Who Owns What?
A Factor Model for Direct Stockholding

Vimal Balasubramaniam John Y. Campbell
Tarun Ramadorai Benjamin Ranish*

November 5, 2020

Abstract

We build a cross-sectional factor model for investors' direct stockholdings, by analogy with standard time-series factor models for stock returns. We estimate the model using data from almost 10 million accounts in the Indian stock market. We find that account characteristics such as the size of the equity portfolio and the number of stocks it contains are important predictors of individual stockholdings. We use our new factor model to analyze the equity coholdings matrix, which measures the degree to which stocks are jointly held in investor portfolios, and to measure the strength of clienteles for stocks with different characteristics. We show that coheld stocks tend to have higher return correlations, and that characteristics with stronger holdings clienteles tend to have more volatile factor returns.

*Balasubramaniam: Queen Mary University of London, and CEPR. Email: v.balasubramaniam@qmul.ac.uk. Campbell: Department of Economics, Littauer Center, Harvard University, Cambridge MA 02138, USA, and NBER. Email: john_campbell@harvard.edu. Ramadorai: Imperial College, London SW7 2AZ, UK, and CEPR. Email: t.ramadorai@imperial.ac.uk. Ranish: Board of Governors of the Federal Reserve System. Email: ben.ranish@frb.gov.

1 Introduction

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

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

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

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

from the time-series factor literature.

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

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

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

As a first step towards a better description of household portfolio choice, we measure the importance of different stock characteristics by the strength of the investor clienteles associated with them. To measure clientele strength, we calculate the holding variance of stock portfolios that load linearly on these characteristics, and then use the contribution of the off-diagonal elements to this holding variance. This procedure 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 the return context, stocks have

broadly similar variances so the diagonal elements diversify away and make a relatively small contribution to total portfolio variance; but this is less true in the holdings context which leads us to measure clientele strength using off-diagonal elements. Among the characteristics we consider, stock price has the strongest investor clientele but stock age, realized volatility, and market capitalization also have strong clienteles. Of course some of these characteristics are cross-sectionally correlated with one another, but we find broadly similar results when we orthogonalize the characteristics across stocks.

Next we include both account and stockholding characteristics in our empirical multi-factor model of household portfolio choice. Despite the existence of investor clienteles, we find that most of the explanatory power in our model comes from account characteristics—particularly the number of stocks held, the market value of the account, and the age of the account—rather than the characteristics of stockholdings. This implies that clienteles of investors who favor particular stock characteristics tend to share certain account characteristics. It is analogous to the finding that political contributions of households can be predicted from their residential addresses, even without knowing the history of their past political contributions. More specifically, we find that high-priced, large-cap, and low-volatility stocks tend to be held by large, well diversified, older accounts.

Among the factors based on stockholding characteristics, the most important factors capture preferences for particular business groups or industries, or for stock characteristics such as share price, stock age, popularity, turnover, or dividend payments. The attributes of stocks that are emphasized in time-series factor models such as the Fama-French (1993) model are relatively less important.

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

stock characteristics for which we estimate strong clienteles also tend to have high return volatilities.

Related literature

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

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

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

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

portfolio demands.

Organization of the paper

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

2 A Factor Model for Stockholding

In this section, we first introduce the concept of the "coholdings matrix" which captures households' combinations of various stocks in their portfolios. We then describe our new approach to estimating and characterizing the sample coholdings matrix.

2.1 The Coholdings Matrix

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

the time dimension into a single period.[1] This eliminates the need for time subscripts in our notation.

We first define a holdings vector Q_h (denoting "quantum" or "quantity") for household h. Q_h is a column vector with N elements Q_{ih}, one for each stock. In our empirical application, we measure holdings as a dummy variable, setting $Q_{ih} = 1$ if household h holds stock i, and $Q_{ih} = 0$ otherwise.[2]

For a given household h, consider the $N \times N$ outer product matrix:

$$\Omega_{Qh} = Q_h Q_h'. \tag{1}$$

We call this the "coholdings matrix for household h". Given that each Q_{ih} is a holdings dummy, the diagonal elements of Ω_{Qh} are 1 for each stock held by household h and 0 for each stock not held, while the off-diagonal elements are 1 for each pair of stocks held by household h.[3]

An object of particular interest is the expectation of Ω_{Qh} across households, which we simply call the "coholdings matrix."

$$\Omega_Q = \mathrm{E}[\Omega_{Qh}], \tag{2}$$

where the expectation in equation (2) is taken cross-sectionally across households.

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

[2] Alternative approaches would be to set Q_{ih} as the portfolio weight of stock i in the portfolio of household h, or as the fraction of the total market capitalization of stock i held by household h.

[3] We do not cross-sectionally demean the holdings measure for each stock, but it would be straightforward to do so by subtracting the vector $\overline{Q}$ from Q_h, where $\overline{Q} = \frac{1}{H}\sum_{h=1}^{H} Q_h$.

To estimate Ω_Q, we can calculate the "sample coholdings matrix":

$$\widehat{\Omega}_Q = \frac{1}{H}\sum_{h=1}^{H}\Omega_{Qh}. \tag{3}$$

This matrix captures the average propensity for households to hold stocks (on the diagonal) or pairs of stocks (on the off-diagonal). Equivalently, one might say that the diagonal elements measure the popularity of each stock among investors, and the off-diagonal elements measure the popularity of each pair of stocks. The matrix $\widehat{\Omega}_Q$ is a useful positive description of how households combine stocks into portfolios, and, as we show later, has multiple uses, including understanding investor clienteles in particular types of stocks.

To develop intuition about the sample coholdings matrix, we note that it is analogous to the familiar sample covariance matrix of stock returns. To construct the sample 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 sample covariance matrix of returns uses time periods where the sample coholdings matrix uses households, but otherwise the two matrices have the same structure.

This analogy gives rise to two more useful observations. First, the sample coholdings matrix must be positive semi-definite whenever $H > N$, just as the sample covariance matrix of returns must be positive semi-definite whenever $T > N$. Second, it would be straightforward to define a holdings correlation matrix following the analogy with the covariance matrix, simply dividing the elements of the sample coholdings matrix by the geometric average of the corresponding diagonal elements. The diagonal elements of the holdings correlation matrix will be 1 and the off-diagonal elements will be between -1 and 1.

2.2 An Empirical Model of the Coholdings Matrix

How can we empirically characterize the coholdings matrix? A naïve approach would be to regress the off-diagonal elements of the sample coholdings matrix onto variables that characterize the similarity of stock pairs in different dimensions such as firm size, age, and so forth. This would be the equivalent of regressing the sample covariances of stock pairs onto measures of the similarity of those stock pairs. An important drawback of this approach is that the fitted values from such a regression do not necessarily imply a positive semi-definite matrix.

The empirical literature on the covariance matrix of returns offers a useful alternative approach. In that literature (e.g., Fama and French 1992, 1993), it is standard to estimate a time-series regression for each stock i in which the stock's return at time t is linear in a set of K factor realizations at time t. This gives the covariance matrix of returns a special low-dimensional structure which is guaranteed to be positive semi-definite.

In studying the coholdings matrix, the equivalent procedure is to estimate, for each stock i, a cross-sectional regression:

$$Q_{ih} = \alpha_i + \sum_{k=1}^{K} \beta_{ik} F_{kh} + \varepsilon_{ih}, \qquad h = 1, ..., H, \tag{4}$$

where β_{ik} is the loading of stock i on factor k, and F_{kh} is the factor realization for household k. In the special case where Q_{ih} has been cross-sectionally demeaned and the factors also have zero cross-sectional means, then by construction $\alpha_i = 0$.

In equation (4), the factors could be attributes of the household, such as account size or account age, which are not affected by the composition of the household's portfolio. Going back to the analogy with factor models with stock returns, these are like time-series factors that are estimated without reliance on the behavior of other stocks, such as shocks to inflation or industrial production. However, the factors could also be attributes of household portfolios, like the average size of other stocks held or the average book-to-

market ratio of other stocks held. This is analogous to using the contemporaneous returns on other stocks to create factors such as HML and SMB in the usual Fama-French-style time-series analysis.[4]

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 tend to hold stock i. We estimate these coefficients freely, stock by stock, but also report weighted averages of the coefficients using important stock characteristics as weights. This enables us to measure the determinants of clienteles not only for individual stocks, but also for stock characteristics.[5]

Turning back to the coholdings matrix, in equation (4), if the assumptions needed for $\alpha_i = 0$ are satisfied, and if in addition the factors are orthogonal to one another, and enough factors are included to make the error terms ε_{ih} uncorrelated across households h for all stocks i, then the diagonal elements of the coholdings matrix Ω_Q take the form:

$$\Omega_{Q,ii} = \sum_{k=1}^{K} \beta_{ik}^2 \sigma_k^2 + \sigma_i^2, \tag{5}$$

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 average coholdings matrix take the form:

$$\Omega_{Q,ij} = \sum_{k=1}^{K} \beta_{ik} \beta_{jk} \sigma_k^2, \tag{6}$$

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

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

[5]An alternative procedure would be to restrict the betas of individual stocks on account and stockholding characteristics to be linear functions of stocks' characteristics, and to estimate the restricted version of equation (4) as a panel regression. We do not pursue this alternative here.

of coholdings.

These properties of the model follow from the linearity of equation (4). A disadvantage of (4) 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 (5) and (6).

3 Data

3.1 Data on Indian Equity Ownership

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

The older and larger of the two depositories, National Securities Depository Limited (NSDL), accounts for 64% of the roughly 9.7 million individual accounts we study in August 2011, with the remainder held at Central Depository Services Limited (CDSL). These two depositories together record almost all trading in and holdings of Indian equity at the account-issue level at a monthly frequency.[6]

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

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

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

Bombay Stock Exchange (BSE) and the National Stock Exchange (NSE) highlight that virtually all stock transactions take place in dematerialized form.

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

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

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

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

the portfolios of institutions or government entities (which we also observe). We also exclude non-public equities, which the typical household may have difficulty acquiring. Furthermore, since there is no requirement in India that publicly listed equities should have a large investor base, we remove de-facto private equities. We define these as stocks in the bottom 25th percentile ranked by the number of shareholders invested at the end of the previous month. This cutoff corresponds to removing equities with fewer than 1,177 investors at the end of July 2011 from the August 2011 cross-section of stocks that we study. After applying these filters, our final sample comprises 3,103 Indian equities and the portfolios of 9.7 million individual accounts that hold at least one of these stocks at the end of August 2011.

3.2 Data on Indian Stock Characteristics

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

We handle missing stock characteristics as follows. For stocks missing an industry assignment, we assign values to their industry dummies equal to the fraction of stocks in the given industry. For other missing (continuous) characteristics, we use all available characteristics in a regression to impute values for any characteristics that are missing.[12] This has little impact on our results as our use of rank-normalized characteristics limits

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

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

the influence of any measurement errors, and characteristics are missing for relatively few stock holdings.[13]

3.3 Summary Statistics

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

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

As documented in Campbell et al. (2019), the account size distribution is dispersed and right-skewed, with a median account size of US$ 780, and a mean account size of over US$ 11,000, close to the 90th percentile value of US$ 13,000. This distribution of account sizes is similar to the United States when accounting for the differences in per-capita GDP between the two countries, as we show in online appendix Figure A.2. Jayaraj and Subramanian (2008) show that the median (wealthiest) deciles of Indian households had average total asset values of about $3,000 ($35,000) in 2008, meaning that the stock

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

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

portfolios we study represent a non-trivial share of wealth for many of the investors in the data.

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

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

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

[15]Details are given in online appendix Figure A.1. We classify states as follows: west India includes states along the western border from Goa to Rajasthan, north India includes states along the northern border or north of Rajasthan plus Madhya Pradesh and Chhattisgarh, east India includes Orissa, Jharkhand and all states further east, and south India includes Karnataka, Andhra Pradesh, Kerala and Tamil Nadu.

skewness; and popularity among retail investors (the average value of Q_{ih} across h).[16]

The table reveals that the median stockholding of retail investors is at the 91st percentile of the size distribution, and is mechanically far more popular than the stock at the median popularity ranking. The other characteristics of the median stockholdings are in line with this tilt towards large, popular stocks, since larger stocks in our sample tend to have higher share prices, lower book-market ratios, and lower past realized returns, volatility, and skewness. However stockholding characteristics vary significantly across accounts, with a standard deviation of close to 0.2 for most characteristics (the support is -0.5 to 0.5). We explore the volatility of these factors in greater detail in section 5.

Details on other stockholding characteristics are reported in the online appendix. Figure A.1 in the appendix shows the average distribution of stockholdings across seven industries and business groups. Manufacturing and oil and gas are the two most common industry groups held, represented in roughly 40% and 19% respectively of all retail stockholdings. Business groups—sets of independently listed companies with a large ownership stake and common control by a single underlying entity—are quite common in developing countries (e.g., Anagol and Pareek 2019), and in our data, 886 of the 3,103 stocks are affiliated with 266 business groups. In the average account, the top 10 business groups account for about 30% of stockholdings, with remaining business groups accounting for a further 20% of stockholdings. We account for both industry group holdings and business group affiliations in our set of factors used to explain stockholdings and coholdings.

Figure A.3 in the online appendix shows correlations between account and stockholding characteristics. Within the set of account characteristics, the most correlated are size, the number of stocks held, and the number of stocks traded, but all correlations are below 0.6. The largest correlations are within the set of stockholding characteristics. Size (market capitalization) is strongly positively correlated with share price (0.81) and popularity

[16]Turnover and return-based stock characteristics are computed over 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.

(0.77), and negatively correlated with book to market (−0.61). Share price is significantly negatively correlated with realized volatility (−0.57). Realized returns and skewness are positively correlated (0.59). Other correlations between stockholding characteristics are less than 0.5 in magnitude. Correlations between account and stockholding characteristics are generally smaller. This suggests that an empirical model including both types of factors will be well-behaved.

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

4 Diversification of Stockholdings

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

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

4.1 Constrained Mean-Variance Optimization under the CAPM

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

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

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

problem. For $N_h = 1$ we simply choose the stock which is maximally correlated with the market.

Panel A of Figure 2 plots the results from this exercise for N_h ranging from 1 to 50. The height of each grey bar in panel A indicates the maximum obtainable Sharpe ratio associated with each value of N_h on the horizontal axis using the LASSO implied portfolio of stocks.[18] This maximum Sharpe ratio doubles as N_h increases from 1 to 5, with more modestly rising values up to around $N_h = 25$ and 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 next consider investors' performance under a popular four-factor model of returns.

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

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

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

Panel B of Figure 2 plots the results from the four-factor exercise in a similar form to the CAPM exercise in Panel A. The unconstrained optimal Sharpe ratio is over twice as high as under the CAPM. However, the constrained optimal Sharpe ratio is only about 90% of the unconstrained optimal Sharpe ratio at $N_h = 50$, as the four-factor tangency portfolio applies negative weights to some stocks. In contrast to the optimal portfolios, households' Sharpe ratios are only modestly higher under the four-factor model and therefore are always even further below the constrained optimal level. This results because few portfolios lean heavily towards factors that have been well compensated historically, aside from the market factor—which accounts for less than half of the tangency portfolio.

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

[19] We use Agarwalla, Jacob, and Varma (2013) data 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.

5 Investor Clienteles

If investors have heterogeneous preferences over particular stock characteristics, that is if they form stock characteristic clienteles, then these preferences should be reflected in the variance of the long-short portfolio holding $\omega' \Omega_Q \omega$, where ω is a vector of stock-level weights constructed from the demeaned ranks for a given characteristic, and Ω_Q is the coholdings matrix. This variance measures the dispersion of a characteristic across investors' portfolios. The analogy in the traditional time-series analysis of returns is the argument that a pervasive characteristic represents a potentially important risk if a long-short portfolio formed by sorting stocks on this characteristic has a relatively high return variance (Kozak, Nagel, and Santosh 2018).

The total variance of portfolios' characteristic factors consists of two components. The contribution from the diagonal elements of Ω_Q measures the popularity of extreme values of the characteristic, whereas the contribution from the off-diagonal elements of Ω_Q measures whether stocks with similar ranks on the characteristic tend to be held together. Our analysis differs from time-series analysis of returns in that some stocks are far more widely held than others, and also in that most investors hold relatively few stocks. The first fact implies that there are some extremely large diagonal elements of Ω_Q, and the second fact implies that many off-diagonal elements of Ω_Q are quite small. To avoid an excessive influence of diagonal elements, we compute the off-diagonal portion of the variance and use it as our measure of clientele strength.

In Panel A of Table 2 we report both of these variance measures for each of our nine continuous stock characteristics in the cross-section of 9.7 million individual accounts in August 2011. Each characteristic is measured by its rank across the cross-section of stocks as in Table 1. We compute these variance measures under three different weighting schemes across the 9.7 million individual accounts in August 2011. In the first two rows, accounts are weighted equally. In the next two rows, accounts are weighted

proportionally to their total stockholding value in August 2011. In the last two rows of the panel, accounts are weighted proportionally to their turnover, giving zero weight to entirely passive accounts.

Share price emerges as the strongest characteristic clientele under equal and turnover-weighting of accounts, and is a close second when value-weighting accounts. Market capitalization clienteles appear strong as well, though market capitalization and share price are significantly correlated in the cross-section. Realized returns and volatility also appear to have significant clienteles as well—particularly when we focus on active investors by turnover-weighting accounts. Given that these characteristics are less persistent, significant portfolio turnover is required to maintain exposure to them. More generally, the significant differences in clientele strength across different account weighting schemes suggests that clientele strength is tied to account characteristics.

Across characteristics, total characteristic factor variance is much larger than our measure of clientele strength based only on the off-diagonal elements of the coholdings matrix, reflecting the wide dispersion in popularity across stocks and the undiversified nature of most accounts. The difference between the two largely reflects the distribution of characteristics of the most widely stocks. Stock age, for example, has a large factor variance because some of the most popular stocks in our data are relatively old, while others are relatively new.

Stock characteristics are correlated with one another, and this complicates interpretation of the clientele effects reported in Panel A. Investors may appear to have preferences for particular characteristics only because they happen to be correlated with other characteristics that investors actually care about. To clarify the relative importance of clienteles, we measure each characteristic sequentially, controlling for correlations with all characteristics that have stronger clienteles. We implement this by first identifying the characteristic with the strongest clientele, which is share price when equally-weighting accounts. Then, we use a kernel regression to orthogonalize all other characteristics to

share price and find the characteristic with the second strongest clientele.[20] We proceed in this manner, using multivariate kernel regressions to normalize remaining characteristics in each round with respect to all of the previously selected characteristics. This process is carried out separately under each weighting scheme for accounts.

In Panel B of Table 2 we report results for our sequentially orthogonalized stock characteristics. The results are the same as in Panel A for the strongest characteristic, but different for all the others. The order of columns in the table represents the order in which each characteristic is selected when accounts are equally weighted. In addition to the total factor variance and clientele strength, we report the order in which characteristics are selected under each weighting scheme.

Orthogonalization does not radically alter the previous pattern of results. The main effects of orthogonalization are to diminish the role of market cap, which is correlated with stock price, but the market cap factor remains moderately important. Orthogonalization also minimizes the importance of realized skewness and the book-market ratio Given the importance of the book-market ratio in academic analysis of value investing, it is striking how little role it appears to play in portfolio construction in our dataset.

6 Factor Models of Stockholdings

Our objective in this section is to model the coholdings matrix Ω_Q by applying the framework developed in Section 2. We begin by briefly describing the construction of our observed factors, then summarize stock-level coefficient estimates and the explanatory power of the different factors for the panel of all stocks. We conclude this section by contrasting the observed multifactor model results with those obtained from an unobserved (principal-component-based) factor analysis of stockholdings.

[20] Our kernel regressions use Euclidean distance, a Gaussian kernel, and are calibrated such that 10% of stocks are within twice the bandwidth parameter. Results are similar if we use a linear regression with quadratic terms including cross-products.

6.1 Observed Multifactor Model

6.1.1 Factor construction

We construct household-specific and household-portfolio-specific factors from the account and stockholding characteristics summarized in Table 1. We add four sets of factors to this set. First, we include dummy variables to capture the four geographical zones in which households are located. Second, we add industry factors which capture the share of the portfolio in each of six industry groups, namely, financial services; food agriculture and textiles; information technology; manufacturing; oil and gas; and other retail. Third, we add business group factors which capture the share of the portfolio in each of 11 business groups.[21] Finally, we add a variable measuring the fraction of each investor's portfolio that pays dividends.

As mentioned earlier, account characteristic factors do not rely on the composition of the investors' stockholdings, and are thus analogous to pure time-series factors (e.g., industrial production or changes in GDP) in the asset-pricing setting. In contrast, stockholding characteristic factors which depend on the composition of households' portfolios are analogous to return-based Fama-French-style factors. However, while there are large numbers of stocks available at each point in time which can be used to generate factors in the standard asset pricing setting, each household's portfolio is often composed of only a few different stocks, generating a sparsity problem. Unless addressed, this can generate a mechanical relationship between estimated betas and stockholding characteristic factors. To insulate ourselves from this issue, we employ a leave-out approach. In particular, when estimating betas for a given stock i, we exclude this stock from the computation of the stockholding characteristic factors employed in the regression. Moreover, our full set of 36 factors also includes a dummy to indicate single-stock accounts.[22] The stock-

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

[22] For such accounts, in the regression corresponding to the particular stock held by these accounts, by construction, all stockholding characteristic factors are undefined given our leave-out approach. We

holdings characteristic factors employed here are constructed using the orthogonalization procedure detailed in Section 5.

6.1.2 Estimation and results

We estimate stockholdings using all $K = 36$ 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 (4), and is run with 9.7 million household observations.

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

For ease of interpretation, we first divide each factor by its unconditional standard deviation in each stock-specific regression, and multiply by 100 for readability. $\tilde{\beta}_{ik}$ is then the percentage point increase in the unconditional holding probability of stock i for a one standard deviation increase in factor k.

Table 3 summarizes the $\tilde{\beta}_{ik}$ estimated from the 3,103 stock-specific estimates of equation (4). 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. By construction, the cross-stock mean $\tilde{\beta}_k$ is mechanically equal to zero (except for the coefficient on N_h) and is therefore uninteresting.[23] The first four columns of the table therefore summarize the cross-stock distribution of $\tilde{\beta}_{ik}$, presenting the cross-stock stan-

simply set these factors to a neutral value of zero in such cases, though the account characteristics continue to be well-defined. The stockholdings characteristics for these single-stock accounts are of course defined in the regressions corresponding to all the other (unowned) stocks in the universe.

[23]The time-series return analog is the fact that the mean capitalization-weighted beta on the market return is one and the mean beta on all other factors is zero. In this sense the N_h factor is analogous to the market return, but in our context the mean beta on this factor is 0.53% given our normalization of loadings.

dard deviation, and the 10th, 50th, and 90th percentiles of the cross-stock distribution of factor betas. The last two columns show the average t-statistic across all 3,103 regressions, and the percentage of estimated β's that are statistically significantly different from zero at the 10% level.

Panel A of Table 3 shows the distribution of $\tilde{\beta}_{ik}$ for the account characteristic-based factors, and Panel B summarizes the distribution of $\tilde{\beta}_{ik}$ for stockholding characteristic-based factors. The final two columns of both panels reveal that the majority of factors have high t-statistics on average, with a few exceptions such as realized skewness and some of the business group factors. 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 appears high on average, they exhibit very different levels of cross-stock variation. A necessary condition for a useful factor is that it helps to predict cross-sectional dispersion in household stockholdings. The equivalent in the standard returns setting is factors such as SMB and HML that exhibit a large cross-sectional spread in normalized factor loadings, and help to explain the time-variation in realized returns across stocks. We later discuss how specific stock characteristics are connected with account and account-stockholding characteristic-based factors, but for now, we simply discuss the magnitude of the cross-stock spread in factor loadings seen in Table 3.

Account-characteristic factors

Based on the standard deviation of $\tilde{\beta}_{ik}$, the account-characteristic factor with the single largest variation in explanatory power is N_h, the number of stocks held in the account. The loadings show that almost every stock is more likely to be held as N_h increases, though this tendency varies enormously. Mechanically, we can show that a one-standard-deviation increase in N_h increases the mean holding probability by about 0.53 percentage points. However, for the median stock this increase is 0.28 percentage points, while it is

0.09 and 1.17 percentage points at the 10th and 90th percentiles respectively.

The next most important account-characteristic factor for predicting stockholdings, based on the cross-stock standard deviation of $\tilde{\beta}_{ik}$, is account size. The percentiles of the distribution of loadings on account size show that roughly 90% of stocks become less frequently held as account size increases, holding constant all of the other factors in the model. Put differently, larger accounts, holding constant other model factors, tend to select stocks from a relatively small subset of the whole universe of Indian stocks. Account age and the number of stocks traded are also quite important factors, but account turnover is less so.

There is suggestive evidence of geography-based stock selection, which, as we show later, is mainly driven by local bias of the sort found by Coval and Moskowitz (1999) for US mutual funds. However, geographical effects are quite weak in comparison with the role of both N_h and account size, at least for the fairly broad geographical account locations that our data capture.

Stockholding-characteristic factors

Panel B of Table 3 turns to factors based on accounts' stockholding characteristics. The table divides factors into five categories, namely the Fama and French (1993) style factors capturing the size and value characteristics of household portfolios; return-based factors based on realized stock returns experienced in the portfolio; behavioral factors capturing revealed preferences through stockholdings for high or low share price, popular, old, high-turnover, or dividend-paying stocks; business group factors; and industry factors.

Once again using the factor loading cross-stock standard deviation as a guide, Panel B reveals that the share price, market cap, and age of other stockholdings are among the most useful factors for predicting whether a household will hold a given stock. However, the single most useful predictor is an indicator for the household's share of other stocks in the Reliance (ADAG) business group. This factor, while not useful for predicting the holdings of most stocks—as seen by the modest spread between the 10th and the 90th

loading percentiles—is extremely useful in predicting holdings of stocks in this business group. It is an example of a factor that is important in a large finite sample of stocks, although it is not pervasive in the sense of Ross (1976) or Connor and Korajczyk (2019).

After these factors, holdings of stocks in the oil and gas industry, stocks in the Reliance (DAG) business group, volatile stocks, and dividend-paying stocks are the next most predictive. In comparison with these characteristics, factors based on stockholdings' book-market ratio, realized skewness, and turnover have relatively weaker power to predict stockholdings.

6.1.3 Explanatory power

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

The approach that Connor and Korajczyk (2019) recommend is to first stack all $3,103$ stocks into a single pooled OLS regression. In our implementation, we regress stockholdings dummies on a set of observable factors F_k, in which there are stock-specific loadings on these factors. In Panel A of Table 4 we report R^2 statistics from such a pooled regression, in which we weight each stockholdings indicator by the inverse of the standard deviation of each stock's aggregate holding probability, i.e., $\sqrt{\hat{Q}_i(1-\hat{Q}_i)}$. This normal-

ization means that the aggregate R^2 apportions equal weight to each stock seen in the data.[24]

The first row of Panel A, Table 4 shows that the R^2 of the full multifactor model on the equally weighted pooled data stands at 1.74%. 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-characteristic factors contribute roughly 82% of the total explanatory power in the equally-weighted case, with stockholdings-characteristic factors accounting for roughly 9% of the total R^2. The two contributions do not add up to 100%, as the underlying factors are not orthogonal to one another.

Turning to the specific account-characteristic factors, Panel A of Table 4 shows that N_h does play a very important role, though it is not the only important factor. Account size and age also play a relatively important role, while geographical factors and turnover are less important. This analysis helps to bring together disparate themes in prior literature on the influence of account characteristics on stockholding propensities into a common framework. For example, account size and wealth have been highlighted as important determinants of stockholdings behavior by Campbell et al. (2019) and Bach et al. (2020), and account age by Campbell et al. (2014) and Betermeier et al. (2017).

Turning to the set of stockholding-characteristic factors, behavioral factors and busi-

[24]Panel A of Appendix Table A.1 experiments with alternative weighting schemes in the pooled regression. The "Unweighted" pooled regression puts emphasis on the model's ability to explain which accounts hold the most widely held stocks (Column 1), as these account for the bulk of the variance in the pooled stockholding data. The variance of our stockholding indicator is maximized when $E[Q_i] = 0.5$. For comparison, the most widely held stock in our sample (Reliance Power) appears in roughly 40% of accounts. The second column of Panel A, Appendix Table A.1 reports the R^2 from a different pooled regression which weights each stock by the inverse variance multiplied by the share of the market capitalization of the stock held by retail investors, given our focus on understanding household portfolio construction—this scheme equal weights stocks but emphasizes explanatory power for stocks with a high share of retail (rather than institutional) ownership.

ness groups appear to contribute the largest amount in this set of factors to the pooled unweighted R^2 despite the typically modest level and cross-sectional spread seen in the $\tilde{\beta}_{ik}$ on these factors in Table 3.

Before moving to a deeper economic understanding of "who owns what," and of the predicted coholdings matrix using the multifactor model, the next subsection applies an unobserved principal components analysis (PCA-based) factor model to the data, and discusses the complementary insights obtained from observed and unobserved multifactor models.

6.2 Unobserved Multifactor Model

We further build our understanding of household portfolio construction with an unobserved multifactor model that is based on PCA. We first compute the principal components of the 3,103 by 3,103 covariance matrix of stockholdings derived from the 9.7 million accounts that we observe. For comparability with the observed factor approach, we normalize the stockholdings data by the inverse standard deviation of the aggregate holding probability of each stock before computing the covariance matrix of stockholdings. This approach prevents the principal components from explaining few popular stocks such as Reliance Power and ascribes equal importance to all stocks.

The first principal component is the eigenvector of this covariance matrix which corresponds to the largest eigenvalue, and subsequent principal components are estimated as the eigenvectors associated with successively smaller eigenvalues of the covariance matrix. By construction, these principal components are orthogonal to one another, and are normalized linear combinations of household stockholdings that together summarize the total variance of stockholdings. They are ordered by the fraction of the total variance that they capture.

6.2.1 Statistical significance of factors

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

Ahn and Horenstein (2013) suggest a procedure to detect the number of natural rate factors in the data. They show that if there are r natural rate factors, the ratio of successive eigenvalues should peak when comparing the eigenvalues of the rth and $r+1$th factors. Online appendix Figure A.5 shows this ratio of eigenvalues for the PCA unobserved factor model, using the Ahn and Horenstein (2013) recommended parameters applied to our empirical setting. This ratio rapidly declines following the first principal component, meaning that the eigenvalue ratio test suggests there is a single unobserved factor.

Panel B of Table 4 shows that a 10-factor PCA model explains 3.7% of the variance of stockholdings, 42% of which is accounted for by the first factor, 15% by the second factor, and 9% by the third factor. This pattern is further illustrated in online appendix Figure A.4. The implication is that a 1-factor PCA model has less explanatory power than our observed multifactor model, while a 2-factor PCA model has slightly greater explanatory power than our model.

The eigenvalue ratio test can have difficulty when data have a handful of dominant, and many smaller natural rate factors, so we also consider an alternative test suggested by Connor and Korajczyk (2019). The marginal explanatory power of factors, measured by the sum of squared β or by marginal R^2 statistics should differ between natural rate and other factors. While the sum of squared βs increases at rate n for natural rate

factors, Connor and Korajczyk's test is based on the assumption that it increases no faster than $n^{1-2\delta}$ for all other factors. With this assumption, the authors derive a test statistic and threshold for natural rate factors based on the marginal R^2 statistic. The test is conservative in the sense that it does not count natural rate factors with inherent explanatory power below this threshold.

In the online appendix we implement Connor and Korajczyk's test for two values of δ, 0.2 as suggested by Connor and Korajczyk (2019) and 0.1. We find that 11 of our 36 observed factors exceed the threshold when $\delta = 0.2$, and 5 exceed it when $\delta = 0.1$. Among the PCA-based factors, 11 exceed the threshold when $\delta = 0.2$ but only the first factor exceeds it when $\delta = 0.1$.

6.3 Comparing Observed and Unobserved Multifactor Models

In this subsection, we attempt to glean insights about the nature of coholdings obtained from both observed and unobserved factor models. Given the discussion in the previous subsection, we focus on a PCA with 10 unobserved factors, which we discuss alongside our full observed factor model.

Figure 3 relates the first 10 principal components to the observed factors in our multifactor model. Each principal component is regressed on the full set of 36 observed factors, and the regression coefficients are represented as a heatmap, in which negative coefficients are in shades of blue, those close to zero are in white, and positive coefficients are in shades of red. Darker shades indicate larger absolute magnitudes.

Panel A of Figure 3 shows that the first principal component of stockholdings is strongly (negatively) related to the number of stocks held and traded, and to a lesser extent, to account size and account age. As conjectured when discussing the cross-sectional dispersion of $\tilde{\beta}_{ik}$, this appears consistent with the stockholdings of large, well-diversified, and established accounts being systematically different from the holdings of small, undiversified, new accounts. Panels B and C show that the first principal component also

has some relationship to the characteristics of household stockholdings including their average market cap and age (panel B), the share of the portfolio held in the Reliance (ADAG) business group (Panel C), and the share of the portfolio invested in the oil and gas industry (Panel D).

The second principal component's loadings on account characteristics are almost perfectly negatively correlated with those of the first component, but this principal component differs in some of its loadings on stockholding characteristics. While the first principal component identifies small undiversified investors who hold large stocks with low volatility and high share price (Panel B), the second component identifies large diversified investors who do the same. The second principal component also appears to load mildly positively on the Tata business group, and mildly negatively on the Reliance ADAG business group. The higher principal components load on a mix of factors across the three panels, with relationships with our observed factors that generally grow weaker as we progress from 3-10.

Table 3 suggests that the observed factors in our model do a good job of capturing the most important drivers of stockholdings, but it also highlights the difficulty of capturing the wide variety of unobserved drivers of stockholdings, many of which likely pertain to the motivations of more obscure clienteles. To explore these relationships further, Figure 4 compares the R^2 statistics of the observed multifactor model with those associated with the unobserved principal-components-based factor model for each stock.

Panel A of the figure examines the comparison between the observed model R^2 (y-axis) against a simple unobserved model based only on the first principal component. The observed multifactor model performs better for the data points represented in light blue, and the darker blue shade shows the stocks for which the first principal component outperforms. The data points in red show the ten most widely-held stocks in the data, which with one principal component, the PCA approach fails to fit well.

Panel B of the figure adds in principal components 2-10 into the unobserved factor

model. Though the higher principal components load on a mix of observed factors, the relationships generally are weaker, and this is reflected in the fact that the PCA approach with ten factors explains many stock holdings better—particularly the holdings of less popular stocks which are harder for either model to explain. Overall, Figure 4 shows that there are complementarities between the observed and unobserved factor approaches, with the benefit of the observed factor model being economically interpretable.[25]

7 Insights from Multifactor Models

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

7.1 Empirical and Model-Implied Coholdings

The elements of the model-predicted coholdings matrix are given by equation (6). To facilitate interpretation, especially as estimated coholdings vary substantially in our empirical setting, we simply convert the predicted and actual coholdings matrices into coholdings correlation matrices by dividing the elements of the sample coholdings matrix and the model predicted coholdings matrix by the geometric average of their corresponding diagonal elements. As highlighted earlier, the diagonal elements of the (actual and predicted)

[25]Online appendix Table A.1 contrasts the overall explanatory power of the observed factor model and the unobserved PCA-based model, using the pooled R^2 approach introduced earlier. When more weight is put on widely held stocks that account for more of the observed variation in stockholdings, the observed factor model ($R^2 = 7.2\%$) has a better fit than the first principal component alone ($R^2 = 2.6\%$), though the 10 principal component model fits better overall ($R^2 = 9\%$).

holdings correlation matrices then equal 1, and the off-diagonal elements range between -1 and 1.

Figure 5 plots coholdings correlations estimated in the data on the vertical axis against their model-implied counterparts on the horizontal axis. Panel A uses our observed multifactor model, while Panel B employs the PCA-based unobserved factor model.

Panel A shows that the observed factor model performs well, since model-implied coholdings correlations are a seemingly unbiased estimate of empirical coholdings correlations. The highest density of observations is on or close to the 45-degree line, and over- and under-predictions are fairly evenly distributed. Virtually all coholdings are positive, driven by positive and significant βs on the number of stocks held by the investor, and most fall in the range of 0.01 to 0.03. Panel B of Figure 5 shows how empirical coholdings correlations stack up against their PCA 1-10 model-based counterparts. In this case, the bulk of estimated coholdings fall on the 45° line, with a tighter fit than for the observed factor model.

7.2 Characteristic Clienteles and Factor Loadings

We first investigate the extent to which the stock characteristic clienteles we identified in Section 5 are associated with particular account characteristics of investors. To do this, we estimate our observed factor model using only the account characteristic factors. Then, we contruct the weighted average $\tilde{\beta}_k$ for each account characteristic, using our various stock characteristics' demeaned ranks as weights. This tells us which types of accounts make up the clientele for each characteristic.

Table 5 shows the results of this exercise, with each column representing results for a different account characteristic $\tilde{\beta}_k$. In Panel A, these stock characteristics are our de-meaned rank characteristics (as in Table 1), with positive and negative weights scaled to sum to +1 and -1 respectively. Panel B performs the same calculation using stock characteristics that have been orthogonalized to characteristics with stronger clienteles

by the procedure applied in Section 5 (while equal-weighting accounts). Results can be interpreted as a weighted-average difference in $\tilde{\beta}_k$ between stocks ranked above and below average in the given characteristic. Colors in the table reflect the sign and magnitude of the results.

The largest clientele effects in Panel A are associated with the number of stocks held in the account. This should not be too surprising as Table 3 showed that this characteristic has by far the most volatile $\tilde{\beta}_k$. A one-standard-deviation increase in the number of stocks held increases the probability of holding large, high-priced stocks with low book-market ratios and low return volatility. It also very strongly increases the probability of holding the most popular stocks.

An increase in account size—all other characteristics held constant—is also associated with an increase in holdings of popular, large, high-priced stocks with low book-market ratios and low return volatility. Thus account size and the number of stocks held are two account characteristics that tend to put investors into similar clienteles. However other characteristics play a different role. An increase in the number of stocks traded is associated with an increased tendency to hold high-turnover stocks with low realized returns and skewness and high realized volatility. An increase in account age is associated with a modest increase in holdings of low turnover, old, small stocks, with low prices and high realized returns. Patterns are qualitatively similar, although somewhat weaker, when we use orthogonalized stock characteristics in Panel B.

Table 6 provides results analogous to Table 5 using our full 36-factor model. The account characteristic effects are generally quite similar between the two tables, though slightly smaller in the full model. In other words, account characteristics retain most of their ability to predict stock clienteles even when accounting for the characteristics of the other stocks held in the portfolio.

Table 7 provides a similar analysis of the stockholding characteristic betas from our full model. Recall that the stockholding characteristic factors used in the estimation of

these betas exclude stock i in factor construction, meaning that there is no mechanical relationship between the stocks' characteristics and their stockholding characteristic betas. Therefore, positive coefficients along the diagonal are evidence of residual clientele effects that have not been fully explained by measured account characteristics. In contrast, negative coefficients suggest that investors seek to diversify the characteristic in question (e.g., they are more inclined to add a large stock to a portfolio that is otherwise heavy on smaller stocks).

Across the stock characteristics in Table 7, the strongest residual clienteles are for share price, stock age and market capitalization. We also see some interesting off-diagonal terms capturing a tendency for investors to select stocks on the basis of several related characteristics. The average share price rank of other stocks held, for example, is a relatively good predictor of holding popular and large-cap stocks, and stocks with a low book-market ratio and low volatility. When characteristics are orthogonalized, the off-diagonal terms generally shrink, but the patterns remain qualitatively similar.

Overall, the results in this section suggest that certain investors seek "stable" or "high-quality" stocks, defined by some combination of high share prices, high market capitalization, high popularity, low book-market ratios, and low volatility. Larger and better diversified accounts are particularly likely to have this tendency. .

7.3 Coholdings and Return Covariances

Figure 6 plots the relationship between return correlations and measures of coholdings. The return correlation estimates are based on weekly Indian stock returns data for the year leading up to August 2011, when we estimate coholdings. The plots use a subsample of observations, which are sampled from the joint distribution of return and coholdings correlations. The empirical density of both return correlations (to the right of each plot) and coholdings correlations (above each plot) are also shown in the figure.

Panel A of the figure estimates the relationship between estimated return correlations

on the vertical axis and "raw" estimated coholdings correlations on the horizontal axis, while Panel B of the figure replaces raw coholdings correlations with observed-factor model-implied coholdings correlations.

The figure shows a clear positive relationship between return correlations and coholdings probabilities for stocks. The R^2 from a linear regression of return correlations on raw coholdings correlations is about 10% for the empirical coholdings in Panel A, implying that the correlation between these two correlations is roughly $\sqrt{0.1}$ or 32%. This rises to $R^2 = 19\%$, or a correlation of roughly 44%, when return correlations are regressed on model-implied coholdings correlations in Panel B.

Complementing our clientele strength exercise in Section 5, we assess stock characteristics' importance in the return dimension by constructing long-short portfolios for each characteristic, and then measuring the variance of the portfolio returns over time. Here, we construct long-short portfolios as $\omega' \Omega_r \omega$ where ω are the stock-level weights constructed as the demeaned ranks for a given stock-characteristic, and Ω_r is the covariance matrix of stock returns. We estimate the covariance matrix using weekly returns over the period April 2002 through August 2011. The portfolio weights in each month are given by the demeaned characteristic ranks at the end of the previous month, scaled so that long- and short-side weights each sum to one. As in Table 2, we consider both the total factor variance as well as the contribution of off-diagonal terms separately.[26]

The top rows of Table 8 present results using non-orthogonalized characteristics, while the bottom rows use the orthogonalized characteristics from Table 2. As with the holding factors, the return factors constructed using share price are the strongest over our sample period. Return factors constructed based on stock size, realized volatility, realized returns and popularity are also relatively strong. Patterns are similar with orthogonalized characteristics, although lower-ordered characteristics naturally show less volatility under

[26]The distinction is less important in this context, as returns are relatively less volatile than holdings Q_{ih}. Due to the changing number of stocks and changing characteristics over time, we measure the variance contribution of diagonal terms as $\Sigma_t \Sigma_i w_{i,t-1}^2 (R_{it} - \bar{R}_i)^2 / T$, where w_k represents weights given by stock characteristic k.

this approach.

Figure 7 compares the strength of these return factors with the holdings factors analyzed in Table 2. Across stock characteristics, there is a strong positive correlation between the volatility of a characteristic's returns and the volatility of its holdings. The notable exception is stock age, which has a strong holdings clientele effect but relatively low volatility of returns. This figure extends the previous result of Figure 6, that coheld stocks tend to move together, to show that characteristics with strong holdings clienteles also have a strong tendency for return comovement.

While these are preliminary observations, the clear positive relationship between coholdings correlations and return correlations is intriguing. If Indian investors were attempting to diversify portfolios with a small number of stocks, they would tend to cohold stocks with relatively low return correlations. On the other hand, if investor clienteles buy and sell 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.

8 Conclusion

In this paper we have suggested that a factor model for investors' stockholdings provides a natural way to understand household portfolio decisions and the structure of investor clienteles for different types of stocks. The model is a cross-sectional 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 stockholdings are the norm at the time of our analysis.

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

Certain characteristics of stocks seem to have strong clientele effects associated with them, meaning that many investors' portfolios load either positively or negatively on these characteristics. The price of a stock has the strongest clientele effect, but market capitalization, stock age, realized volatility, and realized returns are also important. We find a relatively weak clientele effect for the book-market ratio, despite the prominence of this characteristic in academic finance.

We use our model to estimate which types of accounts hold which stocks and make up the clienteles for these characteristics. The most important factor in the model is the number of stocks held, implying that concentrated and diverse portfolios hold different types of stocks. Other account characteristics such as account size and age (time in the market) are also important. By including all these account characteristics 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] We find that account characteristics help to explain clienteles: specifically, larger and better diversified accounts have a tendency to hold stocks with high share prices, high market capitalization, high popularity, low book-market ratios, and low volatility.

The characteristics of other stocks held in an account are somewhat less important in our factor model. Despite the existence of characteristic clienteles, the membership

[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).

of these clienteles is often well predicted by account characteristics. However, we do find that stockholding-characteristic factors matter when they are constructed from the characteristics that have the strongest clienteles.

Finally, we present two preliminary findings on the relation between coholdings and comovement of stock returns. Stocks that are more commonly coheld tend to correlate more strongly with one another, and stock characteristics with stronger clienteles have more volatile long-short portfolio returns. These patterns run counter to the view that investors optimally diversify their portfolios conditional on a constraint on the number of stocks held, but they reinforce the idea that clientele effects, captured by coholdings propensities, contribute to common variation in stock returns.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Fama, E. F. and K. R. French (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.

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

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

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

Table 1
Summary Statistics

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

Variable Name	Mean	Std. Dev.	P25	Median	P75	P90
Account Characteristics						
Age	61.30	36.89	39.00	52.00	84.00	124.00
Size ('000s USD)	11.54	533.43	0.14	0.78	3.54	13.01
Turnover	0.38	1.17	0.00	0.02	0.18	0.71
No. Stocks	8.45	16.48	1.00	4.00	9.00	20.00
No. Stocks Traded	4.74	11.24	0.00	1.00	5.00	13.00
Stockholding Characteristics						
Share Price	0.16	0.20	0.06	0.21	0.30	0.38
Stock Age	-0.09	0.23	-0.27	-0.09	0.05	0.20
Realized Volatility	-0.13	0.17	-0.24	-0.16	-0.04	0.09
Market Capitalization	0.34	0.18	0.28	0.41	0.47	0.48
Realized Returns	-0.07	0.18	-0.19	-0.07	0.03	0.14
Turnover	0.09	0.17	0.01	0.09	0.20	0.31
Popularity	0.41	0.12	0.38	0.45	0.49	0.50
Book/Market	-0.10	0.17	-0.20	-0.12	-0.02	0.12
Realized Skewness	-0.16	0.17	-0.28	-0.17	-0.08	0.04

Table 2
Clientele Strength

Each column of this table presents the variance of a given stock characteristic factor (mean stock holding characteristic) across accounts in August 2011. Clientele strength represents the portion of factor variance contributed by accounts' coholdings (stock holding covariance terms), as most accounts hold few stocks. Variance terms are computed alternately equally, value- and turnover-weighting accounts. Panel A uses de-meaned rank characteristics (-0.5 to 0.5) as reported in Table 1. Panel B uses (de-meaned rank) characteristics that have been orthogonalized via multivariate kernel regression to characteristics with stronger clientele strength. An iterative procedure is used to order characteristics by clientele strength.

Panel A: De-meaned Characteristic Ranks

	Share Price	Stock Age	Realized Volatility	Market Capitalization	Realized Returns	Stock Turnover	Popularity	Book/Market	Realized Skewness
Equal-weighting accounts	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
Clientele Strength (×100)	1.24	0.77	0.54	1.05	0.42	0.38	0.40	0.50	0.28
Total Factor Variance (×100)	4.15	5.34	2.83	3.26	3.09	2.87	1.55	2.96	2.76
Value-weighting accounts									
Clientele Strength (×100)	0.93	0.96	0.48	0.72	0.57	0.24	0.54	0.49	0.39
Total Factor Variance (×100)	1.63	2.73	1.21	1.15	1.58	1.01	1.15	1.25	1.60
Turnover-weighting accounts									
Clientele Strength (×100)	1.52	0.46	0.88	1.05	0.74	0.47	0.46	0.67	0.36
Total Factor Variance (×100)	3.67	3.73	2.92	2.39	2.88	2.19	1.31	2.61	2.36

Panel B: Orthogonalized Characteristic Ranks

	Share Price	Stock Age	Realized Volatility	Market Capitalization	Realized Returns	Stock Turnover	Popularity	Book/Market	Realized Skewness
Equal-weighting accounts									
Clientele selection order	1	2	3	4	5	6	7	8	9
Clientele Strength (×100)	1.24	0.66	0.38	0.31	0.29	0.15	0.13	0.08	0.05
Total Factor Variance (×100)	4.15	4.75	2.71	1.69	3.25	2.84	1.58	2.45	2.38
Value-weighting accounts									
Clientele selection order	2	1	8	6	4	5	3	7	9
Clientele Strength (×100)	0.78	0.96	0.07	0.17	0.29	0.19	0.40	0.13	0.02
Total Factor Variance (×100)	1.55	2.73	0.73	0.73	1.28	1.22	1.00	0.94	1.14
Turnover-weighting accounts									
Clientele selection order	1	5	2	4	3	6	8	7	9
Clientele Strength (×100)	1.52	0.30	0.54	0.35	0.44	0.15	0.14	0.13	0.05
Total Factor Variance (×100)	3.67	3.33	2.49	1.32	2.71	1.93	1.30	2.11	1.85

Table 3
Multifactor Regression Estimates

For each stock i we run the regression specification in Equation (4) of the paper over the set of 9.7 million individual investors in August 2011. This table presents the coefficients summarized for all 3103 stocks in sample. We multiply the coefficent by 100, so the units of beta represent the percentage point change in holding probability per standard deviation change in the factor. Each row in Panel A corresponds to an account characteristic factor, and each row in Panel B corresponds to a stockholding characteristic factor. Columns show the standard deviation, 10^{th}, 50^{th}, 90^{th} percentiles of the cross-sectional distribution, respectively. The last two columns present the average of the absolute values of the $t-$statistic, and the percent of stocks with β_{ik} that are statistically significant at the 5% level.

Panel A: Account Characteristics

	Std. Dev	10%	50%	90%	Avg. |t-stat|	% Significant (5% level)
Age	0.16	-0.04	-0.01	0.03	21.74	96.13
Size	0.35	-0.09	-0.03	0.01	38.00	96.49
Turnover	0.08	-0.02	0.00	0.02	10.30	83.50
No. Stocks	0.83	0.09	0.28	1.17	265.75	100.00
No. Stocks Traded	0.25	-0.09	-0.01	0.12	40.86	95.65
Geographic Region						
Southern	0.10	-0.03	0.00	0.03	11.44	76.89
Northern	0.09	-0.02	0.00	0.02	8.56	74.61
Western	0.15	-0.04	0.00	0.04	12.80	86.37

Panel B: Stockholding Characteristics

	Std. Dev	10%	50%	90%	Avg. |t-stat|	% Significant (5% level)
Fama-French factors						
Market Capitalization	0.27	-0.07	-0.03	0.03	26.77	97.78
Book/Market	0.06	-0.02	0.00	0.01	8.54	82.66
Return-based factors						
Realized Volatility	0.12	-0.03	0.01	0.04	15.03	91.23
Realized Returns	0.09	-0.02	0.00	0.03	12.04	83.34
Realized Skewness	0.04	-0.01	0.00	0.01	4.90	55.72
Behavioral factors						
Share Price	0.33	-0.14	-0.02	0.10	35.45	95.62
Popularity	0.09	-0.04	-0.01	0.03	15.63	92.10
Stock Age	0.20	-0.05	0.00	0.05	18.56	86.14
Turnover	0.06	-0.01	0.00	0.02	7.84	77.80
Dividend Paying	0.14	-0.03	0.01	0.04	10.10	88.53
Business Group Holdings						
Reliance (ADAG)	0.38	-0.03	0.01	0.03	17.96	95.39
Tata	0.05	-0.02	0.00	0.01	5.52	72.70
Reliance (DAG)	0.15	-0.02	0.01	0.03	14.66	94.46
Birla Aditya	0.05	-0.01	0.00	0.00	4.72	62.26
Jaypee	0.08	-0.01	0.00	0.01	5.61	71.83
Jindal	0.05	-0.01	0.00	0.00	4.64	57.75
Mahindra	0.05	-0.01	0.00	0.01	6.09	80.66
Suzlon	0.07	-0.02	0.00	0.01	6.71	77.22
Vedanta	0.04	-0.01	0.00	0.00	6.13	80.70
Adani	0.09	-0.02	0.00	0.00	5.54	60.01
Others	0.09	-0.02	0.00	0.03	11.11	84.69
Industry Holdings						
Financial Services	0.10	-0.05	0.01	0.03	16.27	93.46
Food, Agri. and Textiles	0.05	-0.01	0.00	0.02	7.69	74.99
Information Technology	0.09	-0.03	0.00	0.01	9.60	82.60
Manufacturing	0.09	-0.01	0.00	0.03	8.76	83.40
Oil and Gas	0.17	-0.02	0.01	0.03	13.21	94.13
Other Retail	0.10	-0.02	0.00	0.01	9.27	86.17

Table 4
Contribution to Explanatory Power: Marginal R-squared

This table presents the relative contribution of different groups of factors to the explanatory power of the regressions summarized in Table 3. The first row in Panel A presents the full model R-squared from a pooled least squares model with stock-specific intercepts and loadings on F_k. In each row following the first, we re-estimate this model excluding factors corresponding to the characteristic(s) listed at left, and report the reduction in R-squared that results (i.e. the marginal R-squard) as a percentage of the full model R-squared. We apply stock-specific weights to equalize each stock's contribution to the pooled R-squared ("Equally Weighted"). Panel B presents pooled R-squareds for our 10 factor unobserved PCA model, as well as the R-squared associated with each of the first three principal components of this model. Stock level weights are used to construct R-squareds is just as in Panel A.

Panel A: Observed Factor Model

	Equally Weighted
Full R-squared	1.74
	Percent of Full R-squared
Account Characteristics based Factors	81.96
No. Stocks	47.87
Size	10.50
Age	5.52
Geographic factors	0.96
Turnover	0.13
Stockholding Characteristics based Factors	8.72
Behavioral factors	1.95
Business group factors	1.55
Industry factors	1.05
Fama-French factors	1.02
Return factors	0.62

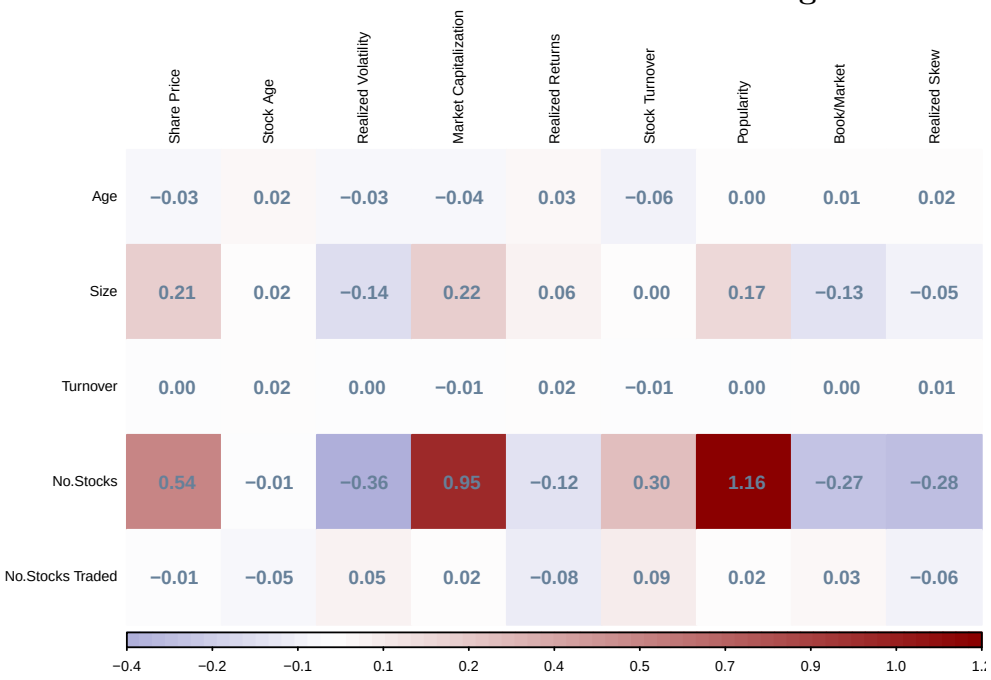
Panel B: Unobserved PCA Factor Model

	Equally Weighted
PCA 1-10 Model	3.69
	Percent of Full R-squared
PC1	42.36
PC2	14.61
PC3	8.64

Table 5
Factor Loadings and Stock Clienteles:
Account characteristics factors

This table presents the stock characteristic rank weighted loadings from a regression of stock holdings Q_{ih} only with account characteristic factors. Each row represents the account characteristic factor, and each column for the various stock clienteles of interest. Panel A weights the loading using demeaned rank weights scaled to be a long-short portfolio, while Panel B weights the loadings using orthogonalized characteristics (as in Table 2) instead, with weights demeaned and scaled, constructed similarly to Panel A.

Panel A: De-meaned characteristic rank weights



Panel B: Orthogonalized characteristic rank weights

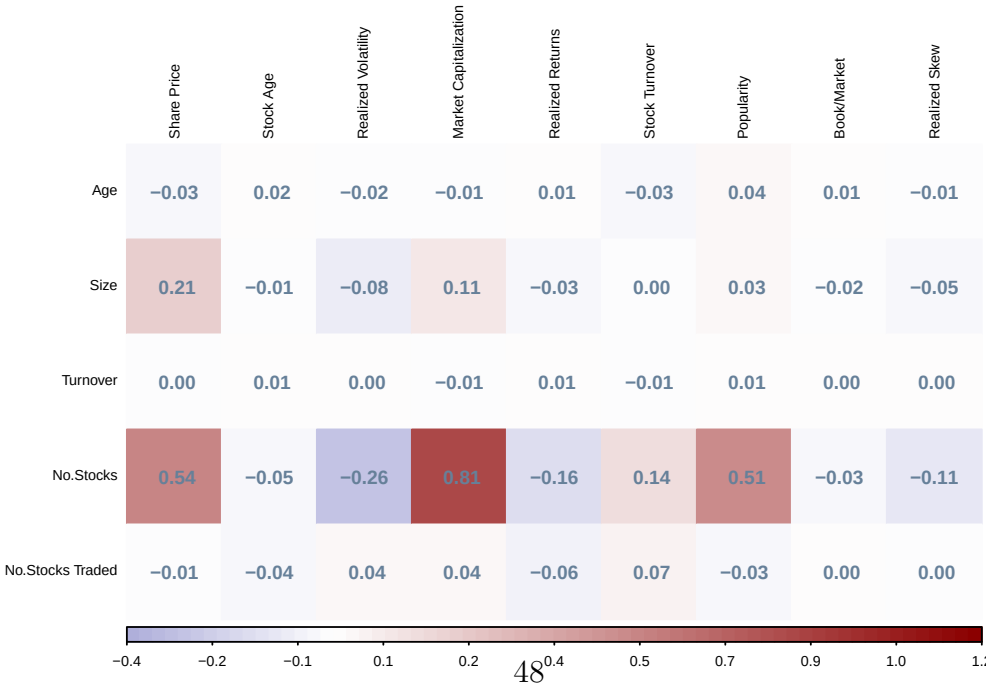
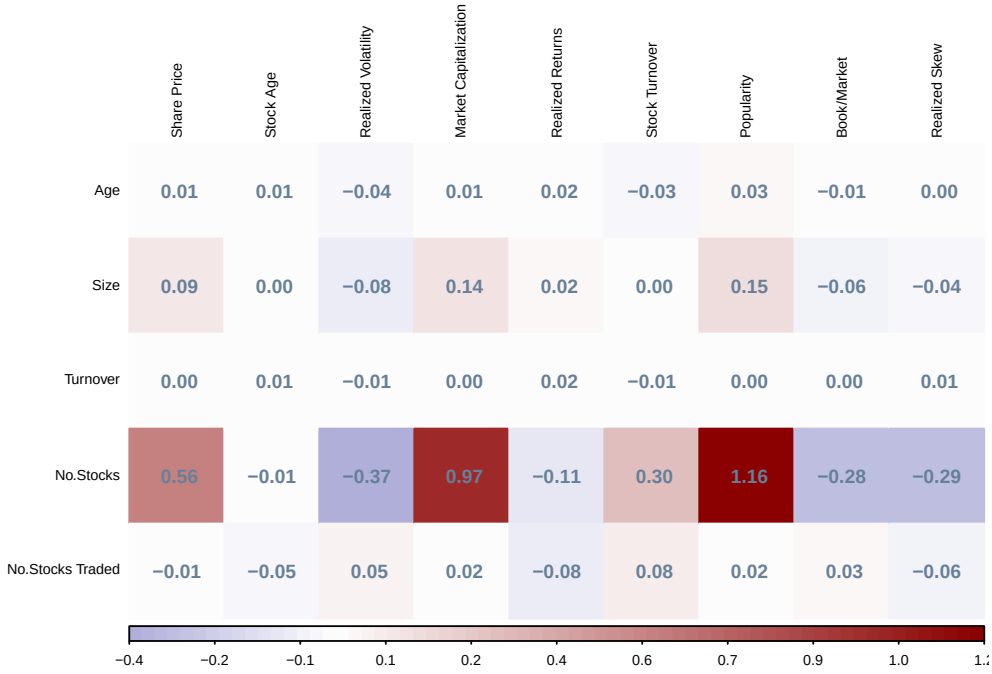


Table 6

Factor Loadings and Stock Clienteles: All factors

This table presents the stock characteristic rank weighted loadings from a regression of Q_{ih} with both account characteristic factors and stockholding characteristic factors. Each row represents the account characteristic factor, and each column for the various stock clienteles of interest. Panel A weights the loading using demeaned rank weights scaled to be a long-short portfolio, while Panel B weights the loadings using orthogonalized characteristics (as in Table 2) instead, with weights demeaned and scaled, constructed similarly to Panel A.

Panel A: De-meaned characteristic rank weights



Panel B: Orthogonalized characteristic rank weights

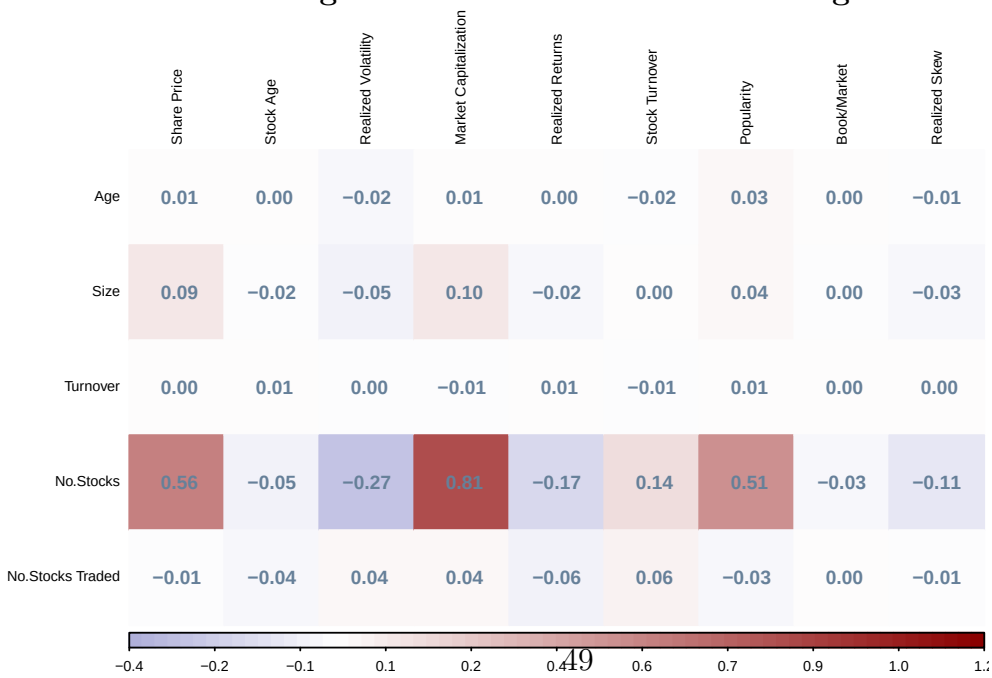


Table 7
Factor Loadings and Stock Clienteles: All factors

This table presents the stock characteristic rank weighted loadings from a regression of Q_{ih} with both account characteristic factors and stockholding characteristic factors. Each row represents the stockholding characteristic factor, and each column for the various stock clienteles of interest. Panel A weights the loading using demeaned rank weights scaled to be a long-short portfolio, while Panel B weights the loadings using orthogonalized characteristics (as in Table 2) instead, with weights demeaned and scaled, constructed similarly to Panel A.

Panel A: De-meaned characteristic rank weights

Panel B: Orthogonalized characteristic rank weights

Table 8
Returns Factor Strength

Each column of this table presents the variance of a given stock characteristic return factor using weekly returns from April 2002 through September 2011. Portfolio weights are de-meaned stock characteristic ranks as of the end of the previous month (non-orthogonalized in Panel A, and orthogonalized as in Table 2 in Panel B), scaled to be a long-short portfolio. Return factor strength represents the portion of factor variance contributed by off-diagonal terms in the returns covariance matrix.

Panel A: De-meaned Characteristic Ranks

	Share Price	Stock Age	Realized Volatility	Market Capitalization	Realized Returns	Stock Turnover	Popularity	Book/Market	Realized Skewness
Return Factor Strength (× 1000)	0.865	0.060	0.479	0.697	0.347	0.157	0.329	0.098	0.130
Total Factor Variance (× 1000)	0.889	0.083	0.506	0.722	0.376	0.182	0.351	0.128	0.152
Total Factor Volatility	2.98%	0.91%	2.25%	2.69%	1.94%	1.35%	1.87%	1.13%	1.23%

Panel B: Orthogonalized Characteristic Ranks

	Share Price	Stock Age	Realized Volatility	Market Capitalization	Realized Returns	Stock Turnover	Popularity	Book/Market	Realized Skewness
Return Factor Strength (×1000)	0.865	0.050	0.198	0.180	0.151	0.046	0.085	0.015	0.014
Total Factor Variance (× 1000)	0.889	0.073	0.223	0.206	0.181	0.074	0.112	0.043	0.039
Total Factor Volatility	2.98%	0.85%	1.49%	1.44%	1.34%	0.86%	1.06%	0.65%	0.62%

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%).

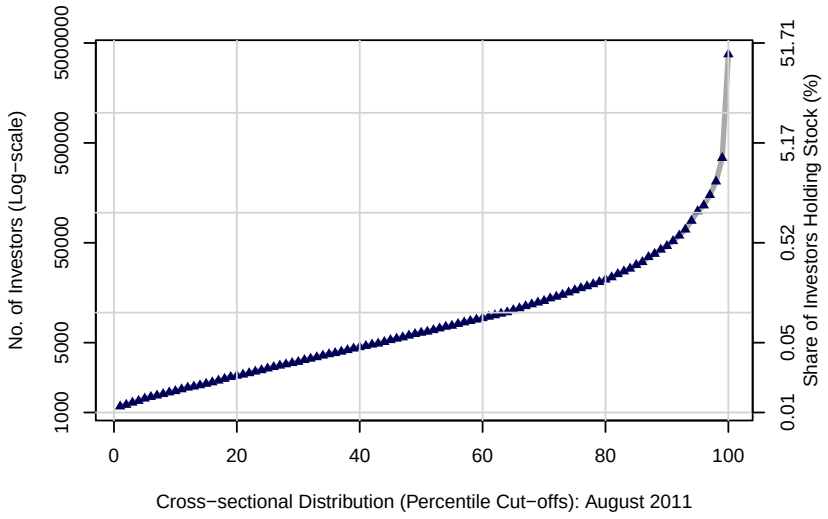
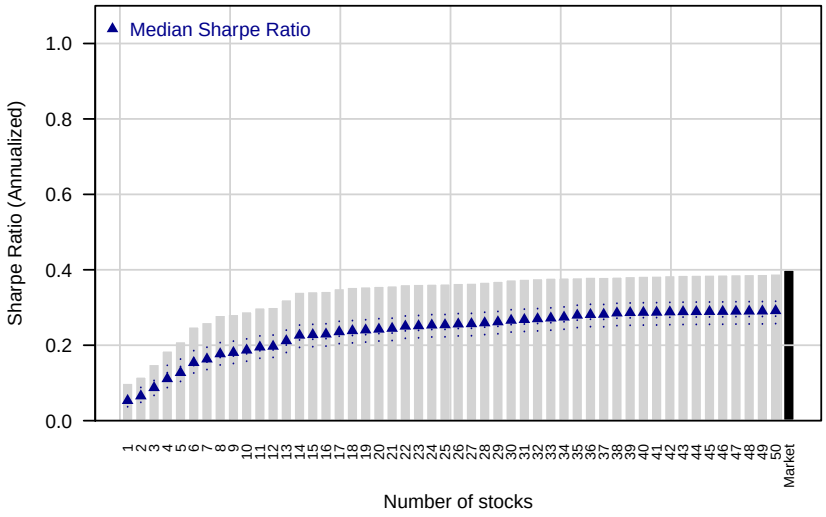


Figure 2
CAPM and Four Factor-Implied Sharpe Ratio

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

Panel A: CAPM-Implied Sharpe Ratio Estimates



Panel B: Four Factor-Implied Sharpe Ratio Estimates

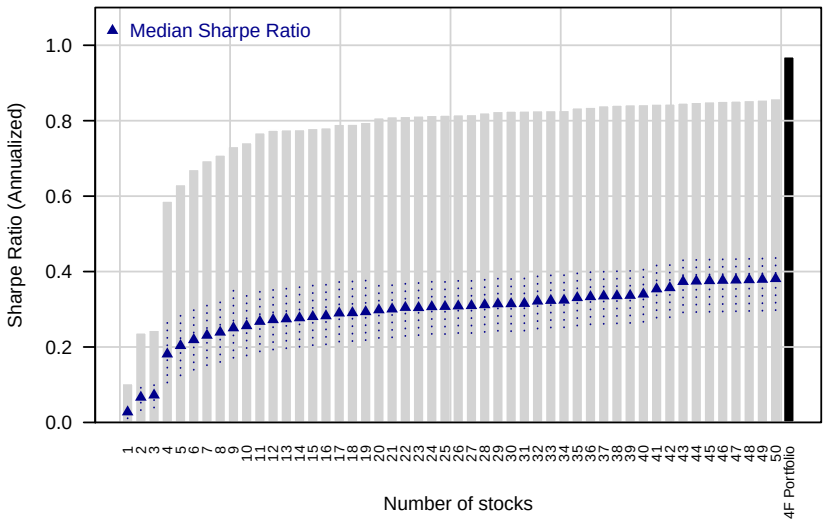
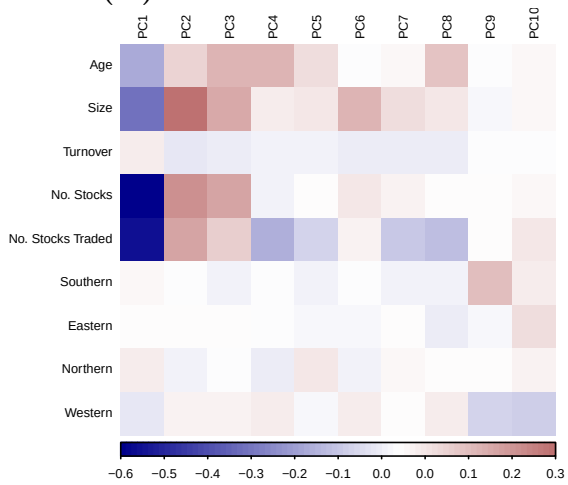


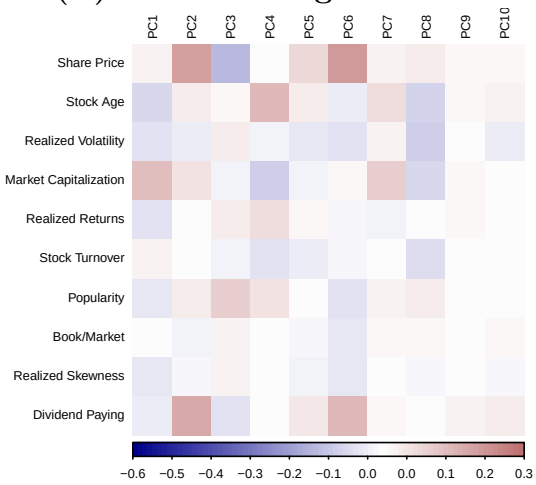
Figure 3
Loadings of PCA 1-10 Factors on Observed Factors

Panels A, B and C present a heatmap of the loadings of PC 1-10 factors (in columns) on observed factors (in rows), all normalized by their standard deviations to allow for comparison. Shades in darkest red and blue represent large positive and negative loadings on the PC factor, with the exception of the loading of "No. stocks" on "PC1", which is −0.8, but rescaled to be the same color as −0.6 for visualization.

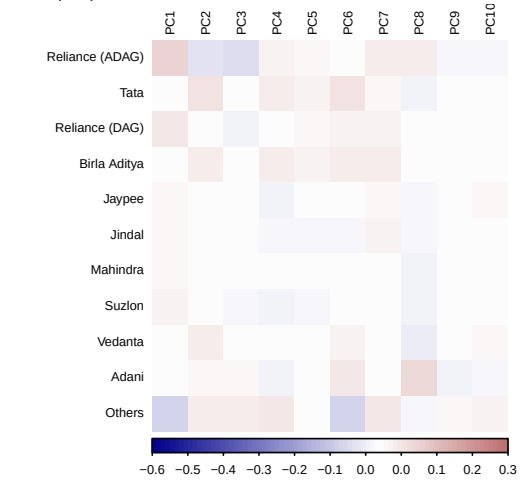
Panel (A): Account Characteristics



Panel (B): Stockholding Characteristics



Panel (C): Stock Industry Characteristics



Panel (D): Stock Business Group Characteristics

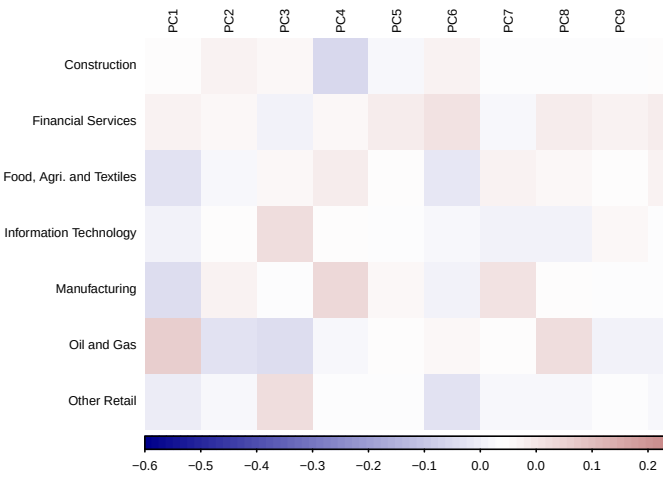
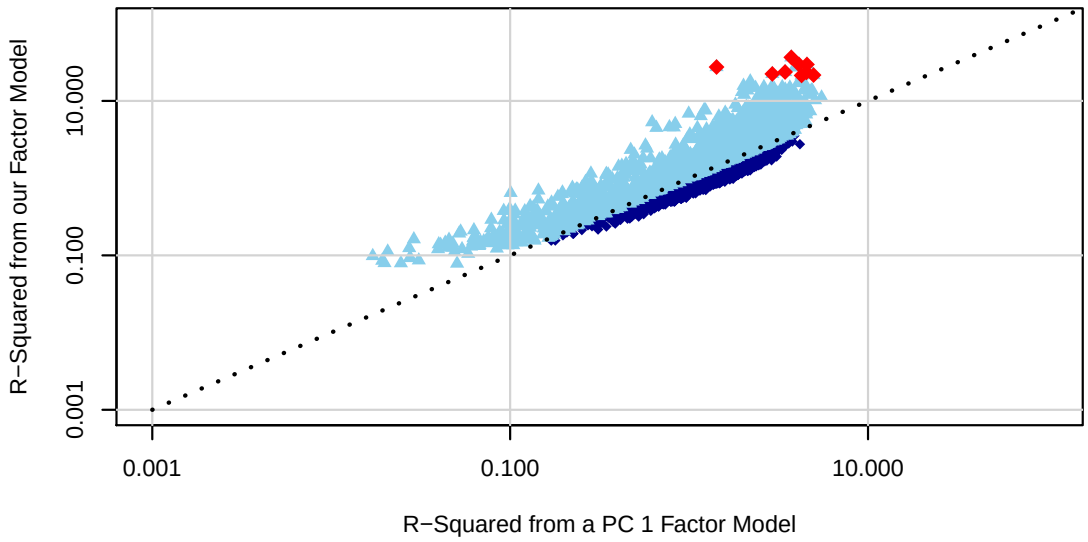


Figure 4
Comparison of Stock Level R-Squareds

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

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



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

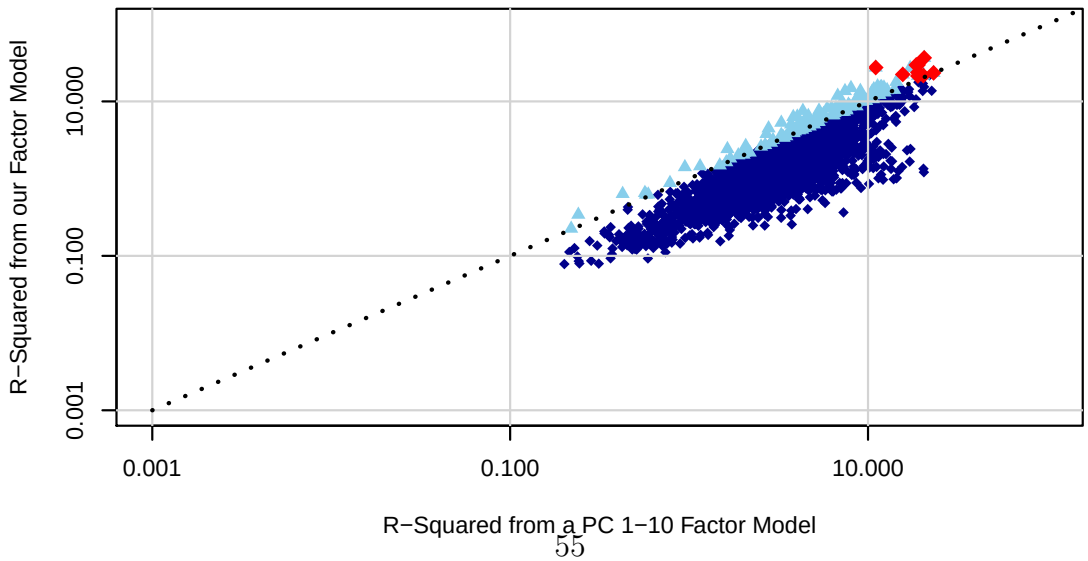


Figure 5
Empirical vs Model Coholdings

This figure plots the empirical coholding likelihood ($y-$axis), against the model-implied coholding likelihood measured as in Equation (1) ($x-$axis), both on logarithmic scale. We exclude 4322 stock pairs that have non-positive coholding estimates ($\leq 0.1\%$ of all pairs in the data). The dashed lines plot the 45-degree line. Panel A plots the observed factor model implied coholdings while Panel B plots the unobserved PCA 1-10 model implied coholdings. The darker regions have greater density of observations.

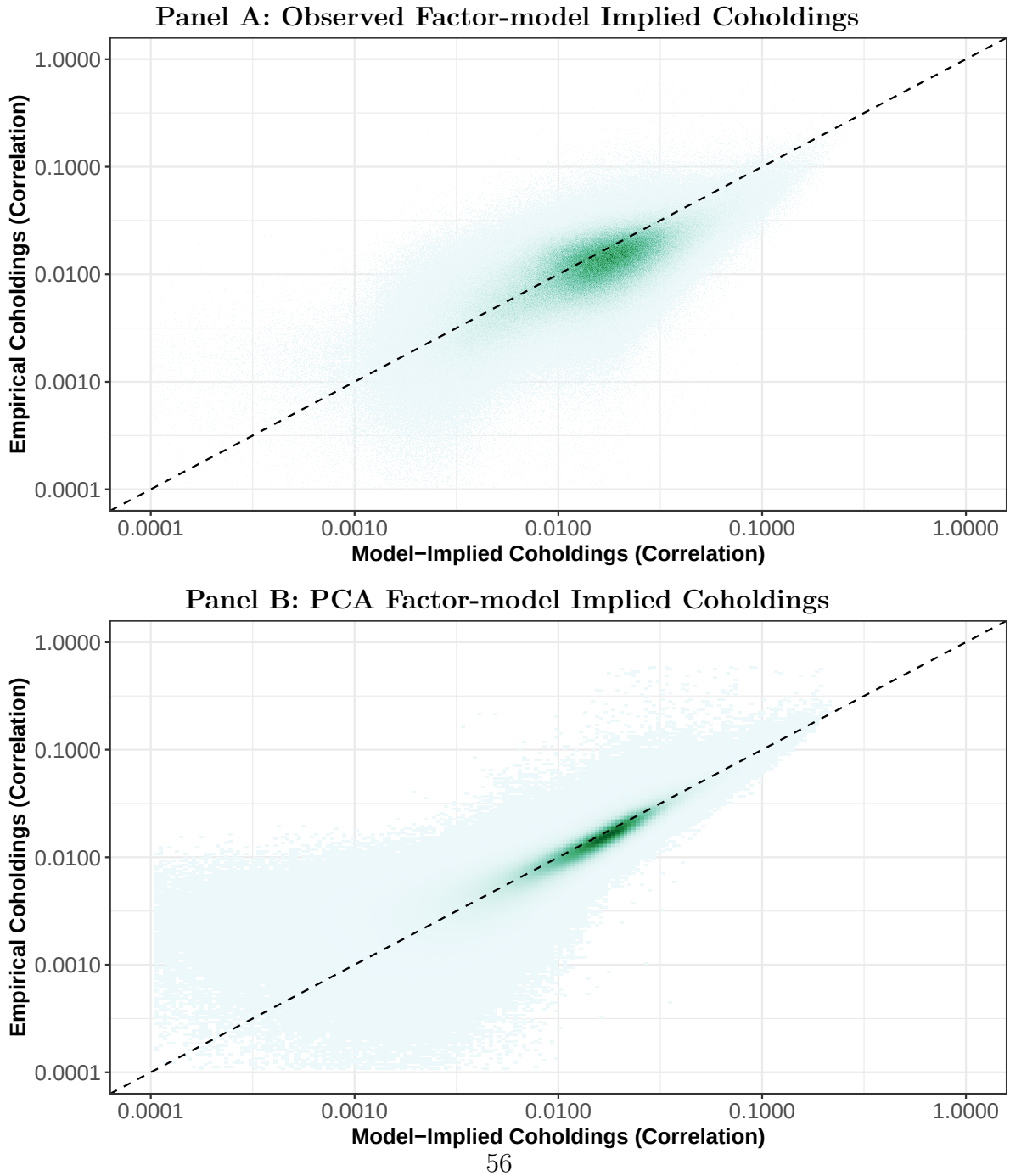
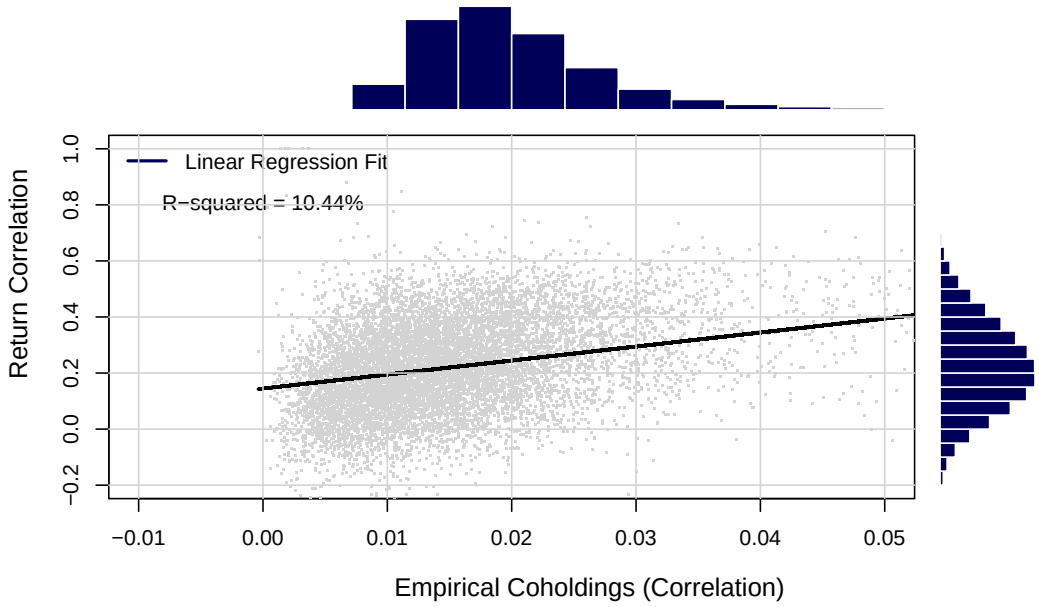


Figure 6
Return Covariance and Coholdings

This figure plots the relationship between cross-stock correlation ($y-$axis) against the coholding (correlation) measure ($x-$axis). Panel A presents the relationship with the empirical coholding estimates on the $x-$axis, while Panel B presents the observed factor model-implied coholding (correlation). The marginal distribution of the two variables are presented as histograms to the top (coholding (correlation)) and right (return correlation) of the scatter plot. The return correlation estimates are based on weekly returns data for a year leading up to August 2011.

Panel A: Empirical Coholdings (Correlation)



Panel B: Observed Factor Model-implied Coholdings (Correlation)

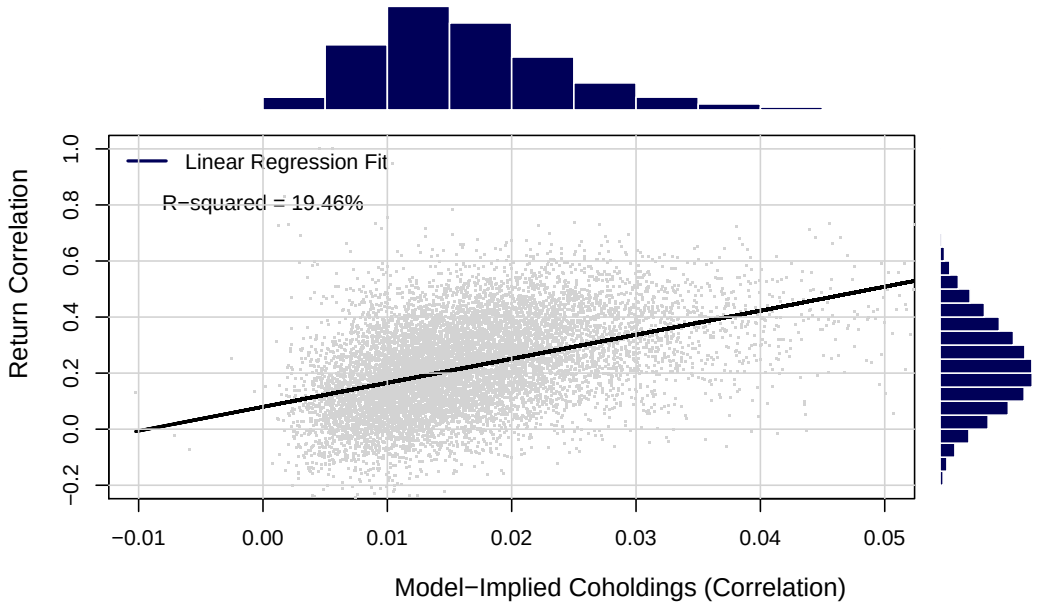
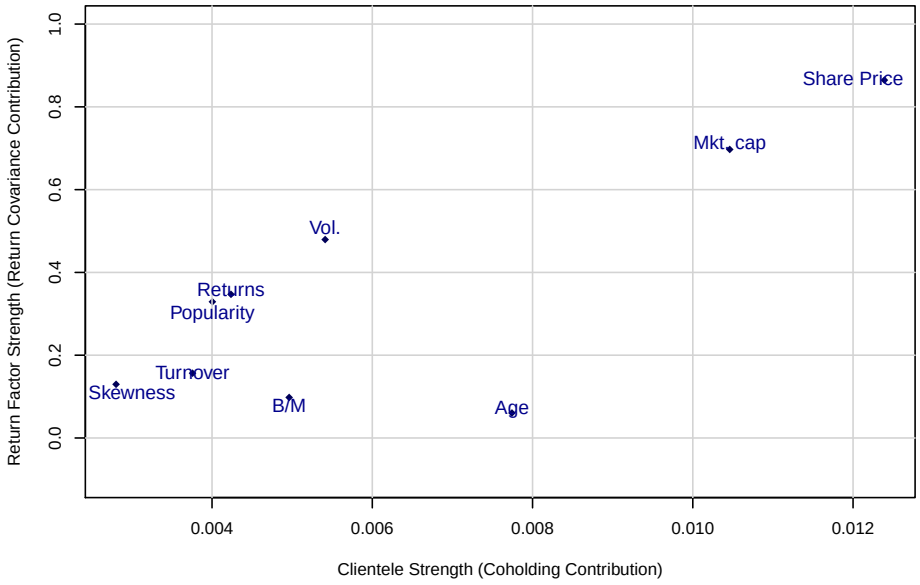


Figure 7

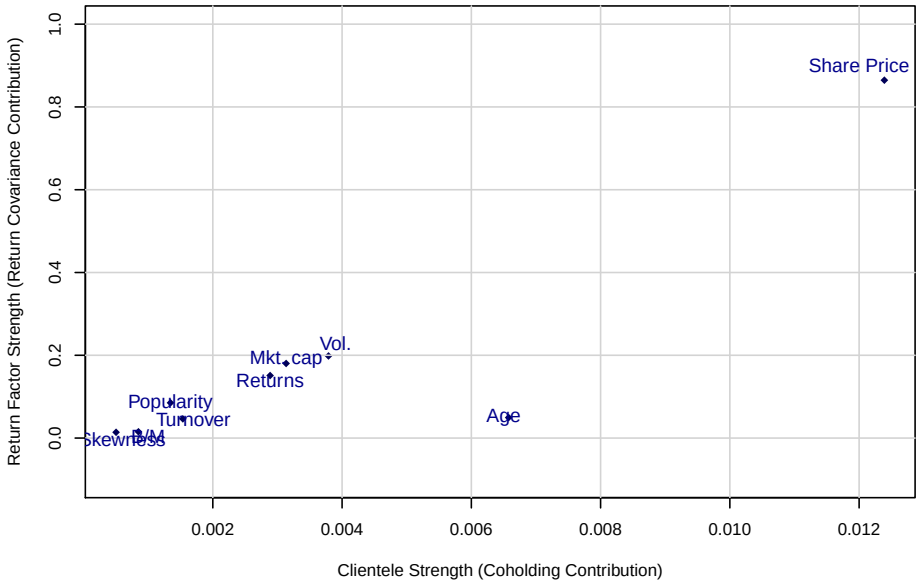
Return Factor Strength and Clientele Strength

These plots compare the clientele strength (see Table 2) with the return factor strength (see Table 8) for each stock characteristic. Panel A use non-orthogonalized characteristics, whereas Panel B uses the orthogonalized characteristics (corresponding to Panel B in Tables 2 and 8).

Panel A: De-meaned characteristic rank



Panel B: Orthogonalized characteristic rank



Online Appendix

Who Holds What?
A Factor Model for Direct Stockholding

Vimal Balasubramaniam John Y. Campbell
Tarun Ramadorai Benjamin Ranish

Table A.1
Contribution to Explanatory Power: Pooled Marginal R-squared
Other Weighting Schemes

This table presents the relative contribution of different groups of factors to the explanatory power of the regressions summarized in Table 3. The first row in Panel A presents the full model R-squared from a pooled least squares model with stock-specific intercepts and loadings on F_k. In each row following the first, we re-estimate this model excluding factors corresponding to the characteristic(s) listed at left, and report the reduction in R-squared that results (i.e. the marginal R-squard) as a percentage of the full model R-squared. In the first column these regressions are all ordinary ("Unweighted") least squares regressions, whereas the second column applies stock-specific weights to make these contributions to the poooled R-squared proportional to the individual investor share of the stocks' ownership ("Individual Share Weighted"). Panel B presents pooled R-squareds for our 10 factor unobserved PCA model, as well as the R-squared associated with each of the first three principal components of this model. The stock level weights used to construct R-squareds are varied across the columns just as in Panel A.

Panel A: Observed Factor Model

	Unweighted	Individual Share Weighted
Full R-squared	7.20	1.13
	Percent of Full R-squared	
Account Characteristics based Factors	59.39	82.15
No. Stocks	26.74	49.32
Size	13.22	1.34
Age	10.03	1.82
Geographic factors	1.54	0.97
Turnover	0.28	0.11
Stockholding Characteristics based Factors	19.18	11.52
Behavioral factors	2.44	2.94
Business group factors	6.41	1.03
Industry factors	1.67	1.10
Fama-French factors	1.24	1.71
Return factors	0.78	0.87

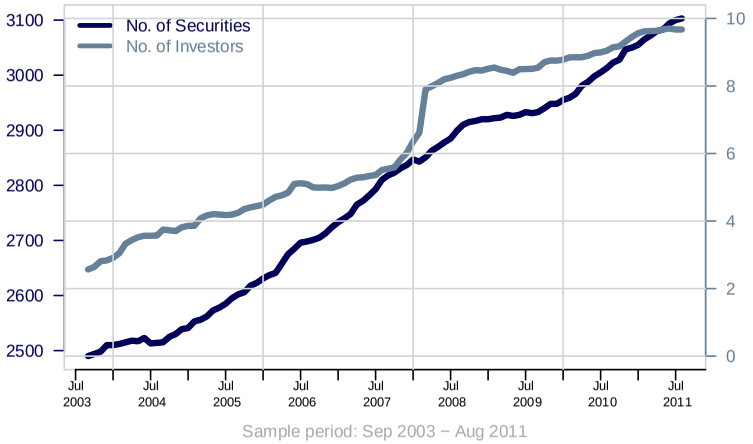
Panel B: PCA Factor Model

	Unweighted	Individual Share Weighted
PCA 1-10 Model	8.96	3.01
	Percent of Full R-squared	
PC1	28.70	37.80
PC2	25.53	13.02
PC3	11.96	13.38

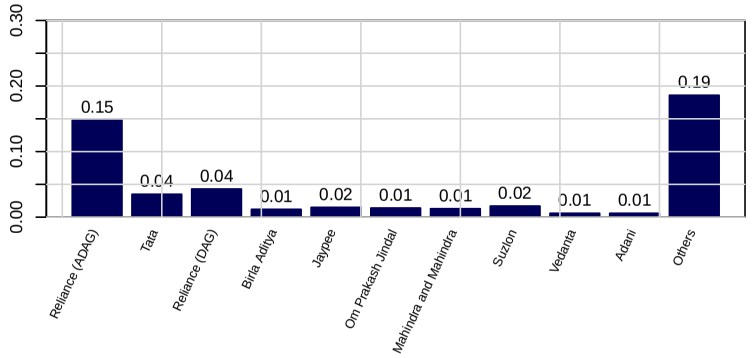
Figure A.1
Summary Statistics

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

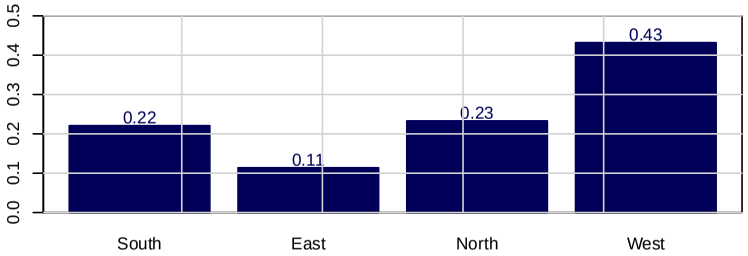
Panel A: Number of Observations



Panel B: Business Groups



Panel C: Geography



Panel D: Industry

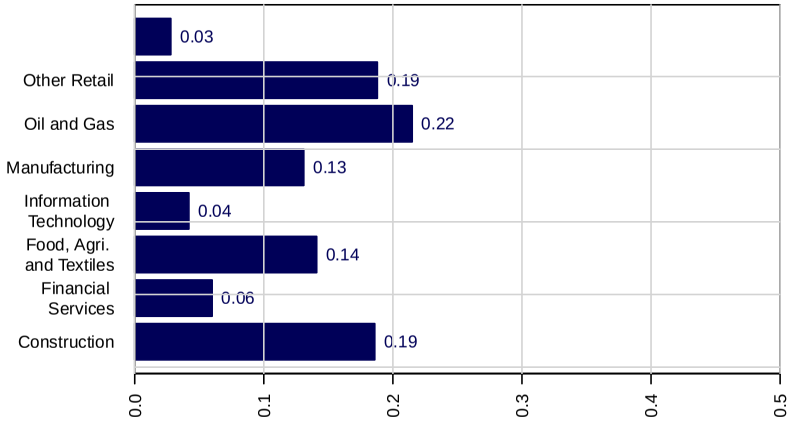


Figure A.2
Comparison of U.S. and Indian Household Stock Wealth

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

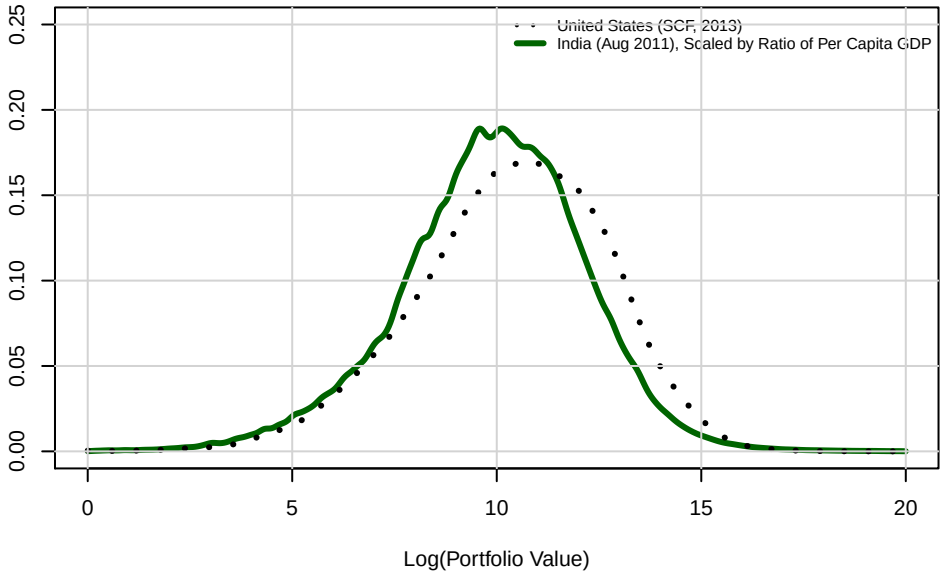


Figure A.3
Correlation Matrix

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

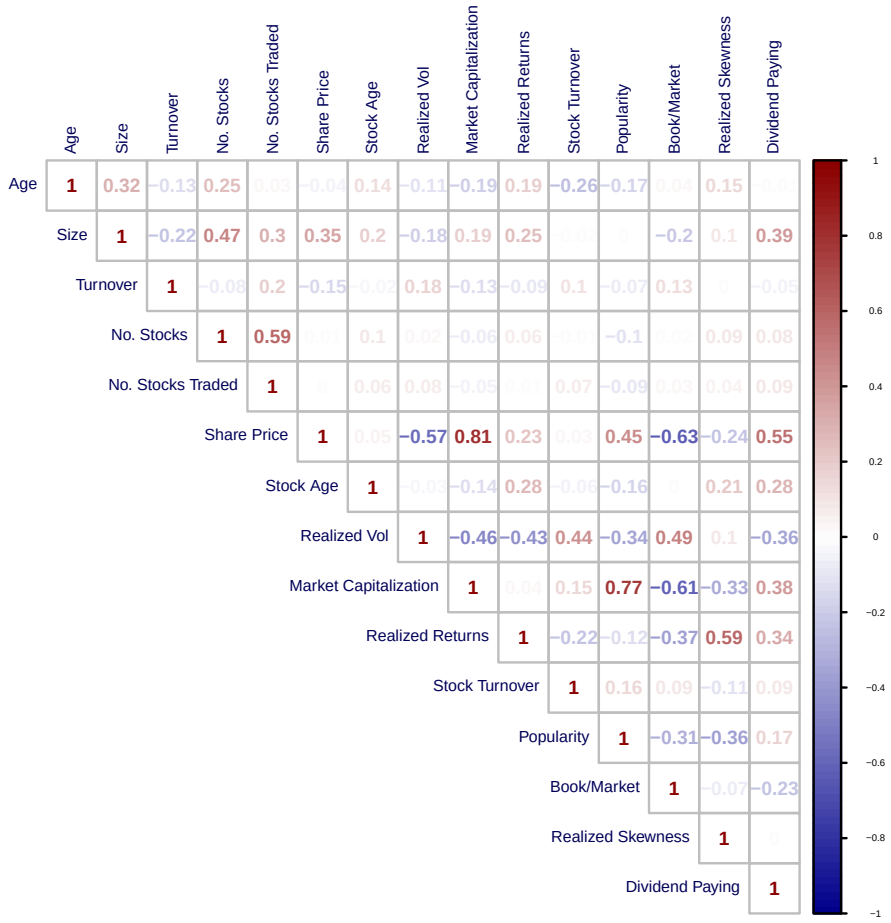


Figure A.4
Unobserved Factor Model: Principal Components

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

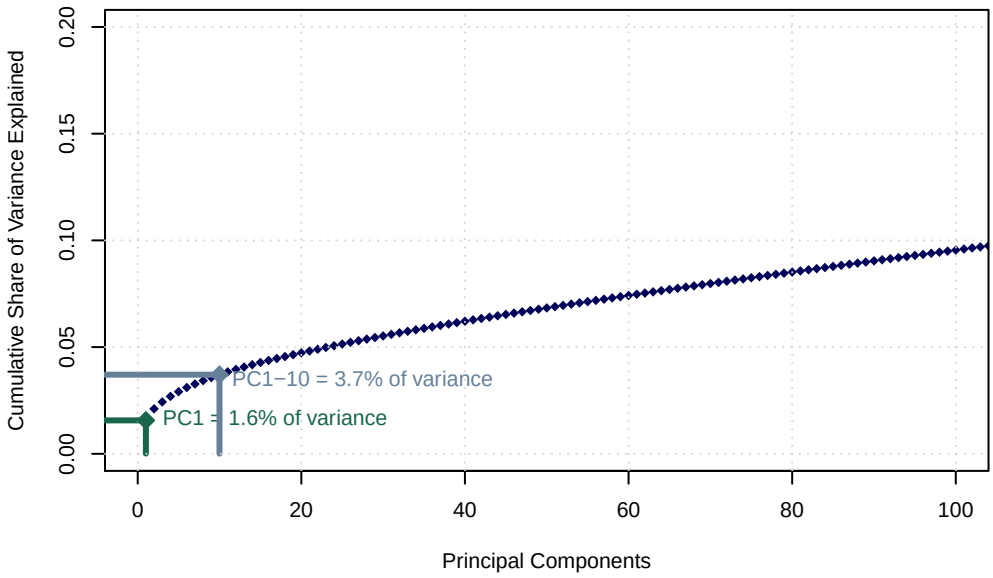


Figure A.5

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

This figure presents the Eigenvalue Ratios of ordered, ratio of adjacent eigenvalues of the matrix, following Ahn and Horenstein (2013). The line presents the ratio of the adjacent eigenvalues for the Q matrix, scaled by the inverse of the within stock standard deviation.

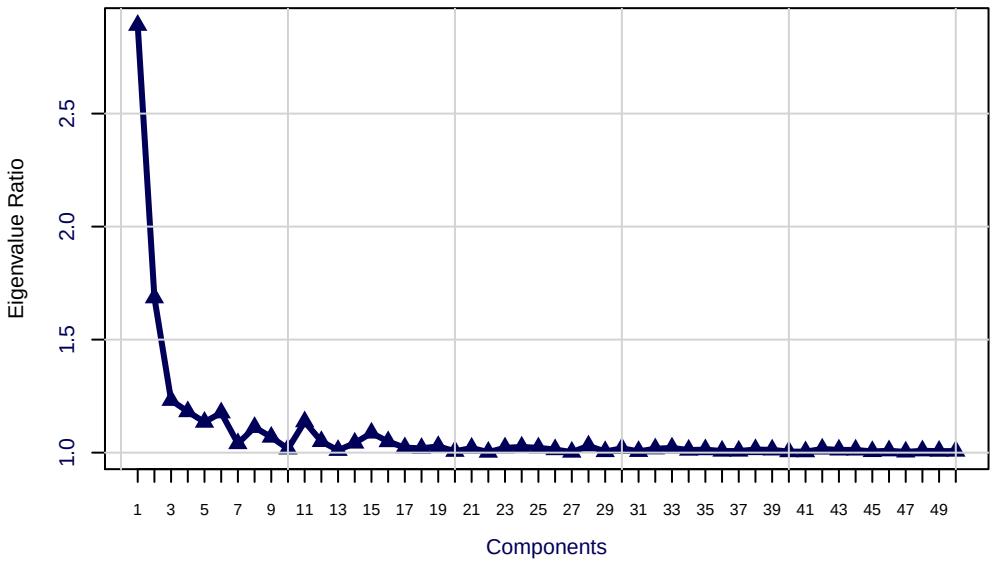
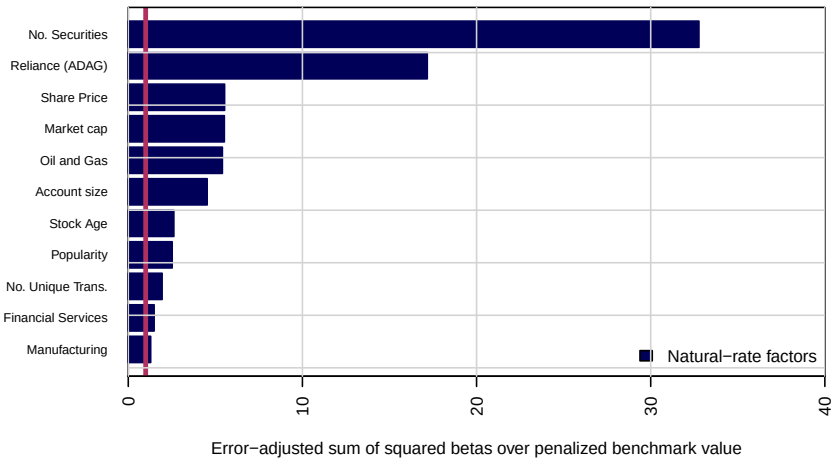


Figure A.6
Natural-rate and Semi-strong Factors

Panel A presents the Connor and Korajczyk (2019) test statistic for each of the observed model's factors as a proportion of the suggested threshold value (based on $\delta = 1/5$), corresponding to a marginal R-squared of 0.11%. Factors statistically significantly above the threshold are identified as natural rate and presented in dark blue. Panel B applies the same analysis to PCA-derived factors.

Panel (A): Observed Multifactor Model



Panel (B): Unobserved PCA Factor Model

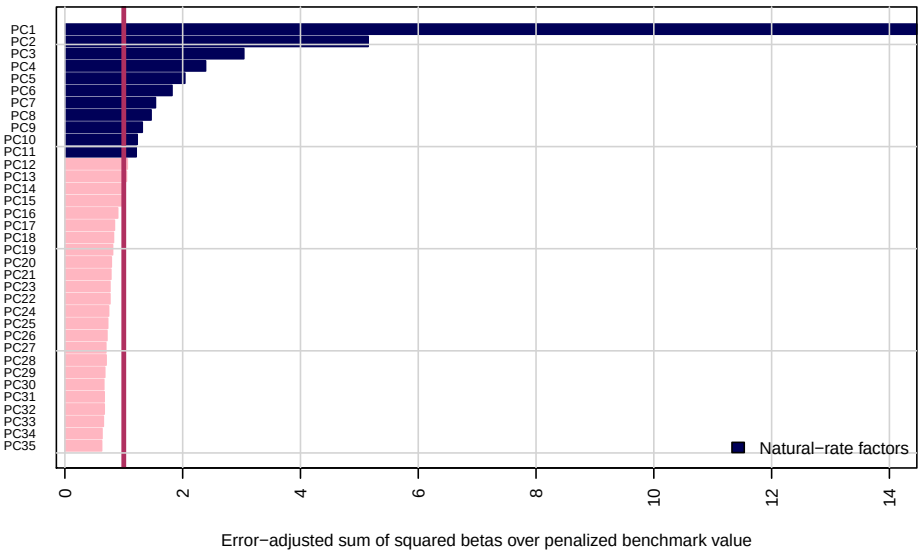
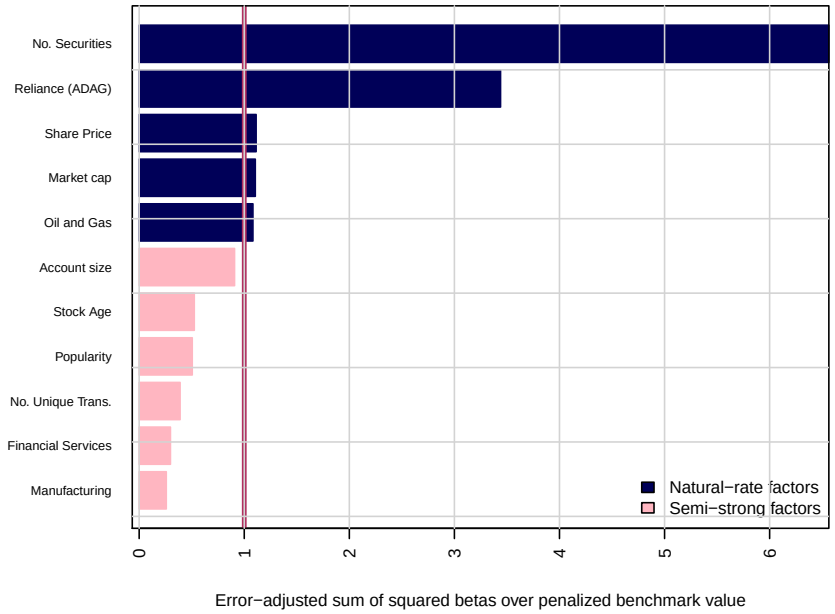


Figure A.7
Natural-rate and Semi-strong Factors

This figure reproduces the same analysis as Figure A.6, but assumes stronger semi-strong factors ($\delta = 1/10$), leading to a marginal R-squared cutoff near 0.5%.

Panel (A): Observed Multifactor Model



Panel (B): Unobserved PCA Factor Model