Who Owns What?
A Factor Model for Direct Stock Holding*†

Vimal Balasubramaniam John Y. Campbell

Tarun Ramadorai Benjamin Ranish

July 15, 2021

Abstract

We build a cross-sectional factor model for investors' direct stock holdings, 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 covariance, 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, Dartmouth, Case Western Reserve, Stanford, CU Hong Kong, London School of Economics, Rutgers, Goethe University, the Istanbul Finance Seminar, Virtual Finance Seminar, 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. The views in this paper are solely the responsibility of the authors and should not be interpreted as those of the Board of Governors of the Federal Reserve System, or of other members of their staff.

1 Introduction

Finance theorists have developed a rich variety of models that describe household portfolio choice. Most famously, the venerable CAPM says that all investors should hold the same risky portfolio, with different degrees of cash dilution to accommodate differences in risk aversion (Markowitz 1952, Tobin 1958, Sharpe 1964, Lintner 1965).

What evidence we have on household portfolios differs substantially from the CAPM prediction, documenting considerable heterogeneity in portfolio composition (for a survey, see Curcuru et al. (2010)). More complex theories to explain this heterogeneity can be broadly classified into two groups. One group focuses on heterogeneous financial circumstances such as non-traded income risks (Mayers 1972, Vissing-Jørgensen 2002, Fagereng et al. 2018), investment horizons (Merton 1973), or liquidity needs (Amihud and Mendelson 1986). Another group emphasizes differences in investor familiarity with firms and beliefs about their returns (Merton 1987, Harris and Raviv 1993, Meeuwis et al. 2018), or heterogeneous preferences for firm characteristics such as ethical and environmental quality (Hong and Kacperczyk 2009, Pástor et al. 2020).

Progress on assessing these explanations for household portfolio heterogeneity requires a more general characterization of the structure of heterogeneity: a parsimonious summary of "who owns what." Our goal in this paper is to provide such a general characterization. We introduce a new framework to organize the empirical evidence in a way that can guide the refinement of theoretical explanations. We apply this framework on a large, detailed, and comprehensive administrative dataset of household stockholdings in India, where we unearth robust evidence of clientele effects which can help to inform theories of household portfolio construction.

This task is challenging for a number of reasons. The conceptual 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 stock holdings across households that is analogous to the classic time-series factor model for stock returns over time (Fama and French 1992). This allows us to exploit numerous insights and methods from the time-series factor literature and, as in that literature, permits "dimension reduction" using a small number of factor portfolios (Fama and French 1993).

Another challenge, evident in the relatively small empirical literature on these important questions, arises from the difficulty of measuring the complete portfolios of households. 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) and Betermeier et al. (2017), but the important role played by mutual funds in these countries makes it hard to interpret individual stock holdings without also looking through mutual fund holdings to the underlying stocks held by funds.[1] We make progress by applying our methods to Indian administrative data on direct stock holdings, exploiting the very limited share of mutual funds in India emphasized by Campbell et al. (2014) and Campbell et al. (2019).

As a first step towards a general description of household portfolio choice, we ask "what" characteristics some investors seek out and others avoid. We measure the clientele strength for each characteristic as the variance of the holdings-weighted characteristic (i.e., portfolio tilt) across households, using each household's portfolio shares as its holdings. 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.

In a refinement of this approach, we distinguish two components of characteristic clientele strength: one coming from the popularity of individual stocks with extreme characteristic values, and one coming from the tendency of investors to hold pairs of stocks with similar characteristics. These two components are analogous to the contributions of individual stock variances and of cross-stock covariances to the return variance of a factor portfolio, but in the holdings context, as a result of severe heteroskedasticity, the variance contributions tend to be more important than they are in the context of returns.

Among the characteristics we consider, firm age (the number of years since listing) has the strongest investor clientele but stock price, turnover, and recent past returns also have strong clienteles. The characteristics of stocks that are emphasized in time-series factor models such as the Fama-French (1993) model are relatively less important. While some of these characteristics are correlated with one another in the cross-section of stocks, we find very similar results when we orthogonalize the characteristics across stocks.

It is striking that the two strongest characteristic clientele effects we find are for firm age and stock price. Firm age is potentially correlated with more fundamental stock characteristics like volatility and market capitalization, but it is more important in our analysis than these other characteristics; and stock price is an arbitrary characteristic,

[1]These problems are less severe for an extensive related literature that studies households' trading behavior and performance, 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).

since firms can reset it as they wish using stock splits. This suggests the potential relevance of a behavioral theory in which investor attention is drawn to certain characteristics whether or not they are fundamentally important. These findings are also of interest to corporate finance, given the extensive literature showing how managers can "cater" to investor clienteles using policies such as stock splits (Baker et al. 2009, Baker and Wurgler 2013).

We also show that stock characteristics form clusters that tend to be coheld by Indian investors. The most important cluster, as identified by principal components analysis of characteristic tilts, consists of old (established) stocks that pay dividends and have high turnover and volatility. Some investors strongly prefer these stocks, while others strongly prefer to hold stocks with the opposite characteristics: young firms (recent IPOs) that do not pay dividends and have low turnover and volatility. A second cluster contains easily traded "lottery-like" stocks: young firms with low share prices, high turnover, high returns, and positive skewness. A third cluster comprises dividend-paying stocks with low beta, low turnover, and high past returns.

To develop our understanding of clienteles further, we turn our attention to the attributes of investors and ask "who" makes up clienteles that favor the same stocks. We find intuitive patterns in which larger, older, and better diversified accounts tend to tilt their portfolios towards the first and third clusters described above—established stocks and dividend-paying low beta stocks—but disfavor young, lottery-like stocks. In contrast, accounts with high turnover and extremely underdiversified accounts have the opposite preferences and favor young, lottery-like stocks. Accounts holding a single stock strongly prefer young, large stocks or "mega-IPOs."

In a rational model of income risk hedging, investor attributes that drive portfolio tilts should line up with income risks. This would be plausible for geographical location attributes, which do have a modest effect in our multifactor analysis, but is less likely for an attribute such as account age. With the caveat that our data only permits measurement of account attributes rather than the attributes or demographic characteristics of the underlying investors (such as their age, income, or sector of employment), we note that behavioral models stress learning from experience (Malmendier and Nagel 2011, Anagol et al. 2021), making account age a very natural attribute under such an interpretation of the evidence.

We develop our analysis further by estimating a factor model at the stock level to predict household stock holdings, telling us in more granular detail "who owns what." We use this model to parsimoniously summarize the coholdings matrix, which captures the average intensity with which households hold stocks and pairs of stocks together in

portfolios. 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. "Account-attribute factors" are attributes of stock holding 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. In contrast, "portfolio-attribute factors" are based on the characteristics of the other stocks held in each account: these factors are analogous to return-based factors in a time-series model such as the Fama-French SMB and HML factors. We estimate the loadings of stocks on these factors using unrestricted cross-sectional regressions, and show that account-attribute factors account for over half the explanatory power of the factor model, with account size playing a particularly important role. The most important portfolio-attribute factors capture the tendency for certain accounts to hold stocks that are members of business groups linked by common ownership.

We conclude our analysis by exploring the relation between measured coholdings and return comovement across stocks. We show that there is a significantly positive relation across stock pairs between coholdings and stock return covariances, a finding which suggests that investor clienteles may be important drivers of stock return comovements. The positive relation between coholdings and covariances is also visible at the level of characteristic portfolios.

These results appear contrary to naïve approaches to portfolio diversification, which would predict coholdings of stocks with low return correlations, and a negative relation between coholdings and covariances. Related to this, we show that Indian 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 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.

Our work is connected to the small but growing literature which uses detailed data to describe household portfolio construction. For example, Dorn and Huberman (2010) 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)

show that Scandinavian households hold stocks that have a more positive correlation with their labor income than average, indicating a tendency to "anti-hedge". Other authors investigate household mutual fund holdings (rather than direct stock holdings), with diverse results. For example, Betermeier et al. (2017) look at portfolio tilts in household mutual fund choices, showing evidence consistent with risk-based theories, while Grinblatt et al. (2016) show that differences in IQ help to predict mutual fund choices, more consistent with the importance of behavioral factors.

Some of our findings on household stock holding behavior have parallels in the literature on institutional stock holding. 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 also 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 measures the strength of investor clienteles over a range of stock characteristics and attributes of investors. Section 5 estimates multifactor models of stock holdings, not only our model with observable factors but also models with unobserved principal-components-based factors. Section 6 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 7 concludes. An internet appendix, Balasubramaniam et al. (2021), provides additional details on the empirical analysis.

2 Factor Structure in Stock Holdings

In this section, we introduce some concepts that we use to structure our empirical investigation of cross-sectional patterns in stock holdings. We first define the holdings matrix, which summarizes the holdings of N stocks by H households. From this, we derive a stock coholdings matrix and show how it can be used to measure the strengths of different types of investor clienteles. We describe how we collapse N stocks into a set of K characteristics to make the analysis lower-dimensional. Later in the paper, we return to the stock level, and describe a cross-sectional factor model for stock holdings, analogous to the familiar time-series factor models used to describe stock returns.

2.1 Stock Holdings and Coholdings

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' stock holding 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 dimensionality of the problem, we begin by collapsing the time dimension into a single period.[2] This eliminates the need for time subscripts in our notation.

The stock holdings matrix

We first define an N by H **stock 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 **stock holdings vector**.

Elements of the stock holdings matrix

There are several ways to define the elements of the stock holdings matrix Q, and we make a choice that has 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. We therefore define the elements of Q in a way that ensures each household has equal weight in our analysis, despite this variability across households. This ensures that our conclusions reflect the behavior of a representative individual stock investor.

Specifically, we set the stock holdings vector for household h, Q_h, to be the vector of portfolio shares for household h. Since portfolio shares add to one, this approach equalizes the sum of the elements of Q across households: $\iota' Q_h = 1$ for all h.

The stock coholdings matrix

Consider the demeaned stock holdings vector for household h:

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

[2] 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. The internet appendix shows the stability of our inferences when we re-estimate using data from other years.

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 stock coholdings matrix Ω_h 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 $T > N$.

The diagonal elements of Ω_h measure the variances of stock holdings across households. Because households have sparse stock holdings vectors Q_h, with many zero elements and a few positive ones, the variability of stock holdings is high for stocks that are more widely held.[3] Thus, we can also say that the diagonal elements of Ω_h capture the popularity or holdings intensity of each stock among individual investors.

Similarly, the off-diagonal elements of Ω_h measure the covariances of stock holdings across households. The same logic as above implies that these off-diagonal elements capture the popularity or coholdings intensity of pairs of stocks among individual investors.

2.2 Characteristics and Clientele Effects

We are interested not so much in specific stocks as in characteristics of stocks. For each continuous stock characteristic of interest, we define c as a zero-mean N-vector of the cross-stock rank of the characteristic on the interval $[-0.5, 0.5]$.

Although each stock characteristic has an equal-weighted average of zero across all stocks, the holdings-weighted average characteristic across stocks 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

[3]The intuition is easiest to see in a simplified example where all households own only a single stock. Then, the elements of Q_h are either zero or one, and the i'th diagonal element of Ω_h is the variance of a binomial random variable that equals one with probability equal to the holding probability p of the stock. The variance of a binomial is $p(1-p)$ which is increasing in p for all $p < 1/2$. Since the largest p in our dataset is about 0.4, the diagonal elements of Ω_h are greater for more widely held stocks.

demand. Accordingly, we focus not on the average characteristic but on the variance of the holdings-weighted characteristic across households. This captures the tendency for a characteristic to be strongly favored by some households and strongly disfavored by others: we say that such a characteristic has a strong clientele effect.

To study the clientele for a particular characteristic we calculate $c'Q_h$, the holdings-weighted characteristic, or equivalently the characteristic-weighted holding, for each household h. Since each characteristic has mean zero in the cross-section of stocks, we can also describe $c'Q_h$ as the **characteristic tilt** of the household's portfolio. 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 object to $c'Q_h$ is $c'R_t$, the characteristic-weighted return, in each period. The analogous approach to our measurement of characteristic clientele strength 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 high time-series variance of returns (Kozak et al. 2018).

The quadratic form in equation (3) can be decomposed into a contribution from the diagonal elements of Ω_h and a contribution from the off-diagonal elements. The diagonal component reflects the extent to which intensely held stocks have extreme characteristic values, while the off-diagonal component reflects the extent to which stocks with extreme characteristic values tend to be held together. Even though we have many stocks, so that one might expect the diagonal component to be modest as it generally is in a time-series context, extreme heteroskedasticity in holdings intensity across stocks and the large number of concentrated household portfolios imply an important role for the diagonal component in our context. In our empirical analysis we look at these two components separately and find that they play somewhat different roles.

Our choice of portfolio share as the elements of Q makes our characteristic clientele strength measure $c'\Omega_h c$ 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 $c'\Omega_h^v c$ using this reduced dataset.

2.3 From Stock Coholdings to Characteristic Coholdings

While we have defined clientele effects for particular stock characteristics, we are also interested in whether investors group stock characteristics into clusters when they construct portfolios.

To permit such analysis, we define Q^* to refer to the K by H **characteristic holdings matrix** which summarizes households' holdings of different stock characteristics. To go from the stock holdings matrix to the characteristic holdings matrix, we aggregate the elements of the stock holdings matrix along each of K characteristic dimensions.

Consider $k = 1, ..., K$ characteristics. For each stock characteristic k of interest, we define c_k as a zero-mean N-vector of the rank of each stock's characteristic on the interval $[-0.5, 0.5]$. For each household, $c_k' Q_h$ then captures the ranked household holding of characteristic k. The K by H matrix Q^* which has $c_k' Q_h$ in each of its entries is then the characteristic holdings matrix. The entries of this matrix describe the holdings-weighted tilts towards K different stock characteristics for each of the H households.[4]

Analogously to the stock coholdings matrix, the empirical **characteristic coholdings matrix**, defined over K stock characteristics, is the $K \times K$ matrix

$$\Omega_h^* = H^{-1} \sum_{h=1}^{H} \tilde{Q}_h^* \tilde{Q}_h^{*\prime}. \tag{4}$$

Here, the diagonal elements of Ω_h^* capture the intensity with which particular stock characteristics are held, and the off-diagonal elements capture the intensity with which particular pairs of stock characteristics are held, averaging across all households.

The characteristic coholdings matrix is analogous to the empirical covariance matrix of stock portfolio returns. To construct this return covariance matrix, we also begin with a single time period and calculate the outer product matrix of (characteristic-weighted) portfolio returns in that period (after time-series demeaning returns), and subsequently average these outer products over time.

To detect characteristic clusters in the data, we extract principal components of the characteristic coholdings matrix Ω_h^*, to decompose investor preferences for characteristics into orthogonalized basis vectors. We then regress observed stock characteristic tilts $c_k' Q_h$ on these unobserved principal components to assign characteristics to clusters.

[4] As we later describe, in our empirical work, we orthogonalize these K characteristics against each other using an ordering procedure informed by our clientele strength analysis. This provides confidence that we capture preferences for different characteristics over and above their correlation with one another in the cross-section of individual stocks.

2.4 Linking Who with What

Thus far, we have discussed measures of clientele strength for stock characteristics, but we have not linked the attributes of investors to the characteristics of the stocks they hold. A simple first step is to regress investors' characteristic tilts $c_k' Q_h$ on investor attributes F_h. These investor attributes might include demographics such as investors' geographical location, but they might also include attributes of their portfolios such as total portfolio size, the extent of time they have been in the market, their portfolio turnover, and so on.

A second step is to combine characteristics into clusters using the principal component analysis discussed above, and regress these principal components, computed at the household level, on investor attributes.

This analysis can be conducted in a univariate fashion, to answer simple questions such as whether investors with larger portfolios, or those who are better diversified, hold particular clusters of characteristics, or in a multivariate fashion in which (orthogonalized) characteristic tilts are regressed simultaneously on a number of investor attributes.

Such analysis can also be conducted at the stock level rather than the characteristic level. As we later describe, we do so by setting up and estimating a factor model (analogous to classic factor models in the literature explaining stock returns) to explain stock holdings, where factors F_h vary across investors rather than across time.

3 Indian Equity Market Data

3.1 Equity Ownership

Our data on Indian stock holdings, 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.

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.[5]

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.[6] 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.[7] 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.[8]

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.[9] 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, 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.

[5]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.

[6]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

[7]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

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

[9]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.

Furthermore, since there is no requirement in India that publicly listed equities 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.[10]

Some of our stock characteristics are categorical variables, and others are continuous. Because the distribution of raw continuous characteristics is often skewed and fat-tailed, we rank stocks by their characteristic values and use the demeaned rank as our stock-level characteristic measure for each continuous characteristic. By construction, this demeaned rank has a uniform distribution ranging from -0.5 to 0.5, with a mean of zero.[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 the given industry. For other missing characteristics which take continuous values, we use all available characteristics in a regression to impute values for the missing characteristics.[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

[10] Where two or more data sources differ, we first select the two sources that are more consistently in close agreement for the stock. From these two, we use the source that is more consistently in close agreement across stocks. 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 an average portfolio share of at least 1%.

[11] 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.

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

stock holdings.[13]

The continuous stock characteristics we consider are share price, stock age (years since listing), realized volatility, market capitalization, realized returns, turnover, market beta, book-market ratio, and realized skewness. 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.

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 over a million new investors into the market.

Table 1 summarizes attributes 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 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 internet 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.

Our empirical work utilizes several other account attributes, including the number of

[13]Specifically, for August 2011, we impute stock age for 6.2%, the book-market ratio for 3.2%, and lagged returns, volatility and skewness for about 0.24% of stock holdings. 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.

stocks held by the account (of the total set of 3,103 stocks that we consider), the number of stocks traded, and portfolio turnover. Table 1 shows that all these attributes 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. The median retail investor holds large stocks, with a portfolio at the 95th percentile of the firm size distribution. The other characteristics of median stock holdings 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 stock characteristic tilts 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 tilts in greater detail in section 4.

Details on other investor attributes are reported in the internet appendix. Figure A.1 in the appendix shows the distribution of accounts across four regions of 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. The figure also shows the distribution of stock holdings 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 (see 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 stock holdings, with remaining business groups accounting for a further 16% of stock holdings.

Figure A.3 in the internet appendix shows correlations between account attributes and characteristic tilts in investor portfolios. Within the set of account attributes, there are positive correlations between account age, account size, the number of stocks held, and the number of stocks traded, but all correlations are below 0.6. The strongest correlations are within the set of characteristic tilts. Portfolio tilts towards larger stocks (measured by market capitalization) are strongly positively correlated with tilts towards

high-share price stocks (0.78) and negatively correlated with tilts towards value stocks with high book-market ratios (−0.57). Share price and market cap tilts are both negatively correlated with realized volatility tilts. Tilts towards stocks with high past returns have strong correlations with other characteristic tilts. These correlations reflect both investor preferences and the correlation of characteristics in the cross-section of stocks, a problem we handle later in the paper by orthogonalizing characteristics across stocks.

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.[15] The characteristics of stock holdings 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.

4 Clientele Effects

In this section we apply our methodology to evaluate the strength of clientele effects in Indian stock holdings.

4.1 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 tilt $c'Q_h$ across households,

[15] 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.

where c is a stock characteristic rank ranging from -0.5 to 0.5, and the i'th element of Q_h is the weight of stock i in investor h's portfolio. The use of portfolio weights ensures that each investor has equal weight in the holdings matrix Q_h, although investors with concentrated portfolios will contribute more strongly to the stock coholdings matrix and to characteristic clientele strength.

Table 2 shows this measure of clientele strength for a range of stock characteristics that we observe in the data. These include the nine continuous stock characteristics described in Table 1, but also include a range of categorical variables such as whether or not the stock belongs to a business group, and dummies for industry membership of seven industry groups. To place continuous and categorical variables on an equal footing, we divide the variances of categorical variables by three to account for the fact that the maximum variance of a binary variable is three times the variance of a uniformly distributed variable with a range of one.[16] We indicate the categorical variables with italic font in the row labels for these variables.

The table has two panels, in the first of which we present "raw" clientele strength measures, and in the second of which we orthogonalize characteristics relative to one another. Orthogonalization helps to ensure that the clienteles we identify for 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.[17]

In the table, stock characteristics are presented in descending order of their clientele strength when the characteristics are orthogonalized. The first column in each panel shows the total clientele strength, with categorical variables renormalized. The second column in each panel shows the percentage of clientele strength that comes from off-diagonal elements or coholdings, rather than diagonal elements or holdings. Off-diagonal contributions are always smaller than diagonal contributions in this table where all investors are included.[18] The ordering of clientele strength across characteristics is similar

[16]The maximum variance of a binomial variable is 1/4, while the variance of a uniformly distributed variable with a range of one is 1/12.

[17]Online Appendix Section A describes the orthogonalization procedure in detail.

[18]The off-diagonal contribution is more important for investors with larger portfolios, as well as for those with well diversified portfolios (a Herfindahl-Hirschman index, or sum of squared portfolio weights, of 0.2 or less, corresponding to the diversification of an equally weighted portfolio containing five or more stocks). These results are shown in Tables A.1 and A.2 in the internet appendix. Shifting focus to well diversified investors or to the off-diagonal share of total variance (Table A.3) do not alter our main

whether we consider total clientele strength or the diagonal and off-diagonal contributions separately, so for simplicity, we focus our discussion on total clientele strength. For reference, the third column in each panel shows the raw standard deviations of characteristic tilts across households without renormalizing categorical tilts.

The strongest clientele effect in Table 2 is associated with stock age. Some Indian individual investors strongly prefer to hold young companies (recent IPOs), while other investors strongly prefer established companies.[19] Looking at other continuous characteristics, the second strongest clientele effect is associated with share price. As noted earlier, share price is strongly correlated (0.78) with market capitalization in the cross-section of Indian stocks; but the clientele strength for share price is noticeably stronger than that for market cap when these characteristics are studied separately without orthogonalization, and share price dominates when we orthogonalize characteristics. 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 relatively uniform (all individual investors tend to hold large companies) which explains the weaker clientele effect for this characteristic.

Among other continuous stock characteristics, both stock turnover and past realized returns also have relatively strong clientele effects in our data. The turnover clientele effect may reflect the tendency for some investors to prefer liquid stocks and others to focus on illiquid stocks in their portfolios. Turnover is positively correlated with volatility and market beta in the cross-section of Indian stocks, and it drives these other characteristics down the orthogonalized ranking of clientele strength. The clientele effect for past realized returns could arise because some investors 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 preferences from preferences for other stock characteristics that happened to do well in the period September 2010–August 2011.

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. One approach is to compare clientele strength to what we would observe under a series of simple

conclusions about clientele strength, so we discuss the measures derived from all investors in what follows.

[19]Tables A.1 and A.2 show that the stock age effect is strongest regardless of whether we look at all investors, investors with large portfolios, or only well diversified investors. The age effect remains strongest even if we exclude the top 10 or 50 most popular stocks from the clientele strength calculation.

alternative models. Table A.4 in the internet appendix reports clientele strength, broken into the holdings and coholdings components, for three alternative models applied to orthogonalized characteristics. If stocks are randomly picked with probability proportional to the free-float capitalization of the stock—which we view as a realistic benchmark, in contrast with, say, equal-weighted random stockpicking—then we obtain different clientele strengths for each characteristic which are reported in the table. Similarly, if we assume that investors hold optimally mean-variance diversified portfolios conditional on the number of stocks they hold (which we estimate using a lasso procedure that we describe later in the paper), then we again obtain different values for each characteristic which are reported in the table. We find that clientele effects in our Indian data are strong relative to all these alternatives. For a few characteristics, mean-variance optimization implies stronger coholdings components of clientele strength than we find in the data, but overall clientele strength is always considerably stronger in the data than can be explained by any of these alternative models.

As we have seen, a range of continuous stock characteristics have strong associated clientele effects. Table 2 also reveals that investor clientele effects for discrete categories of stocks are very strong. For example, whether or not a stock pays dividends exhibits the second strongest clientele effect of all stock characteristics observed in the data, showing that some investors prefer dividend-paying stocks while others eschew them. There is also a strong clientele for stocks that belong to business groups, with some investors attracted to such stocks and others avoiding them.[20] The fact that there are clienteles for these categorical stock characteristics appears consistent with theories of classification of risky assets into "styles" (Barberis and Shleifer 2003).

4.2 Stock Characteristic Clusters

In this subsection we assess whether there are clienteles for clusters of stock characteristics in the Indian equity market. To reduce the dimensionality of the analysis, we consider only our nine continuous stock characteristics together with the categorical dividend paying characteristic which has a particularly strong clientele effect. We apply the approach discussed in section 2, using orthogonalized characteristic tilts to create the K by H characteristic holdings matrix Q^* which has $c_k' Q_h$ in each of its entries. This matrix describes the holdings-weighted average tilts towards K different orthogonalized stock characteristics for each of the H households. We use this to compute the empirical

[20] Internet appendix Table A.1 shows that this business group clientele effect is even stronger among large individual investors.

characteristic coholdings matrix Ω_h^*, and finally extract principal components (PCs) of Ω_h^*. The first three PCs account for a large fraction of the total variance of household tilts to the characteristics with the strongest clienteles, as reported in Figure 3, and we focus on these three PCs in our analysis of characteristic clusters.

Figure 3 visualizes characteristic clusters by regressing observed stock characteristic tilts (i.e, the rows of the characteristic holdings matrix) simultaneously on three principal components, which are simply linear combinations of the stock characteristic tilts and by construction are orthogonal to one another. Figure 3 plots the loadings from these regressions, and reveals the types of stock characteristics that are grouped together into these three distinct clusters.

PC1, shown in red, captures an investor clientele for old (established) stocks that pay dividends and have high turnover and volatility. Since clientele effects are defined by variance, this can also be interpreted as a preference by some investors to hold young (IPO), non-dividend-paying stocks with low turnover and volatility, i.e, the characteristic cluster defines a spectrum along which investors are located. PC2, shown in blue, captures an investor clientele for (or against) young stocks with low share prices, high turnover, high returns, and positive skewness. We view this as a clientele effect for "lottery-like" stocks. PC3, shown in green, shows that dividend-paying stocks with low beta, low turnover, and high past returns constitute the third characteristic cluster in individual portfolios.

Table 3 verifies the clusters revealed by the principal components analysis in a simpler fashion. Panel A of the table shows how tilts towards each of ten characteristics, ordered by their clientele strength, are correlated across investor portfolios. The table essentially plots Ω_h^*, but normalized by the variances along the diagonal to convert covariances into correlations. The table color-codes these correlations using a heatmap, with deeper shades of red representing positive correlations, and blue indicating negative correlations. In Panel B of the table, we re-order these characteristics based on the principal components analysis. The first four characteristics are dominant in PC1, the next three are dominant in PC2 while playing a smaller role in PC1, and the last three are relatively important in PC3. We see positive correlations among the characteristic tilts within each block, which are strongest for PC1 and weakest for PC3 reflecting the declining importance of successive PCs in the structure of the characteristic coholdings matrix Ω_h^*.

In the internet appendix, we assess the robustness of these conclusions to variations in the weighting of individual accounts and the characteristics we consider. Table A.5 repeats Table 3 weighting accounts by their size, showing broadly similar correlations among characteristic tilts although the characteristics included in PC1 are less strongly

correlated among large accounts relative to those in PC2 and PC3. Figure A.4 reconstructs Figure 3 after applying size weights to investors. These PCs capture clientele effects within the set of investors holding larger portfolios, and are thus a bit different. For this group, PC1 continues to capture a strong clientele effect for stocks of extreme ages. PC2 and PC3 capture contrarian preferences for stocks with poor realized returns– but the PC2 clientele seeks negative skewness, low price, and low beta, while the PC3 clientele prefers high beta and high turnover. Figure A.5 repeats Figure 3 including all characteristic tilts (including discrete categories of stocks), and shows a similar structure for PC1, although somewhat different patterns for PC2 and PC3 when all characteristics are included.

4.3 Who Owns What Characteristics?

Having identified strong clientele effects both for individual stock characteristics and for clusters of characteristics, our next step is to link preferences for particular characteristics with investor attributes. We do so by regressing tilts to characteristics and PCs onto a range of account attributes that we observe in the data.

We focus on ten account attributes that can be calculated without knowledge of the particular stocks held: continuous measures of account value, age, turnover, the number of stocks held, and the number of stocks traded; a dummy variable that captures whether an investor is extremely underdiversified, i.e., holding only a single stock in their account; and dummy variables for the four regions where investors may be located. Much as we do for stock characteristics, we transform continuous account attributes into ranks ranging from -0.5 to 0.5.[21]

Table 4 shows coefficients from univariate regressions of stock characteristic tilts onto these investor attributes. The table reveals that large accounts hold established dividend-paying stocks with high share prices, volatility, and past returns but low betas. These accounts also prefer growth stocks with relatively lower market cap that are not part of a business group. Well diversified and older accounts have many of the same preferences.[22] Accounts with high turnover, by contrast, tilt towards small, low-share-price,

[21]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, such as account turnover (which is often zero) and the number of stocks held, are shared by a significant fraction of all households, this assignment of ranks reduces the variance of the ranked attribute across households by less than 10% while preserving the mean of zero.

[22]These patterns are broadly consistent with the findings of Campbell et al. (2014) in a study focusing exclusively on account age, and of Campbell et al. (2019) in a study focusing exclusively on account size.

value stocks with high turnover, volatility, beta, and skewness (i.e., the easily traded "lottery-like" stocks detected earlier in PC2), and single-stock accounts hold young, non-dividend-paying stocks, affiliated with business groups, that have high market capitalization and low volatility (which we can characterize as "mega-IPOs").

These findings are reinforced by Table 5, which repeats these univariate regressions on account attributes, but using the three principal components as left-hand side variables. The table reaffirms that larger, older, and better diversified accounts tend to tilt their portfolios towards PC1 and PC3, i.e., established stocks and dividend-paying low-beta stocks, but disfavor PC2, i.e., young, lottery-like stocks. In contrast, accounts with high turnover and extremely underdiversified accounts tend to favor young, lottery-like stocks, and load negatively on both PC1 and PC3. The negative loading on PC1 is most pronounced for accounts holding a single stock, reflecting the fact that this one stock tends to be a young, large stock, i.e., a mega-IPO.[23]

4.4 Robustness and Stability Over Time

The internet appendix extends these results in several directions. Tables A.6 to A.8 report regressions like those of Table 4 for discrete firm characteristics such as business group affiliation, industry group, and firm headquarters location. Table A.8 shows that there is a pronounced home bias evident in the data, with investors located in the South, West, and North of the country preferring stocks headquartered in those locations. Tables A.9 and A.10 show that the patterns in Tables 4 and 5 are robust to weighting accounts by their size, and Tables A.11 and A.12 show that these patterns are robust to using multivariate regressions on all account attributes simultaneously rather than univariate regressions. Table A.13 shows that the first three PCs from the full set of characteristic tilts have similar relationships with investor attributes as when the PCs are extracted from the subset used in our main analysis.

It is natural to ask whether the clientele effects we have identified are stable over time. The internet appendix also explores this question. Table A.14 shows that the ordering of characteristic clientele strength is largely stable when comparing August cross-sections in each year from 2002 through 2010 with the August 2011 cross-section we discuss in the body of the paper. Table A.15 presents the average correlation of investors' characteristic tilts estimated using the August cross-sections in each year from 2003 through 2010, and shows that the resulting clusters are similar to those in Table 3, with the PC1 and

[23]Internet appendix Table A.6 provides supporting evidence by showing that the single-stock dummy has a strong positive loading on the Reliance ADAG business group dummy. This business group contains Reliance Power, the firm which had a mega-IPO in 2008.

PC2 clusters very stable over time, and some shifts in PC3. Table A.16 utilizes average factor loadings from regressions identical to those in Table 4, but estimated on the 2003-2010 August cross-sections, and shows that the relations between account attributes and characteristic tilts estimated in Table 4 also appear in earlier time periods. In Table A.17, we fix the composition of the three PCs to be the same as in 2011, but estimate it over 2003-2010. Reassuringly, Table A.17 shows that the resulting portfolios have very similar clienteles as those observed in Table 5, which uses the August 2011 cross-section. Overall, these exercises reassure us that our inferences are not merely an artifact of the 2011 time period, and are stable over time.

5 Factor Models of Stock Holdings

Our focus so far has been on stock characteristics rather than on individual stocks. For some applications, such as individual corporations learning about the clientele that holds their stocks, going down to the more granular stock level may be necessary. In this section we develop a methodology for doing this by estimating a factor model of stock holdings, and we contrast results from observed and unobserved factor models.

5.1 Observed Multifactor Model: Theory

Our approach is to estimate a factor model, a cross-sectional analog of the models commonly used to describe the variation in stock returns over time. Focusing on the portfolio share holdings matrix Q, for each stock i we can estimate a cross-sectional regression:

$$Q_{ih} = \alpha_i + \sum_{k=1}^{K} \beta_{ik} F_{kh} + \varepsilon_{ih}, \qquad h = 1, ..., H, \tag{5}$$

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

In equation (5), the factors can 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 can 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.

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 use a "leave-out" approach that excludes own holdings when we construct portfolio attribute factors. This implies that portfolio attribute factors are missing for all accounts that hold only a single stock.

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.

The factor model (5) simplifies the structure of the stock coholdings matrix Ω_h. Consider a situation where $\alpha_i = 0$, as will be the case if equation (5) is estimated using household-demeaned holdings $\tilde{Q}_h$ 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 take the form:

$$\Omega_{h,i,i} = \sum_{k=1}^{K} \beta_{ik}^2 \sigma_k^2 + \sigma_i^2, \tag{6}$$

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 take the form:

$$\Omega_{h,i,j} = \sum_{k=1}^{K} \beta_{ik} \beta_{jk} \sigma_k^2, \tag{7}$$

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 (5). A disadvantage of (5) 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 (6) and (7).

5.2 Observed Multifactor Model: Implementation

To estimate an observed multifactor model for stock holding, we construct 15 account-attribute and portfolio-attribute factors from the account and portfolio attributes summarized in Table 1 and discussed previously in the paper. To this we add several other sets of factors. First, we include 3 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, construction; 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 10 large business groups. Fourth, we add 3 dummy variables for the broad geographical zones in which the firms are headquartered.[24] Finally, we add a dummy variable for single-stock accounts, for which portfolio attribute factors are unavailable given our "leave-out" factor construction. The factors enter the model in raw form, without orthogonalization. In all, we have 38 factors in our observed multifactor model.

We estimate stock holdings 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 (5), 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' stock holdings, 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 6 summarises the $\tilde{\beta}_{ik}$ estimated from the 3,103 stock-specific estimates of equation (5). 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.

[24] To avoid collinear factors, we exclude the other retail industry, and the eastern geographical zone.

The cross-stock mean $\overline{\beta}_k$ 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 stock holdings. 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 absolute $t-$statistic across all 3,103 regressions, and the percentage of estimated $\tilde{\beta}_{ik}$'s that are statistically significantly different from zero at the 5% level.

Panel A of Table 6 shows the distribution of $\tilde{\beta}_{ik}$ for the account-attribute factors, and Panel B summarizes the distribution of $\tilde{\beta}_{ik}$ for investors' portfolio characteristic tilts. 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 stock holdings. 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-attribute and portfolio-attribute factors, but for now, we simply discuss the magnitude of the cross-stock spread in factor loadings seen in Table 6.

Account-attribute factors

The account-attribute factor with the highest cross-sectional standard deviation of factor loadings is account size. The next most important account-attribute factor, again looking at the standard deviation of $\tilde{\beta}_{ik}$ across stocks, 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. Of the continuous account attributes, turnover and account age follow in order of importance. 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.

As discussed earlier, there is some evidence of geography-based stock selection, 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 6 turns to portfolio-attribute factors based on portfolio characteristic tilts. The table divides these factors into six 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 stock holdings for high or low share price, old, high-turnover, or dividend-paying stocks; business group factors; industry factors, and geographical factors capturing the location of the headquarters of the firms held in investors' portfolios.

The loadings for many of the 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.

Comparison of explanatory power

Connor and Korajczyk (2019) introduce a way to assess the performance of specific groups of factors in multifactor models. The approach they 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 7 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 stock holding data.

The first row of Table 7 shows that the R^2 of the full multifactor model is 3.96%. 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 57% of the total explanatory power in the equally-weighted case, with portfolio-

attribute factors accounting for roughly 36% 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, account size is the most important and the single-stock dummy is the next most important. This analysis helps to bring together disparate themes in prior literature on the influence of account characteristics on stock holding propensities into a common framework. For example, account size and wealth have been highlighted as important determinants of stock holdings 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, business group factors are the most important. Industry factors are next most important, followed by Fama-French factors.

Stock characteristics and factor loadings

Akin to the analysis in Table 4, we can aggregate individual factor loadings across stocks up to the level of characteristics, using the characteristic ranks constructed for each stock to generate weighted average loadings, using the demeaned ranks as weights.

Tables A.18 and A.19 in the internet appendix show the results of this exercise, which delivers very similar results to the analysis of stock characteristics described earlier. For example, the first row of the top panel of Table A.18 verifies 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. Table A.19 shows how investors' portfolio attributes are related to orthogonalized 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. Since we have orthogonalized the characteristics in the cross-section of stocks, any such association is not mechanical but reflects investor behavior.

Unsurprisingly, the diagonal elements are all positive and tend to be the strongest effects in this part of the table, but there are some interesting off-diagonal effects indicating that stock characteristics cluster into groups, with similar investor clienteles holding constellations of these characteristics simultaneously, as we saw in the characteristic-level analysis. Our multifactor model shows once again that investor clienteles form around related stock characteristics.

5.3 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) applied at the individual stock level. To construct such a model, we compute the principal components (PCs) of the 3,103 by 3,103 covariance matrix of stock holdings derived from the 9.7 million accounts that we observe. The first PC is the eigenvector of this covariance matrix which corresponds to the largest eigenvalue, and subsequent PCs are estimated as the eigenvectors associated with successively smaller eigenvalues of the covariance matrix. By construction, these PCs are orthogonal to one another, and are normalized linear combinations of household stock holdings that together summarize the total variance of stock holdings. They are ordered by the fraction of the total variance that they capture.

In internet appendix Table A.20 we report the pooled R^2 statistic for a 10-factor PCA model and for the first, second, and third PCs. Since the PCs 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 A.20 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 (4.0%). 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 is almost entirely eliminated because the first PCA factor is chosen to explain the holdings of that one stock. Once we eliminate the ten most widely held stocks from the sample, the first three PCA factors have negligible explanatory power. If we eliminate the fifty most widely held stocks, the observed multifactor model clearly dominates even the ten-factor PCA model. This reflects the fact that PCA methodology applied to stock holdings concentrates on explaining patterns in the large number of holdings of a very few stocks. Internet appendix Figure A.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).

6 Coholdings and Return Covariances

In this section we study the relation between the clientele effects that we have identified for particular stock characteristics, and the return variances and covariances of portfolios of stocks formed on the basis of these characteristics. Table 8 shows results when we regress the return variances of these portfolios (formed using the c_k weighting scheme for each of the ten characteristics k) on their total holding variances (clientele strengths), and on the diagonal and off-diagonal components of clientele strength separately. We do this both for the variances of these portfolios (columns 1-3) as well as for the covariances of these portfolios with one another (columns 4-6). We find little effect of total clientele strength, but when we decompose clientele strength into diagonal and off-diagonal contributions in columns 3 and 6, we find that the off-diagonal contribution which captures coholdings propensity has a strong positive effect on portfolio variances and covariances, while the diagonal contribution which captures the tendency for widely held stocks to have extreme characteristic values has a negative effect. Although we only observe ten variances and 45 covariances, the explanatory power of the regressions in columns 3 and 6 is impressive at 50% and 69% respectively. Table A.21 confirms that this pattern holds true when we size-weight all investors while constructing the characteristic tilt estimates.

We interpret these findings as follows. Widely held stocks tend to have low return variances, so characteristics that load on these stocks also tend to have low variances. However, characteristics that attract coholdings have volatile returns, either because investor clienteles form around characteristics whose stock prices move together, or because investor clienteles themselves move stock prices as money flows in and out of favored characteristics. Similarly, characteristics that are coheld by overlapping clienteles also tend to move together.

We verify these insights using regressions of stock return volatility and correlations on stock-level holdings variances and coholdings. In Table A.22 in the internet appendix, we regress return variance on holdings variance, both in levels and in cross-sectional ranks. We find weak evidence of a negative relationship in levels and strong and significant evidence of a negative relationship when both left- and right-hand side variables are measured in ranks, as well as a strong negative relationship when return volatility is in natural units and holdings variance is measured in ranks. This reflects the fact that widely held stocks, whose holdings variance is high, tend to be well diversified stocks with low variance. Figure 4 Panel A plots this relationship, binning holdings variance into deciles on the horizontal axis, and plotting the average return volatility in each bin on the vertical axis.

In Table A.23 we ask how stock return covariances are related to holdings covariances, the off-diagonal elements of the coholdings matrix, in a regression with roughly 4.8 million observations. We obtain a significant positive relationship although the explanatory power is modest (about 0.4% in levels and 1-1.3% in ranks). Figure 4 Panel B plots this relationship, binning holdings covariance into deciles on the horizontal axis, and plotting the average pairwise return correlation in each bin on the vertical axis. The plot shows a clear positive relationship: stocks that are held together tend to move together. Online appendix Tables A.24 and A.25 show that these patterns documented using the August 2011 cross-section continue to remain strong when estimated using other cross-sections of the data.

6.1 The Failure of Diversification

The tendency for coheld stocks to move together suggests that Indian investors are not following the prescription of classical finance theory to diversify portfolio risk. To show this quantitatively, in this section we contrast the data on Indian investor portfolios with the solution to a constrained diversification 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 here.

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 5 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.[25] 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 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 also consider investors' performance under a popular four-factor model of returns. 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.[26] 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 5 reveals that the four factor benchmark makes the optimal diversification conjecture even more tenuous. Few portfolios lean heavily towards factors that

[25]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.

[26]We use the dataset of Agarwalla et al. (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.

have been well compensated historically, aside from the market factor—which accounts for less than half of the tangency portfolio. This exercise shows little evidence of constrained mean-variance optimization.

Overall, the positive relationship between return covariances and coholdings 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 coheld 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.

7 Conclusion

In this paper we have suggested that a factor model for investors' stock holdings 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 analog 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 stock holdings are the norm at the time of our analysis. While direct stock holdings 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' stock holdings 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 weaker 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.[27] Also, we find that characteristics form clusters with similar clienteles, even after we orthogonalize those characteristics in the cross-section of Indian stocks. Established, dividend-paying stocks have similar clienteles, as do easily traded lottery-like stocks.

Finally, we explore the relation between coholdings and the covariances of stock returns. Stocks and characteristic portfolios that are more commonly coheld tend to correlate 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.

[27]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).

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.

Baker, M., R. Greenwood, and J. Wurgler (2009). Catering through nominal share prices. *The Journal of Finance 64*(6), 2559–2590.

Baker, M. and J. Wurgler (2013). Behavioral corporate finance: An updated survey. In *Handbook of the Economics of Finance*, Volume 2, pp. 357–424. Elsevier.

Balasubramaniam, V., J. Y. Campbell, T. Ramadorai, and B. Ranish (2021). 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.

Barberis, N. and A. Shleifer (2003). Style investing. *Journal of Financial Economics 68*(2), 161–199.

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.

Curcuru, S., J. Heaton, D. Lucas, and D. Moore (2010). Heterogeneity and portfolio choice: Theory and evidence. In Y. Ait-Sahalia and L. P. Hansen (Eds.), *Handbook of Financial Econometrics: Tools and Techniques*, pp. 337–382. San Diego: North-Holland.

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.

Fagereng, A., L. Guiso, and L. Pistaferri (2018). Portfolio choices, firm shocks, and uninsurable wage risk. *The Review of Economic Studies 85*(1), 437–474.

Fama, E. F. and K. R. French (1992). The cross-section of expected stock returns. *Journal of Finance 47*(2), 427–465.

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.

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.

Harris, M. and A. Raviv (1993). Differences of opinion make a horse race. *The Review of Financial Studies 6*(3), 473–506.

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.

Malmendier, U. and S. Nagel (2011). Depression babies: Do macroeconomic experiences affect risk taking? *The quarterly journal of economics 126*(1), 373–416.

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

Meeuwis, M., J. A. Parker, A. Schoar, and D. I. Simester (2018). Belief disagreement and portfolio choice. Technical report, National Bureau of Economic Research.

Merton, R. (1987). A simple model of capital market equilibrium with incomplete information. *Journal of Finance 42*(3), 483–510.

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*.

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.

Tobin, J. (1958). Liquidity preference as behavior towards risk. *Review of Economic Studies 25*(2), 65–86.

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.

Vissing-Jørgensen, A. (2002). Towards an explanation of household portfolio choice heterogeneity: Nonfinancial income and participation cost structures. Technical report, National Bureau of Economic Research.

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. Account Turnover is the investors' average monthly value of trades over the past year divided by the portfolio size as on August 2011. 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. Characteristic tilts are the value weighted average stock characteristic of investors' portfolios, where the stock characteristics (except for "Dividend Paying") 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. Stock Turnover and Realized Volatility, Returns, Skewness, and Market Beta are measured over the previous year, using weekly data. Dividend paying equals one for stocks that paid a cash dividend in the previous year and zero for stocks that did not.

Variable Name	Mean	Std. Dev.	P10	P25	Median	P75	P90
Account Attributes							
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
Account 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
Characteristic Tilts							
Stock Age	-0.06	0.27	-0.43	-0.30	-0.08	0.15	0.34
Dividend Paying	0.68	0.38	0.00	0.44	0.86	1.00	1.00
Share Price	0.22	0.21	-0.06	0.13	0.26	0.38	0.44
Stock Turnover	0.08	0.19	-0.14	-0.02	0.07	0.22	0.33
Realized Returns	-0.02	0.20	-0.28	-0.16	0.00	0.10	0.22
Market Beta	0.11	0.18	-0.12	-0.02	0.12	0.23	0.34
Realized Skewness	-0.15	0.19	-0.34	-0.30	-0.17	-0.05	0.12
Realized Volatility	-0.17	0.18	-0.35	-0.30	-0.21	-0.09	0.09
Book/Market	-0.14	0.18	-0.33	-0.25	-0.19	-0.07	0.08
Market Capitalization	0.38	0.17	0.18	0.37	0.45	0.48	0.49

Table 2
Characteristic Clientele Strength

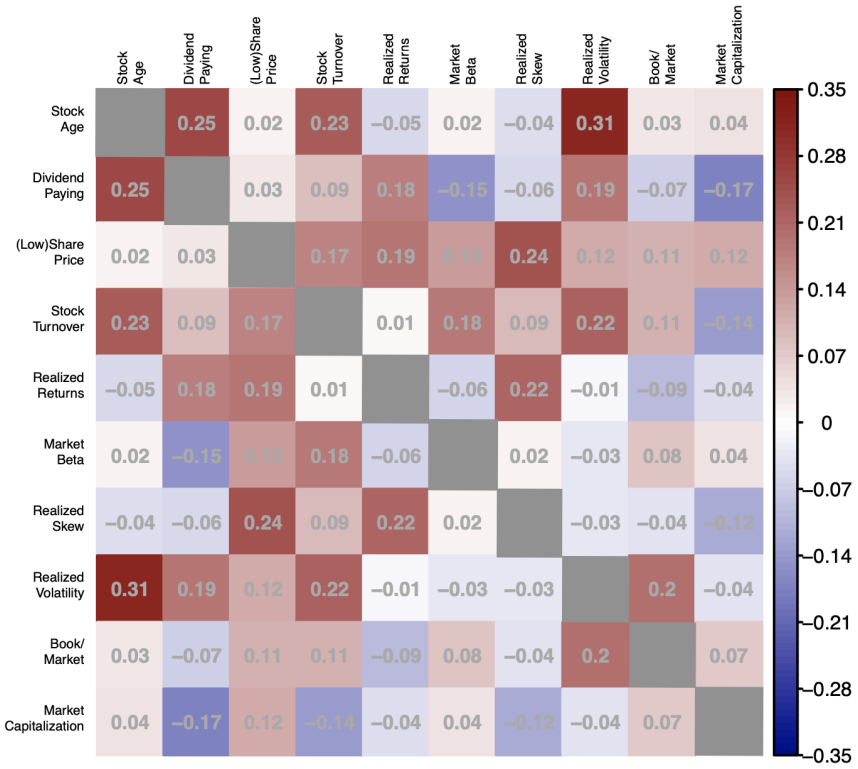
The shaded rows of this table present the variance of investors' characteristic tilts (characteristic clientele strength) in August 2011, using both raw and orthogonalized stock characteristics produced following the procedure described in section 2.2. Characteristics are presented in descending order of this variance (using orthogonalized characteristics). The second of each set of columns presents the percentage of the variance that can be attributed to off-diagonal elements of the coholdings matrix, i.e. the tendency for investors to cohold stocks with similar characteristics. The third and final columns convert the variances into standard deviations. The variances, but not standard deviations, of discrete stock characteristics (*italicized* row labels) are divided by 3 to account for the fact that the maximum variance of a binary variable is three times as large as one that is uniformly distributed with a range of one. All statistics weight investors equally. As the set of industry and zone dummies are collinear, the weakest one is dropped in the orthogonalization.

	Raw			**Orthogonalized**		
Characteristic	Variance	% Off-diagonal	Std. Dev	Variance	% Off-diagonal	Std. Dev
Stock Age	0.074	11.8%	0.273	0.074	11.77%	0.273
Dividend Paying	0.047	12.5%	0.375	0.049	11.88%	0.384
Share Price	0.044	22.4%	0.211	0.045	12.85%	0.213
Stock Turnover	0.035	8.5%	0.187	0.045	7.01%	0.211
Business Group: Any	0.048	6.8%	0.378	0.042	5.93%	0.354
Realized Returns	0.040	13.5%	0.201	0.040	7.36%	0.201
Zone: West	0.041	5.8%	0.350	0.040	5.20%	0.344
Beta	0.033	11.7%	0.181	0.035	6.02%	0.186
Industry: Financial Services	0.030	6.5%	0.298	0.033	6.97%	0.313
Realized Skew	0.035	7.5%	0.187	0.030	3.88%	0.173
Realized Volatility	0.033	15.9%	0.182	0.030	5.45%	0.172
Industry: Manufacturing	0.038	5.4%	0.339	0.026	1.74%	0.279
Book/Market	0.032	13.0%	0.180	0.026	3.87%	0.161
Market Capitalization	0.029	21.2%	0.171	0.025	6.89%	0.159
Industry: Oil & Gas	0.032	10.0%	0.310	0.022	9.02%	0.256
Business Group: Reliance (ADAG)	0.024	11.9%	0.266	0.021	11.52%	0.250
Zone: South	0.024	4.6%	0.267	0.018	3.03%	0.232
Industry: IT	0.023	3.3%	0.263	0.013	1.39%	0.198
Industry: Construction	0.011	2.6%	0.182	0.012	9.09%	0.193
Public Sector Enterprise	0.030	10.1%	0.302	0.012	5.80%	0.193
Business Group: Reliance (DAG)	0.012	6.8%	0.192	0.012	6.67%	0.193
Industry: Food, Agro. & Textiles	0.008	3.7%	0.158	0.010	4.67%	0.177
Zone: North	0.027	3.9%	0.286	0.008	0.92%	0.153
Business Group: Tata	0.007	4.2%	0.147	0.007	2.83%	0.141
Business Group: Suzlon	0.003	1.6%	0.094	0.003	1.44%	0.094
Business Group: Mahindra	0.003	1.1%	0.089	0.003	0.77%	0.090
Business Group: Jaypee	0.003	7.8%	0.087	0.002	7.29%	0.087
Business Group: Jindal	0.002	1.5%	0.081	0.002	1.35%	0.084
Business Group: Birla Aditya	0.002	1.7%	0.075	0.002	1.28%	0.079
Business Group: Adani	0.001	2.9%	0.057	0.002	5.58%	0.075
Business Group: Vedanta	0.001	2.1%	0.058	0.001	2.51%	0.060
Industry: Other	0.005	4.0%	0.120			
Zone: East	0.011	2.2%	0.178			

Table 3
Correlations of Investors' Characteristic Tilts

The panels below present the correlations of investors' characteristic tilts (using orthogonalized characteristics, excluding business group, industry and geographic zone) in August 2011, equally weighting each investor. The shading illustrates the sign and magnitude of the correlations indicated in gray text. In Panel A, characteristics are presented in order of their clientele strength (Table 2), whereas Panel B reorders the characteristics to illustrate the presence of clusters of (coheld) characteristics.

Panel A: Ordered by Characteristic Clientele Strength



Panel B: Ordered by Cluster

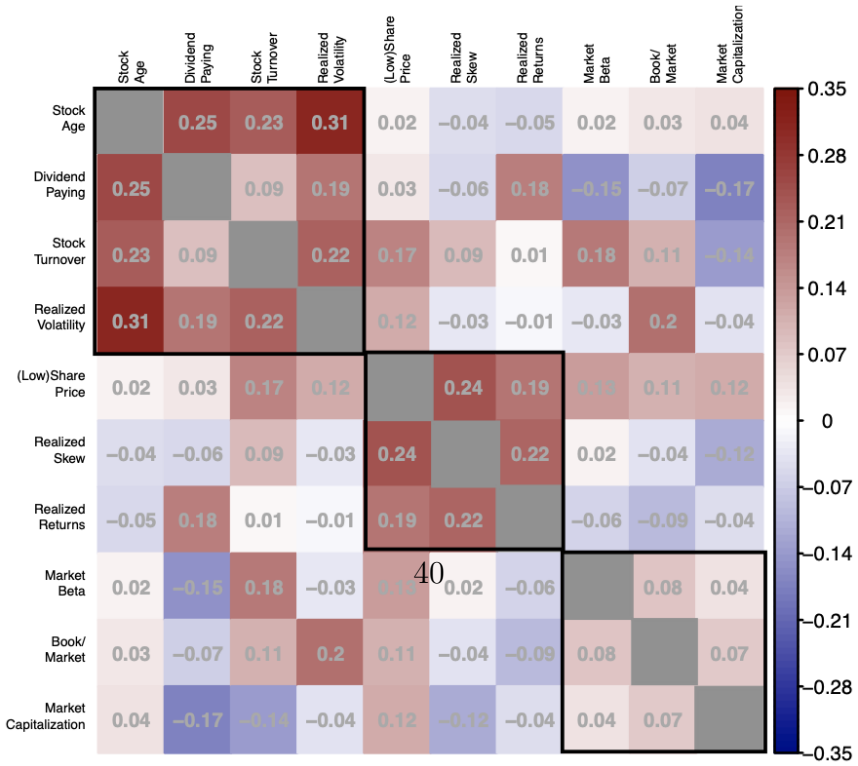


Table 4
Who Owns What: Characteristic Tilts
Single Factor Model Estimates

This table presents coefficients, along with the corresponding standard errors and R-squared, from univariate regressions of investors' characteristic tilts (in columns, using orthogonalized characteristics) on account attributes (in rows). Statistics equally weight all investors in August 2011. Shading of coefficients corresponds to their sign and magnitude.

	Stock Age	**Dividend Paying**	**Share Price**	**Turnover**	**Bus. Gp. Any**	**Realized Returns**	**Market Beta**	**Realized Skew**	**Realized Volatility**	**Book /Market**	**Market Cap.**
Size	7.11	9.65	5.84	0.13	-4.22	2.89	-1.80	-0.64	1.71	-1.79	-2.29
	(0.008)	(0.006)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)
	6.78	18.89	7.53	0.00	4.28	2.06	0.94	0.14	0.98	1.24	2.06
No.Stocks	3.36	3.08	1.17	0.09	-1.21	1.45	-0.52	0.48	0.81	-0.47	-1.09
	(0.009)	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)
	1.51	1.93	0.30	0.00	0.35	0.52	0.08	0.08	0.22	0.09	0.47
Age	4.25	1.75	1.00	-2.93	-0.77	2.57	-1.91	0.92	1.39	0.63	-1.73
	(0.009)	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)
	2.43	0.62	0.22	1.93	0.14	1.64	1.05	0.28	0.65	0.15	1.19
Turnover	-1.26	-1.93	-3.75	2.36	-0.14	-1.13	0.93	1.08	1.29	0.95	0.14
	(0.009)	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)
	0.21	0.75	3.09	1.25	0.00	0.32	0.25	0.39	0.56	0.35	0.01
Single Stock Dummy	-5.17	-5.75	-0.48	-0.93	1.85	-2.17	0.58	-0.12	-2.07	0.65	0.61
	(0.009)	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)	(0.005)
	3.59	6.70	0.05	0.19	0.82	1.16	0.10	0.01	1.45	0.17	0.15
No. Stocks Traded	1.67	1.89	-0.15	1.53	-1.19	0.00	0.39	0.50	1.10	-0.06	-0.86
	(0.009)	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)
	0.37	0.72	0.01	0.52	0.34	0.00	0.04	0.08	0.41	0.00	0.29
Eastern	0.42	0.38	-0.21	0.66	-0.22	0.08	0.45	-0.16	0.19	0.11	0.31
	(0.009)	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)
	0.02	0.03	0.01	0.10	0.01	0.00	0.06	0.01	0.01	0.00	0.04
Southern	0.08	1.00	-0.52	1.13	-0.72	0.69	0.86	0.00	0.30	0.00	-0.59
	(0.009)	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)
	0.00	0.20	0.06	0.29	0.13	0.12	0.22	0.00	0.03	0.00	0.14
Western	1.04	-0.10	0.53	-0.83	0.26	0.05	-1.25	0.45	0.25	-0.18	-0.29
	(0.009)	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)
	0.15	0.00	0.06	0.16	0.02	0.00	0.45	0.07	0.02	0.01	0.03
Northern	1.04	-1.15	0.05	-0.63	0.57	-0.79	0.28	-0.40	-0.72	0.13	0.68
	(0.009)	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)
	0.35	0.27	0.00	0.09	0.08	0.16	0.02	0.05	0.18	0.01	0.18

Table 5
Who Owns What: Characteristic Clusters
Single Factor Model Estimates

This table reports the factor loadings, the standard error, and the R^2 from univariate regressions of the first three principal components of investors' characteristic tilts (using orthogonalized characteristics and excluding business group, industry and geographical zone) in August 2011 on account attributes. The account attributes (in rows) are normalized to each have variance of one. Shading of coefficients corresponds to their sign and magnitude. The estimates equally weight all investors in the data.

	PC1	PC2	PC3
Size	0.0922	-0.0549	0.0828
	(0.000)	(0.000)	(0.000)
	9.21	5.00	12.50
No.Stocks	0.0402	-0.0105	0.0270
	(0.000)	(0.000)	(0.000)
	1.75	0.18	1.33
Age	0.0351	-0.0169	0.0350
	(0.000)	(0.000)	(0.000)
	1.33	0.47	2.23
Turnover	-0.0012	0.0396	-0.0313
	(0.000)	(0.000)	(0.000)
	0.00	2.60	1.79
Single Stock Dummy	-0.0720	0.0066	-0.0410
	(0.000)	(0.000)	(0.000)
	5.62	0.07	3.07
No. Stocks Traded	0.0292	0.0044	0.0050
	(0.000)	(0.000)	(0.000)
	0.92	0.03	0.04
Eastern	0.0079	0.0033	-0.0026
	(0.000)	(0.000)	(0.000)
	0.07	0.02	0.01
Southern	0.0102	0.0117	0.0035
	(0.000)	(0.000)	(0.000)
	0.11	0.23	0.02
Western	0.0051	-0.0099	0.0067
	(0.000)	(0.000)	(0.000)
	0.03	0.16	0.08
Northern	-0.0219	-0.0023	-0.0093
	(0.000)	(0.000)	(0.000)
	0.52	0.01	0.16

Table 6
Stock-level Estimates: Multi Factor Model Summary

For each stock, we estimate the observed factor model in using the equally-weighted cross-section of 9.7 million investors in August 2011. This table summarizes the factor loadings 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 attribute, and each row in Panel B corresponds to a characteristic tilt. Characteristic tilts used in the regression exclude any holdings of the selected stock in their construction. 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 for which the factor loading is statistically significantly different from zero at the 5% level.

Panel A: Account Attributes

	Std. Dev	10%	50%	90%	Avg. |t-stat|	Sig.(5% level)
Age	3.53	-0.40	0.05	0.63	13.70	85.92
Size	14.85	-0.77	-0.01	0.96	18.12	90.36
Turnover	5.90	-0.10	0.01	0.34	6.41	62.26
No. Stocks	2.17	-0.15	0.00	0.30	4.49	57.20
No. Stocks Traded	1.81	-0.16	0.01	0.28	4.69	54.46
Single Stock Dummy	10.19	-0.46	-0.07	0.08	8.19	80.89
Geographic Region						
Southern	2.89	-0.27	0.00	0.31	6.20	57.43
Northern	3.90	-0.27	-0.01	0.16	4.50	50.66
Western	5.06	-0.34	0.04	0.34	6.37	71.16

Panel B: Portfolio Characteristic Tilts

	Std. Dev	10%	50%	90%	Avg. |t-stat|	Sig.(5% level)
Fama-French factors						
Book/Market	1.17	-0.02	0.07	0.30	6.38	84.85
Market Capitalization	2.17	-0.44	-0.14	-0.02	10.06	94.55
Market Beta	1.66	-0.22	-0.02	0.20	5.39	63.81
Return-based factors						
Realized Returns	1.67	-0.34	0.01	0.14	5.12	56.94
Realized Volatility	1.42	-0.06	0.07	0.30	6.67	84.53
Realized Skewness	1.47	-0.06	0.02	0.15	3.90	48.08
Behavioral factors						
Share Price	1.89	-0.77	-0.16	-0.02	12.21	95.55
Stock Age	2.61	-0.34	0.05	0.38	9.09	82.02
Turnover	1.03	-0.22	-0.02	0.31	6.12	72.22
Dividend Paying	1.83	-0.83	-0.17	0.00	13.11	93.23
Business Group Holdings						
Reliance (ADAG)	12.64	-0.14	0.00	0.18	4.96	49.69
Tata	1.32	-0.18	0.00	0.07	3.30	46.70
Reliance (DAG)	2.24	-0.57	-0.06	0.09	4.56	56.82
Birla Aditya	2.39	-0.11	-0.01	0.02	2.42	29.91
Jaypee	2.35	-0.23	-0.01	0.04	3.41	49.47
Jindal	1.22	-0.15	-0.01	0.02	2.70	38.58
Mahindra	0.81	-0.18	-0.01	0.03	2.62	35.48
Suzlon	1.39	-0.27	-0.03	0.02	4.13	60.33
Vedanta	0.78	-0.12	-0.01	0.04	2.40	34.87
PSE	2.40	-0.13	0.01	0.11	3.98	49.79
Industry Holdings						
Construction	2.02	-0.35	-0.04	0.04	3.95	50.63
Financial Services	1.67	-0.63	-0.07	0.03	7.52	76.64
Food, Agri. and Textiles	1.13	-0.12	0.01	0.21	4.06	51.11
Information Technology	1.51	-0.11	0.00	0.14	4.22	52.76
Manufacturing	1.07	-0.07	0.03	0.30	5.38	65.23
Oil and Gas	3.36	-0.14	0.00	0.11	4.03	45.60
Geography						
Southern	0.79	-0.10	0.00	0.11	3.58	44.80
Northern	0.82	-0.09	0.00	0.06	2.81	35.74
Western	1.00	-0.16	-0.02	0.09	3.96	50.50

Table 7
Stock-level Estimates: Explanatory Power

This table presents the relative contribution of different groups of factors to the explanatory power of the regressions summarized in Table 6. The first row in Panel A presents the full model R-squared from a pooled least squares model with stock-specific intercepts and loadings. In each row following the first, we re-estimate this model excluding factors corresponding to the attributes or tilts listed at left, and report the reduction in R-squared that results (i.e. the marginal R-squared) as a percentage of the full model R-squared.

Full R-squared	3.96
	Percent of Full R-squared
Account Attributes	57.15
Size	15.31
One Stock Accounts	9.89
Turnover	4.36
Age	1.49
No. Stocks	0.35
Geographic factors	0.26
Characteristic Tilts	36.23
Business group	13.66
Industry factors	2.11
Behavioral factors	1.14
Fama-French factors	0.89
Return factors	0.82

Table 8
Characteristic Return and Tilt Variances and Covariances

This table presents regressions of characteristic return factor variances (columns 1 through 3) and covariances (columns 4 through 6) on characteristic tilt variances and covariances, or the contribution to these characteristic tilt (co)variances due to stock holdings (diagonal) and coholdings (off-diagonal). All variables are constructed using orthogonalized characteristics and equally weighting investors, and excludes the business group, industry and geographical zone characteristics. Regressions use the August 2011 cross-section, with return factor (co)variances constructed using weekly returns from March 2002 to August 2011.

Dep. Var: Return Factor (Co)Variance	Variances			Covariances		
	(1)	(2)	(3)	(4)	(5)	(6)
Holding Factor (Co)variance	0.003			0.0001		
	(0.004)			(0.002)		
Of which, coholding contribution		0.028	0.120***		0.038***	0.063***
		(0.018)	(0.038)		(0.008)	(0.006)
Of which, holding contribution			-0.020***			-0.010***
			(0.008)			(0.001)
Constant	Y	Y	Y	Y	Y	Y
Adj. R-squared	-0.047	0.133	0.499	-0.023	0.359	0.685
N	10	10	10	45	45	45

* p < 0.10, ** p< 0.05, *** p < 0.01

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%).

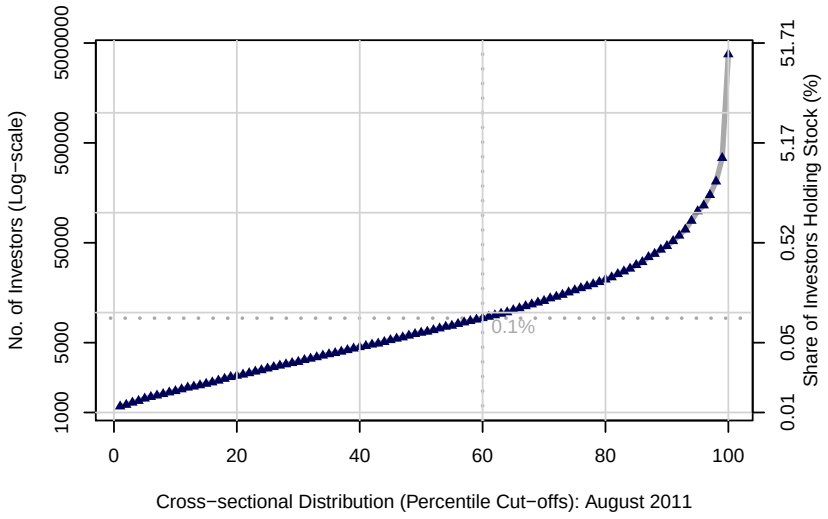


Figure 2
Average Portfolio Share

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).

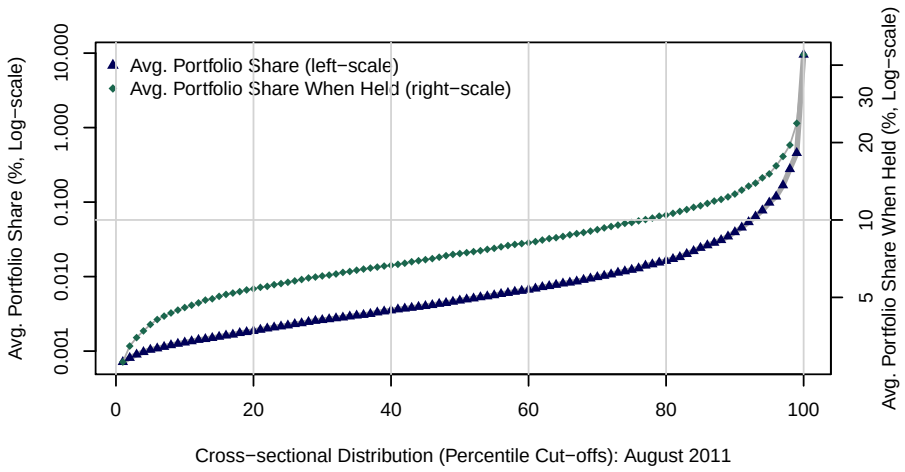


Figure 3
Stock Characteristic Clusters
Principal Component Analysis

This figure presents the loadings of investors' characteristic tilts (using orthogonalized characteristics) on the first three principal components extracted from the 9.7M x 10 matrix of characteristic tilts (excluding business group, industry and geographical zones). Characteristics appear along the horizontal axis in decreasing order of their clientele strength. All statistics are constructed by equally weighting investors.

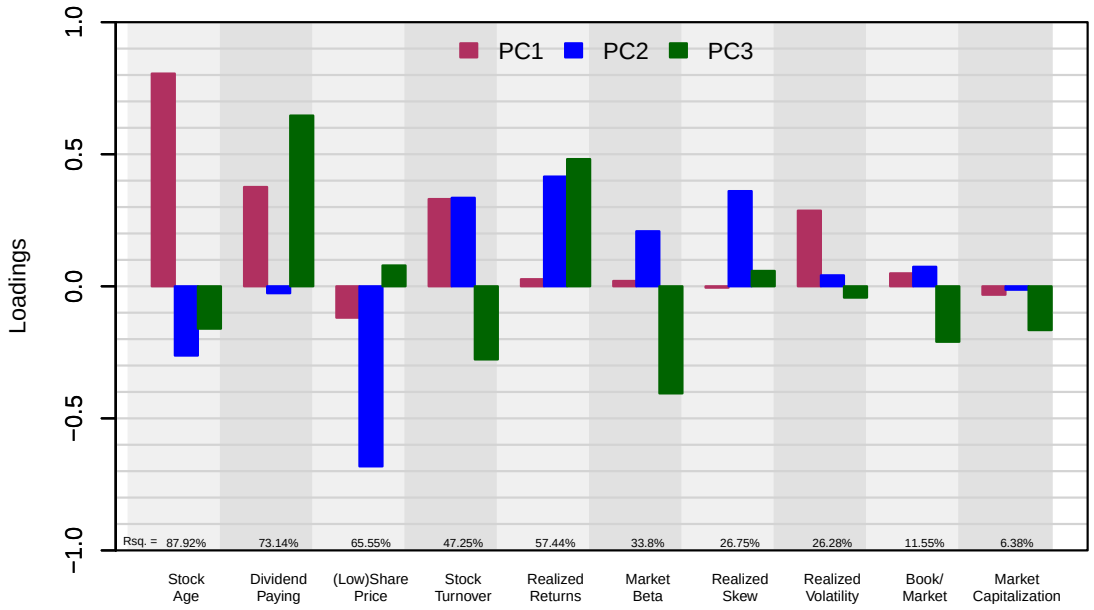
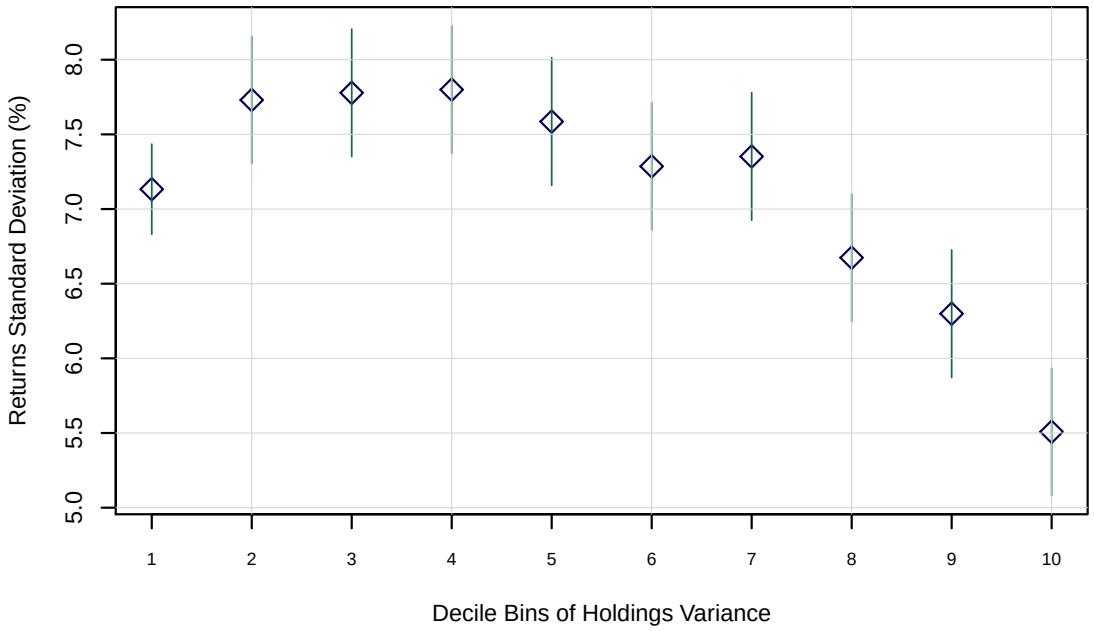


Figure 4
Stock Returns and Holdings (Co)Variances

This figure presents the relationship between stock return volatility and decile bins of holdings variance for 3103 stocks (Panel A) and between stock return correlation and decile bins of holdings covariances for all stock pairs (Panel B). Standard errors are presented vertical lines for each bin.

Panel A: Stock Return Volatility and Holdings Variance



Panel B: Stock Return Correlation and Holdings Covariance

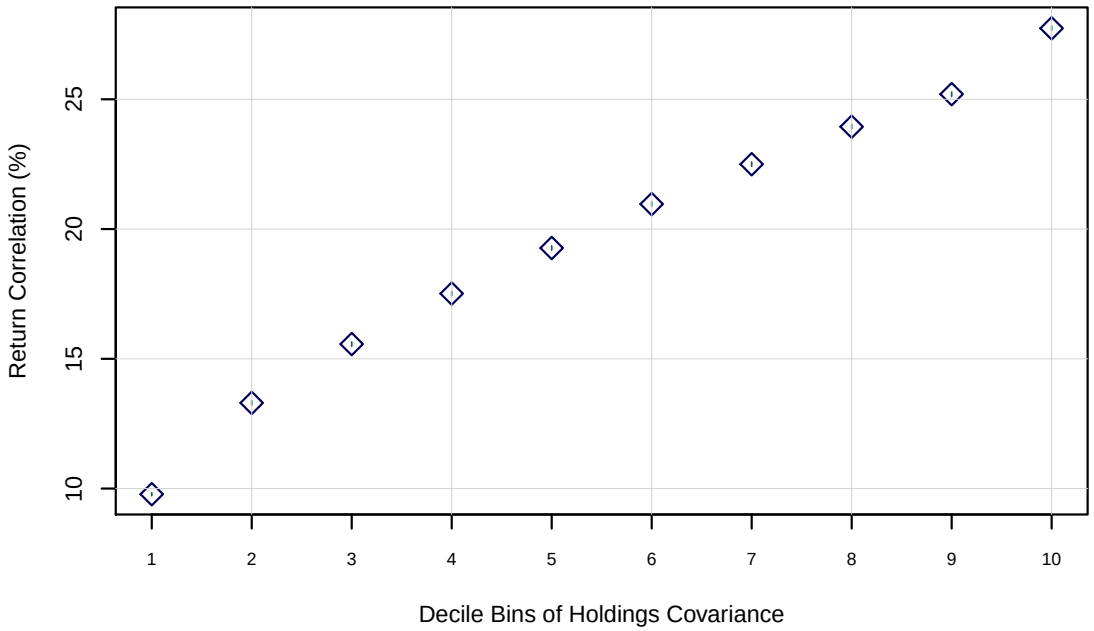
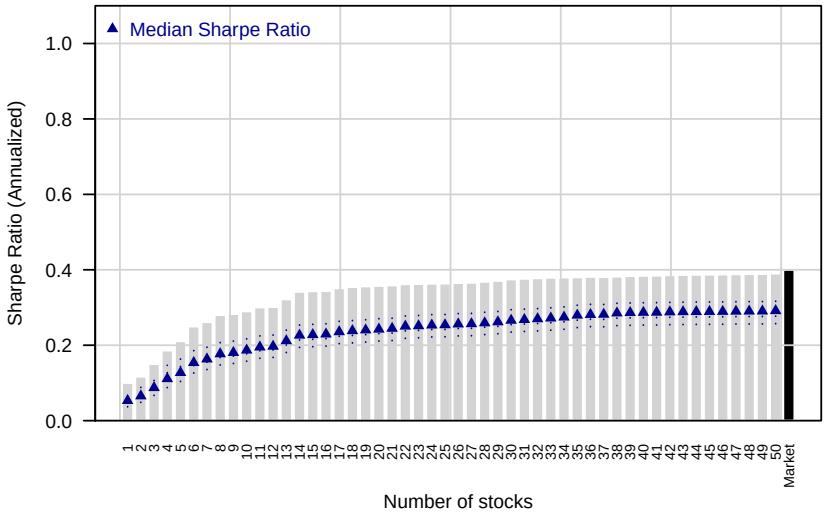


Figure 5
CAPM and Four Factor-Implied Sharpe Ratios

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