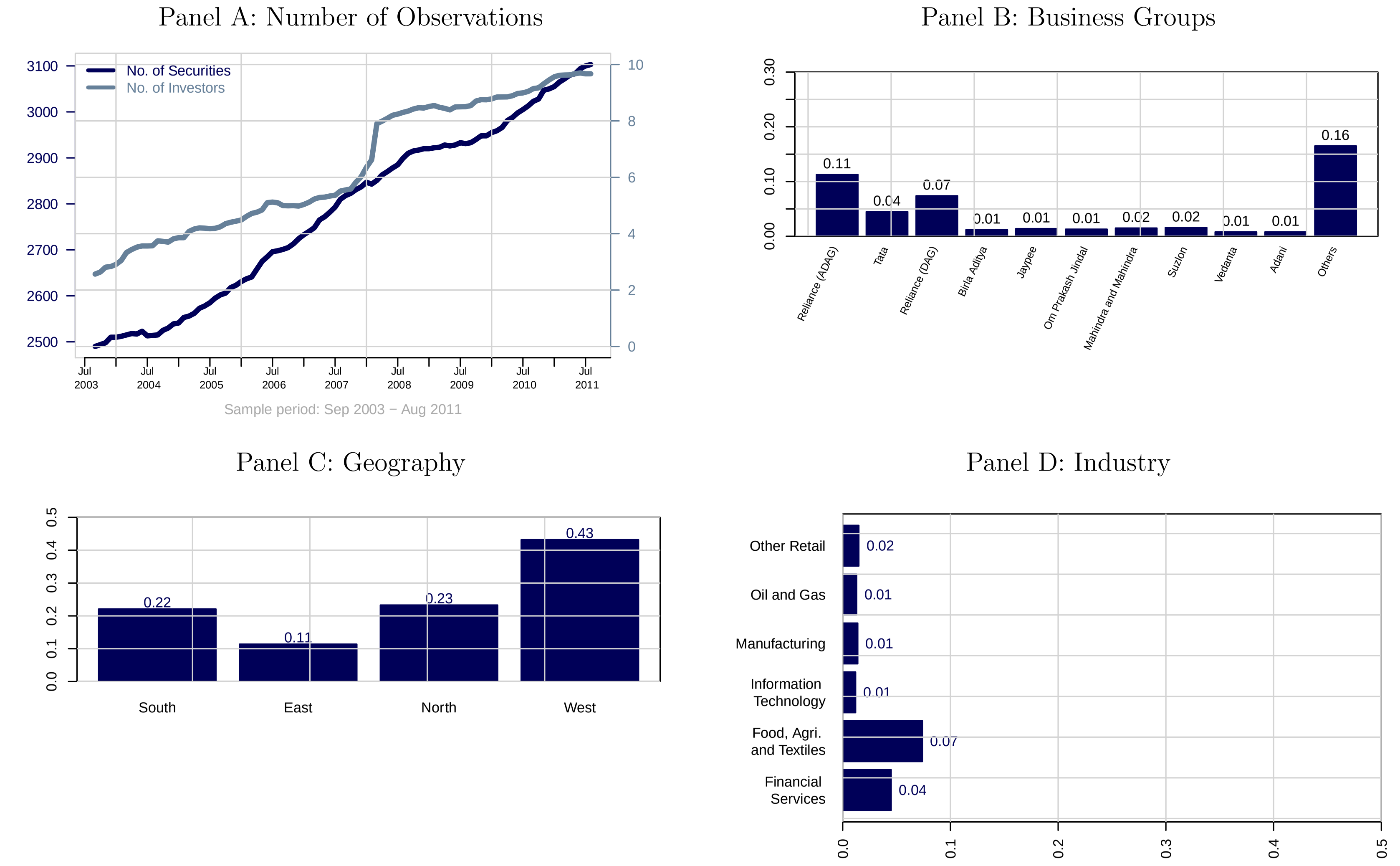
Who Holds What?
A Factor Model for Direct Stockholding

Vimal Balasubramaniam John Y. Campbell
Tarun Ramadorai Benjamin Ranish

Figure A.1

Summary Statistics

Panel A plots the number of investors in our data (right axis) in millions, and the number of stocks in our data (left axis) over time. Panel B plots the share of each business group (x-axis) in the average investor's stockholdings. Panel C plots the geographic region of the investor; Panel D summarizes the presence of each industry (y-axis) in the average investors' stockholdings.



Panel A: Number of Observations

Sample period: Sep 2003 − Aug 2011

Panel B: Business Groups

Panel C: Geography

Panel D: Industry

Figure A.2

Comparison of U.S. and Indian Household Stock Wealth

This figure presents the empirical kernel density plot of d the distribution of the logarithmic value of all equity investments in US dollars in the United States (black dashed line) from the Survey of Consumer Finances (SCF), 2013 and in Indian depository accounts in August 2011. The Indian portfolio value distribution is scaled by the ratio of per capita GDP in India to the United States.

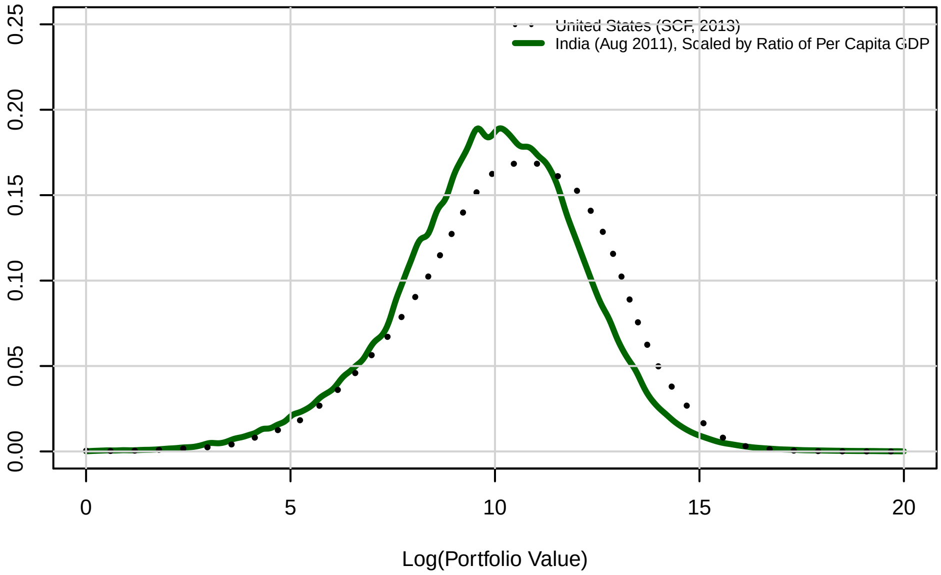
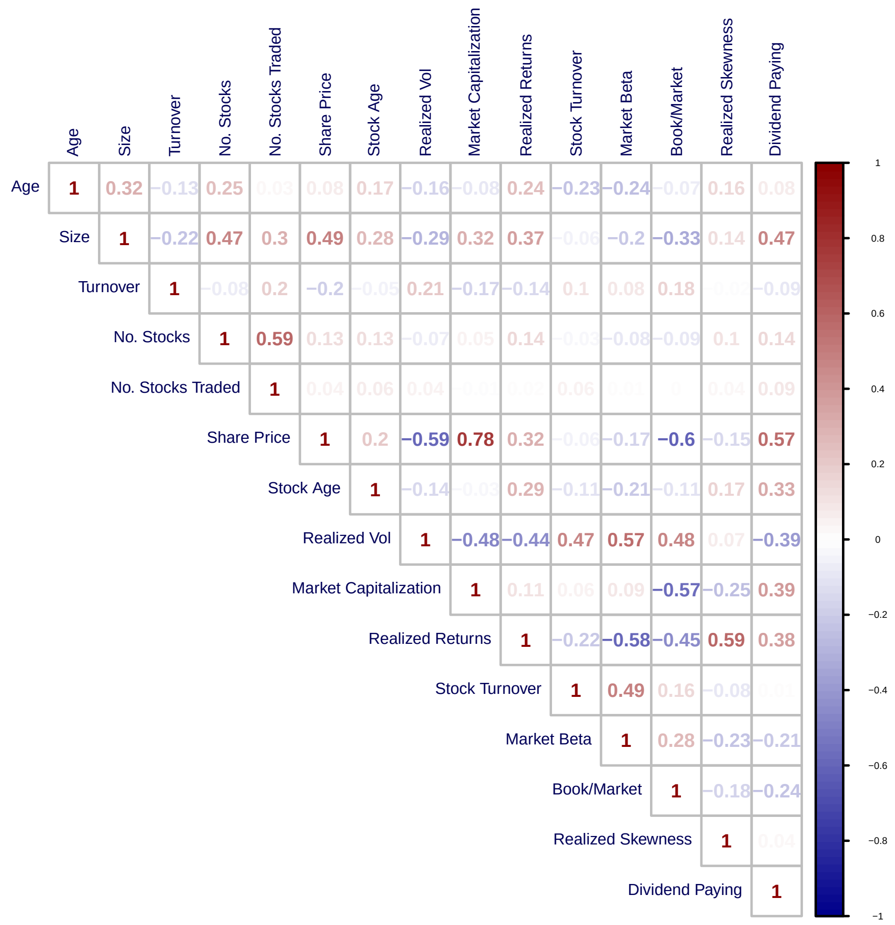


Figure A.3
Correlation Matrix

This figure plots the correlation between the main observed factor variables of interest, constructed in the same way as documented in Table 1.



	Age	Size	Turnover	No. Stocks	No. Stocks Traded	Share Price	Stock Age	Realized Vol	Market Capitalization	Realized Returns	Stock Turnover	Market Beta	Book/Market	Realized Skewness	Dividend Paying
Age	1	0.32	−0.13	0.25	0.03	0.08	0.17	−0.16	−0.08	0.24	−0.23	−0.24	−0.07	0.16	0.08
Size		1	−0.22	0.47	0.3	0.49	0.28	−0.29	0.32	0.37	−0.06	−0.2	−0.33	0.14	0.47
Turnover			1	−0.08	0.2	−0.2	−0.05	0.21	−0.17	−0.14	0.1	0.08	0.18	−0.02	−0.09
No. Stocks				1	0.59	0.13	0.13	−0.07	0.05	0.14	−0.03	−0.08	−0.09	0.1	0.14
No. Stocks Traded					1	0.04	0.06	0.04	−0.01	0.02	0.06	0.01	0	0.04	0.09
Share Price						1	0.2	−0.59	0.78	0.32	−0.06	−0.17	−0.6	−0.15	0.57
Stock Age							1	−0.14	−0.03	0.29	−0.11	−0.21	−0.11	0.17	0.33
Realized Vol								1	−0.48	−0.44	0.47	0.57	0.48	0.07	−0.39
Market Capitalization									1	0.11	0.06	0.09	−0.57	−0.25	0.39
Realized Returns										1	−0.22	−0.58	−0.45	0.59	0.38
Stock Turnover											1	0.49	0.16	−0.08	0.01
Market Beta												1	0.28	−0.23	−0.21
Book/Market													1	−0.18	−0.24
Realized Skewness														1	0.04
Dividend Paying															1

Figure A.4

Unobserved Factor Model: Principal Components

This figure presents the proportion of variance explained by each of the principal components of an equally weighted portfolio of all stocks.

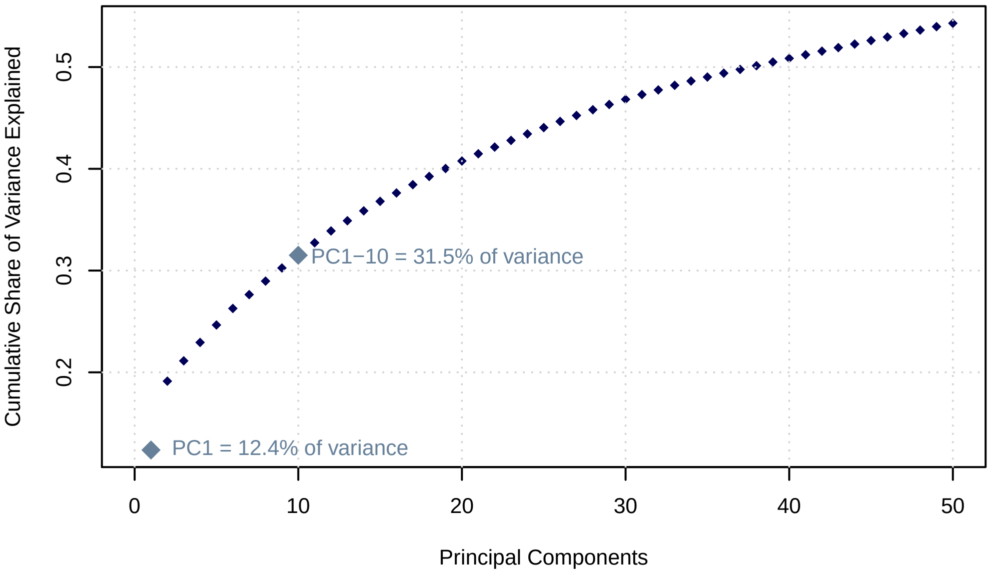


Figure A.5

Unobserved Factor Model: Ahn and Horenstein (2013) Eigenvalue Ratio Test

This figure presents the Eigenvalue Ratios of ordered, ratio of adjacent eigenvalues of the matrix, following Ahn and Horenstein (2013).

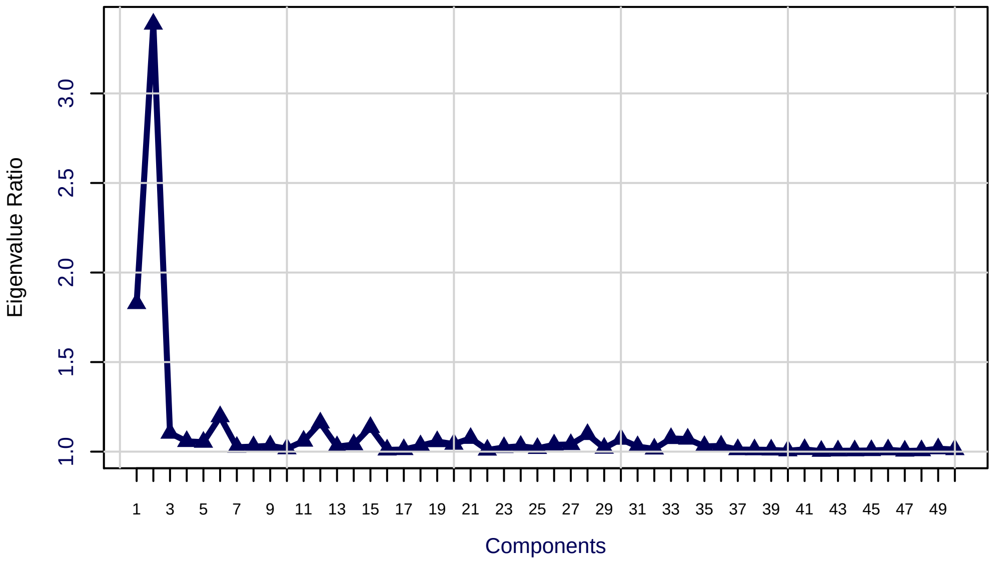


Figure A.6

Unobserved Factor Model: Bai and Ng (2002) PC_{p2} test

The loss function is defined as $PC_{p2}(k) = V(k, \hat{F^k}) + k\hat{\sigma}^2 \left(\frac{N+H}{NH}\right) lnC^2_{NH}$, where $V(k, \hat{F^k})$ is the mean sum of squared residuals from a k factor model. Bai and Ng (2002) propose that $\hat{\sigma}^2$ can be replaced by $V(kmax, F^{\hat{kmax}})$, and C^2_{NH} with $min(N, H)$, which is 3103 in our case. We assume a $kmax$ of 35 for the estimates.

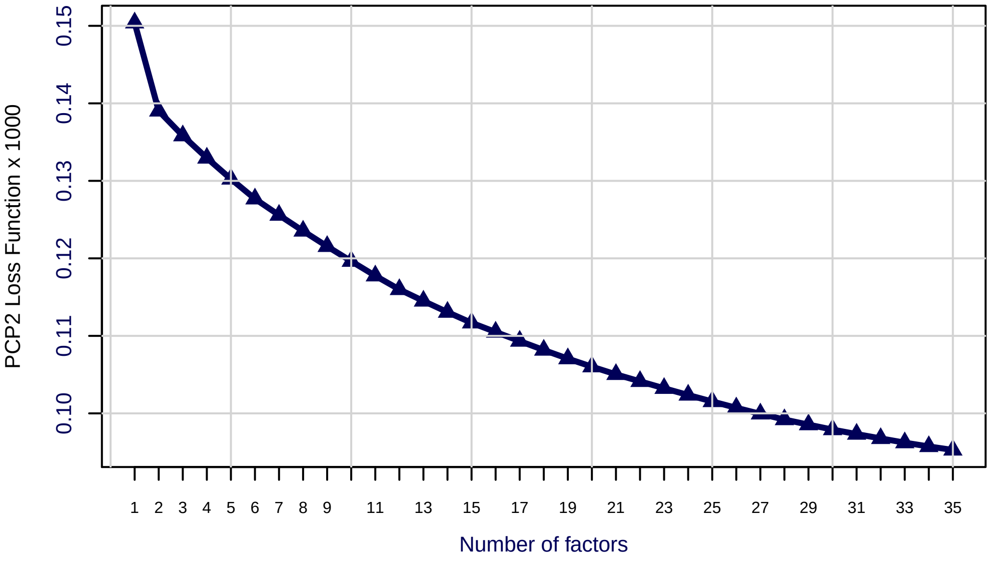
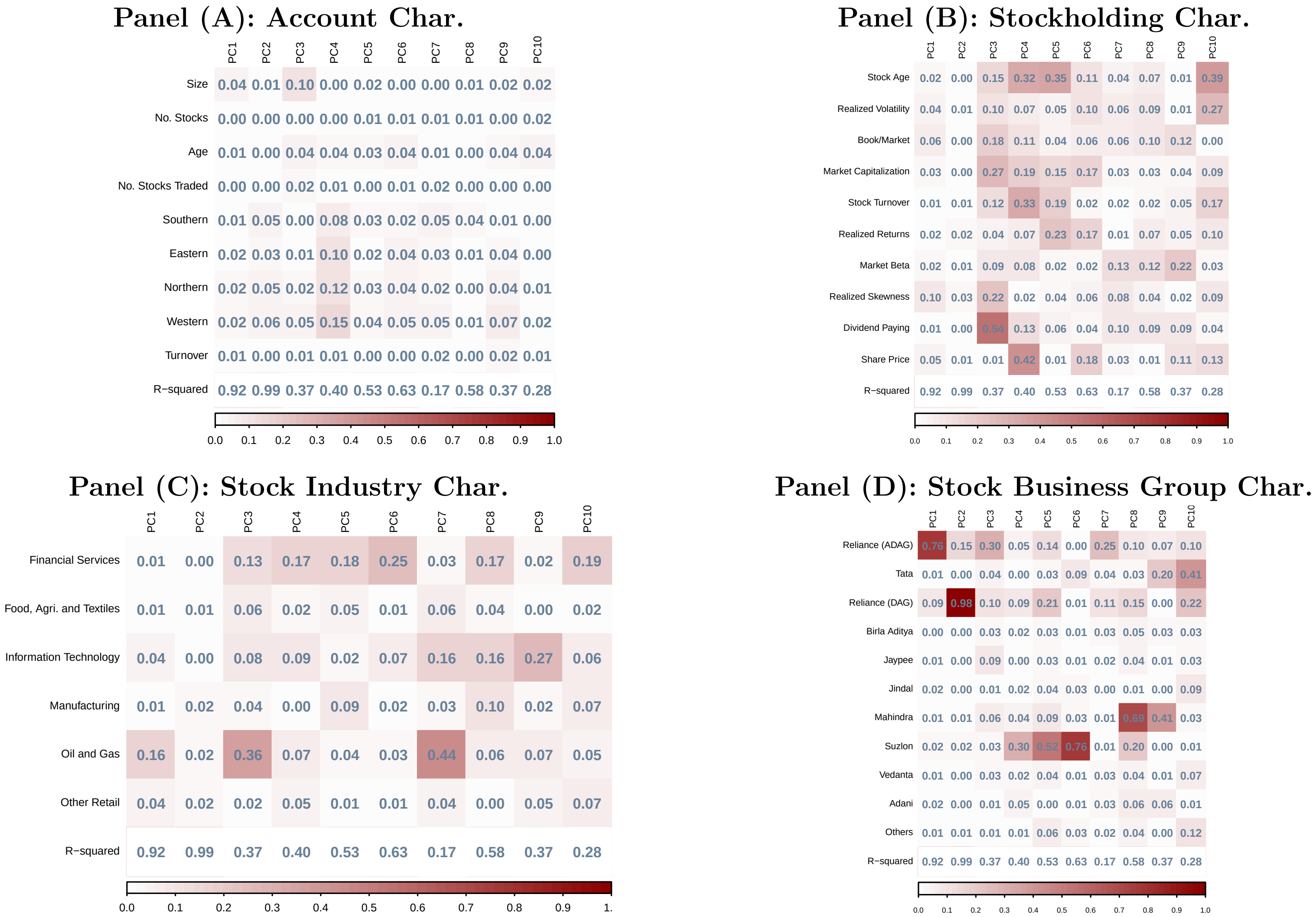


Figure A.7

Loadings of PCA 1-10 on Observed Factors (Unscaled Qv)

Panels A, B and C present a heatmap of the absolute value of loadings of PC 1-10 factors (in columns) on observed factors (in rows), all normalized by their standard deviations to allow for comparison. Darkest red shade indicates large loadings on the PC Factor.



Panel (A): Account Char.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Size	0.04	0.01	0.10	0.00	0.02	0.00	0.00	0.01	0.02	0.02
No. Stocks	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.00	0.02
Age	0.01	0.00	0.04	0.04	0.03	0.04	0.01	0.00	0.04	0.04
No. Stocks Traded	0.00	0.00	0.02	0.01	0.00	0.01	0.02	0.00	0.00	0.00
Southern	0.01	0.05	0.00	0.08	0.03	0.02	0.05	0.04	0.01	0.00
Eastern	0.02	0.03	0.01	0.10	0.02	0.04	0.03	0.01	0.04	0.00
Northern	0.02	0.05	0.02	0.12	0.03	0.04	0.02	0.00	0.04	0.01
Western	0.02	0.06	0.05	0.15	0.04	0.05	0.05	0.01	0.07	0.02
Turnover	0.01	0.00	0.01	0.01	0.00	0.00	0.02	0.00	0.02	0.01
R−squared	0.92	0.99	0.37	0.40	0.53	0.63	0.17	0.58	0.37	0.28

Panel (B): Stockholding Char.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Stock Age	0.02	0.00	0.15	0.32	0.35	0.11	0.04	0.07	0.01	0.39
Realized Volatility	0.04	0.01	0.10	0.07	0.05	0.10	0.06	0.09	0.01	0.27
Book/Market	0.06	0.00	0.18	0.11	0.04	0.06	0.06	0.10	0.12	0.00
Market Capitalization	0.03	0.00	0.27	0.19	0.15	0.17	0.03	0.03	0.04	0.09
Stock Turnover	0.01	0.01	0.12	0.33	0.19	0.02	0.02	0.02	0.05	0.17
Realized Returns	0.02	0.02	0.04	0.07	0.23	0.17	0.01	0.07	0.05	0.10
Market Beta	0.02	0.01	0.09	0.08	0.02	0.02	0.13	0.12	0.22	0.03
Realized Skewness	0.10	0.03	0.22	0.02	0.04	0.06	0.08	0.04	0.02	0.09
Dividend Paying	0.01	0.00	0.54	0.13	0.06	0.04	0.10	0.09	0.09	0.04
Share Price	0.05	0.01	0.01	0.42	0.01	0.18	0.03	0.01	0.11	0.13
R−squared	0.92	0.99	0.37	0.40	0.53	0.63	0.17	0.58	0.37	0.28

Panel (C): Stock Industry Char.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Financial Services	0.01	0.00	0.13	0.17	0.18	0.25	0.03	0.17	0.02	0.19
Food, Agri. and Textiles	0.01	0.01	0.06	0.02	0.05	0.01	0.06	0.04	0.00	0.02
Information Technology	0.04	0.00	0.08	0.09	0.02	0.07	0.16	0.16	0.27	0.06
Manufacturing	0.01	0.02	0.04	0.00	0.09	0.02	0.03	0.10	0.02	0.07
Oil and Gas	0.16	0.02	0.36	0.07	0.04	0.03	0.44	0.06	0.07	0.05
Other Retail	0.04	0.02	0.02	0.05	0.01	0.01	0.04	0.00	0.05	0.07
R−squared	0.92	0.99	0.37	0.40	0.53	0.63	0.17	0.58	0.37	0.28

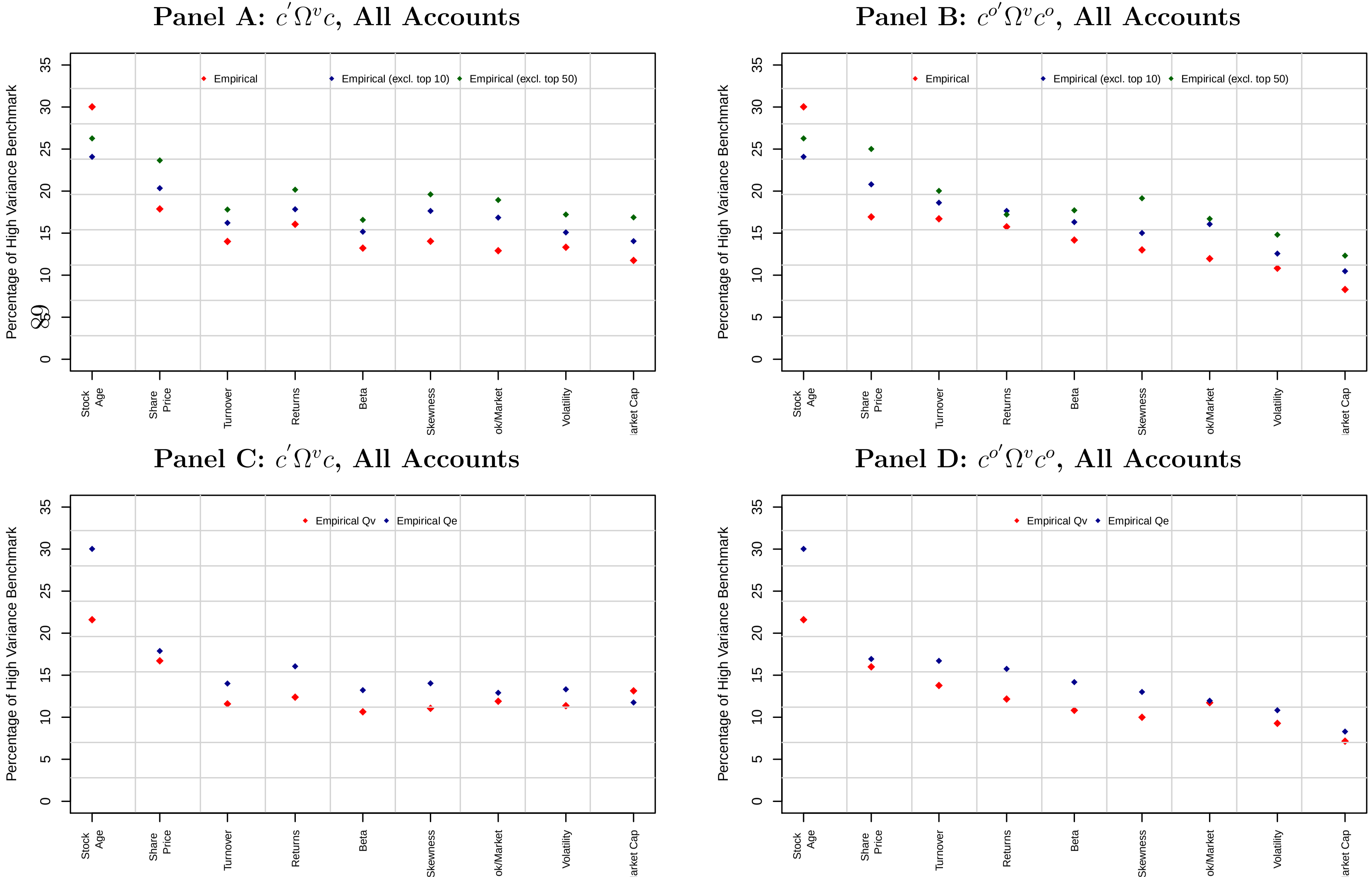
Panel (D): Stock Business Group Char.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Reliance (ADAG)	0.76	0.15	0.30	0.05	0.14	0.00	0.25	0.10	0.07	0.10
Tata	0.01	0.00	0.04	0.00	0.03	0.09	0.04	0.03	0.20	0.41
Reliance (DAG)	0.09	0.98	0.10	0.09	0.21	0.01	0.11	0.15	0.00	0.22
Birla Aditya	0.00	0.00	0.03	0.02	0.03	0.01	0.03	0.05	0.03	0.03
Jaypee	0.01	0.00	0.09	0.00	0.03	0.01	0.02	0.04	0.01	0.03
Jindal	0.02	0.00	0.01	0.02	0.04	0.03	0.00	0.01	0.00	0.09
Mahindra	0.01	0.01	0.06	0.04	0.09	0.03	0.01	0.69	0.41	0.03
Suzlon	0.02	0.02	0.03	0.30	0.52	0.76	0.01	0.20	0.00	0.01
Vedanta	0.01	0.00	0.03	0.02	0.04	0.01	0.03	0.04	0.01	0.07
Adani	0.02	0.00	0.01	0.05	0.00	0.01	0.03	0.06	0.06	0.01
Others	0.01	0.01	0.01	0.01	0.06	0.03	0.02	0.04	0.00	0.12
R−squared	0.92	0.99	0.37	0.40	0.53	0.63	0.17	0.58	0.37	0.28

Figure A.8

Stock Characteristic Clientele Strength: Robustness

This figure presents variations on the analysis of empirical variance of stockholdings characteristics in Figure 4 (Panels A and B). Panels A and B of this figure compare the empirical variance against the same measure constructed after excluding either the top 10 or 50 stocks (by mean Q_v). Panels C and D compare the empirical variance against a version constructed using equal (Q_e) rather than value (Q_v) weights for investors' stocks.



Panel A: $c^{'}\Omega^{v}c$, All Accounts

Panel B: $c^{o'}\Omega^{v}c^{o}$, All Accounts

Panel C: $c^{'}\Omega^{v}c$, All Accounts

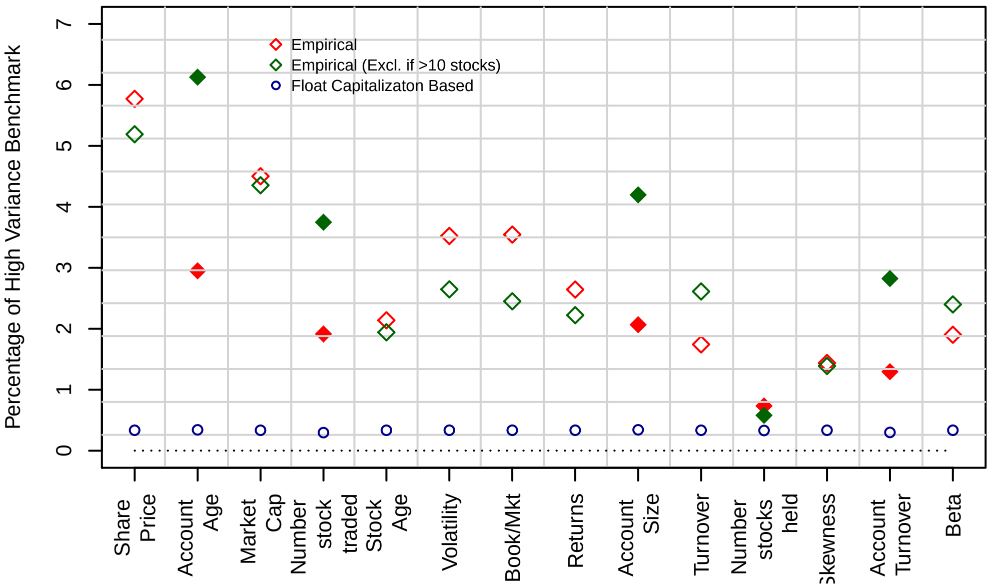
Panel D: $c^{o'}\Omega^{v}c^{o}$, All Accounts

Figure A.9

Investor Attribute Clientele Strength: Robustness

This figure compares the variance of investor bases' characteristics (from Figure 5) against a version computed after excluding investors holding more than 10 stocks (green diamonds).

Panel A: $a^{'}\Omega^{s}a$



Panel B: $a^{o'}\Omega^{s}a^{o}$

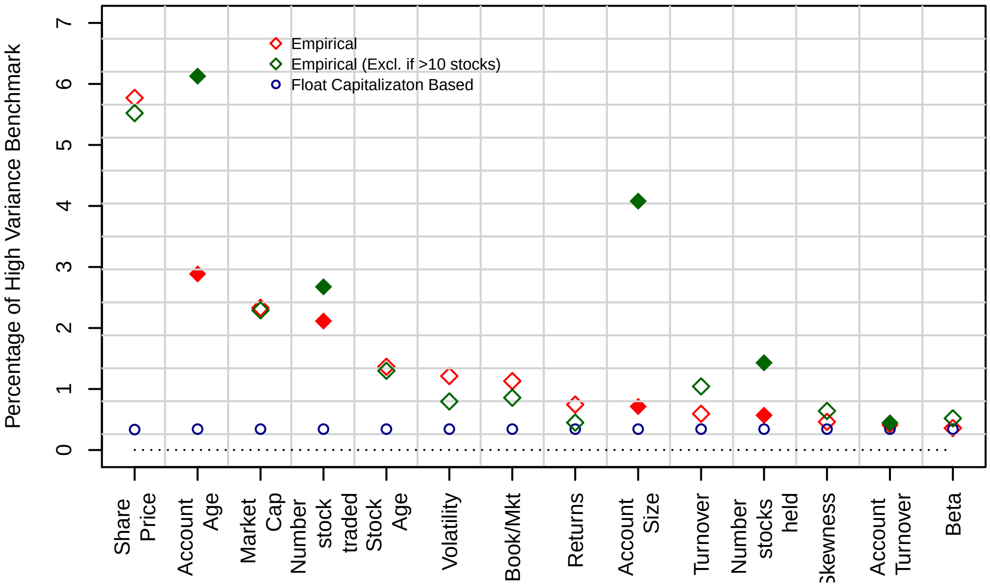
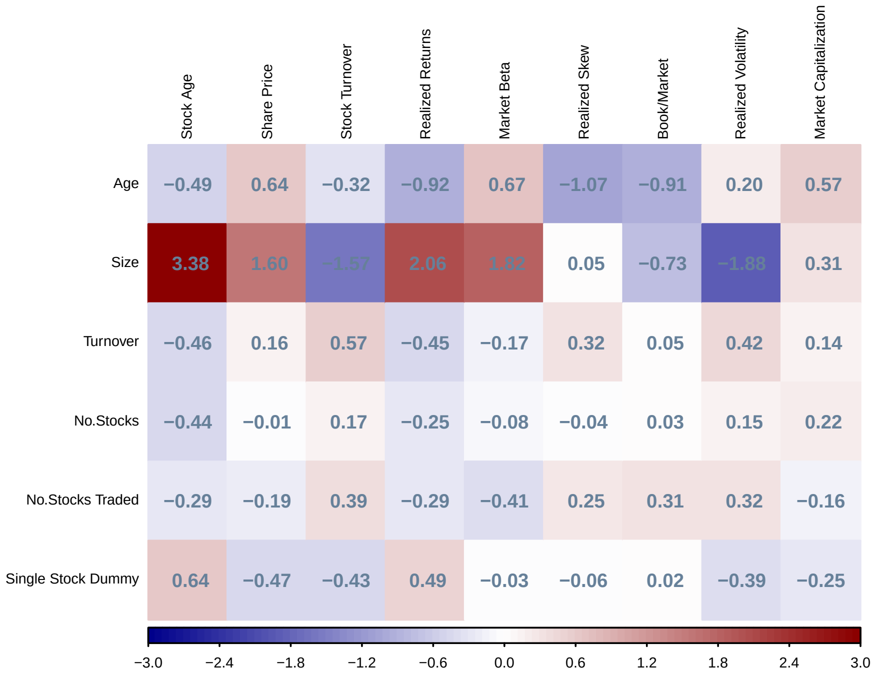


Table A.1
Factor Loadings and Stock Characteristic Clienteles
Robustness Check for Account characteristics factors

This table presents a variation on the analysis in Table 5 using weights based on non-orthogonalized stock characteristic ranks.

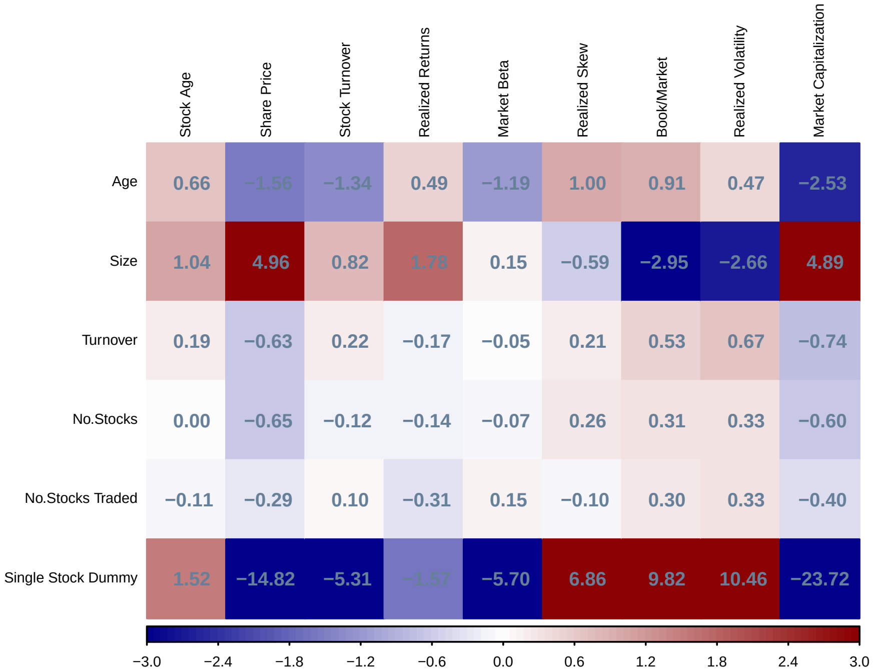


	Stock Age	Share Price	Stock Turnover	Realized Returns	Market Beta	Realized Skew	Book/Market	Realized Volatility	Market Capitalization
Age	−0.49	0.64	−0.32	−0.92	0.67	−1.07	−0.91	0.20	0.57
Size	3.38	1.60	−1.57	2.06	1.82	0.05	−0.73	−1.88	0.31
Turnover	−0.46	0.16	0.57	−0.45	−0.17	0.32	0.05	0.42	0.14
No.Stocks	−0.44	−0.01	0.17	−0.25	−0.08	−0.04	0.03	0.15	0.22
No.Stocks Traded	−0.29	−0.19	0.39	−0.29	−0.41	0.25	0.31	0.32	−0.16
Single Stock Dummy	0.64	−0.47	−0.43	0.49	−0.03	−0.06	0.02	−0.39	−0.25

Table A.2
Factor Loadings and Stock Clienteles
Robustness Check for All Factors

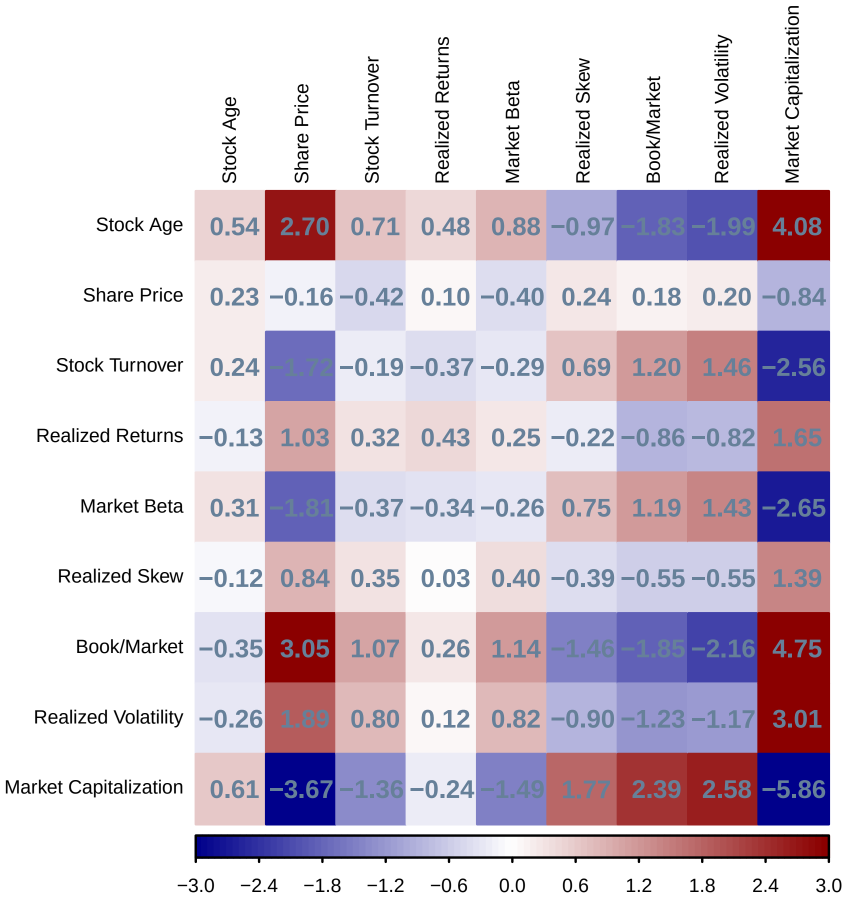
This table presents a variation on the analysis in Table 6 using weights based on non-orthogonalized stock characteristic ranks.

Panel A: Account Characteristics



	Stock Age	Share Price	Stock Turnover	Realized Returns	Market Beta	Realized Skew	Book/Market	Realized Volatility	Market Capitalization
Age	0.66	−1.56	−1.34	0.49	−1.19	1.00	0.91	0.47	−2.53
Size	1.04	4.96	0.82	1.78	0.15	−0.59	−2.95	−2.66	4.89
Turnover	0.19	−0.63	0.22	−0.17	−0.05	0.21	0.53	0.67	−0.74
No.Stocks	0.00	−0.65	−0.12	−0.14	−0.07	0.26	0.31	0.33	−0.60
No.Stocks Traded	−0.11	−0.29	0.10	−0.31	0.15	−0.10	0.30	0.33	−0.40
Single Stock Dummy	1.52	−14.82	−5.31	−1.57	−5.70	6.86	9.82	10.46	−23.72

Panel B: Stockholding Characteristics

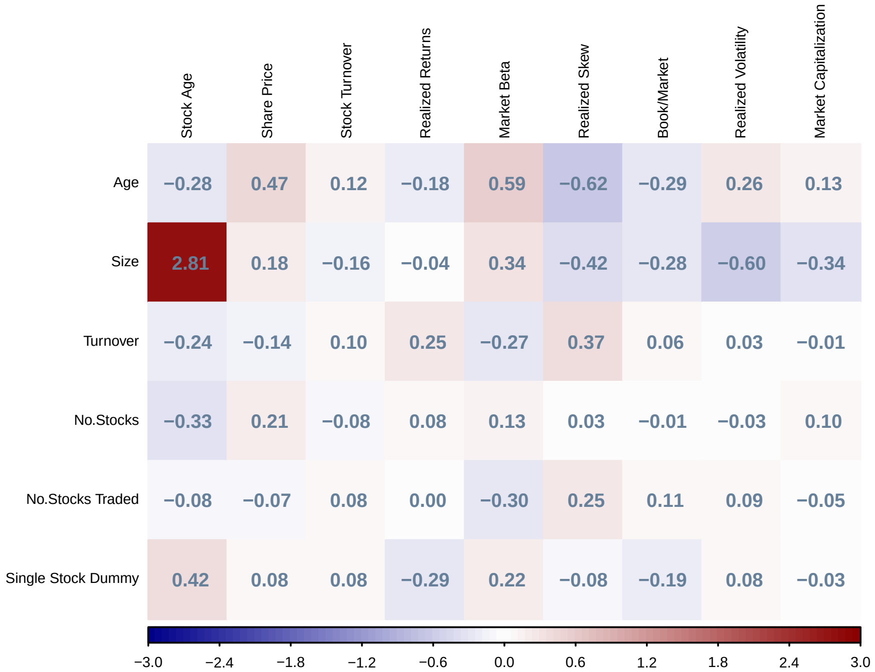


	Stock Age	Share Price	Stock Turnover	Realized Returns	Market Beta	Realized Skew	Book/Market	Realized Volatility	Market Capitalization
Stock Age	0.54	2.70	0.71	0.48	0.88	−0.97	−1.83	−1.99	4.08
Share Price	0.23	−0.16	−0.42	0.10	−0.40	0.24	0.18	0.20	−0.84
Stock Turnover	0.24	−1.72	−0.19	−0.37	−0.29	0.69	1.20	1.46	−2.56
Realized Returns	−0.13	1.03	0.32	0.43	0.25	−0.22	−0.86	−0.82	1.65
Market Beta	0.31	−1.81	−0.37	−0.34	−0.26	0.75	1.19	1.43	−2.65
Realized Skew	−0.12	0.84	0.35	0.03	0.40	−0.39	−0.55	−0.55	1.39
Book/Market	−0.35	3.05	1.07	0.26	1.14	−1.46	−1.85	−2.16	4.75
Realized Volatility	−0.26	1.89	0.80	0.12	0.82	−0.90	−1.23	−1.17	3.01
Market Capitalization	0.61	−3.67	−1.36	−0.24	−1.49	1.77	2.39	2.58	−5.86

Table A.3

Factor Loadings and Stock Characteristic Clienteles
W/o Top 10 stocks, Account characteristics factors

This table presents a variation on the analysis in Table 5 in which the largest 10 stocks (in terms of mean Q_v) are excluded.



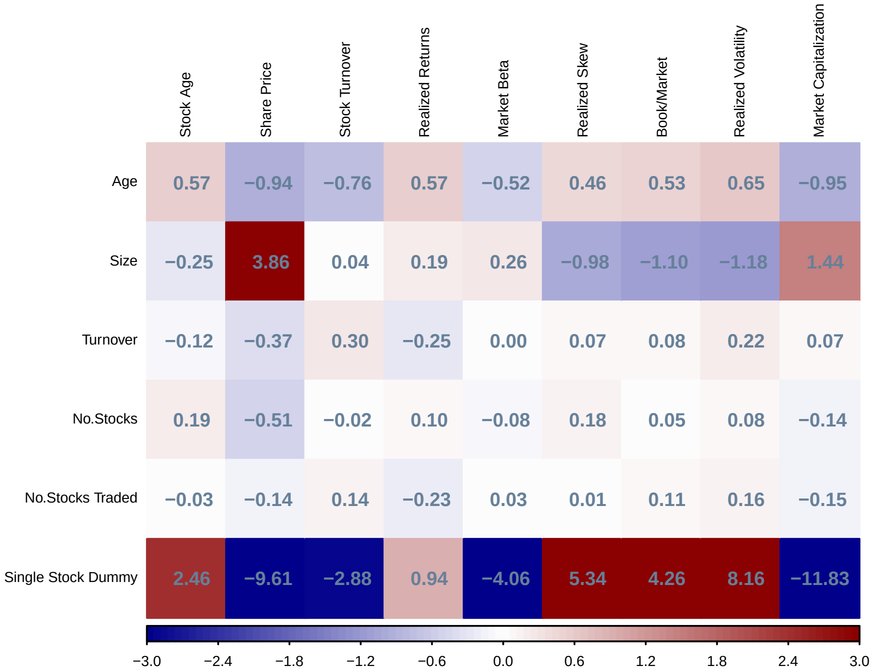
	Stock Age	Share Price	Stock Turnover	Realized Returns	Market Beta	Realized Skew	Book/Market	Realized Volatility	Market Capitalization
Age	−0.28	0.47	0.12	−0.18	0.59	−0.62	−0.29	0.26	0.13
Size	2.81	0.18	−0.16	−0.04	0.34	−0.42	−0.28	−0.60	−0.34
Turnover	−0.24	−0.14	0.10	0.25	−0.27	0.37	0.06	0.03	−0.01
No.Stocks	−0.33	0.21	−0.08	0.08	0.13	0.03	−0.01	−0.03	0.10
No.Stocks Traded	−0.08	−0.07	0.08	0.00	−0.30	0.25	0.11	0.09	−0.05
Single Stock Dummy	0.42	0.08	0.08	−0.29	0.22	−0.08	−0.19	0.08	−0.03

Table A.4

Factor Loadings and Stock Characteristic Clienteles
W/o Top 10 stocks, All Factors

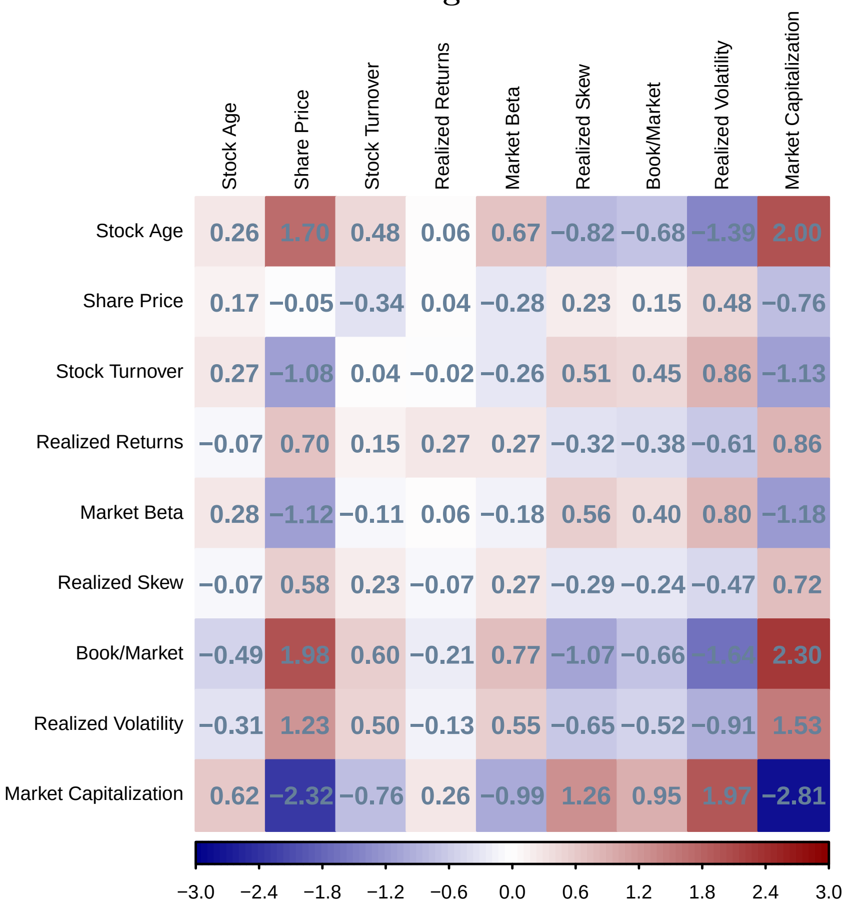
This table presents a variation on the analysis in Table 6 in which the largest 10 stocks (in terms of mean Q_v) are excluded.

Panel A: Account Characteristics



	Stock Age	Share Price	Stock Turnover	Realized Returns	Market Beta	Realized Skew	Book/Market	Realized Volatility	Market Capitalization
Age	0.57	−0.94	−0.76	0.57	−0.52	0.46	0.53	0.65	−0.95
Size	−0.25	3.86	0.04	0.19	0.26	−0.98	−1.10	−1.18	1.44
Turnover	−0.12	−0.37	0.30	−0.25	0.00	0.07	0.08	0.22	0.07
No.Stocks	0.19	−0.51	−0.02	0.10	−0.08	0.18	0.05	0.08	−0.14
No.Stocks Traded	−0.03	−0.14	0.14	−0.23	0.03	0.01	0.11	0.16	−0.15
Single Stock Dummy	2.46	−9.61	−2.88	0.94	−4.06	5.34	4.26	8.16	−11.83

Panel B: Stockholding Characteristics

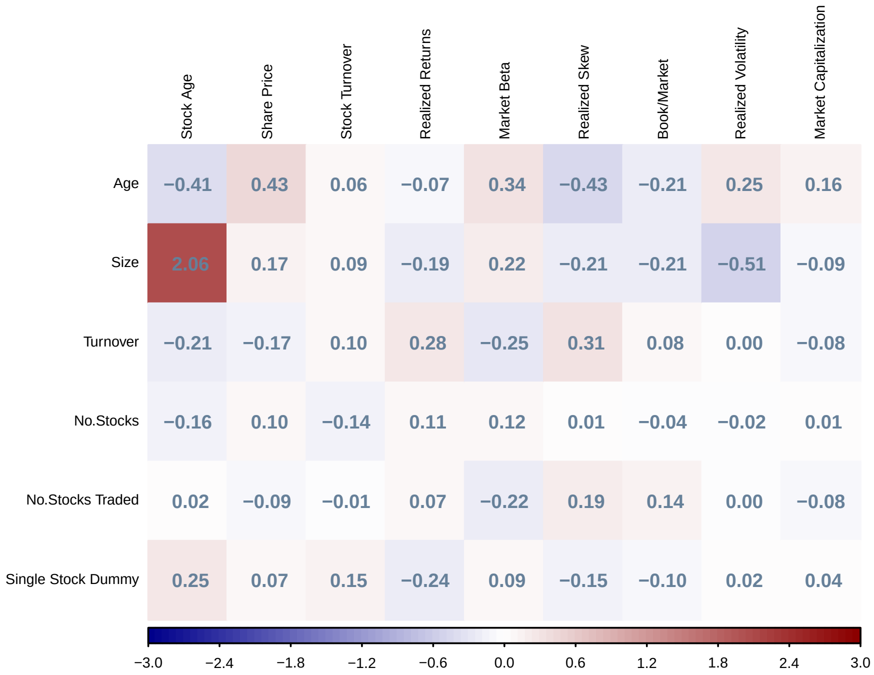


	Stock Age	Share Price	Stock Turnover	Realized Returns	Market Beta	Realized Skew	Book/Market	Realized Volatility	Market Capitalization
Stock Age	0.26	1.70	0.48	0.06	0.67	−0.82	−0.68	−1.39	2.00
Share Price	0.17	−0.05	−0.34	0.04	−0.28	0.23	0.15	0.48	−0.76
Stock Turnover	0.27	−1.08	0.04	−0.02	−0.26	0.51	0.45	0.86	−1.13
Realized Returns	−0.07	0.70	0.15	0.27	0.27	−0.32	−0.38	−0.61	0.86
Market Beta	0.28	−1.12	−0.11	0.06	−0.18	0.56	0.40	0.80	−1.18
Realized Skew	−0.07	0.58	0.23	−0.07	0.27	−0.29	−0.24	−0.47	0.72
Book/Market	−0.49	1.98	0.60	−0.21	0.77	−1.07	−0.66	−1.64	2.30
Realized Volatility	−0.31	1.23	0.50	−0.13	0.55	−0.65	−0.52	−0.91	1.53
Market Capitalization	0.62	−2.32	−0.76	0.26	−0.99	1.26	0.95	1.97	−2.81

Table A.5

Factor Loadings and Stock Characteristic Clienteles
W/o Top 50 stocks, Account characteristics factors

This table presents the orthogonalized stock characteristic rank weighted loadings from a regression of stock holdings Q_{ih} only with account characteristic factors. Each row represents the account characteristic factor, and each column for the various stock clienteles of interest. The sample of stocks considered leaves out the top 50 stocks.



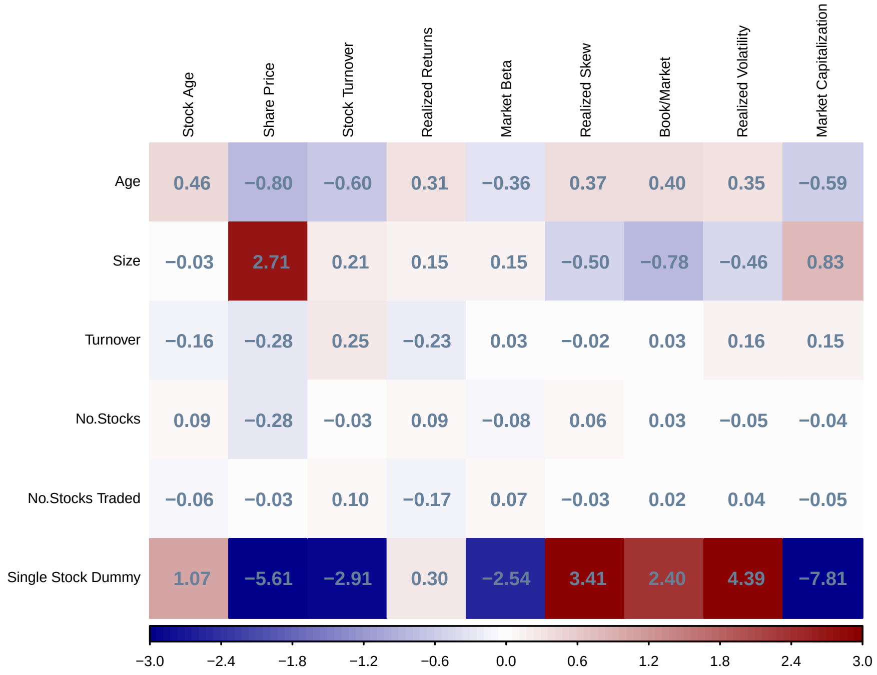
	Stock Age	Share Price	Stock Turnover	Realized Returns	Market Beta	Realized Skew	Book/Market	Realized Volatility	Market Capitalization
Age	−0.41	0.43	0.06	−0.07	0.34	−0.43	−0.21	0.25	0.16
Size	2.06	0.17	0.09	−0.19	0.22	−0.21	−0.21	−0.51	−0.09
Turnover	−0.21	−0.17	0.10	0.28	−0.25	0.31	0.08	0.00	−0.08
No.Stocks	−0.16	0.10	−0.14	0.11	0.12	0.01	−0.04	−0.02	0.01
No.Stocks Traded	0.02	−0.09	−0.01	0.07	−0.22	0.19	0.14	0.00	−0.08
Single Stock Dummy	0.25	0.07	0.15	−0.24	0.09	−0.15	−0.10	0.02	0.04

Table A.6

Factor Loadings and Stock Characteristic Clienteles
W/o Top 50 stocks, All Factors

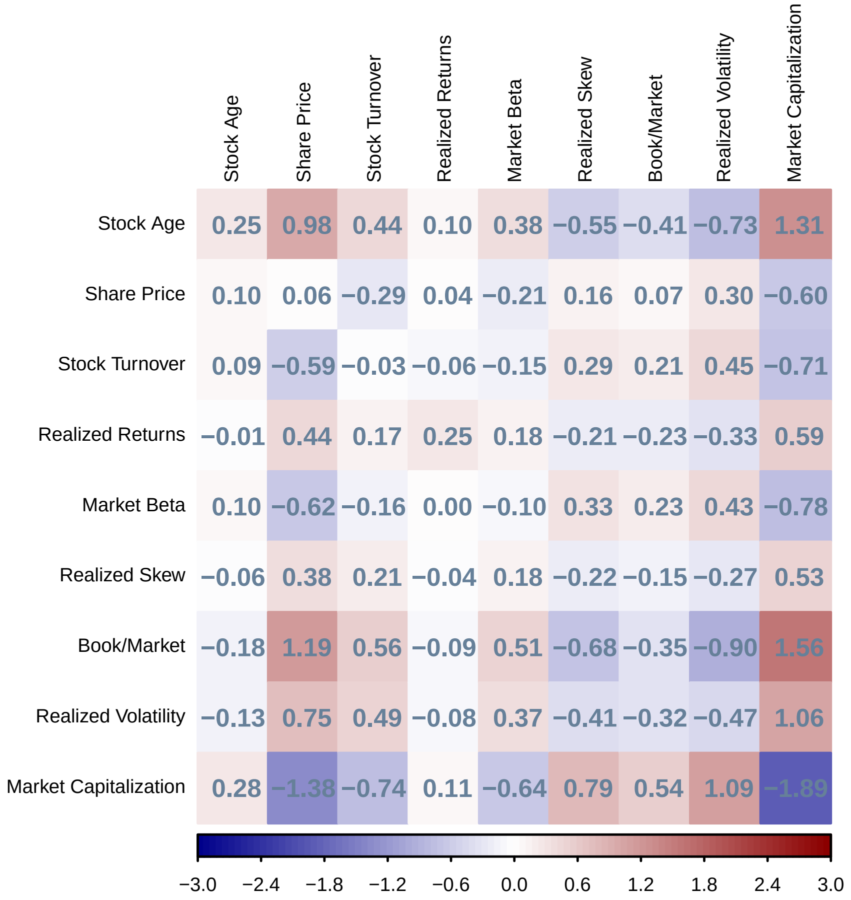
This table presents the orthogonalized stock characteristic rank weighted loadings from a regression of Q_{ih} with both account characteristic factors and orthogonalized stockholding characteristic factors. Each row represents the account characteristic factor, and each column for the various stock clienteles of interest. The sample of stocks considered leaves out the top 50 stocks.

Panel A: Account Characteristics



	Stock Age	Share Price	Stock Turnover	Realized Returns	Market Beta	Realized Skew	Book/Market	Realized Volatility	Market Capitalization
Age	0.46	−0.80	−0.60	0.31	−0.36	0.37	0.40	0.35	−0.59
Size	−0.03	2.71	0.21	0.15	0.15	−0.50	−0.78	−0.46	0.83
Turnover	−0.16	−0.28	0.25	−0.23	0.03	−0.02	0.03	0.16	0.15
No.Stocks	0.09	−0.28	−0.03	0.09	−0.08	0.06	0.03	−0.05	−0.04
No.Stocks Traded	−0.06	−0.03	0.10	−0.17	0.07	−0.03	0.02	0.04	−0.05
Single Stock Dummy	1.07	−5.61	−2.91	0.30	−2.54	3.41	2.40	4.39	−7.81

Panel B: Stockholding Characteristics



	Stock Age	Share Price	Stock Turnover	Realized Returns	Market Beta	Realized Skew	Book/Market	Realized Volatility	Market Capitalization
Stock Age	0.25	0.98	0.44	0.10	0.38	−0.55	−0.41	−0.73	1.31
Share Price	0.10	0.06	−0.29	0.04	−0.21	0.16	0.07	0.30	−0.60
Stock Turnover	0.09	−0.59	−0.03	−0.06	−0.15	0.29	0.21	0.45	−0.71
Realized Returns	−0.01	0.44	0.17	0.25	0.18	−0.21	−0.23	−0.33	0.59
Market Beta	0.10	−0.62	−0.16	0.00	−0.10	0.33	0.23	0.43	−0.78
Realized Skew	−0.06	0.38	0.21	−0.04	0.18	−0.22	−0.15	−0.27	0.53
Book/Market	−0.18	1.19	0.56	−0.09	0.51	−0.68	−0.35	−0.90	1.56
Realized Volatility	−0.13	0.75	0.49	−0.08	0.37	−0.41	−0.32	−0.47	1.06
Market Capitalization	0.28	−1.38	−0.74	0.11	−0.64	0.79	0.54	1.09	−1.89