Who Owns What?
A Factor Model for Direct Stockholding

Vimal Balasubramaniam John Y. Campbell
Tarun Ramadorai Benjamin Ranish*

September 17, 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 on monthly direct equity holdings and trades using data from almost 10 million accounts in the Indian stock market. We find that household-specific factors such as the size of the equity portfolio and the number of stocks it contains are important predictors of individual stockholdings. Factors that capture the popularity, value, and size tilts of investors' portfolios also help predict the stocks they hold. We use our new factor model to analyze the empirical equity "coholdings matrix", which measures the degree to which stocks are jointly held in investor portfolios and thereby captures investor clientele effects.

*Balasubramaniam: Queen Mary University of London. 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.

As a first step in our analysis, we consider two very simple models of portfolio choice and show that neither fits our data. Both models take as given the number of stocks held by each household. The first model assumes that households randomly select stocks with probabilities equal to their overall popularity in household portfolios. We find that contrary to this model, certain stocks are disproportionately held in single-stock portfolios while other stocks are disproportionately held in portfolios containing large numbers of stocks. This implies that the number of stocks held in a portfolio is an important factor to include in our observable factor model.

The second model assumes that households select stocks to 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.

Our empirical multifactor model provides a much better description of household portfolio choice. Within this model, the most important factors are based on 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.

Among the factors based on stockholding characteristics, the most important factors capture preferences for particular business groups or industries, for return characteristics such as volatility, high recent returns, or skewness, or for stock attributes such as 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.

One important use of our factor model is to estimate the coholdings propensity—the tendency for a pair of stocks to be held in the same investors' portfolios—for any pair of stocks. We stack the coholdings propensities for all stock pairs into a "coholdings matrix", which we argue is a useful tool for measuring investor clientele effects on stock prices. We show that there is a positive correlation between coholdings propensities and stock return correlations, which is another way to see that investors' portfolios are not optimally diversified and which suggests that investor clientele effects may be important drivers of stock return comovements.

Related literature

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

Among those papers that do study household portfolio choice, it is common to study choices of mutual funds rather than direct stockholdings. For example Grinblatt et al. (2016) highlight the impacts of IQ on mutual fund choice by Finnish investors using

detailed data on mutual fund choices alongside less detailed information on direct equity investment. Betermeier et al. (2017) use Swedish data to estimate value and growth tilts in household portfolios, but they estimate these tilts directly for mutual funds and do not attempt to look through to the implied weights on individual stocks.

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

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

Organization of the paper

The organization of our paper is as follows. Section 2 lays out the factor model that we use to organize our empirical research. Section 3 describes our Indian dataset. Section 4 considers two simple "straw man" models of portfolio choice, one in which households choose stocks randomly with a probability equal to their overall popularity, and one in which households seek maximal diversification conditional on the number of stocks they hold. Section 5 estimates multifactor models of stockholdings, considering models with observable factors as well as those with unobserved principal-components-based factors. Section 6 compares empirically observed coholdings with those predicted by the factor models, relates the estimated coholdings matrix to stock characteristics,

and relates coholdings with return covariances. Section 7 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, a column vector with N elements Q_{ih}, one for each stock. In our empirical application, we simply employ a holdings dummy, i.e., $Q_{ih} = 1$ if household h holds stock i, and $Q_{ih} = 0$ otherwise.[2]

[1] In our empirical application, we mainly study one particular month in the data, August 2011, which–as the last month of data in our sample–provides us the maximum past history for each investor. We also re-run our analysis across all of the months in the dataset to check the persistence of the relationships that we estimate. More generally, another possibility would be to average all the periods observed in the raw data and conduct the analysis on empirical time-averages of holdings.

[2] Other possible choices include Q_{ih} defined as the portfolio weight of stock i in the portfolio of household h, or Q_{ih} defined as the fraction of the total market capitalization of stock i held by household h.

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.

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.

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). It 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),[3] and subsequently

[3]Any holdings measure can also be cross-sectionally demeaned by subtracting the vector $\overline{Q}$ from Q_h, where $\overline{Q} = \frac{1}{H} \sum_{h=1}^{H} Q_h$.

average these outer products over time. Thus, the sample covariance matrix of returns uses time periods where the sample coholdings matrix uses households, 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 (e.g., similarity of firm size, beta, age, etc.) This would be the equivalent of regressing the sample covariances of stock pairs onto measures of the similarity of those stock pairs. In addition to being a somewhat awkward way to model the coholdings matrix given the difficulty of exhaustively conceptualizing dimensions along which stocks are similar, an additional 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 returns offers us some guidance about an alternative, useful 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. The β_{ik} coefficients inform us about the average attributes of the investor clientele for each stock i.[4] In other words, these coefficients can be interpreted as telling us about which types of households like stock i and tend to hold it.

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

[4]While we do not do this in the current draft, as with stock return factor models, it is also possible to restrict the β_{ik} rather than estimating them freely for each stock i. This is analogous to estimating time-series betas for characteristic-sorted portfolios, or estimating restricted stock-level models in which betas are linear in stock-level characteristics.

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 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, we focus primarily on month-end August 2011. This is the last month

of data in our sample, and consequently, provides us the maximum past history for each investor and correspondingly more precise estimates of the factors. The older and larger of the two depositories, National Securities Depository Limited (NSDL), accounts for about 64% of the roughly 9.7 million individual accounts we study in August 2011, with the remainder held at Central Depository Services Limited (CDSL).[5] Our data therefore record almost all trading in and holdings of Indian equity at the account-issue level at a monthly frequency.

In general, we do not observe data on investors' equity derivatives or mutual fund holdings. However, unlike in many other economies, over the period that we study, derivatives and mutual funds do not play a role in the typical individual investor's portfolio of Indian equities. While single-stock futures markets are quite active in India (Martins et al., 2012; Vashishtha and Kumar, 2010), a minority of accounts invest in equity derivatives over our sample period.[6] Moreover, while mutual funds have grown in popularity in India, the typical investor that holds individual equities in our sample has no bonds or mutual funds.[7] Additionally, we estimate that 89% of individuals' aggregate equity holdings in 2011 were direct, as opposed to holdings of equity mutual funds, unit trusts and unit-linked insurance plans.[8]

The sensitive nature of 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

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

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

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

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

of the state, and the type of investor. We use investor type to identify individual investor accounts.[9] A given individual investor can hold multiple accounts, so we aggregate accounts that share the same Permanent Account Number (PAN)—a unique identifier issued to all taxpayers by the Income Tax Department of India. This aggregation may not always correspond to household aggregation if a household has several PAN numbers, for example, if children or spouses have separate PANs. In addition, we are unable to link accounts by PAN between NSDL and CDSL. However, conversations with our data provider suggest that few retail investors have multiple depository relationships.

Given our interest in household portfolio construction, we restrict our current analysis to the portfolios of retail investors in the market, and do not at this stage consider the portfolios of institutions or government entities (which we also observe). We also exclude non-public equities, which the typical household may have difficulty acquiring. 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.[10]

3.2 Data on Indian Stock Characteristics

We match our data on Indian equity holdings to data on returns, market capitalization, book value, turnover, and the age, industry, location, and business group affiliation of the

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

[10] As a result of corporate actions, some equities are associated with multiple International Securities Identification Numbers (ISINs), but one ISIN is prevalent at each point in time.

firm. These data are primarily drawn from the CMIE Prowess database, with Datastream and Compustat Global used to supplement and validate these data.[11] In addition, we attempt to manually fill otherwise 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. Our data on stock and firm characteristics cover 97% or more of total stock positions for all but two attributes, with data on company age and returns extending back at least a year available for 93% of total stock positions in the data.

3.3 Summary Statistics

In the early 21st Century, equity market participation in India underwent dramatic expansion. This is illustrated in the top-left panel of online appendix Figure A.1, which shows that 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.[12] 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.[13]

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

[11] We use only market capitalization, book value, and returns data that we are able to validate through comparison between at least two data sources. We also manually validate the 25 largest and smallest percentage returns observed in the data.

[12] This does not reflect increases in dematerialization, as even at the beginning of our sample period, most Indian stocks were held in dematerialized form.

[13] While subscriptions to IPOs are investor choices, a related issue is that corporate actions can sometimes cause "supply-driven" changes to investor portfolios which we might interpret as factors capturing investor decisions. Our analyses are not sensitive to a small set of extreme coholdings correlation, and so we do not believe this materially affects our inferences, but we highlight this potential caveat in the interests of transparency.

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. Online appendix Figure A.2 shows that this distribution of account sizes is similar to the United States when accounting for the differences in per-capita GDP between the two countries. Jayaraj and Subramanian (2008) show that the median (wealthiest) deciles of Indian households had average total asset values of about $3,000 ($35,000) in 2008, meaning that the stock portfolios we study represent a non-trivial share of wealth for many of the investors in the data.

The accounts in our dataset are spread out across India. Online appendix Figure A.1 shows that the wealthier west of India accounting for 43% of all accounts, the east of India accounting for roughly 11% of all accounts, and the remaining accounts divided roughly equally between the north and south of India.[14]

Our empirical work also utilizes other broad "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

[14]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.

right-skewed.

The bottom half of Table 1 summarizes the characteristics of the stocks held in investor portfolios. With the exception of the variable "dividend paying," which simply measures the fraction of dividend-paying stocks in the portfolio, we compute stockholding characteristics as average ranks. We do so by first ranking all the stocks in the available universe on each characteristic from 0 to 100 (all turnover and return-based stock characteristics are computed using data over the year to August 2011), and then for each investor, we take an equal-weighted average of the rank on each characteristic across all stocks held.

The table reveals that the median stockholding of retail investors is at the 90th percentile of the size distribution, and is (mechanically) also far more popular (widely-held by retail investors) than the stock at the median popularity ranking. We orthogonalize this measure of stock popularity (i.e., the total number of accounts which hold the stock) with respect to market capitalization, with which it is otherwise highly correlated. The other characteristics of the median stockholdings are in line with this tilt towards large, popular stocks, since larger stocks tend to pay dividends, have lower book-market,[15] and over the sample period have lower volatility, lower realized returns, and lower realized skewness.

It is also worth noting that the characteristics of accounts and stockholdings vary significantly across accounts, with a standard deviation of close to 20 (the support is 0 to 100). The wide range observed in these "factors" can potentially help to identify the determinants of stockholdings and coholdings given the framework outlined in the previous section.

Online appendix Figure A.1 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

[15] B/M is computed using the standard Fama-French methodology applied to Indian stocks.

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.

Online appendix Figure A.3 shows that the account and stockholding characteristics we study are generally not very highly correlated with one another. With the exception of the correlation between realized returns and realized skewness (0.85) over the year to August 2011, all other correlations are below 0.6 in magnitude. Moreover, all correlations between an account characteristic and a stockholding characteristic are below 0.4 in magnitude, which suggests that empirical models 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 in this month 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.[16] The characteristics of stockholdings in the summary statistics, therefore, heavily reflect holdings of popular stocks. This distribution also suggests that explanations of stockholding patterns that focus on understanding the composition of investor portfolios may differ considerably from explanations that focus on understanding a given stock's investor base.

We now use a series of factor models to understand the patterns in these data at a

[16]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.

deeper level.

4 Single-Factor Models of Stockholdings

In this section we explore the ability of two simple theoretically motivated single-factor models to explain investors' stockholdings. We show that neither model fits our data well, and this motivates our use of a richer multifactor model.

4.1 Random and Popularity-Based Stock Selection

A primitive in our setup is N_h, the number of stocks held by each household h. We do not attempt to model the determinants of N_h, but proceed as if each household h has a preference for holding a particular number of stocks. Variation across households in the numbers of stocks held could arise from cognitive or real frictions associated with holding and trading multiple stocks, or simply from a lack of financial sophistication.

Mechanically, just as the capitalization-weighted average CAPM beta equals one, the average beta from stock-specific regressions of the stockholding indicator Q_{ih} on N_h (which we denote as β_i) equals $\frac{1}{N}$, where $N = 3,103$ is the total number of stocks in our sample.[17] Thus, analogous to the CAPM (Sharpe, 1964), in place of a (capitalization-weighted) average of excess stock returns at each t, the basic single-factor model in the stockholdings case has $N_h = \sum_i Q_{ih}$ on the right-hand-side, i.e., the number of stocks held by household h.

How do households select the constituents in the set of N_h stocks? One naïve hypothesis is that stocks are randomly selected from the universe of N stocks, with equal probabilities, until N_h is reached. In this "random selection" model, the estimated stockholding probability $\hat{Q}_i^{Rnd}$ (i.e., the estimated fraction of all households which hold the

[17] To see this, note that $Q_{ih} = \alpha_i + \beta_i N_h + \varepsilon_{ih}$. Summing both sides over i, the sum of the β_i equals the single-factor beta from a regression of the sum of Q_{ih} on N_h. Since $\Sigma_i Q_{ih} = N_h$, the average beta must equal $\frac{1}{N}$.

stock) will be identical across stocks, with $\hat{Q}_i^{Rnd} = \sum_h N_h/NH$. We can immediately dismiss this model given the extreme variation in stockholding probabilities across stocks, illustrated in Figure 1.

A second, more plausible hypothesis along similar lines is that given N_h, investors select stocks as a result of a single (unobserved) factor which determines their observed popularity in investor portfolios. To understand the implications of this "pure popularity-based" stock selection model, we use the observed frequencies $\hat{Q}_i = \frac{1}{H}\sum_h Q_{ih}$ with which households pick stocks to estimate their underlying popularity, and use these frequencies and the observed values of N_h to simulate the model's implied holding probabilities $\hat{Q}_{ih}^{Pop}$.

Table 2 provides a first simple assessment of this model. The rows of the table list different stocks, ranked by their unconditional average $\hat{Q}_i$. The most popular stock is Reliance Power ($\hat{Q}_i = 0.399$), which occupies the top row, and then we consolidate the top percentile of stocks ranked by $\hat{Q}_i$ in the second row (with average $\hat{Q}_i = 0.072$), followed by the 90-99 percentile stocks, 50-90 percentile stocks and the bottom half of stocks, all ranked by $\hat{Q}_i$. The columns of the table correspond to numbers of stocks held by households, N_h, which we group from households with single-stock accounts up to the group of households holding more than 51 stocks. Each cell of the table documents the observed stockholding probability divided by the popularity-based model's prediction $\hat{Q}_{ih}^{Pop}$ for each group of stocks listed in the rows.

Table 2 shows interesting deviations from the pure popularity-based selection model. Households with small N_h tend to hold popular stocks more frequently than the pure popularity-based model would predict (this is especially evident with holdings of Reliance Power when $N_h = 1$). This over-representation of more popular stocks in accounts with small N_h is generally mirrored in the under-representation of less popular stocks, although this tendency does not appear monotonic in popularity. The table also reveals that as N_h rises, more popular stocks are generally less prevalent in accounts than the pure popularity-based model would predict, and there is a countervailing tendency to hold

more "obscure" stocks than the model would predict.

Figure 2 expands the scope of Table 2, with N_h ranging from 1 to ≥ 201 along the x-axis, and stocks ranging from most to least popular down the y-axis. The height of each row corresponds to the unconditional holding probability $\hat{Q}_i$ associated with the percentile of stocks in question, and the cells are shaded according to the magnitude of $\hat{Q}_{ih}/\hat{Q}_{ih}^{Pop}$. Reliance Power, which is seen in 40% of accounts, is shown separately as the very top (largest) row in the figure, followed by the top percentile of stocks ranked by popularity excluding Reliance Power, the 99th percentile of stocks, and so on. Given the significant reduction in popularity as the ranks decline, the top of the figure occupies significantly greater space than the bottom. The pattern seen in Table 2 is more clearly evident here, with relatively high values of the ratio in the top-left-hand corner, moving down to the middle of the table as N_h increases, and towards the bottom for accounts with the very largest values of N_h.

Overall, this analysis shows that investors holding just a few stocks are even more likely to hold popular stocks than predicted by the pure popularity-based model, and correspondingly, obscure stocks appear substantially more likely to be held in portfolios with many stocks. By way of analogy, if stocks were tourist destinations in France, and N_h represented the level of traveller experience, less experienced travellers appear to visit Paris more frequently than the average number of tourist visits to Paris would predict. Conversely, more frequent travellers tend to eschew Paris, and are often seen travelling to destinations that are off the beaten path. Just as Paris is an entry-level destination for first-time visitors to France, similarly stocks such as Reliance Power are entry-level holdings for small investors who hold only one or two stocks.

Widely held stocks tend to be larger stocks with lower idiosyncratic volatility, suggesting that these deviations from the pure-popularity-based model could arise from a desire to mean-variance optimize subject to the constraint of having to hold just a few stocks. Our next step is therefore to assess whether mean-variance optimization in the presence

of such constraints plays a significant role in investors' portfolio construction.

4.2 Constrained Mean-Variance Optimization

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. However, in light of the patterns in stockholdings detected in the previous section, 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, whilst operating under a constraint on the number of stocks N_h that they hold. As before, cross-household variation in N_h 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. In this regression, returns on the market portfolio are regressed on individual stock returns, 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 resulting portfolios associated with each N_h trade off the regression

goodness of 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.

Figure 3 plots the results from this exercise, implementing the procedure using weekly total realized returns over the period September 2010 through August 2011, and stopping at $N_h = 50$. The height of the grey bars in panel A of the figure show the maximum obtainable Sharpe ratio associated with each value of N_h on the x-axis. The plot shows that the constrained optimum Sharpe ratio benchmark derived from the lasso procedure doubles as N_h increases from 1 to 5, with more modestly rising values up to around $N_h = 24$ and small gains beyond that point. For optimal portfolios of 25 or more stocks, the Sharpe ratio is very close to or virtually identical to that of the market portfolio, which is shown as a green bar. For the 2% of accounts with 50 or more stocks, we therefore assume that the market Sharpe ratio is the constrained optimal Sharpe ratio.

The red triangles in panel A show the location of the median estimated Sharpe ratio of investors' actual stock portfolios observed in the data over the same time period. These values are below the empirical benchmark estimated using the lasso approach for all values of N_h, and the gap is especially large when N_h is low.

Following the approach of Calvet et al. (2007), but taking N_h as given, we define the Relative Sharpe Ratio statistic for household h (RSR_h) as the ratio of the empirically observed Sharpe ratio for household h and the constrained optimum Sharpe ratio associated with N_h. In Panel B of Figure 3, we plot the entire distribution of RSR_h statistics for each value of N_h. The figure shows that RSR_h tends to rise with N_h, meaning that holding just a few stocks is associated with substantially worse performance than the constrained optimum Sharpe ratio, and that holding larger numbers of stocks is associated with performance that is closer to the constrained optimum. This result could reflect the role of financial sophistication in jointly determining performance and N_h, or simply reflect underlying heterogeneity in investors' preferences for taking idiosyncratic risk.

A second feature evident from the plot is that there is also substantial variation in RSR_h within groups of households that share the same N_h. This suggests that models of the type we have considered, which rely on N_h as the principal explanatory factor, will struggle to explain the full range of variation in the data. Interestingly, as N_h rises, the cross-sectional variation in RSR_h appears to grow smaller, in addition to the tendency for RSR_h to increase that was noted earlier.

To better explain this cross-sectional variation in portfolio construction and performance we now estimate multifactor models of household portfolio choice..

5 Multifactor Models of Stockholdings

In this section, we apply the framework developed in Section 2 to build our understanding of how households choose stocks and construct portfolios. We describe our multifactor analysis of the cross-section of households observed in August 2011, our latest complete cross-section of data, and later verify the stability of our inferences when re-estimated on earlier periods of the data.

We begin this section by briefly describing the construction of the factors that we use in our empirical analysis before presenting a first set of results from a richer multifactor model. We conclude this section by contrasting the observed multifactor model results with those obtained from an unobserved (principal-component-based) factor analysis of stockholdings.

5.1 Observed Multifactor Model

5.1.1 Factor construction

We construct household-specific and household-portfolio-specific factors from the account and stockholding characteristics summarized in Table 1. We add three 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.[18]

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 35 factors also includes a dummy to indicate single-stock accounts.[19]

5.1.2 Estimation and results

We estimate stockholdings using all $K = 35$ 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.

[18]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.

[19]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 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.

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 small stocks to have positive β on the factor that measures the average size rank of households' stockholdings. This allows us to capture fairly complex patterns of portfolio construction, which we discuss in greater detail below. For ease of interpretation, we first divide each factor by its unconditional standard deviation in each stock-specific regression, and then scale the estimated $\hat{\beta}_{ik}$ by dividing it by the estimated holding probability of the stock $\hat{Q}_i$ and expressing the result as a percentage, i.e. $\tilde{\beta}_{ik} = 100\hat{\beta}_{ik}/\hat{Q}_i$. $\tilde{\beta}_{ik}$ is then the percentage 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.[20] The first four columns of the table therefore summarize the cross-stock distribution of $\tilde{\beta}_{ik}$, presenting the cross-stock standard deviation, and the 10th, 50th, and 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 such cases, the fraction of coefficients that are

[20] In the asset pricing context, this is analogous to the mean capitalization-weighted multifactor betas on factors other than market returns being zero.

statistically significant at the 5% level also far exceed 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 loadings show that every stock is more likely to be held as N_h increases, though this tendency varies enormously. Stocks at the 10th percentile of $\tilde{\beta}_{ik}$ when k is N_h are roughly 175% more likely to be held, while those at the 90th percentile are 747% more likely to be held for a one-standard deviation in N_h. For reference, Table 1 shows that the cross-household standard deviation of N_h is 16.48, which is almost twice the cross-household average N_h.

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, have portfolios that are more concentrated in a smaller number of stocks.

There is suggestive evidence of geography-based stock selection, which, as we show later, is mainly driven by local bias à la Coval and Moskowitz (1999). However, this is

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.[21] The table divides factors into five categories, namely, Fama and French (1993) style 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 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 size (market capitalization) and popularity (ranking based on $\hat{Q}_i$) of other stockholdings are the two most useful factors for predicting whether a household will hold a given stock.

Holdings of stocks with high turnover and high past-year realized returns are the next most predictive factors. In comparison with these characteristics, stockholdings-based factors based on book-to-market and paid dividends have relatively weaker power to predict stockholdings. Finally, stockholding concentrations in particular business groups and industries exhibit modest variation in cross-sectional predictive power, but as we will see in the next subsection, these factors are helpful at identifying groups of investors that are more closely focused on these attributes.

5.1.3 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

[21] Once again, since our regressions use stockholding characteristics that exclude stock i when estimating β_{ik}, there is no mechanical relationship between stocks' characteristic k and their estimated $\hat{\beta}_{ik}$.

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 normalization means that the aggregate R^2 apportions equal weight to each stock seen in the data.[22]

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 1.75%. 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

[22]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.

express the contribution as a percentage of the full-model R^2. For example, the table shows that account-characteristic factors contribute roughly 80% 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 an important role, though it is not the only important factor. Account size and account age are also relatively 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, business 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. These business group factors are particularly effective at predicting the holdings of popular business-group-affiliated stocks, but relatively less effective at predicting stockholdings of less popular business-group-affiliated stocks.

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.

5.2 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 once again normalize the stockholdings data by the inverse standard deviation of the aggregate holding probability of each stock before computing the covariance matrix of stockholdings.

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.

5.2.1 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 $r+$

1th 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 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 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.

Figure A.6 in the online appendix shows the factors that are counted as "natural rate" under the parameter value $\delta = \frac{1}{5}$ suggested by Connor and Korajczyk (2019). The plotted values represent marginal R-squareds relative to the natural rate factor threshold (which is shown as a red line), and bars colored in blue (pink) correspond to factors that are statistically significantly above (not statistically significantly above) this threshold at the 5% level of significance.

Panel A of this figure shows that ten of our 35 observed factors significantly exceed this threshold. However, intuition suggests that many of the factors that do not meet this threshold are also likely natural rate. Specifically, our analysis below, in Section

6.2, suggests that stockholding characteristic factors are relevant as they reflect clientele effects based on pervasive stock characteristics.[23]

Panel B of the figure shows that 11 of the PCA-based factors lie above the natural rate factor threshold. However, these conclusions are sensitive to the assumption about δ. Figure A.7 in the online appendix shows that only the first principal component is classified as a natural rate factor when $\delta = \frac{1}{10}$ and the marginal R-squared threshold rises to 0.5%. Under the same threshold, four of our observed factors are considered natural rate.

5.3 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.

Table 5 relates the first 10 principal components to the observed factors in our multifactor model. Each principal component is regressed on the full set of 35 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 Table 5 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, while the second principal component is positively related to the same set of account attributes. As conjectured when discussing the cross-sectional dispersion of $\tilde{\beta}_{ik}$, this appears consistent with the stockholdings of large, well-diversified,

[23] Aside from market capitalization, popularity, age, and three of the industry factors, these characteristic clienteles are too weak (marginal R-squareds not statistically significantly above 0.11%) to be detected as natural rate. Conversely, the second-largest marginal R-squared is associated with the Reliance (ADAG) factor, which seems an unlikely natural rate factor as it derives much of its significance from its ability to predict holdings specifically in the Reliance (ADAG) business group.

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 business groups and industries, loading positively on the Reliance (ADAG) business group share of household portfolios (Panel C), as well as the share of the portfolio invested in the oil and gas industry (Panel D)

As mentioned above, the second principal component loads (positively) on account size as well as the number of stocks held and traded (Panel A), but also seems to load on stockholdings characteristics such as the size of stocks held, realized returns of stocks held, and the propensity to hold dividend-paying stocks (Panel B). This 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 5 suggests that the observed factors in the model do a good job of capturing the most important drivers of stockholdings, but also highlights the difficulty of capturing the wide variety of unobserved idiosyncratic 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, and shows that the PCA approach finds it worthwhile to spend several of its factors on identifying accounts that specifically hold such widely-held stocks.[24] In contrast, the observed factor model does not include factors specifically geared to match stockholdings in specific stocks, though it is fair to say that the business-group factors may serve a similar purpose. 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]

6 Insights from Multifactor Models

In this section, to better understand the structure of coholdings, we first compare empirically observed coholdings with those generated by our factor models. We then relate the estimated factor loadings to the characteristics of particular stocks to learn about "who owns what" in the Indian data. We conclude the section by relating Indian stocks' coholdings matrix with the covariance matrix of stock returns.

6.1 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 diag-

[24]This pattern is even stronger when the PCA approach is "unweighted," i.e., with the most widely-held stocks contributing disproportionately to the total variance of coholdings.

[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 ($R^2 = 7.3\%$) has a better fit than the first principal component alone ($R^2 = 2.6\%$), though the 10 principal component model fits better overall ($R^2 = 9\%$). However, in this case, the observed multifactor model fits better for most stocks, while the unobserved factor model focuses on explaining stockholdings in the most widely-held stocks.

onal elements. As highlighted earlier, the diagonal elements of the (actual and predicted) 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 (y-axis) against their model-implied counterparts (x-axis). Panel A of figure does this when the model is the 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.

6.2 Stock Characteristics in Relation to Factor Loadings

Our factor model βs are unconstrained estimates, but we now attempt to understand their relationship with stock characteristics to enable a deeper interpretation of the sources of stockholding predictability. Our approach is to regress each of our set of 3,103 $\tilde{\beta}_{ik}$'s (scaled as described in Table 3) on the full set of characteristics used to construct stockholding characteristic factors.

Table 6 shows the results of this exercise. Each column of the table shows regression coefficients from a cross-sectional multiple regression where each observation is a particular stock. The dependent variable in each such regression is the stock-specific factor beta on the factor listed in each column label, and the rows show the regressors, which are

stock-specific characteristics. Panel A of the table presents these results for account-characteristic factors, and panel B for stockholdings-characteristic factors.

The independent variables in these regressions are transformed into ranks between 0 to 1. Consequently each coefficient represents the difference in $\tilde{\beta}_{ik}$ between stocks with the highest and lowest characteristic listed in the rows. Put differently, these coefficients capture how the sensitivity of the stockholding probability to a one standard-deviation increase in the factor varies across stocks of different types. The signs and magnitudes of these coefficients are illustrated with colors ranging from deep red (largest positive) to deep blue (largest negative). While most coefficients are highly statistically significant, those which are statistically insignificant at the 10 percent level are presented in light gray font. The table also shows that the R^2 statistics in these cross-sectional regressions are high—with the majority above 40%, which shows that coholdings factor loadings are well-explained by underlying stock characteristics.

The importance of this exercise can be more clearly seen when discussing specific coefficients. For example, Panel A shows that $\tilde{\beta}_{ik}$ on N_h, the number of stocks held in the account, is strongly predicted by the attributes of the stock under consideration. In particular, stocks with lower market capitalization, lower popularity, higher book-market, which are older, lower turnover, and with higher realized returns and lower realized volatility tend to be held more by accounts with high N_h, controlling for other account and stockholding characteristics.

To interpret magnitudes, consider moving from the smallest to the largest stock in the sample, controlling for all other stock attributes. Table 6, first row, fourth column shows that the decrease in the baseline probability that the largest stock is held as N_h increases by one standard deviation is 290%. Similarly, moving from the least to the most popular stock decreases its baseline holding probability by 468% for a one standard deviation increase in N_h. This relates to our observation in Figure 2 that the most widely held stocks have relatively lower loadings on N_h than would be expected if stocks

were randomly selected.[26] The table also shows that holding N_h and other account characteristics constant, accounts that are larger or older are more likely to hold large and popular stocks.

Panel B presents coefficients from regressions of stockholding characteristic $\tilde{\beta}_{ik}$ coefficients on underlying stock characteristics. 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. Any positive coefficients along the diagonal are therefore evidence of clientele effects associated with the stock characteristic (e.g., investors who hold large stocks are more likely to wish to own other large stocks). 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).

The panel shows strong evidence in favor of clientele effects, with statistically significant and positive coefficients along the diagonal for every stock characteristic in Panel B. These clientele effects are particularly strong for stock market capitalization and stockholding popularity, and to a lesser extent, for recent realized stock returns, stock age, and stock turnover. For each standard deviation decrease in the market capitalization of investors' (other) stockholdings, the probability of holding the smallest stock relative to the largest stock increases by 520% (relative to its baseline holding probability). In contrast, evidence suggests that relatively few investors seek to build value-oriented portfolios, as proxied by either book-market ratios or whether the stock pays a dividend.

Off-diagonal elements can be interpreted as evidence of clienteles for portfolios of stocks with several attributes that investors regard as related. For example, Table 6 Panel B shows that high market capitalization and low volatility stocks are more frequently held in portfolios with high market capitalization stockholdings and low realized volatility

[26]Recall that our popularity measure represents the number of stocks holding the account, orthogonalized to market capitalization.

stockholdings, suggesting that there may be a clientele of investors seeking "stable" stock investments.

Online appendix Table A.2 shows the coefficients from regressions of stockholding business group $\tilde{\beta}_{ik}$ on business group indicators. Interestingly, while there is a significant clientele effect for a few of the large business groups (such as Reliance ADAG, Mahindra, and Jindal), the clientele effect is quite small for many of the other business groups. Most off-diagonal elements are small, but tend to be positive, suggesting that some investors have a preference for business-group-affiliated stocks over non-business-group stocks rather than a preference for some business group affiliations over others.

Finally, online appendix Table A.3 shows coefficients of similar regressions of $\tilde{\beta}_{ik}$ on account location and industry stockholdings dummies on stock headquarter location (panel A) and stock industry group dummies (panel B) respectively. Panel A shows that account location betas reflect a proximity bias of the sort documented by Coval and Moskowitz (1999). For example, compared to an investor in East India, an investor located in South India has a 74% higher probability of holding a South Indian headquartered stock than an East Indian headquartered stock. This finding illustrates the way in which our approach integrates insights about portfolio construction that have hitherto been documented in isolation from one another.

Panel B of online appendix Table A.3 shows that stockholding industry betas are related to industry-focused clienteles. A one standard deviation increase in the share of the portfolio invested in oil and gas, which reduces the share invested in construction (the omitted dummy), results in a 25% increase in the probability of holding an oil and gas stock relative to a construction stock.

6.3 Stock Return Covariance and Coholdings

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 (y-axis) and "raw" estimated coholdings correlations (x-axis), while Panel B of the figure replaces raw coholdings correlations with observed-factor model-implied coholdings correlations.

The figure shows that there is 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 (Panel A), i.e., the correlation between these two correlations is roughly $\sqrt{0.1}$ or 32%. This rises to $R^2 = 20\%$, or a correlation of roughly 45%, when return correlations are regressed on model-implied coholdings correlations (Panel B).

While these are very preliminary observations, the clear positive relationship observed between return correlations and coholdings 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. More generally, in equilibrium asset pricing models holdings and returns are jointly determined, and different models have different implications for the relationship between them. These plots warrant further investigation, as they are a first step to more deeply understanding the empirical relationships between holdings and returns.

7 Conclusion

In this paper we have suggested that a factor model for investors' 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.

The single most important factor in our model is the number of stocks held, implying that concentrated and diverse portfolios hold different types of stocks. Stocks that are generally popular (widely held) are particularly popular "entry-level" stocks for concentrated portfolios. The extreme example of this phenomenon is Reliance Power, a stock that is held by 40% of Indian equity accounts. Other account characteristics such as size (market value) 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]

Stockholding characteristics are less important for portfolio choice, but we do detect investor clienteles for small-cap stocks, popular stocks, and stocks with high turnover and past realized returns. Interestingly, value appears to have only a weak clientele effect

[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).

whether measured by the book-market ratio or by dividend payments.

The loadings of stocks on cross-sectional factors have a generally plausible relationship to the characteristics of those stocks. Since popular stocks are particularly popular in concentrated portfolios, they have a low loading on the factor representing the number of stocks held in an account. Large and popular stocks are more likely to be held in larger, older accounts. Clientele effects are revealed by a tendency for stocks with a particular characteristic to load on factors that measure the same characteristic in the other stocks held by an account.

Finally, we present a preliminary finding that stocks that tend to be coheld also tend to correlate more strongly with one another. This runs counter to the view that investors optimally diversify their portfolios conditional on a constraint on the number of stocks held, but it reinforces the idea that clientele effects, captured by coholdings propensities, contribute to common variation in stock returns.

References

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. *The 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? *The 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. *The Journal of Finance 55*(2), 773–806.

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

Betermeier, S., L. E. Calvet, and P. Sodini (2017). Who are the value and growth investors? *The 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.

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).

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. *The 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? *The Journal of Finance 66*(3), 1011–1041.

Fama, E. F. and K. R. French (1992). The cross-section of expected stock returns. *The 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*.

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

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. *The 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.

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. *The 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: Journal of the Econometric Society*, 867–887.

Odean, T. (1998). Are investors reluctant to realize their losses? *The 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. *The Review of Financial Studies 23*(2), 705–739.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The 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 current (August 2011) portfolio size. Turnover is winsorized at the 99th percentile. No. Stocks is the number of stocks in our sample held by the investor. No. Stocks Traded is the number of unique stocks traded by the investor over the past year. Stockholding characteristics (except Dividend Paying) represent the average normalized (0 to 100) rank of the given characteristic across the set of stocks in our sample held by the investor. Popularity is the fraction of households that hold the stock, orthogonalized to the market capitalization of the stock. Stock Age is the number of months since the stock began publicly trading. 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. Dividend Paying represents the investors' share of stockholdings that paid a dividend over the past year.

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						
Popularity	81.92	15.73	73.89	84.55	94.35	99.11
Stock Age	40.16	23.98	20.44	40.59	55.23	70.19
Book/Market	38.53	18.90	27.65	36.33	47.88	62.21
Market Capitaliztaion	83.94	18.31	77.90	90.55	96.92	98.46
Turnover	59.23	17.73	50.92	59.32	69.94	81.52
Realized Volatility	41.32	19.14	29.48	41.94	51.75	64.94
Realized Returns	45.30	19.46	32.37	46.61	58.24	68.76
Realized Skewness	41.45	18.12	29.69	41.83	52.18	63.33
Dividend Paying	0.73	0.32	0.57	0.82	1.00	1.00

Table 2

Observed vs. Expected Holding Probability with Popularity-based Selection

This table shows the ratio of observed stockholding probability to the expected stockholding probability from a popularity-based selection model. The rows group stocks by their stockholding probability, Q_i, the average of which is presented in the first column. All other columns present the ratio of stockholding probabilities grouped by the number of stocks held in the account.

	Average Q_i	One stock held	2-10 stocks	11-25 stocks	26-50 stocks	51+ stocks
Reliance Power	0.399	2.315	1.063	0.760	0.729	0.792
Top percentile	0.072	0.882	1.161	1.055	0.895	0.718
90-99th percentile	0.012	0.947	0.983	1.035	1.046	0.931
50-90th percentile	0.002	0.678	0.832	0.955	1.071	1.218
Bottom half	0.0003	0.880	0.874	0.893	0.964	1.279

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 standardize the coefficients as follows: $\frac{\hat{\beta}_{ik}}{Q_i} \times 100$, where $\hat{\beta}_{ik}$ is the coefficient from the regression with each factor scaled by its unconditional standard deviation, and Q_i is the estimated holding probability of stock i. 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 statistically significant at the 5% level.

Panel A: Account Characteristics

	Std. Dev	10%	50%	90%	Avg. |t-stat|	% Significant (5% level)
Age	33.36	-60.04	-18.91	23.33	21.93	95.97
Size	73.91	-178.26	-60.69	0.90	40.53	97.71
Turnover	14.56	-16.72	2.08	19.20	10.66	83.69
No. Stocks	225.36	174.82	397.43	747.04	265.56	100.00
No. Stocks Traded	71.00	-97.78	-16.92	72.92	40.29	95.46
Geographic Region						
Southern	30.78	-18.38	3.95	36.29	11.47	77.60
Northern	25.13	-15.58	3.96	23.60	8.60	76.02
Western	32.13	-27.68	4.12	34.52	12.78	86.01

Panel B: Stockholding Characteristics

	Std. Dev	10%	50%	90%	Avg. |t-stat|	% Significant (5% level)
Fama-French factors						
Market Capitalization	207.05	-488.20	-183.48	45.29	31.77	97.39
Book/Market	34.73	-50.95	-5.73	24.65	9.24	79.05
Return-based factors						
Realized Volatility	36.59	-13.74	20.27	67.82	10.02	86.56
Realized Returns	52.24	-38.08	11.20	83.74	7.57	78.92
Realized Skewness	31.04	-36.38	2.02	27.51	5.36	63.36
Behavioral factors						
Popularity	132.58	-262.20	-90.54	67.26	22.71	94.52
Stock Age	40.31	-45.65	5.18	47.40	15.84	81.86
Turnover	74.36	-99.10	21.03	68.85	13.54	91.30
Dividend Paying	32.36	-30.56	7.63	49.92	9.67	83.66
Business Group Holdings						
Reliance (ADAG)	23.26	-11.49	18.90	48.36	15.03	92.91
Tata	8.21	-4.73	4.38	15.14	6.15	78.86
Reliance (DAG)	22.11	-4.06	21.03	52.99	15.06	95.04
Birla Aditya	6.30	-5.66	0.87	8.01	4.49	59.62
Jaypee	7.87	-4.61	4.99	14.10	6.22	81.99
Jindal	6.52	-5.12	-0.33	5.49	3.95	49.53
Mahindra	7.41	-3.60	5.16	14.03	5.90	78.63
Suzlon	10.85	-7.29	6.66	21.03	8.06	85.95
Vedanta	5.72	-10.18	-3.62	4.36	5.26	74.64
Adani	5.40	-5.64	-1.35	4.37	4.76	52.85
Others	21.78	-29.63	-4.26	24.32	12.56	88.59
Industry Holdings						
Financial Services	17.59	-2.64	19.36	37.85	10.53	94.07
Food, Agri. and Textiles	15.93	-10.82	4.75	18.65	8.14	75.15
Information Technology	18.25	-22.23	0.75	20.74	8.34	78.15
Manufacturing	26.56	-21.58	13.01	41.68	8.83	76.96
Oil and Gas	29.77	-3.00	38.50	70.40	14.86	93.75
Other Retail	15.37	-18.03	-3.24	8.72	7.56	79.44

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. In the column, 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. The stock level weights used to construct R-squareds is just as in Panel A.

Panel A: Observed Factor Model

	Equally Weighted
Full R-squared	1.75
	Percent of Full R-squared
Account Characteristics based Factors	79.96
No. Stocks	18.15
Size	15.05
Age	5.35
Turnover	0.51
Geographic factors	0.36
Stockholding Characteristics based Factors	8.98
Business group factors	1.58
Industry factors	1.35
Behavioral factors	1.25
Return factors	1.28
Fama-French factors	0.96

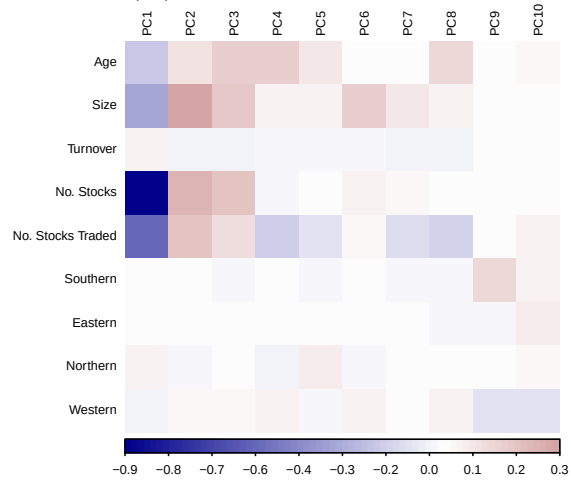
Panel B: Unobserved PCA Factor Model

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

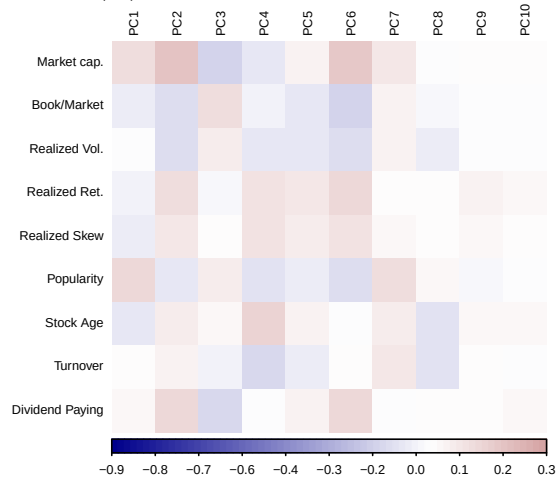
Table 5
Loadings of Observed Factors on PCA 1-10 Factors

Panels A, B and C present a heatmap of the loadings of observed factors (in rows) on a multiple regression with PC 1-10 factors (in columns), 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.

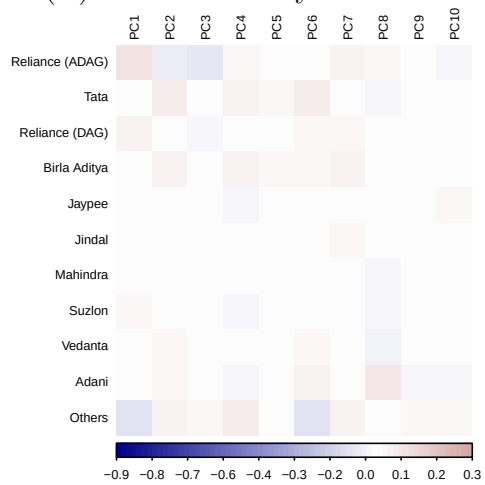
Panel (A): Account Characteristics



Panel (B): Stockholding Characteristics



Panel (C): Stock Industry Characteristics



Panel (D): Stock Business Group Characteristics

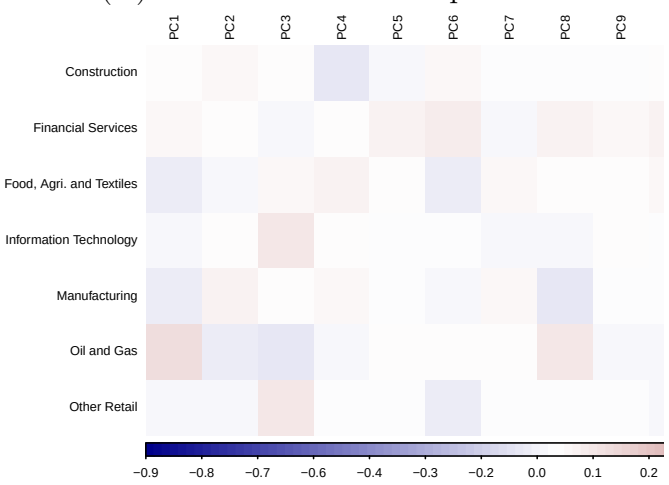
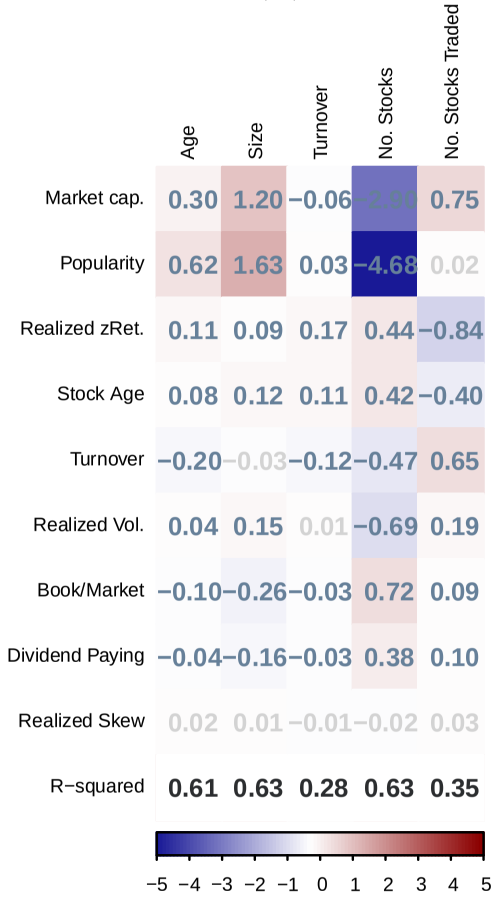


Table 6

Stock characteristics and Observed Multifactor Betas

Each column in Panel A and B presents coefficients from the regression of an account characteristic factor beta (Panel A) or stockholding characteristic factor beta (Panel B) on stock characteristics. The factor betas (dependent variables) are scaled just as in Table 3. In addition to the rows of these tables, stock characteristics in the regressions include indicators for the stocks' business group affiliation (if any), headquarters zone, industry, and dummies to indicate missing characteristics. Stock characteristics are rank normalized, so coefficients represent the change in (scaled) factor betas associated with moving from the lowest to highest in the given stock characteristic. The colors represent effect size, with effects that are statistically insignificant at the 10 percent level presented in light gray.

Panel (A) Account characteristics



Panel (B) Stockholding characteristics

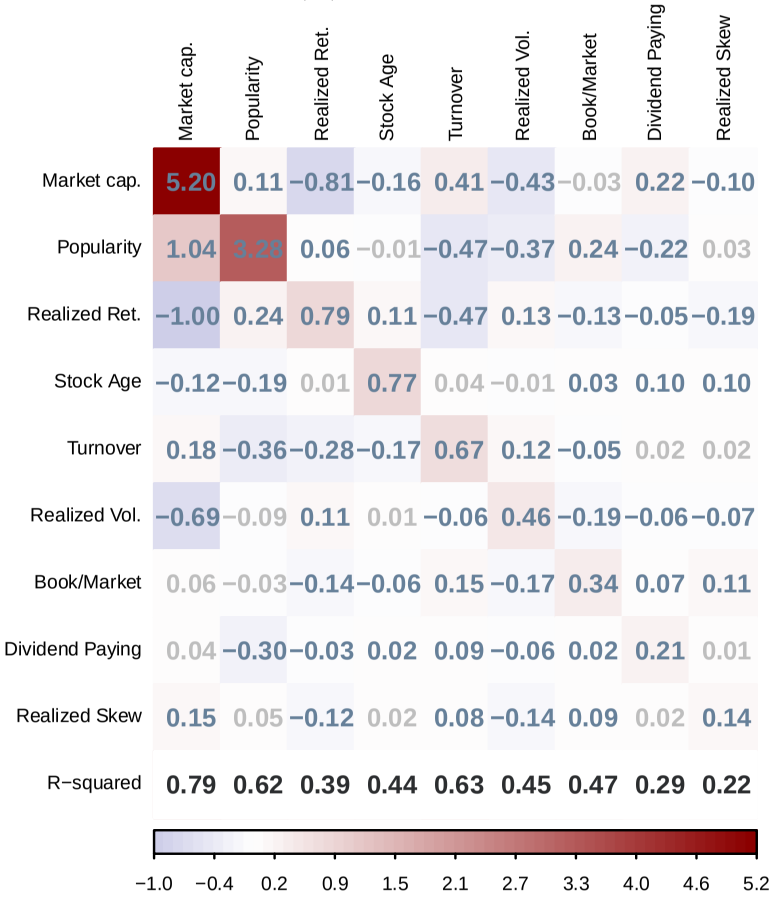


Figure 1
Number of Investors per Stock

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%).

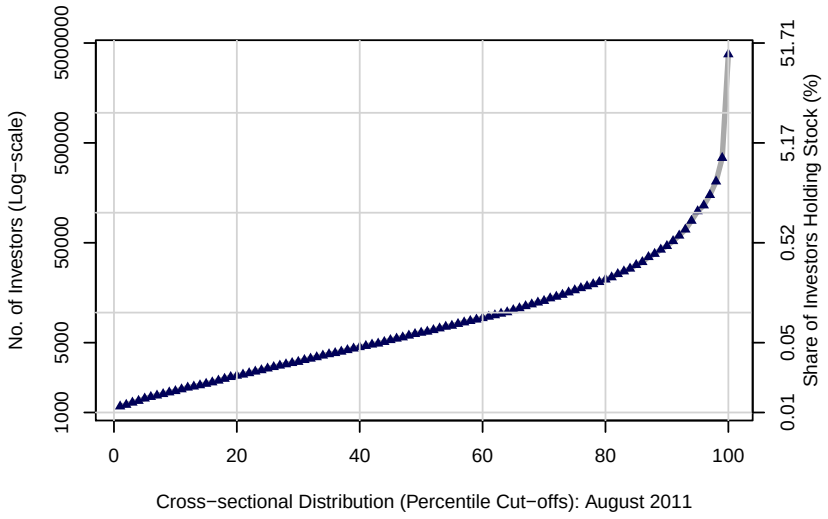


Figure 2
Observed Over Expected Stockholding Probabilities
Popularity-based Stock Selection

This heatmap presents the ratio of observed over expected stockholding probabilities for accounts grouped by the number of stocks held (columns) and stocks grouped by the percentile of the unconditional stockholding probability distribution (rows). Cells in white refer to the ratio being 1, while cells in various intensities of red (blue) mark the deviation above (below) 1. The top row splits the top percentile into all stocks except Reliance Power, and includes a separate row for the Reliance Power stock. The height for each row corresponds to the corresponding percentiles of the unconditional stockholding probabilities.

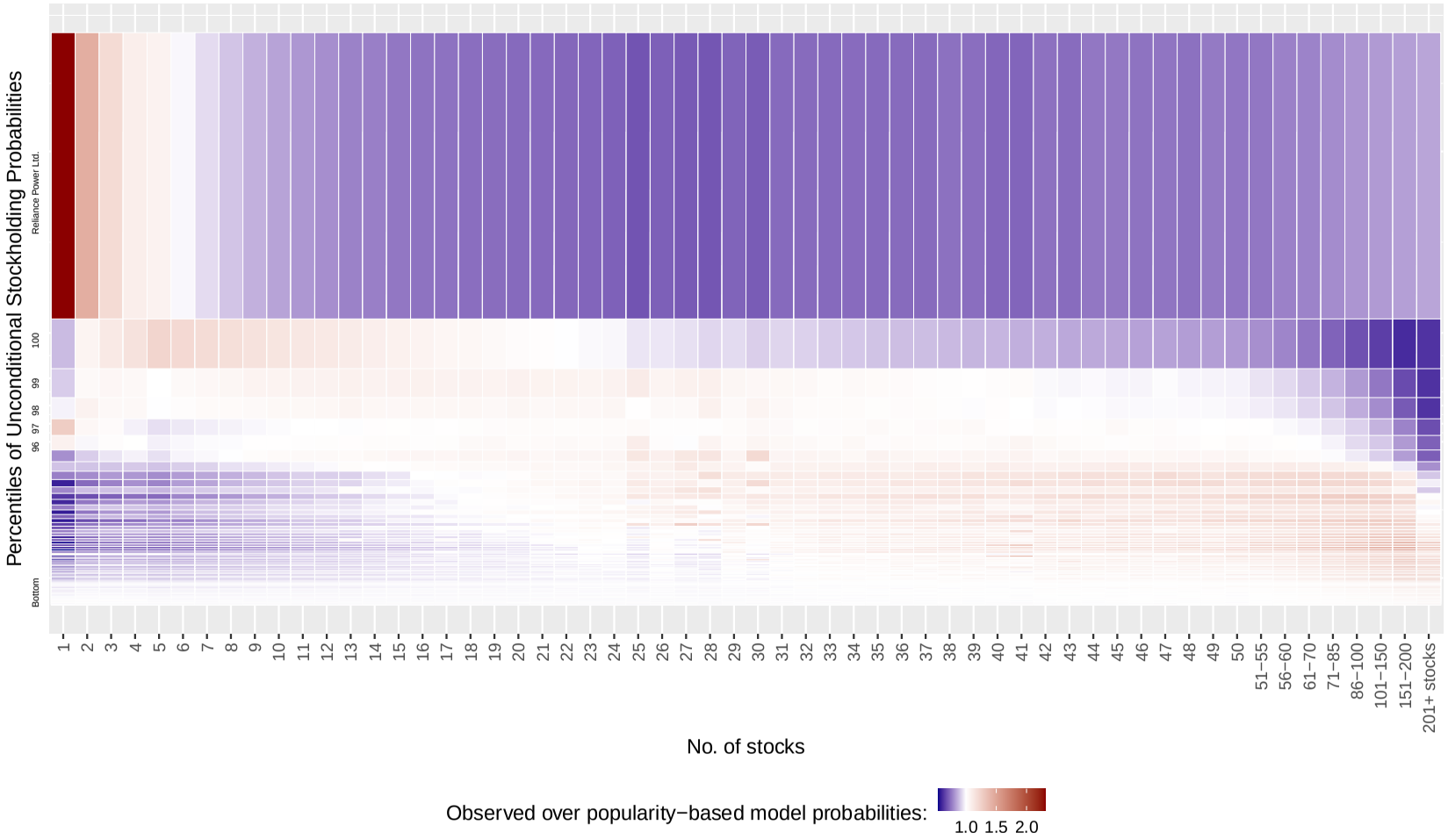
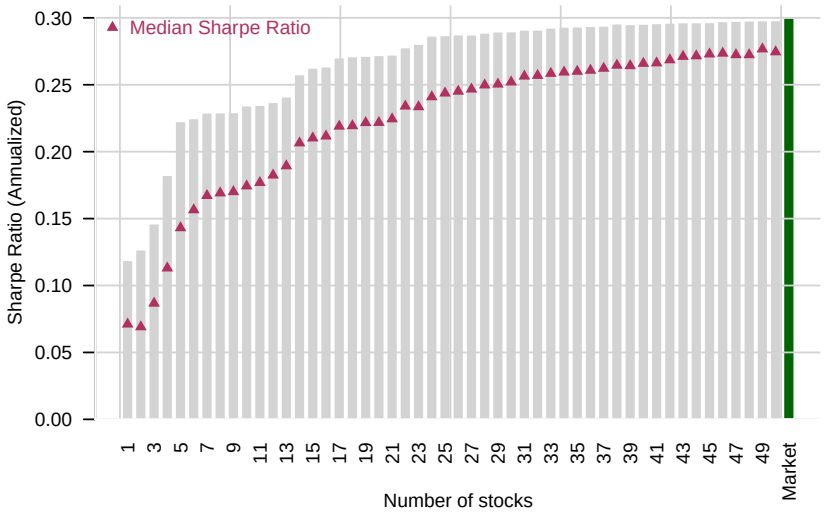


Figure 3
Relative Sharpe Ratio (RSR)

Panel A presents the annualized sharpe ratio from the best N stock portfolio as described in Section XX of the paper. 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 a year leading up to August 2011. The triangle plots the median CAPM implied Sharpe ratio for households in our data, for the same time period. Panel B plots the relative sharpe ratio (RSR), the CAPM implied sharpe ratio of household h (S_h) is scaled by the sharpe ratio of the best N stock portfolio, $\frac{S_h}{S_{N_h}}$. We restrict this analysis to accounts with no more than 50 stocks. We present the cross-section distribution in the value of the relative sharpe ratio as a heat-map, with the red (black) line marking the median (mean) of the cross-sectional RSR distribution for each value of N_h.

Panel A: Best Sharpe Ratio Estimates



Panel B: RSR Distribution by Number of Stocks

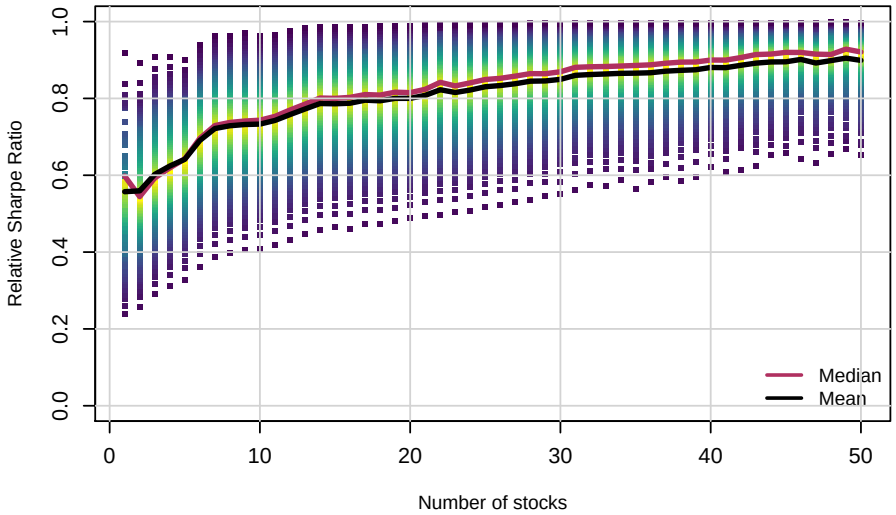
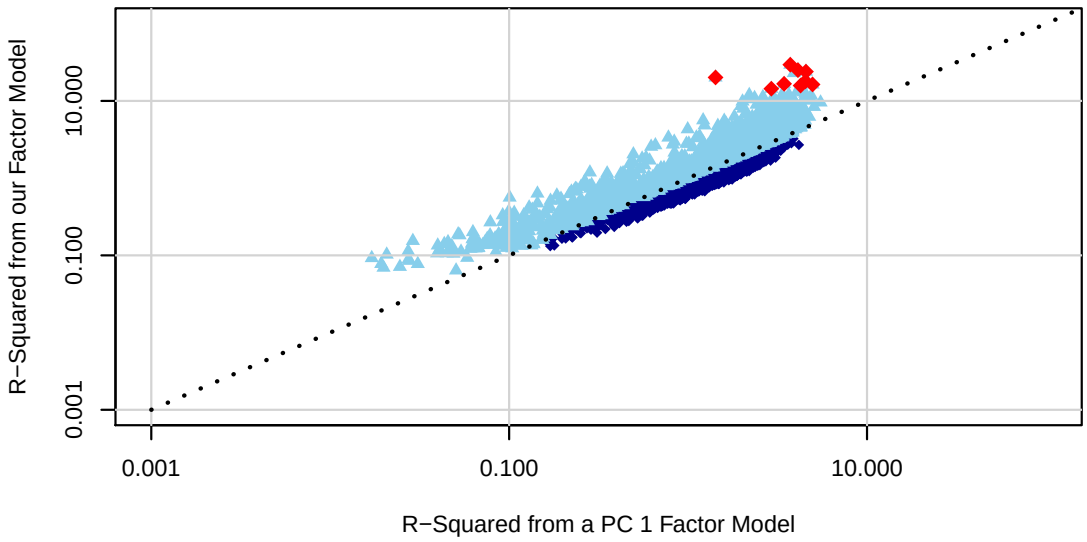


Figure 4
Comparison of Stock Level R-Squareds

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

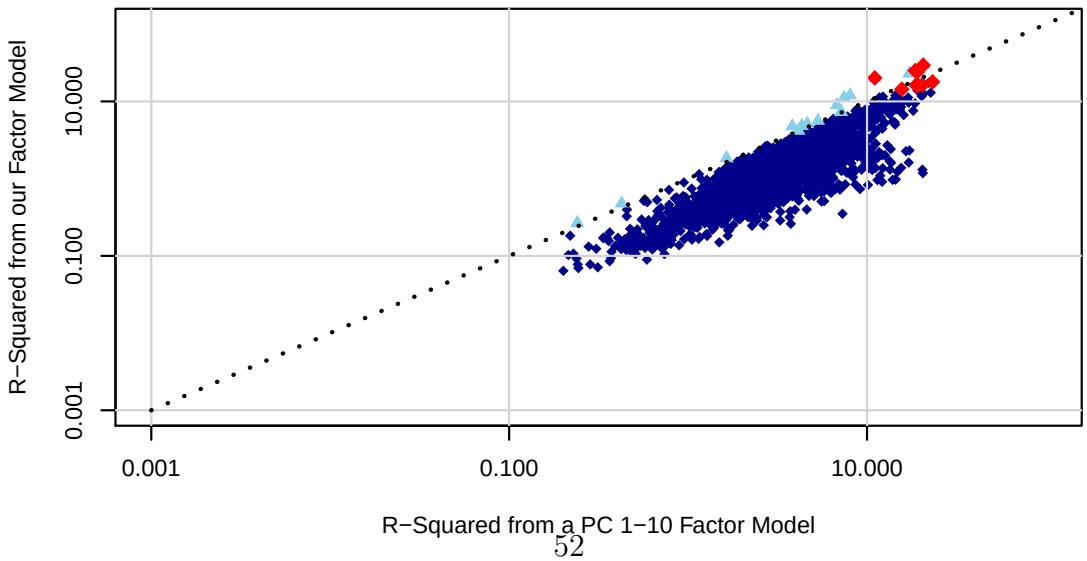
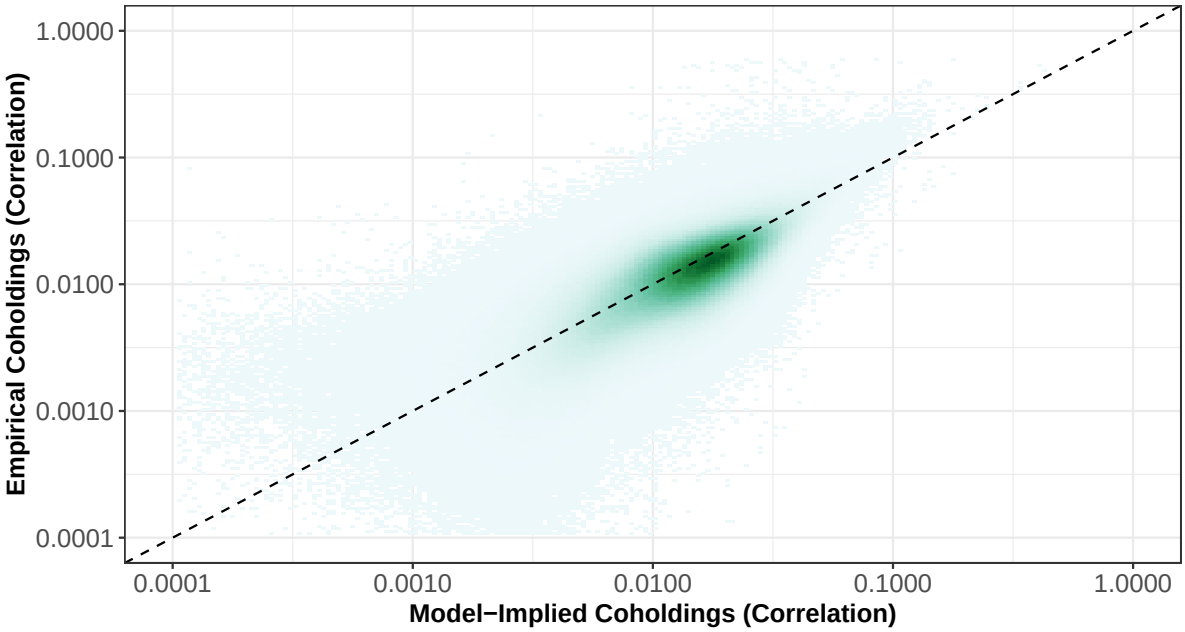


Figure 5
Empirical vs Model Coholdings

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.

Panel A: Observed Factor-model Implied Coholdings



Panel B: PCA Factor-model Implied Coholdings

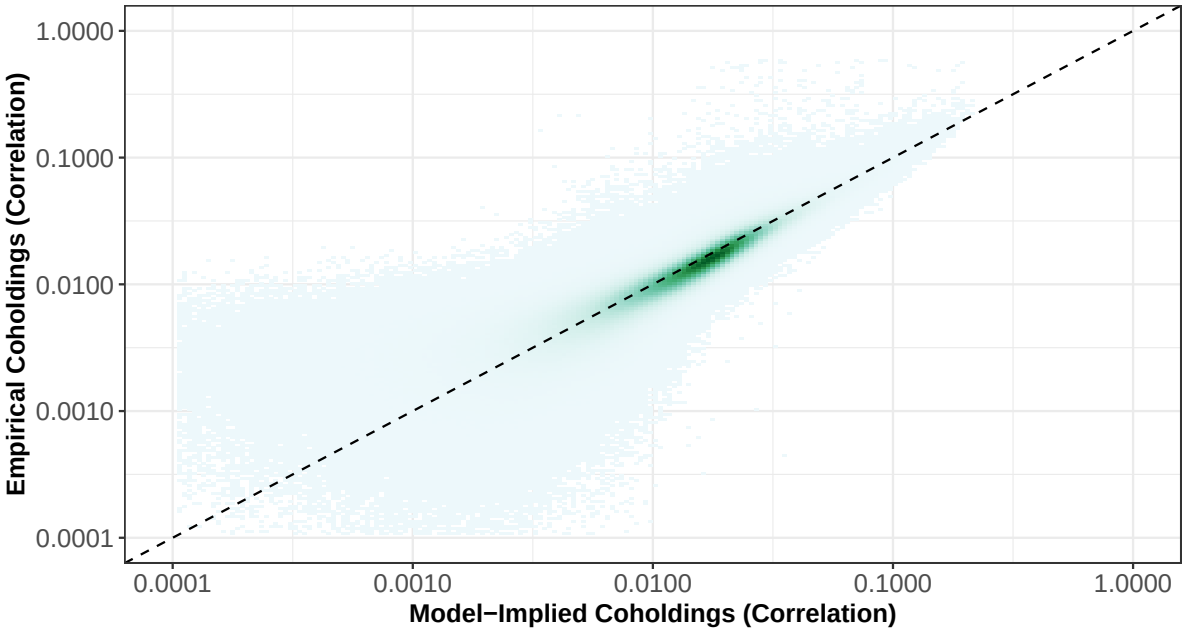
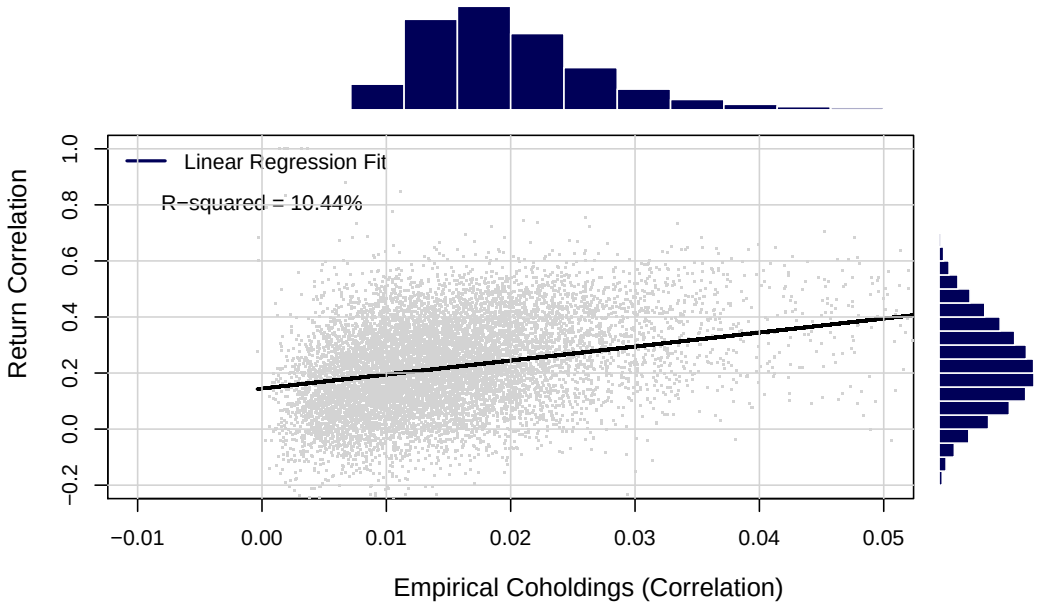


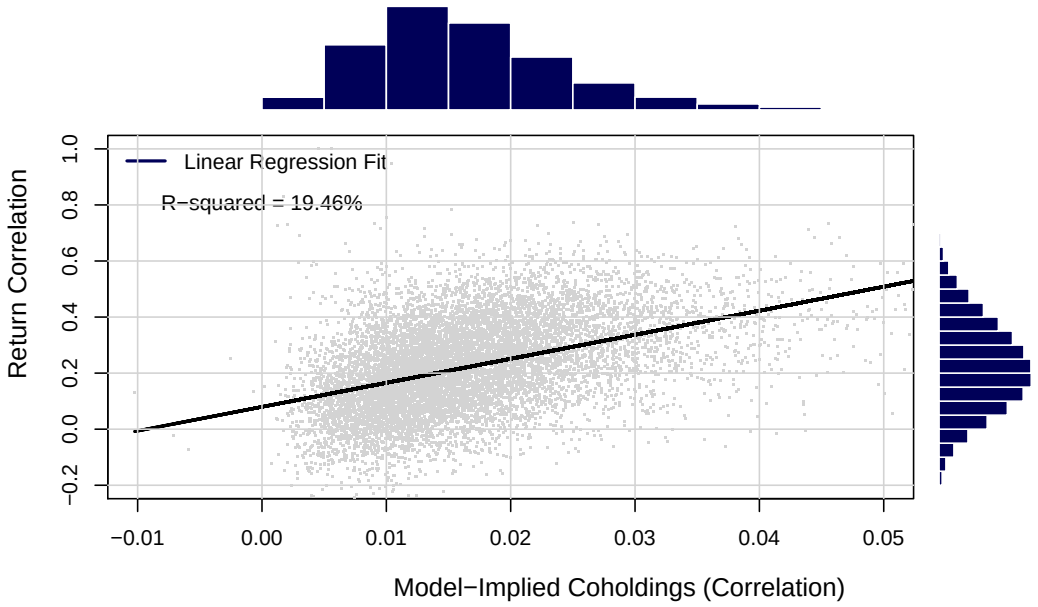
Figure 6
Return Covariance and Coholdings

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)



Online Appendix

Who Holds What?
A Factormodel for Direct Stockholdings

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	7.27	1.13
	Percent of Full R-squared	
Account Characteristics based Factors	58.59	78.21
No. Stocks	38.46	21.35
Size	35.16	8.28
Age	9.68	6.28
Turnover	2.71	0.22
Geographic factors	0.48	0.41
Stockholding Characteristics based Factors	19.89	11.87
Business group factors	5.69	1.05
Industry factors	2.12	1.40
Behavioral factors	2.03	1.32
Return factors	1.68	2.75
Fama-French factors	1.52	0.97

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

Table A.2

Business Groups, and Business Group Factor Betas

Each column of this table presents coefficients from the regression of stock portfolio business group betas on stocks' own characteristics. Rows display only coefficients on stocks' business group affiliation dummies to focus on industry clienteles–coefficients on all other characteristics are suppressed. Reported coefficients represent the increase in relative holding probability associated with stocks in the (row) business group–relative to stocks not in any business groups–as the share of (other) stocks in the portfolio shift from including none of the (column) business group to only stocks in that business group. Other aspects of the analysis and table construction follows Table 6.

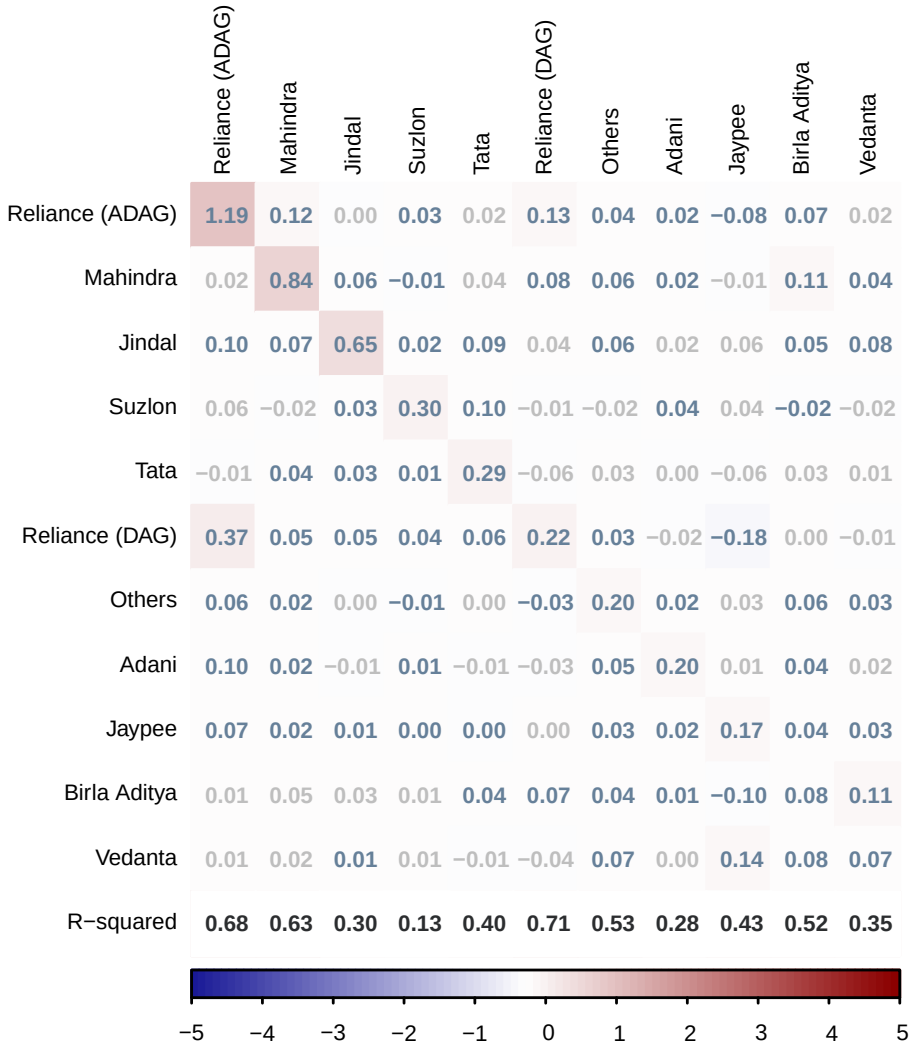


Table A.3

Geographic Zone, Industry Groups and Zone and Industry Factor Betas

Each column of the tables below presents coefficients from the regression of account zone betas (Panel A) and stockholding industry betas (Panel B) on stocks' own characteristics. Rows display only those coefficients on stocks' headquarters zone (Panel A) or industry (Panel B) dummies to focus on local bias and industry clienteles respectively. The omitted dummies correspond to Eastern India and the construction industry respectively. Reported coefficients represent the increase in relative holding probability associated with stocks in the (row) zone/industry, relative to the omitted category, when the account is in the column zone or other stockholdings are invested exclusively in the column industry. Other aspects of the analysis and table construction follows Table 6.

Panel (A) Geography

	South	West	North
South	0.74	0.47	0.38
West	0.38	0.63	0.38
North	0.38	0.45	0.59
R−squared	0.37	0.26	0.30

0.26 0.31 0.36 0.41 0.45 0.50 0.55 0.59 0.64 0.69 0.74

Panel (B) Industry

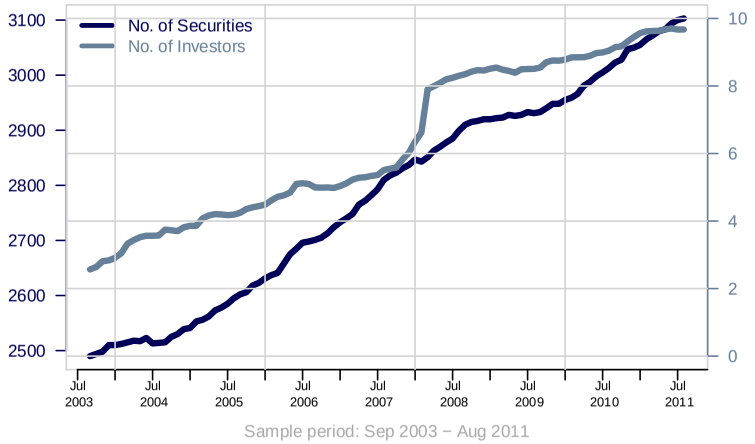
	Oil and Gas	Information Technology	Manufacturing	Financial Services	Food, Agri. and Textiles	Other Retail
Oil and Gas	0.25	0.18	0.08	0.11	0.10	0.04
Information Technology	0.16	0.24	0.06	0.07	0.15	0.00
Manufacturing	0.21	0.17	0.20	0.08	0.11	0.03
Financial Services	0.19	0.15	0.05	0.20	0.13	0.08
Food, Agri. and Textiles	0.15	0.13	0.04	0.10	0.20	0.06
Other Retail	0.19	0.11	0.07	0.16	0.12	0.14
R−squared	0.35	0.48	0.22	0.30	0.44	0.28

−5 −4 −3 −2 −1 0 1 2 3 4 5

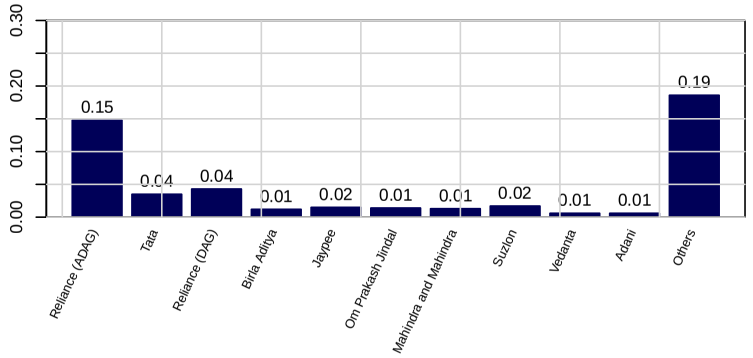
Figure A.1
Summary Statistics

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.

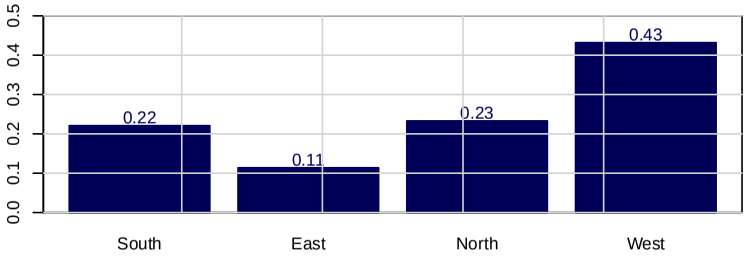
Panel A: Number of Observations



Panel B: Business Groups



Panel C: Geography



Panel D: Industry

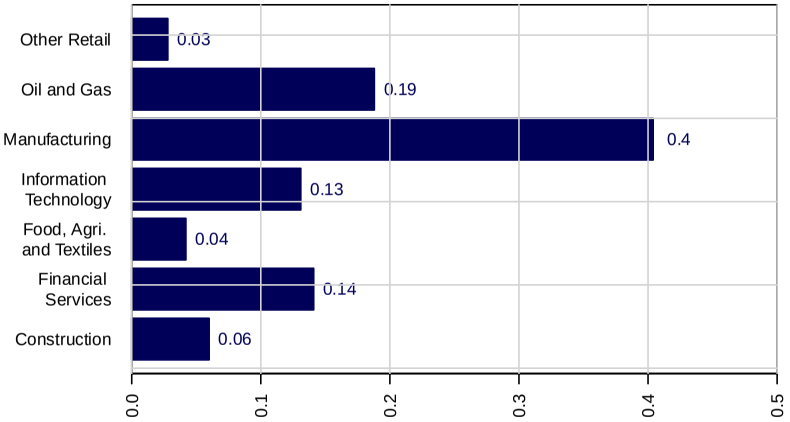


Figure A.2
Comparison of U.S. and Indian Household Stock Wealth

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.

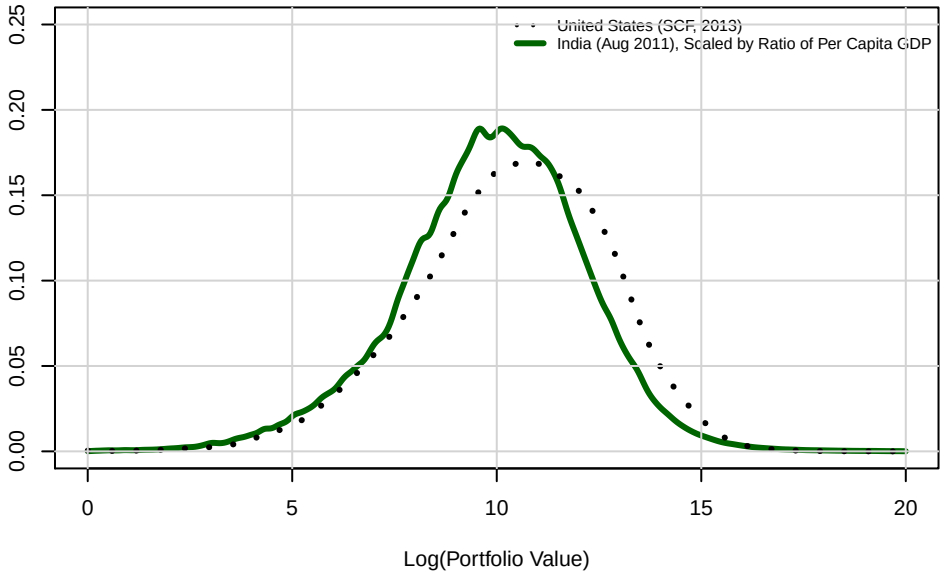


Figure A.3
Correlation Matrix

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

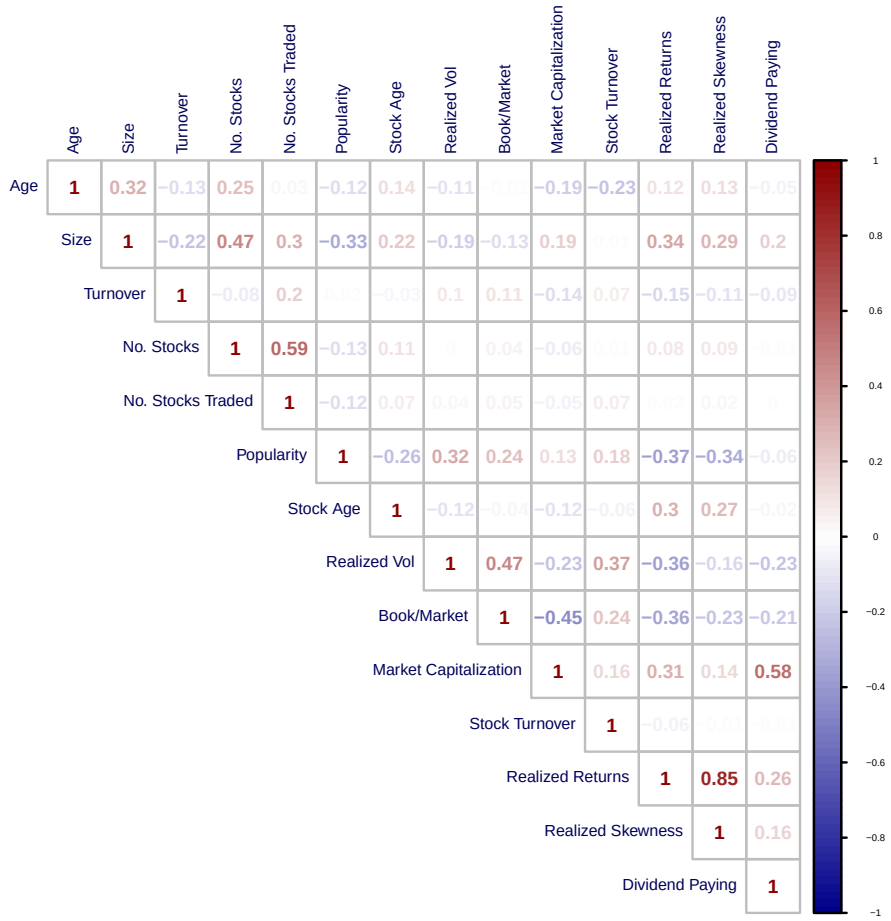


Figure A.4
Unobserved Factor Model: Principal Components

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

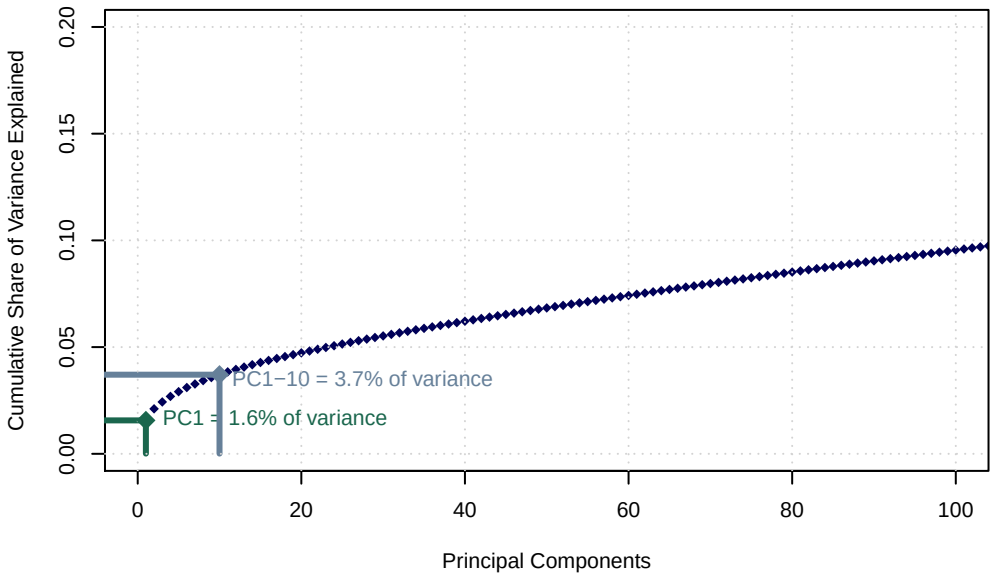


Figure A.5

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.

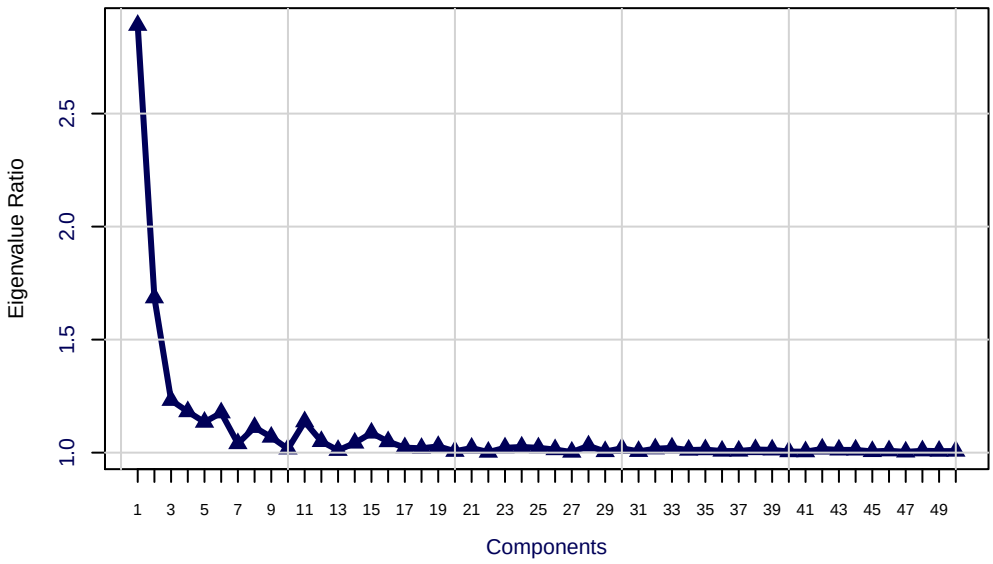
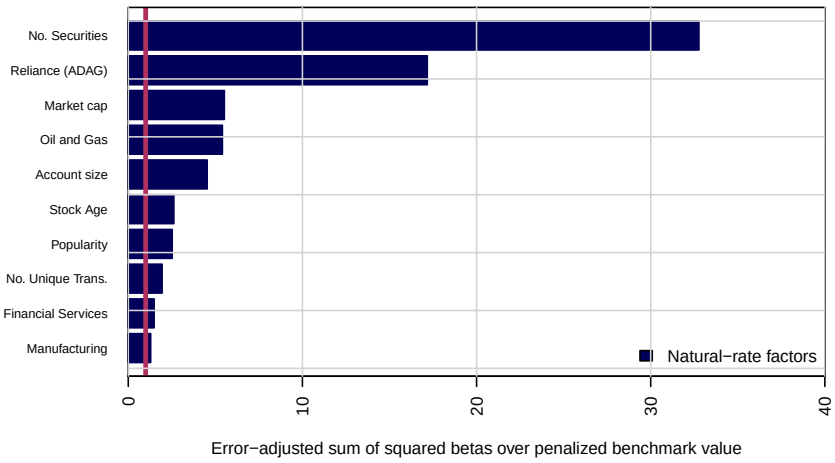


Figure A.6
Natural-rate and Semi-strong Factors

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

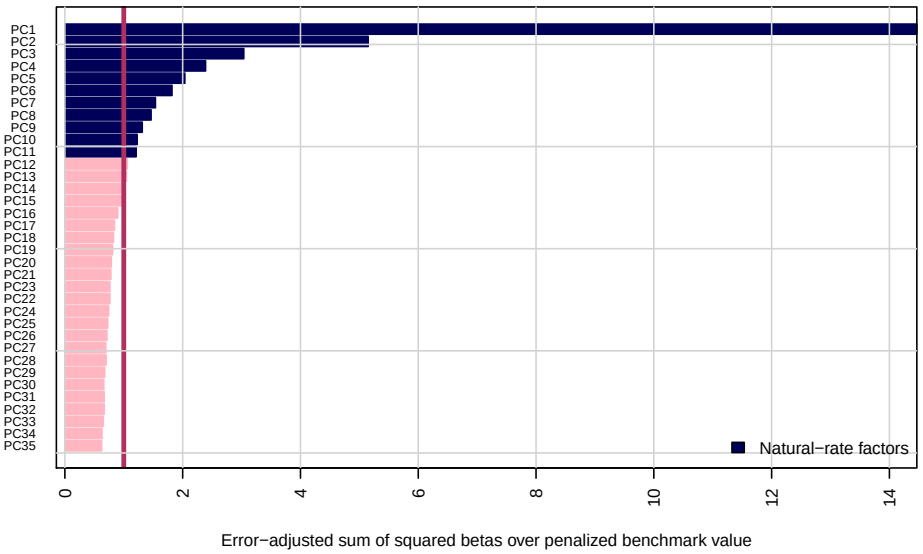
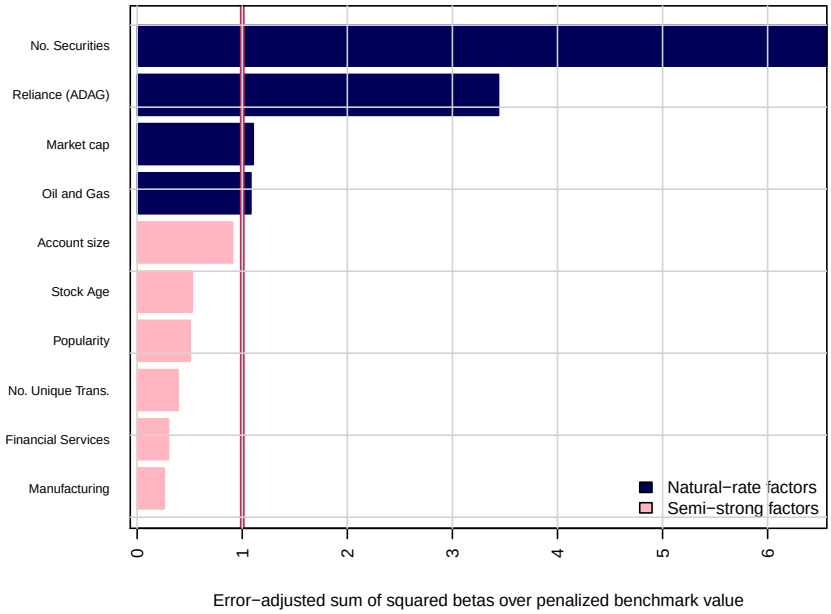


Figure A.7
Natural-rate and Semi-strong Factors

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