Costly short sales and nonlinear asset pricing*

Theodoros Evgeniou† Julien Hugonnier‡ Rodolfo Prieto§

December 2019

Abstract

We study a dynamic general equilibrium model with costly-to-short stocks and heterogeneous beliefs. The model is solved in closed-form and shows that costly short sales drive a wedge between the valuation of assets that promise identical cash flows but are subject to different lending fees. The price of an asset is given by the risk-adjusted present value of its future cash flows, which include both dividends and an endogenous yield derived from lending fees. This pricing formula implies that asset returns satisfy a modified capital asset pricing model which includes a negative adjustment for lending fees and, thus, provides a theoretical foundation for the recent findings on the role of lending fees as an explanatory variable of stock returns. Empirical results are consistent with the theory proposed.

Keywords: Shorting fees; Endogenous participation; Equity lending; Heterogeneous beliefs; Rational agents; Dynamic equilibrium.

JEL: D51, D52, G11, G12.

[Preliminary draft]

*We thank Pierre Collin-Dufresne, Jerome Detemple, Bernard Dumas, Philip Dybvig, Steve Ross and seminar participants at Boston University, Universidad de los Andes, 9th AMaMeF Conference (Paris), and Panos Mavrokonstantis for excellent data assistance. Support from SFI, and the hospitality of the Southwest University of Finance and Economics (SWUFE-IFS) where part of this paper was written is gratefully acknowledged.

†INSEAD, France. E-mail: theodoros.evgeniou@insead.edu

‡EPFL, SFI, Switzerland, and CEPR, UK. E-mail: julien.hugonnier@epfl.ch

§INSEAD, France. E-mail: rodolfo.prieto@insead.edu

1 Introduction

Despite the extensive literature on the impact of short sales constraints on stock returns (see for example D'Avolio (2002), Diether et al. (2002a), Mitchell et al. (2002), Cohen et al. (2007) and recently Drechsler and Dreschler (2016)) there are, surprisingly, very few theoretical studies that analyze the role of costly short sales in the price formation process, and the value or return differential associated with by the possibility of shorting an asset. If the possibility of shorting a stock is valuable to investors in the market, then these different trading rules are sufficient to drive a wedge between the risk–return tradeoffs offered by the two investment vehicles despite the fact that they load on the same risks and pay the same cash flows.[1] Our paper develops a theoretical model to study the determinants and magnitude of this wedge.

We consider a continuous-time Lucas exchange economy that includes a riskless asset, and two risky assets that are a claim to a fraction $(\eta, 1-\eta)$ of the aggregate dividend which is modeled as a diffusion process. The economy is populated by two groups of investors who have homogenous logarithmic utility and heterogenous beliefs about the growth rate of the economy. The first risky asset can be shorted at a cost and, while the second asset cannot be shorted.

An investor who wants to short Stock 1 needs to borrow it from another investor who is long, and pays in exchange an exogenously set fee per unit of time as long as the position is maintained. Specifically, we assume that shorting one share of the Stock 1 incurs a flow cost given by $\phi \cdot$Stock 1 where $\phi \geq 0$ is the lending fee. This setup captures in a simple way the rebate rates that are the most common measure of short selling fees (see for example Duffie et al. (2002)). To insure that investors behave competitively in all markets, the lending fees received by the long investor are taken into account as part of the return rather than as a separate source of income. This extra yield is determined endogenously in equilibrium by introducing an additional clearing condition that accounts for the aggregate positions on the short market. This approach is novel to the literature and allows us to easily integrate costly short sales into the model.

If Stock 1 could be shorted at no cost ($\phi = 0$), as is usually assumed in models where short sales are allowed, then the short sale constraint on Stock 2 would have no impact and, as a result, pricing would be linear in the sense that Stock 1 and Stock 2 would represent constant fractions η and $1-\eta$ of the market portfolio. By contrast, our

[1] These questions naturally arise in the context of various markets, such as ETFs and Mutual Funds, Equity carve-outs and Siamese twin stocks.

analysis will show that costly short sales drive a wedge between the valuation of the two investment vehicles and thereby results in *nonlinear* pricing in the sense that the value of Stock 1 represents a fraction of the market portfolio that varies across times and states to reflect the impact of the frictions at play in the model.

The analysis of the equilibrium sheds light on the sources of nonlinear pricing and reveals that short selling frictions may explain the dynamics of asset returns. Its main implications are summarized as follows.

First, our results formalize the intuition in Cochrane (2002), Cherkes et al. (2013) according to which the valuation of assets that can be shorted incorporates not only the present value of its future dividends but also the present value of future cash flows generated by lending fees.[2] Importantly, the model allows to study the extent and determinants of this price correction by deriving it from first principles within a general equilibrium model.

Second, and related, the model provides a rational explanation for the mispricing observed in certain stock carve-outs (see Lamont and Thaler (2003)), since price differentials between freely traded stocks and those kept by the parent company can be interpreted as stocks 1 and 2 respectively in our economy. Furthermore, if fees are sufficiently large, the equilibrium is observationally equivalent to one with pure short-sale constraints and we show both stocks prices correspond to their risk adjusted present value of dividends. This result counters the overpricing hypothesis of Miller (1977) that states that because short-sale constraints hold negative opinions off the market, making pessimists seat on the sidelines, optimistic investors will end up holding *overpriced* assets.

Third, we develop an empirical analysis that leads to a number of results that support the model predictions. First, the model provides a theoretical backdrop for the recent empirical findings of Drechsler and Dreschler (2016) and Beneish et al. (2015) who document that stocks with higher lending fees exhibit low average excess returns that cannot be explained by common risk factors. In particular, Drechsler and Dreschler (2016) argues that negative excess returns are compensation for the systematic risk borne by the small fraction of investors who account for most of the shorting activity, and refer to this finding as the shorting premium. This premium is proxied by a long-short portfolio containing cheap minus expensive to short stocks (CME), which in conjuction with Fama-French factors, would capture the short sellers wealth's portfolio, and get pricing errors

[2] See also Atmaz and Basak (2019) and Muravyev et al. (2018) for similar applications of this idea in the valuation of European options.

in check. We offer an alternative explanation that exploits the long agent's pricing kernel from our model. Using a data sample similar to Drechsler and Dreschler (2016), we adjusts costly to short stock returns by a term that can be identified from lending fees and short interest and which stems directly from our equilibrium pricing equations. Adjusted returns do not exhibit the mispricing reported in Drechsler and Dreschler (2016) in our sample. We complement this result by testing the adjustment as a characteristic in the cross section and find compelling evidence. Second, we develop an alternative measure for predicting aggregate returns using a transform of the short interest measure, aggregated across securities, developed by Rapach et al. (2016). Overall, the predictive power of our measure, at the monthly horizon, seems to be on par with the best individual predictors from the literature.

This paper relates to several branches of the literature. By integrating the market for borrowing stocks into the price formation process, the proposed framework relates to a handful of models with an explicit market for stock shorting. Lending fees are studied in static models by Duffie (1996), Krishnamurthy (2002), and Blocher et al. (2012), while Duffie et al. (2002) develop a dynamic model with a single stock in which search costs and bargaining over loan fees generate a deterministic price process that includes lending fees as a cash stream accrued to long agents. Vayanos and Weill (2008) extend the search model of Duffie et al. (2002) to include two assets with different lending fees and show that the resulting equilibrium produces price differentials between these assets that are in line with the on-the-run premium. As in Duffie et al. (2002), prices are deterministic. By contrast, we show that search frictions are not necessary to sustain an equilibrium with positive loan fees when agents are risk averse and that shorting activity is an important determinant of asset prices, return and volatility dynamics. Recently, Nutz and Scheinkman (2019) propose a continuous-time model of trading with heterogeneous beliefs, risk-neutral agents face quadratic costs-of-carry on positions. Their model is inspired in Harrison and Kreps (1978), and features a single risky asset in partial equilibrium as they assume the existence of a riskless technology in infinitely elastic supply.

In a related contribution Basak and Croitoru (2000) use a model with a risky stock in positive supply, a derivative in zero net-supply and costless short selling to show that mispricing can arise between two securities that carry the same risk if all agents are subject to a portfolio constraint that prevents them from exploiting the induced arbitrage opportunity. Banerjee and Graveline (2014) recently obtained similar conclusions in a static CARA-Normal model with quasi-redundant assets and costly short sales. By contrast, we propose to study the implications of costly shorting in a dynamic setting where all risky assets are in positive supply. This allows us to identify not only expected returns but also volatilities endogenously, and to relate the price deviation between assets to measures of beliefs heterogeneity and the level of lending fees.

The case of pure short sales constraint is nested in our model and has been the focal point of an extensive literature, see for example Miller (1977), Jarrow (1980), Nielsen (1989) for earlier contributions, and Reed (2013) for a recent survey. In a related contribution Gallmeyer and Hollifield (2008) show that the imposition of a short sale constraint increases the equilibrium interest rate, and may either raise or lower asset prices, depending on the assumed risk preferences. When all agents have homogenous logarithmic preferences, they find that the stock price is not influenced by a short sale constraint and volatility is flat. The equilibrium that we derive implies that this result no longer holds when short sales are allowed but costly. In particular, we show that costly short sales generate endogenous wealth transfers between agents, and that these transfers result in nonlinear pricing. In our model prices are still represented by a risk-adjusted present value formula of discounted cash flows, yet costly to short stocks are priced at a premium because they earn additional endogenous cash flows generated by lending fees accrued by long agents.

On the empirical side, and in addition to Drechsler and Dreschler (2016) and Beneish et al. (2015), this paper relates to the numerous empirical studies that investigate the effect of short selling frictions. Blocher and Whaley (2015) show that in order to enhance their returns ETF managers tend to tilt their portfolios toward stocks with higher lending fees. Similarly, the findings of Prado (2015) suggest that institutional investors buy shares in response to an increase in lending fees and, thus, provides an indirect test of the prediction of our model according to which stock prices reflect the expected future income associated with the potential of lending the asset.

2 The model

2.1 Information and preference structure

Time is continuous and runs forever. Uncertainty is generated by a single Brownian motion $(B_t^{(1)})_{t\geq 0}$ and the aggregate dividend process evolves according to a geometric

Brownian motion

$$de_t = \mu_e e_t dt + \sigma_e e_t dB_t^{(1)}$$

for some constants $\mu_e \in \mathbb{R}$, $\sigma_e > 0$. The economy is populated by two (groups of) agents with dogmatic beliefs about the evolution of the dividend process.[3] The information available to agents is summarized by the filtration $\mathbb{F} = \mathbb{F}^e$ generated by the aggregate dividend process. As a result, the value of σ_e is common knowledge but not the value of the growth rate μ_e and we assume that agents have different beliefs about this growth rate. Specifically, we assume that Agent 1 believes that the growth rate is μ_e. Agent 2 believes that it is given by $\mu_e - \bar{\mu}$ for some constant $\bar{\mu} \geq 0$, so that the aggregate dividend evolves according to

$$de_t = (\mu_e - \bar{\mu}) e_t dt + \sigma_e e_t dB_t^{(2)},$$

where $(B^{(2)})_{t\geq 0}$ is a standard Brownian motion under Agent 2's beliefs. The preferences of Agent $k = \{1, 2\}$ are represented by

$$U^{(k)}(c) \equiv E^{(k)} \left[\int_0^\infty e^{-\rho t} \log(c_t) dt \right],$$

for some subjective discount rate $\rho > 0$ and $E^{(k)}$ denotes the expected value under Agent k's beliefs.

2.2 Securities

The financial market consists of three assets: a locally riskless bond in zero net supply whose price satisfies

$$S_{0t} = 1 + \int_0^t S_{0u} r_u du,$$

Asset 1 in positive supply of one share which is the claim to a fraction $\eta \in (0, 1]$ of the aggregate dividend and follow dynamics given by

$$S_{1t} + \int_0^t \eta e_u du = S_{10} + \int_0^t S_{1u} \left(\mu_{1u}^{(k)} du + \sigma_{1u} dB_u^{(k)} \right). \tag{1}$$

[3] We use dogmatic beliefs to reduce the number of state variables. Our solution method can be generalized to allow for time varying beliefs.

Asset 2, also in positive supply, is the claim to the remainder fraction $1-\eta$ of the aggregate dividend and has dynamics

$$S_{2t} + \int_0^t (1-\eta)e_u du = S_{20} + \int_0^t S_{2u}\left(\mu_{2u}^{(k)} du + \sigma_{2u} dB_u^{(k)}\right). \tag{2}$$

The quantities $\left\{r_t, \left(S_{i0}, \mu_{it}^{(k)}, \sigma_{it}\right)\right\}$ for $i = \{1, 2\}$ are determined in equilibrium.

2.3 Shorting frictions

We assume that shares of Asset 2 cannot be shorted, whereas shares of Asset 1 can be shorted at a cost. These are the only frictions in the model and are in line with institutional regularities observed in various markets (see e.g., ETF-mutual funds, equity carve outs, siamese twin stocks). Both stocks are thus substitute assets from the vantage point of a long-only investor.

An investor who wants to short asset 1 needs to borrow it from another investor who is long, and pays in exchange a fee as long as the position is maintained. Specifically, we assume that shorting one share of Asset 1 at date $t \geq 0$ incurs a flow cost given by

$$\phi_t S_{1t} dt$$

where $\phi_t \geq 0$ is the lending fee rate and S_{1t} is the price of Asset 1. This setup captures in a simple way the rebate rates that are the most common measure of short selling fees.[4] We assume that the fee is given by

$$\phi_t = \phi \sigma_{1t}$$

where σ_{1t} is the endogenously determined return volatility of Asset 1 and ϕ is a constant. This functional form allows us to easily control the amount of liquidity in the short market by setting the level of the constant ϕ and capture in reduced form the empirical evidence in Drechsler and Dreschler (2016) and Blocher and Whaley (2015) who show that lending fees are proportional to the volatility of the asset being shorted.[5]

[4]The short-seller must leave collateral with the lender in order to borrow the shares, in turn, the lender pays the short-seller the rebate rate on this collateral. The spread between the interest rate on cash funds and the rebate rate is a direct cost to the short-seller, and is often referred to as the loan fee.

[5]Atmaz and Basak (2019) partial equilibrium model of costly-shorting applied to European options analyzes a similar reduced-form.

In order to model the shortselling interaction competitively, both type of agents take as a given the fact that in order to short the stock they have to pay a shorting fee ϕ_t per dollar of short and that when long they may receive lending fees at rate $\varphi_t \sigma_{1t}$ per dollar of long. Lending fees that are received by the non short agent are taken into account as part of the return rather than as a separate source of income. The rate $\varphi_t \sigma_{1t}$ is determined endogenously in equilibrium by introducing an additional clearing condition that accounts for the aggregate positions on the short market.

The dynamic budget constraint of Agent $k = \{1, 2\}$ evolves according to[6]

$$
\begin{aligned}
dX_{kt} = & \left(r_t X_{kt} - c_{kt} + \varphi_t \sigma_{1t} \pi_{1t}^{(k)+} - \phi \sigma_{1t} \pi_{1t}^{(k)-} \right) dt \\
& + \pi_{1t}^{(k)} \sigma_{1t} \left(dB_t^{(k)} + \theta_{1t}^{(k)} dt \right) + \pi_{2t}^{(k)} \sigma_{2t} \left(dB_t^{(k)} + \theta_{2t}^{(k)} dt \right)
\end{aligned}
$$

where X_{kt} is the wealth of Agent k, with $X_{k0} = \frac{1}{2}(S_{10} + S_{20}) > 0$, c_{kt} is the consumption rate, $\pi_{it}^{(k)}$ is the amount invested in asset $i = \{1, 2\}$, the market price of risk $\theta_{it}^{(k)}$ is defined by

$$
\sigma_{it} \theta_{it}^{(k)} \equiv \mu_{it}^{(k)} - r_t \tag{3}
$$

for $i = \{1, 2\}$, and $\pi^+ \equiv \max(0, \pi)$ and $\pi^- \equiv \max(0, -\pi)$ denote the positive and negative part, respectively. The respective market prices of risk under each agent's beliefs are related by the non-arbitrage condition

$$
\theta_{it}^{(2)} = \theta_{it}^{(1)} - \Delta, \qquad i = \{1, 2\}.
$$

where we have set $\Delta \equiv \bar{\mu}/\sigma_e$. We use the beliefs of Agent 1 as reference beliefs, so that $B \equiv B^{(1)}$.

2.4 Definition of equilibrium

The concept of equilibrium that we use is similar to that of equilibrium of plans, prices and expectations introduced by Radner (1972). An additional equation balances out the short market for Asset 1 and identifies the yield φ:

[6] For other examples of budget constraints with non-linear terms see e.g., Cuoco and Cvitanić (1998) and Cuoco and Liu (2000).

Definition 1. *An equilibrium is a pair of security price processes* (S_{0t}, S_{1t}, S_{2t}) *and an array* $\{c_{kt}, (\pi_{1t}^{(k)}; \pi_{2t}^{(k)})\}_{k=1}^{2}$ *of consumption plans and trading strategies such that:*

1. *Given* (S_{0t}, S_{1t}, S_{2t}) *the consumption plan* c_{kt} *maximizes* $U^{(k)}$ *over the feasible set of Agent* k *and is financed by the trading strategy* $(\pi_{1t}^{(k)}, \pi_{2t}^{(k)})$.

2. *Markets clear:*

$$
\begin{aligned}
\sum_{k=1}^{2} c_{kt} &= e_t, && \textit{Consumption market,} \\
\sum_{k=1}^{2} \pi_{it}^{(k)} &= S_{it}, && \textit{Stock} \quad i = \{1, 2\}, \\
\varphi_t \sum_{k=1}^{2} {\pi_{1t}^{(k)}}^{+} &= \phi \sum_{k=1}^{2} {\pi_{1t}^{(k)}}^{-}, && \textit{Short market Asset 1.}
\end{aligned} \tag{4}
$$

The additional market clearing condition (4) is novel to the literature and allows us to easily integrate costly short sales into the model. Similar to the notion of a central pot used in models of equilibrium with transaction costs (see e.g., Buss and Dumas (2017)) this condition can be understood by imagining that all shares of Asset 1 are made available for borrowing and that the proceeds from any short selling activity are equally split among the agents who are long.

We consider equilibria where $\sigma_i > 0$, so that a positive shock to aggregate output increases stock prices, and investors short Asset 1 when its expected return is low enough. These conditions put a restriction on the parameters of the economy and we ensure they hold in our numerical examples.

3 Optimality and equilibrium

3.1 Individual optimality

Relying on techniques developed for optimization problems with nonlinear wealth dynamics[7] we solve the individual optimization problem in closed form. A direct calculation using the results of Cuoco and Cvitanić (1998) show that the model is arbitrage free if and only if

$$
\max\left\{\theta_{2t}^{(k)}, \theta_{1t}^{(k)} + \varphi_t\right\} \leq \theta_{1t}^{(k)} + \phi. \tag{5}
$$

[7] See, e.g., models with large investor/price impact in Cuoco and Cvitanić (1998) and Cuoco and Liu (2000).

This condition is intuitive: the left hand side gives the maximal excess return that can achieved by going long while the right hand side gives the excess return that can be achieved by going short. The latter has to be larger because otherwise it would be possible to generate a strictly positive riskless excess return by going simultaneously long and short in the risky securities. Note also that, as a direct implication of no arbitrage in (5), we have that the yield φ_t is bounded by the constant ϕ

$$\varphi_t \leq \phi.$$

Since the ranking of excess returns from long positions i.e. $\theta_{2t}^{(k)}$ against $\theta_{1t}^{(k)} + \varphi_t$ is the same for all agents irrespective of their beliefs it must be the case that

$$\theta_{2t}^{(k)} = \theta_{1t}^{(k)} + \varphi_t, \tag{6}$$

for otherwise no agent would ever agree to buy the asset with the lower market price of risk and, as a result, markets would never clear. Using the above conjecture we have that the optimal consumption is explicitly given by

$$c_{kt} = \frac{c_{k0}}{e^{\rho t}\xi_{kt}} \tag{7}$$

where the agent-specific state price density ξ_k evolves according to

$$d\xi_{kt} = -\xi_{kt}\left(x_{kt}dB_t^{(k)} + r_t dt\right), \quad \xi_{k0} = 1, \tag{8}$$

with

$$x_{kt} = \mathbf{1}_{\{\Theta_{kt} \subseteq \mathbb{R}_+\}}\left(\theta_{1t}^{(k)} + \varphi_t\right) + \mathbf{1}_{\{\Theta_{kt} \subseteq \mathbb{R}_-\}}\left(\theta_{1t}^{(k)} + \phi\right)$$

and $\Theta_{kt} \equiv \left[\theta_{1t}^{(k)} + \varphi_t, \theta_{1t}^{(k)} + \phi\right]$. The next proposition summarizes the above result. The optimal consumption policy is given by a constant marginal propensity to consume equal to the discount rate, whereas portfolio positions can be characterized as one of three types: Long on both stocks, Non participation on both stocks, and Short on Asset 1.

Proposition 1. *Optimal consumption is given by*

$$c_{kt} = \rho X_{kt}.$$

Let $\hat{\pi}_i^{(k)} \equiv \pi_i^{(k)}/X_k$. Optimal portfolio allocation is described by three regions as follows:

1. *Long positions on both stocks*

$$\mathcal{L}_k \equiv \left\{\theta_{1t}^{(k)} + \varphi_t \geq 0\right\}, \quad \hat{\pi}_{1t}^{(k)}\sigma_{1t} + \hat{\pi}_{2t}^{(k)}\sigma_{2t} = \theta_{2t}^{(k)}. \tag{9}$$

2. *Non Participation on both stocks*

$$\mathcal{N}_k \equiv \left\{-\phi \leq \theta_{1t}^{(k)} \leq 0, \theta_{2t}^{(k)} \leq 0\right\}, \quad \hat{\pi}_{1t}^{(k)} = 0, \hat{\pi}_{2t}^{(k)} = 0. \tag{10}$$

3. *Short position on Asset 1, Non participation on Asset 2,*

$$\mathcal{S}_k \equiv \left\{\theta_{1t}^{(k)} \leq -\phi, \theta_{2t}^{(k)} \leq 0\right\}, \quad \hat{\pi}_{1t}^{(k)}\sigma_{1t} = \theta_{1t}^{(k)} + \phi, \hat{\pi}_{2t}^{(k)} = 0. \tag{11}$$

The optimal portfolio takes the form of a mean variance policy that reflects the presence of agent specific beliefs and shorting costs. As agents behave competitively, the yield φ enters the optimal allocation in (9) only through the no arbitrage condition in (6). In region $\mathcal{L}_k$, Agent k is indifferent between going long in either stock because they provide the same return. They are perfect substitutes. This implies portfolios are undetermined as any admissible pair $(\hat{\pi}_1^{(k)}, \hat{\pi}_2^{(k)})$ that satisfies (9) represent optimal allocations. In region $\mathcal{N}_k$, Agent k does not hold either stock, as net returns of holding a long position on both stocks are negative, whereas shorting Asset 1 while paying the fee is suboptimal, as $-\sigma_{1t}\theta_{1t}^{(k)} - \phi_t \leq 0$. In region $\mathcal{S}_k$, Agent k goes short on Asset 1, as the net return of the corresponding short position is positive $-\sigma_{1t}\theta_{1t}^{(k)} - \phi_t > 0$.

3.2 State variables and trading regions

We derive the equilibrium in closed form in terms of aggregate consumption and an endogenous state variable that tracks the consumption share of optimistic agents, $s \equiv c_o/e \in (0,1)$. It follows a diffusion process

$$ds_t = s_t m(s_t)dt + s_t v(s_t)dB_t,$$

where the pair (m, v) is determined in equilibrium.

The price of the market portfolio is invariant to the frictions in the model and is given explicitly by the usual logarithmic valuation of the aggregate dividend

$$X_{1t} + X_{2t} = S_{1t} + S_{2t} = e_t/\rho \equiv P_t,$$

where

$$X_{1t} = s_t P_t, \qquad X_{2t} = (1 - s_t) P_t. \tag{12}$$

Agent 1 holds long positions in both risky assets in all states, i.e., $\mathcal{L}_1 \equiv s \in (0, 1)$, so that the state space is fully characterized by the three trading regions of Agent 2 described in Proposition 1. The equilibrium dynamic budget constraints are given by

$$dX_{1t}/X_{1t} = (r_t - \rho)dt + \theta_{2t}^{(1)} \left(dB_t + \theta_{2t}^{(1)} dt \right), \tag{13}$$

$$dX_{2t}/X_{2t} = \begin{cases} (r_t - \rho)dt + (\theta_{2t}^{(1)} - \Delta) \left(dB_t + \theta_{2t}^{(1)} dt \right), & \mathcal{L}_2, \\ (r_t - \rho)dt, & \mathcal{N}_2, \\ (r_t - \rho)dt + \left(\theta_{1t}^{(1)} - \Delta + \phi \right) \left(dB_t + (\theta_{1t}^{(1)} + \phi)dt \right), & \mathcal{S}_2, \end{cases} \tag{14}$$

An application of Itô's lemma to (12) gives

$$dX_{1t}/X_{1t} = (\mu_e + m(s_t) + \sigma_e v(s_t))\, dt + (v(s_t) + \sigma_e)\, dB_t, \tag{15}$$

$$dX_{2t}/X_{2t} = \left(\mu_e - \frac{s_t(m(s_t) + v(s_t)\sigma_e)}{1 - s_t} \right) dt + \left(\sigma_e - \frac{s_t v(s_t)}{1 - s_t} \right) dB_t, \tag{16}$$

Matching dynamics in (14) and (16) gives a system for $(m, v, r, \theta_2^{(1)})$ with a unique solution as functions of (s, φ), as we detail next.

Proposition 2. *The dynamics of the consumption share of Agent 1 are determined by*

$$(m(s), v(s)) = \begin{cases} \left(v(s)^2, \Delta(1 - s)\right), & \mathcal{L}_2, \\ \left(v(s)^2, \sigma_e(1 - s)/s\right), & \mathcal{N}_2, \\ \left(v(s)(\sigma_e + s(\phi - \varphi(s)) + v(s)) - \bar{\mu}(1 - s), (1 - s)\left(\Delta - \phi + \varphi(s)\right)\right), & \mathcal{S}_2. \end{cases} \tag{17}$$

The interest rate and market price of risk are in turn given by

$$r(s) = \begin{cases} \bar{r}(s) \equiv \rho + \mu_e s + (\mu_e - \bar{\mu})(1-s) - \sigma_e^2, & \mathcal{L}_2, \\ \rho + \mu_e - \frac{\sigma_e^2}{s}, & \mathcal{N}_2, \\ \bar{r}(s) + s(1-s)\left(\phi - \varphi(s)\right)\left(\Delta - \phi + \varphi(s)\right), & \mathcal{S}_2, \end{cases} \tag{18}$$

and

$$\theta_2^{(1)}(s) = \begin{cases} \bar{\theta}(s) \equiv \sigma_e + \Delta(1-s), & \mathcal{L}_2, \\ \frac{\sigma_e}{s}, & \mathcal{N}_2, \\ \bar{\theta}(s) - (1-s)\left(\phi - \varphi(s)\right), & \mathcal{S}_2. \end{cases} \tag{19}$$

We write $\varphi(s)$ as the yield depends on the consumption share s in equilibrium. Eq. (17) shows that the consumption share of Agent 1 is positively correlated with dividends, as $v(s) \geq 0$, therefore it tends to increase (decrease) following sequences of positive (negative) cash flow shocks. The intuition for this result is clear, as Agent 1 benefits from sequences of positive shocks due to his long-only strategy.

We use results in (17) and (19) to identify the equilibrium regions as follows. The Long region follows from the inequality in (9) $\mathcal{L}_2 \equiv \{\sigma_e + \Delta(1-s) \geq \Delta\}$, so that

$$\mathcal{L}_2 \equiv s \in (0, s^*] \quad \text{and} \quad s^* \equiv \frac{\sigma_e}{\Delta}. \tag{20}$$

Note that the equilibrium will be unaffected by costly short sales if $s^* \geq 1$, an intuitive property since a large s^* corresponds to small heterogeneity in beliefs. Similarly, the Non Participation region follows from (10), $\mathcal{N}_2 \equiv \left\{\frac{\sigma_e}{s} \geq \Delta - \phi\right\}$ so that

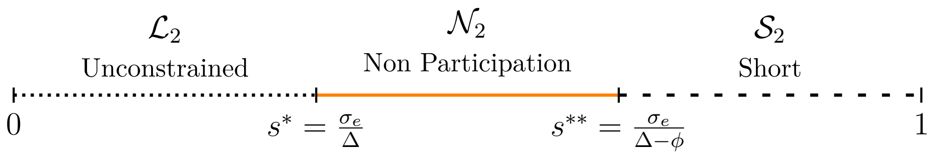
$$\mathcal{N}_2 \equiv s \in [s^*, s^{**}], \quad \text{and} \quad s^{**} \equiv \frac{\sigma_e}{\Delta - \phi}. \tag{21}$$

Intuitively, $s^{**}/s^* = \Delta/(\Delta - \phi)$ shrinks as the heterogeneity of beliefs increases and grows as the fee increases. Finally, note from (21) that if $\phi < \Delta$ then $s^* < s^{**}$, it follows from (11) that the region over which the pessimistic agent holds a short position on Asset 1 is non empty if and only if $s^{**} < 1$,

$$\mathcal{S}_2 \equiv s \in [s^{**}, 1).$$

Figure 1 displays the state space in an equilibrium with shorting. Note that the share η plays no role on the determination of the boundaries across regions.

Figure 1: State space of consumption share s in an equilibrium with pessimistic agents shorting Asset 1.



4 Analysis

4.1 Non linear pricing

We introduce q as the fraction in the market portfolio of Stock 1, so that

$$q_t \equiv S_{1t}/P_t. \tag{22}$$

Definition 2. *Pricing is nonlinear when the share of Stock 1 on the market portfolio is different from the share of the aggregate dividend it pays out, i.e.,* $q \neq \eta$*. It is priced at a premium (discount) when* $q_t > (<)\eta$*.*

The next theorem establishes the main pricing result.

Theorem 1. *Assume there is shorting in equilibrium. Stock prices are given by*

$$S_{1t} = \eta E_t\left[\int_t^\infty \xi_{1t,u} e_u du\right] + E_t\left[\int_t^\infty \xi_{1t,u}\varphi_u \sigma_{1u} S_{1u} du\right], \tag{23}$$

$$S_{2t} = (1-\eta)E_t\left[\int_t^\infty \xi_{1t,u} e_u du\right], \tag{24}$$

where $\xi_{1t,u} = \xi_{1u}/\xi_{1t}$*. Asset 1 (2) is priced at a premium (discount) since*

$$\eta < q_t \leq 1.$$

The discounted marginal utility of the (long only) optimist prices both stocks, in the sense that asset prices in (23) and (24) correspond to the risk-adjusted present value of future cash flows. The novelty is that cash flows for the costly-to-short stock are endogenously determined. The shorting premium

$$\frac{S_{1t}}{\eta} - \frac{S_{2t}}{1-\eta} = \frac{1}{\eta}\Lambda_t$$

with

$$\Lambda_t \equiv E_t\left[\int_t^{\infty} \xi_{1t,u}\varphi_u\sigma_{1u}S_{1u}du\right]$$

shows that costly shorting is the source of nonlinear pricing. The term Λ_t can be thought of as the value of the *convenience yield* that accrues to long investors of Asset 1.[8] It is non zero only if shorting Asset 1 is not only allowed but costly. In our next result, we highlight how heterogeneity in beliefs determines trading regions and thus the existence of nonlinear pricing.

Proposition 3. *The following bounds on Δ determine the corresponding trading regimes:*

$$\Delta \leq \sigma_e, \qquad \Lambda_t = 0, \qquad \textit{Long only activity} \quad (s^* \geq 1) \tag{25}$$

$$\Delta \in (\sigma_e, \sigma_e + \phi], \qquad \Lambda_t = 0, \qquad \textit{No shortselling activity} \quad (s^* < 1, s^{**} \geq 1) \tag{26}$$

$$\Delta > \sigma_e + \phi, \qquad \Lambda_t > 0, \qquad \textit{Shortselling activity} \quad (s^{**} < 1) \tag{27}$$

The equilibrium in (25) is frictionless, that is, it is observationally equivalent to the limiting case $\phi = 0$, despite the fact Asset 2 cannot be sold short. The equilibrium in (26) is the same as one where agents cannot short sell but non participation of would-be short agents arises in an endogenous way.

Remark 1. *As $\Lambda_t = 0$ in (26) and both prices correspond to the risk adjusted expected present value of their dividends, our theory counters the overpricing hypothesis of Miller (1977) that states that because short-sale constraints hold negative opinions off the market, making pessimists seat on the sidelines, optimistic investors will end up holding overpriced assets.*[9]

Finally, from (27), a sizable difference in beliefs is critical in supporting an equilibrium with shorting activity. In other words, (27) shows that the maximal fee that can be

[8]We use the term convenience yield as in Brennan (1991), Pindyck (1992), Cherian et al. (2004), Cochrane (2002), in the sense of being a benefit due to holdings in an inventory.

[9]See Duffie et al. (2002) and Vayanos and Weill (2008) for a similar result but where prices are deterministic.

imposed while still generating some shorting activity is given by

$$\overline{\phi} \equiv \Delta - \sigma_e.$$

Remark 2. *The model provides a rational explanation for the mispricing observed in certain stock carve-outs. A well-known example of such mispricing occurred after the spin-off of Palm from 3Com. After the carve-out and IPO of 5% of Palm shares, the parent company 3Com still owned the remaining 95% of Palm. The mispricing documented by Lamont and Thaler (2003) came from the fact that extrapolating the market valuation of the traded Palm shares to the remaining 95% of Palm resulted in a valuation that exceeded the market valuation of 3Com. This exercise does not take into account the fact that while the 5% of freely traded Palm shares could be lent to investors wanting to go short, the shares held by the parent company 3Com could not. In other words, the freely traded Palm shares are akin to Asset 1 so that their price should include a shorting premium as in* (23) *while the remaining shares held by 3Com are akin to Asset 2 whose price only reflects the present value of future dividends as in* (24)*.*

4.2 Computation of q

From the non-arbitrage condition in (5) and an application of Itô's lemma in (22), the sharing rule q is a function of the consumption share only s and solves the ODE

$$\left(\mu_e - r(s) - \sigma_1(s)\theta_1^{(1)}(s)\right) q(s) + s(m(s) + \sigma_e v(s))q'(s) + \frac{1}{2}s^2 v(s)^2 q''(s) + \eta\rho = 0. \quad (28)$$

This ODE has an explicit solution in the Unconstrained and No Participation regions, as we show next:

Proposition 4. *The function $q(s)$ for $s \in (0, s^{**}]$ is given by*

$$q(s) = \begin{cases} \eta + a(1-s)^{\frac{1}{2}-\frac{1}{2}\sqrt{1+\frac{8\rho}{\Delta^2}}} s^{\frac{1}{2}+\frac{1}{2}\sqrt{1+\frac{8\rho}{\Delta^2}}}, & \mathcal{L}_2, \\ \eta + b_1(1-s)^{\frac{1}{2}+\frac{1}{2}\sqrt{1+\frac{8\rho}{\sigma_e^2}}} + b_2(1-s)^{\frac{1}{2}-\frac{1}{2}\sqrt{1+\frac{8\rho}{\sigma_e^2}}}, & \mathcal{N}_2, \end{cases} \quad (29)$$

where (a, b_1, b_2) are constants detailed in the Appendix.

To solve for q over the region $[s^{**}, 1)$ we proceed as follows. First, we compute the compensation term φ as a function of σ^1 and q. From the clearing condition in (4) and the equilibrium outcome for $(\theta_1^{(1)}, \sigma_1)$, a linear equation for $\varphi(s)$ obtains, so that

$$\frac{\varphi(s)}{\phi} = \frac{(\sigma_1(s) - \sigma_e)\frac{q(s)}{q'(s)} - (1-s)\sigma_e}{\sigma_1(s)q(s) + (\sigma_1(s) - \sigma_e)\frac{q(s)}{q'(s)} - (1-s)\sigma_e}. \tag{30}$$

Next, we solve for σ_1 combining

$$\sigma_1(s) = \sigma_e + \frac{q'(s)}{q(s)} sv(s) \tag{31}$$

with (17) and (30), a result we describe in the next section, and use (29) to compute $q(s^*)$ and $q'(s^*)$. At s^*, we have then two boundary conditions. The function q over $[s^{**}, 1)$ and the identification of constants (a, b_1, b_2) are then obtained by solving the ODE in (28) shooting towards the terminal condition $q(1) = \eta$.

Figure 2 shows q increases with the liquidity of the asset as measured by the constant ϕ (panel (a)) and decreases with the supply of the shortable asset η (panel (b)) and the heterogeneity of beliefs among investors Δ (panel (c)).

Insert Figure 2 here

Figure 3 shows the yield $\varphi_t\sigma_{1t}$ generated by the lending fees increases with both the liquidity of the asset as measured by the constant ϕ (panel (a)) and the heterogeneity of beliefs among investors Δ (panel (c)) and decreases with the supply of the shortable asset η (panel (b)).

Insert Figure 3 here

As the term φ appears in good times (high s) and disappears in bad times (low s), it is related (negatively) to momentum and long term reversal.

4.3 Volatility

The fact that the ratio $S_{1t}/P_t = q(s_t)$ is time-varying implies that, in equilibrium, the two risky assets have stochastic volatility despite the fact that, as result of the assumption of logarithmic utility, the volatility of the market portfolio is constant and equal to that of aggregate dividends.

Combining our explicit formula for the equilibrium stock prices using the solution for q, and the equilibrium dynamics of the consumption share in (31) leads to the following characterization of σ_1.

Proposition 5. *The volatility of Asset 1 is a positive and continuous function of the consumption share s given by*

$$\sigma_1(s) = \begin{cases} \sigma_e + \Delta \frac{q'(s)}{q(s)} s(1-s), & \mathcal{L}_2, \\ \sigma_e + \sigma_e \frac{q'(s)}{q(s)}(1-s), & \mathcal{N}_2 \\ \sigma_e + \frac{q'(s)}{q(s)} \left(\frac{\sqrt{c_1(s)^2 - 4(1+q'(s))c_2(s)} - c_1(s)}{2(1+q'(s))} + (1-s)\sigma_e \right), & \mathcal{S}_2. \end{cases} \tag{32}$$

where $(c_1(s), c_2(s))$ are functions detailed in the Appendix.

The three-region structure that emerges in equilibrium goes in the right direction in explaining clustering and persistence in volatility. The rich relation between trading activity and changes in the price levels may help explain the complex empirical patterns between them documented in Gallant et al. (1992).

Insert Figure 4 here

In the two extremes, i.e., when $s = 0, 1$, the market is dominated by one of the agents and it approaches the volatility of the dividend process, σ_e. Costly-to-short assets endogenously have higher levels of volatility in bad times and lower levels in good times. As seen in panel (a) as ϕ increases, Asset 1's volatility deviates further from σ_e. In other words, using ϕ as a measure of liquidity, then the excess volatility generated by shorting increases as liquidity decreases. Panel (b) shows that as η increases, the behavior of Asset 1 resembles the market portfolio and thus its volatility approaches σ_e. In other words, the excess volatility generated by shorting dampens as supply increases. Finally, panel (c) depicts as Δ increases, the size of the Non Participation region decreases, increasing shorting volume and stock volatility.

5 Empirical analysis

5.1 Cross Section

The model provides a theoretical backdrop for the recent empirical findings of Drechsler and Dreschler (2016) and Beneish et al. (2015) who document that stocks with higher lending fees exhibit low average excess returns that cannot be explained by common risk factors. In particular, Drechsler and Dreschler (2016) argues that negative excess returns are a compensation for the systematic risk borne by the small fraction of investors who account for most of the shorting activity, and refer to this finding as the shorting premium. This premium is proxied by a long-short portfolio containing cheap minus expensive to short stocks (CME), which in conjuction with Fama-French factors, would capture the short sellers wealth's portfolio, and hence, the proper SDF, and get pricing errors in check.

We offer an alternative explanation which exploits the long agent's pricing kernel. It follows directly from (23) and (24) that the expected excess returns are given by

$$\frac{1}{dt}E_t\left[\frac{dS_{it}+e_{it}dt}{S_{it}}\right]-r_t=\sigma_{it}\theta_{2t}^{(1)}-\mathbf{1}_{\{i=1\}}\varphi_t\sigma_{1t} \tag{33}$$

where e_{it} is the dividend paid by asset $i=\{1,2\}$. This relation shows that a one-factor CAPM obtains in all states where there is no shorting activity. In states with shorting activity the expected return of Asset 1 is adjusted down to account for the yield generated by lending.

Data

Data on stock lending fees is from Markit Securities Finance (MSF). We match the MSF data to the CRSP database to obtain returns data and obtain accounting information by matching to Compustat. We retain only common stocks (share codes 10 and 11 in the CRSP database). We limit our study to a subset of US equities. The data cover the sample period from January 2004 to December 2014.

Insert Table 1 here

Table 1 reports equal-weighted averages of the monthly decile portfolio returns and characteristics. Decile 1 contains the cheapest-to-short stocks while decile 10 contains the most expensive-to-short stocks. Our data sample captures features similar to Drechsler and Dreschler (2016) (see Table 2) with fewer number of firms (270 per decile in our sample versus 336 in theirs). In particular, we highlight that the mean fee per decile and size (market cap) are very close, as well as measures of volatility and realized returns.

We reproduce the positive relationship of fees with realized volatility and the negative relationship of fees with firm size and realized returns. In addition, fees are negatively related with Institutional Ownership (IO), and positively related with the dispersion of beliefs (DISP), the Short Interest ratio (SIR) and Utilization (Util).[10]

Results

For each portfolio decile $i = 1, \dots, 10$ we compute alphas from the empirical counterpart to (33),

$$r_{it} - r_{ft} + \hat{\varphi}_{it} = \alpha_i + \sum_{k=1}^{K} \beta_{ik} f_{kt} + \epsilon_{it} \tag{34}$$

where $f_t = (\text{mkt}_t, \text{smb}_t, \text{hml}_t, \dots)$ are empirical risk factors and $t = 1, \dots, T$. From (4), the term $\hat{\varphi}_t = \varphi_t \sigma_{1t}$ can be computed as

$$\hat{\varphi}_t = \phi \sigma_{1t} \frac{\pi_{1t}^{(2)^-}}{S_{1t} + \pi_{1t}^{(2)^-}} \frac{1/S_{1t}}{1/S_{1t}} = \phi_t \left(\frac{\text{SIR}_t}{1 + \text{SIR}_t} \right) \tag{35}$$

where ϕ_t is the lending fee rate that applies to the asset and the variable SIR_t is the fraction of the existing shares of the asset currently on loan, i.e., the short interest ratio. In our analysis below, we also use a second data proxy for $\hat{\varphi}_t$

$$\hat{\varphi}_t = \phi_t \left(\frac{\text{Util}_t}{1 + \text{Util}_t} \right) \tag{36}$$

where Util_t is a measure provided by Markit defined as the percentage of value of assets on loan divided by the total lendable assets (inventory).

Insert Table 2 here

Table 2 shows the time series estimates for α for the ten portfolios grouped by fee size. Note that Fama-French 3-factor (FF3) alphas fluctuate between -1 and 53 bps and

[10] Definitions are displayed at the bottom of the Table 1.

average lending fees between 4.6 and 52 bps in deciles $1-9$ (see Table 1), whereas decil 10 is featured by a pair of fees and FF3 alpha given by $(642, -32)$ bps. When alphas are computed from the excess returns equation in (34), we apply the adjustments described in (35) (Model 1) and (36) (Model 2) above. For Model (1), alphas decrease in (absolute) size and become statistically insignificant in some decile portfolios. In particular, decile 10's alpha, although still negative at -26 bps, is not statistically significant. For Model (2), decile 10's alpha estimate is 23 bps and not longer significant. What is reported in Drechsler and Dreschler (2016) as returns in decile 10 exhibiting low average excess returns that cannot be explained by common risk factors bears a new interpretation in light of our model. This observation is reproduced in Panel (b) where we report Fama-French 4-factor (FF3+mom) alphas.

We also report alphas for long-short portfolios, similar to the CME factor in Drechsler and Dreschler (2016). The table reproduces that unadjusted returns for long-short portfolios 1-10b generate significant FF3 and FF4 positive alphas. However, once excess returns are adjusted as in (34), alphas are not longer statistically significant.

Next, we test the adjustment as a characteristic using Fama-MacBeth regressions. We split the sample in two equal halves. In the first pass we run time series regressions $t = 1, \dots, T'$ for each portfolio i

$$r_{it} - r_{ft} = a_i + \sum_{k=1}^{K} \beta_{ik} f_{kt} + \epsilon_{it}$$

we get point estimates for β_k, labeled $\hat{\beta}_{ik}$. In the second pass for each $t = T', \dots, T$ we compute cross-section estimates using

$$r_{it} - r_{ft} = \alpha_{it} + \sum_{k=1}^{K} \hat{\beta}_{ik} \lambda_{kt} + \gamma_t \hat{\varphi}_{it}$$

where $\hat{\varphi}_{it}$ is the adjustment term, in (35) or (36), written as a characteristic. The point estimates for risk premia (e.g. slope coefficients modeled as $\lambda_k = E[f_{kt}]$) are then computed as $\bar{\alpha} = \frac{1}{T-T'} \sum_{t=T'}^{T} \hat{\alpha}_t$, $\bar{\lambda}_k = \frac{1}{T-T'} \sum_{t=T}^{T} \hat{\lambda}_{kt}$, and $\bar{\gamma} = \frac{1}{T-T'} \sum_{t=T'}^{T} \hat{\gamma}_t$ and t-stats are time averages of the cross sectional statistics.

Insert Table 3 here

Table 3 shows that the adjustment estimates are negative and statistically significant, as the equation in (33) suggests. Further note that in panel (b), the magnitude of $\bar{\gamma}$ estimate is closer to -1, as in (33).

5.2 Aggregate predictability

Rapach et al. (2016) show that a measure of short interest, aggregated across securities, is the strongest predictor of the equity risk premium and its predicting ability stems predominantly from a cash flow channel. Aggregate returns representation in our model, given by

$$\frac{1}{dt}E_t\left[\frac{dP_t + e_t dt}{P_t}\right] - r_t = \sigma_e \theta_{2t}^{(1)} - \varphi_t \sigma_{1t}\frac{S_{1t}}{P_t}, \tag{37}$$

establishes the negative relationship between expected excess returns and the term $\varphi_t \sigma_{1t} q(s_t)$. Notice that the later depends on an increasing function of short interest and is a measure of market pessimism since it is higher when there is a higher mass of short sellers, matching the results in Rapach et al. (2016).

Insert Table 4 here

Table 4 reports the ordinary least squares estimate of b and the R^2 statistic for the bivariate predictive regression

$$r_{t+1} = a + bx_t + e_{t+1} \quad \text{for} \quad t = 1, \ldots, T-1$$

where r_t is the S&P 500 log excess return for month t, and x_t is the predictor variable in the first column. At the monthly horizon, two of the Welch and Goyal (2008) predictors display significant predictive ability at conventional levels: RVOL and LTY. SII also exhibits significant predictive ability in the second column and importantly so does VARPHI. The estimates for VARPHI has the expected sign (recall that we take the negative of VARPHI). Because monthly returns inherently contain a large unpredictable component, the R^2 statistics will necessarily be small. Overall, the predictive power of VARPHI at the monthly horizon seems to be on par with the best individual predictors from the literature.

The VARPHI–PC entry corresponds to a multiple predictive regression model that includes an intercept and four predictors: VARPHI and the first three principal components extracted from the non-VARPHI variables. Principal components provide an effective way for incorporating the information from a large number of economic variables in predictive regression models for stock returns. Comparing the VARPHI and VARPHI–PC rows of Table 4, shows that including the principal components in the predictive regression has some effect on the predictive ability of VARPHI. The estimated slope coefficients for VARPHI remain sizable when the principal components are included in the predictive regression, and the partial R^2 statistics indicate that VARPHI retains marginal predictive power in the presence of the principal components.

6 Concluding remarks

We study a dynamic general equilibrium model with costly-to-short stocks and heterogeneous beliefs. The model is solved in closed-form and shows that costly short sales drive a wedge between the valuation of assets that promise identical cash flows but are subject to different lending fees. The price of an asset is given by the risk-adjusted present value of its future cash flows, which include both dividends and an endogenous yield derived from lending fees. This pricing formula implies that asset returns satisfy a modified capital asset pricing model which includes a negative adjustment for lending fees and, thus, provides a theoretical foundation for the recent findings of Drechsler and Dreschler (2016) and Beneish et al. (2015) on the role of lending fees as an explanatory variable of stock returns.

In particular, Drechsler and Dreschler (2016) argues that negative excess returns are compensation for the systematic risk borne by the small fraction of investors who account for most of the shorting activity, and refer to this finding as the shorting premium. We offer an alternative explanation that exploits the long agent's pricing kernel from our model. Using a data sample similar to Drechsler and Dreschler (2016), we adjusts costly to short stock returns by a term that can be identified from lending fees and short interest and which stems directly from our equilibrium pricing equations. Adjusted returns do not exhibit the mispricing reported in Drechsler and Dreschler (2016) in our sample. We complement this result by testing the adjustment as a characteristic in the cross section and find compelling evidence. We also develop an alternative measure for predicting aggregate returns using a transform of the short interest measure, aggregated across securities, developed by Rapach et al. (2016). Overall, the predictive power of our measure, at the monthly horizon, seems on par with the best individual predictors from the literature.

A Additional theory results

A.1 Portfolios and indeterminacy

Propositions 1 and 2 allow us to derive in closed-form the trading strategies employed in equilibrium by each agent.

Proposition 6. *Let $\phi < \bar{\phi}$, so there is shorting in equilibrium. Optimal portfolios are given by*

$$\left(\pi_1^{(1)}, \pi_2^{(1)}\right) = \begin{cases} \left(q(s)P - \pi_1^{(2)}, \frac{\sigma_e}{\sigma_2(s)}\left(1 + \frac{\Delta}{\sigma_e}(1-s)\right)sP - \frac{\sigma_1(s)}{\sigma_2(s)}\left(q(s)P - \pi_1^{(2)}\right)\right), & \mathcal{L}_2, \\ \left(q(s), 1-q(s)\right)P, & \mathcal{N}_2, \\ \left(\frac{1}{\sigma_1(s)}\left(\sigma_e + (1-s)\left(\Delta - \phi + \varphi(s)\right)\right)sP - \frac{\sigma_2(s)}{\sigma_1(s)}(1-q(s))P, (1-q(s))P\right) & \mathcal{S}_2. \end{cases}$$

$$\left(\pi_1^{(2)}, \pi_2^{(2)}\right) = \begin{cases} \left(\pi_1^{(2)}, \frac{\sigma_e}{\sigma_2(s)}\left(1 - \frac{\Delta}{\sigma_e}s\right)(1-s)P - \frac{\sigma_1(s)}{\sigma_2(s)}\pi_1^{(2)}\right), & \mathcal{L}_2, \\ (0,0) & \mathcal{N}_2, \\ \left(\frac{1}{\sigma_1}\left(\sigma_e - s\left(\Delta - \phi + \varphi(s)\right)\right)(1-s)P, 0\right), & \mathcal{S}_2. \end{cases}$$

$$\pi_0^{(2)} = \begin{cases} \left(1 - \frac{\sigma_e}{\sigma_2}\left(1 - \frac{\Delta}{\sigma_e}s\right)\right)(1-s)P + \frac{\sigma_1(s) - \sigma_2(s)}{\sigma_2(s)}\pi_1^{(2)}, & \mathcal{L}_2, \\ (1-s)P & \mathcal{N}_2, \\ \left(1 - \frac{1}{\sigma_1(s)}\left(\sigma_e - s\left(\Delta - \phi + \varphi(s)\right)\right)(1-s)\right)P, & \mathcal{S}_2. \end{cases}$$

Combining eqs. (13) and (15), there is a negative relationship between the interest rate and the market price of risk in $\mathcal{L}_2 \cup \mathcal{N}_2$,

$$r(s) = \rho + \mu_e - \sigma_e \theta_2^{(1)}(s)$$

so that a high market price of risk induces Agent 1 to hold a levered position in stocks, while at the same time a lower interest rate induces him to borrow. As s increases after a sequence of good shocks, making Agent 2 poorer, the market price of risk from Agent 2's point of view becomes negative, prompting her to short Asset 1.

As it is suboptimal to short for the given level of lending fees, she stays out of equity markets and fully invests in the money market account. Another way to think about this portfolio movement is the following thought-experiment. For markets to clear when a short-sale constraint is imposed, security prices must change to reduce the optimist's stock demand. Two changes can lead to a reduction in the optimist's stock demand: the market price of risk can drop, and the cost of borrowing at the riskless rate can rise. Both changes occur.

Agent 2's wealth is decreasing in s despite the fact the interest rate rises, in other words, increases in r are not enough to keep up with the growth of optimists in good times. As this movement unfolds, the pessimistic agents reaches a point where the net returns of shorting the stock are positive, so she enters the stock lending market by paying a fee.

Remark 3. *Portfolio positions in all individual assets are completely determined in $\mathcal{N}_2 \cup \mathcal{S}_2$ due to market clearing conditions, however, positions in all individual assets are undetermined* *in $\mathcal{L}_2$. Quantities are parametrized by the Agent 2's position on Asset 1 $\pi_1^{(2)}$ and are only constrained by equilibrium clearing conditions.*

Remark 4. *If $\phi \geq \bar{\phi}$, then there is no shorting premium and $\sigma_i = \sigma_e$, $i = \{1, 2\}$. Indeterminacy in both risky assets remain, but unlike the model with costly shorting, the borrowing amount is uniquely determined by $\pi_0^{(2)} = \frac{\Delta}{\sigma_e} s(1-s)P$ and it is increasing in heterogeneity of beliefs Δ.*

B Proofs

Proof of Proposition 1. Following Cuoco and Cvitanić (1998) and assuming the no arbitrage condition in (5), the corresponding dual problem can be reduced to the quadratic optimization problem given by

$$\inf_{x \in \Theta_{kt}} |x|^2 \quad \text{with} \quad \Theta_{kt} \equiv \left[\theta_{1t}^{(k)} + \varphi_t, \theta_{1t}^{(k)} + \phi\right].$$

Solving this problem gives

$$x_{kt}^* = \mathbf{1}_{\{\Theta_t \subseteq \mathbb{R}_+\}} \left(\theta_{1t}^{(k)} + \varphi_t\right) + \mathbf{1}_{\{\Theta_{kt} \subseteq \mathbb{R}_-\}} \left(\theta_{1t}^{(k)} + \phi\right)$$

and it follows that the optimal consumption is explicitly given by (7), optimal wealth is determined by

$$X_{kt} = E_t^{(k)} \left[\int_t^\infty \xi_{kt,u} c_{ut} du\right] = c_{kt}/\rho$$

where the agent specific state price density evolves according to (8). Optimal portfolio allocation is described in (9), (10) and (11). ■

Proof of Proposition 2. Matching dynamics in (14) and (16) gives a system for $(m, v, r, \theta_2^{(1)})$ with a unique solution as functions of (s, φ) as detailed in (17). The interest rate in (18) and market price of risk perceived by the optimist in (19) follow by using $m(s)$ and $v(s)$ and completing the result using the fact that

$$\begin{aligned} r(s) &= \rho + \mu_e + m(s) - v(s)(\sigma_e + v(s)) - \sigma_e^2, \\ \theta_2^{(1)}(s) &= \sigma_e + v(s) \end{aligned}$$

■

Proof of Theorem 1. Let $\xi_t = e^{-\rho t} \left(\frac{s_0 e_0}{s_t e_t}\right)$ with

$$-\frac{d\xi_t}{\xi_t} = r_t dt + \theta_{2t} dB_t$$

denote the marginal utility process of the optimist in equilibrium. By construction we have that the processes

$$M_{it} = \xi_t S_{it} + \int_0^t \xi_u \left(\eta_i e_u + \mathbf{1}_{\{i=1\}} \varphi_u \sigma_{1u} S_{1u}\right) du, \qquad i \in \{1, 2\} \tag{38}$$

are local martingales under $\mathbb{P}$. By Lemma 1 below we have that these processes are true martingales. In particular, we have that

$$M_{it} = E_t \left[\xi_T S_{iT} + \int_0^T \xi_u \left(\eta_u e_u + \mathbf{1}_{\{i=1\}} \varphi_u \sigma_{1u} S_{1u} \right) du \right]$$

for all $T > 0$ and, therefore,

$$M_{it} = \lim_{T \to \infty} E_t \left[\xi_T S_{iT} \right] + E_t \left[\int_0^\infty \xi_u \left(\eta_i e_u + \mathbf{1}_{\{i=1\}} \varphi_u \sigma_{1u} S_{1u} \right) du \right]$$

by the monotone convergence theorem. To complete the proof it now remains to show that the limit is equal to zero. Let $\lambda_t = 1/s_t - 1$. As shown in the proof of Lemma 1 below we have that

$$\begin{aligned} \xi_T S_{iT} \leq \xi_T (S_{1T} + S_{2T}) = \xi_T P_T = e^{-\rho T} \frac{s_0}{s_T} P_0 \\ = e^{-\rho T} P_0 \left(\frac{1 + \lambda_T}{1 + \lambda_0} \right) \leq e^{-\rho T} P_0 \left(\frac{1 + N_T}{1 + \lambda_0} \right) \end{aligned}$$

for some $\mathbb{P}-$martingale N_t with initial value λ_0 and therefore

$$\lim_{T \to \infty} E_t \left[\xi_T S_{iT} \right] \leq \lim_{T \to \infty} \frac{e^{-\rho T} P_0}{1 + \lambda_0} \left(1 + E_t \left[N_T \right] \right) = \lim_{T \to \infty} e^{-\rho T} P_0 = 0$$

where the last equality uses the assumption that $\rho > 0$. Since $\xi_T S_{iT} \geq 0$ this in turn implies that the limit is zero and the proof is complete. The bound on q follows from the definition of η,

$$\eta < q_t \equiv \frac{\eta S_{2t} + (1 - \eta) \Lambda_t}{S_{2t} + (1 - \eta) \Lambda_t} \leq 1.$$

■

Lemma 1. *The processes* $(M_{it})_{i=1}^2$ *defined by* (38) *are martingales under* $\mathbb{P}$.

Proof. Let $T < \infty$ be fixed. By construction we have that

$$0 \leq M_{it} \leq M_t \equiv M_{1t} + M_{2t} = \xi_t (e_t / \rho) + \int_0^t \xi_u \left(e_u + \varphi_u \sigma_{1u} S_{1u} \right) du$$

and it is thus sufficient to show that the process M_t is a martingale under $\mathbb{P}$ over the time interval $[0, T]$. Since $S_{it} \sigma_{it} \geq 0$ we have that

$$S_{it} \sigma_{it} \leq \sum_{j=1}^2 S_{jt} \sigma_{jt} = P_t \sigma_e = (e_t / \rho) \sigma_e$$

and combining this inequality with the definition of ξ_t and the fact that $\varphi_t \leq \phi$ we deduce that there are constants $(C_0(T), C_1(T))$ such that

$$
\begin{aligned}
|M_t| &\leq \xi_t(e_t/\rho) + \int_0^t \xi_u e_u (1+\phi\sigma_e/\rho)du \\
&= s_0 P_0 \left\{ e^{-\rho t}(1+\lambda_t) + \int_0^t e^{-\rho s}(1+\lambda_s)\left(1+\phi\sigma_e/\rho\right) ds \right\} \\
&\leq C_0(T) + C_1(T) \sup_{\tau\in[0,T]} \left(e^{-\rho\tau}\lambda_\tau\right) \qquad (39)
\end{aligned}
$$

for all $t \in [0,T]$ where the process λ_t is defined by $\lambda_t \equiv 1/s_t - 1$. Using the dynamics of the consumption share process and applying Itô's lemma shows that this process evolves according to

$$
\frac{d\lambda_t}{\lambda_t} = -\psi(s_t,\varphi_t)dt - \Xi(s_t,\varphi_t)dB_t
$$

for some functions $\psi, \Xi : [0,1]\times[0,\phi] \to \mathbf{R}$ such that $\psi(s,\varphi) \geq 0$ and $|\Sigma(s,\varphi)| \leq \Delta$. Therefore, Novikov's condition implies that the process

$$
N_t = e^{\int_0^t \psi(s_u,\varphi_u)du}\lambda_t = \exp\left(-\int_0^t \Xi(s_u,\varphi_u)dB_u - \frac{1}{2}\int_0^t |\Xi(s_u,\varphi_u)|^2 du\right)
$$

is a true martingale under $\mathbb{P}$ over $[0,T]$ and it now follows from Doob's maximal inequality and the definition of N_t that we have

$$
\begin{aligned}
E\left|\sup_{\tau\in[0,T]} \left(e^{-\rho\tau}\lambda_\tau\right)\right|^{1+\epsilon} &= E\left|\sup_{\tau\in[0,T]} \left(e^{-\int_0^\tau (\rho+\psi(s_u,\varphi_u))du} N_\tau\right)\right|^{1+\epsilon} \\
&\leq E\left|\sup_{\tau\in[0,T]} N_\tau\right|^{1+\epsilon} \leq C_\epsilon E\left[N_T^{1+\epsilon}\right]
\end{aligned}
$$

for any $\epsilon > 0$. Now, since $|N_t \Xi(s_t,\varphi_t)| \leq N_t\Delta$ the mean comparison results of Hajek (1985) shows that we have

$$
E\left[N_T^{1+\epsilon}\right] \leq N_0^{1+\epsilon} E\left[\exp\left(-\frac{1}{2}\Delta^2 T + \Delta B_T\right)^{1+\epsilon}\right] = e^{\frac{\epsilon}{2}(1+\epsilon)\Delta^2 T}\lambda_0^{1+\epsilon}
$$

where the equality follows from basic properties of Brownian motion. This in turn implies that the random variable on the right of (39) is $\mathbb{P}-$integrable and the martingale property now follows from the dominated convergence theorem. ∎

Proof of Proposition 3. Examining the bound definition for $\mathcal{L}_2$ in (20), we have that $s^* \geq 1 \Leftrightarrow \Delta \leq \sigma_e$, in which case $\mathcal{L}_2 \equiv s \in (0,1)$ and $\Lambda = 0$. Next, assume the Non Participation region $\mathcal{N}_2$ defined in (21) extends all the way to $s = 1$. This outcome would be the result of the net excess return of shorting Stock 1 being negative,

$$
-\sigma_{1t}\theta_{1t}^{(2)} - \phi_t \leq 0. \qquad (40)
$$

We look for the minimum value of ϕ so that (40) holds in $[s^*, 1]$. Note that if the state space is determined by $\{\mathcal{L}_2, \mathcal{N}_2\}$, $\Lambda = 0$ and the sharing rule is $q = \eta$ and the volatility of Stock 1 is

$\sigma_1 = \sigma_e$. This back in (40) gives

$$-\sigma_{1t}\theta_{1t}^{(2)} - \phi_t = -\frac{\sigma_e^2}{s_t} + \Delta\sigma_e - \phi\sigma_e \leq 0.$$

This is equivalent to

$$\phi \geq \overline{\phi} \equiv \Delta - \sigma_e$$

■

Proof of Proposition 4. The ODE in (28) stems from the no arbitrage dynamics of the stock and an application of Itô's lemma to (22)

$$\begin{aligned}\frac{dS_{1t} + \eta e_t}{S_{1t}} &= (r_t + \sigma_{1t}\theta_{1t}^{(1)})dt + \sigma_{1t}dB_t \\ &= \left(\mu_e + \frac{\eta\rho}{q(s)} + \frac{q'(s_t)}{q(s_t)}s_t(m(s_t) + \sigma_e v(s_t)) + \frac{1}{2}\frac{q''(s_t)}{q(s_t)}s_t^2 v(s_t)^2\right)dt + \sigma_{1t}dB_t\end{aligned}$$

so that

$$\mu_e + \frac{\eta\rho}{q(s)} + \frac{q'(s)}{q(s)}s(m(s) + \sigma_e v(s)) + \frac{1}{2}\frac{q''(s)}{q(s)}s^2 v(s)^2 = r(s) + \sigma_1(s)\theta_1^{(1)}(s)$$

holds at all times. An application of Itô's lemma to (22) and matching the dynamics in (1) (and (2)) gives the volatility representation in (31). The ODE in $s \in (0, s^{**})$ is given by

$$\begin{aligned}\eta\rho - \rho q(s) + \frac{1}{2}\Delta^2(1-s)^2 s^2 q''(s) = 0, &\qquad \mathcal{L}_2, \\ \eta\rho - \rho q(s) + \frac{1}{2}\sigma_e^2(1-s)^2 q''(s) = 0, &\qquad \mathcal{N}_2\end{aligned}$$

The function q is available in closed form over the region $(0, s^{**}]$.

$$q(s) = \begin{cases} \eta + a_1(1-s)^{\frac{1}{2}(1+\ell)}s^{\frac{1}{2}(1-\ell)} + a_2(1-s)^{\frac{1}{2}(1-\ell)}s^{\frac{1}{2}(1+\ell)}, & \mathcal{L}_2, \\ \eta + b_1(1-s)^{\frac{1}{2}(1+r)} + b_2(1-s)^{\frac{1}{2}(1-r)}, & \mathcal{N}_2, \end{cases}$$

with $\ell \equiv \sqrt{1 + \frac{8\rho}{\Delta^2}} \leq r \equiv \sqrt{1 + \frac{8\rho}{\sigma_e^2}}$, as $\ell \leq r$ is equivalent to $s^* \leq 1$. Notice that since at $s = 0$, $|q(0)| < \infty$, we have that $a_1 = 0$ and thus from Theorem 1,

$$a \equiv a_2 > 0.$$

Explicit forms for $\{b_1, b_2\}$ follow from the following conditions: $\lim_{\uparrow s^*} q(s) = \lim_{\downarrow s^*} q(s)$, $\lim_{\uparrow s^*} q'(s) = \lim_{\downarrow s^*} q'(s)$,

$$\begin{aligned}b_1(a) =& -(a/2r)(1 + \ell - (1+r)s^*)(1-s^*)^{-\frac{1}{2}(\ell+r)}(s^*)^{-\frac{1}{2}(1-\ell)}, \\ b_2(a) =& (a/2r)(1 + \ell - (1-r)s^*)(1-s^*)^{-\frac{1}{2}(\ell-r)}(s^*)^{-\frac{1}{2}(1-\ell)},\end{aligned}$$

Since $1 + \ell - (1-r)s^* \geq 1 + \ell - (1+r)s^* \geq 0$, it follows that

$$b_1(a) < 0, \qquad b_2(a) > 0.$$

$\blacksquare$

Proof of Proposition 6. Equilibrium portfolios follow from using the results in Proposition 1 and the equilibrium dynamics of wealth in eqs. (15) and (16) in conjunction with the market prices of risk in (19). In $\mathcal{L}_2$ we have the following system of equations:

$$\underbrace{\begin{bmatrix} \sigma_1 & \sigma_2 & 0 & 0 \\ 0 & 0 & \sigma_1 & \sigma_2 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}}_{A} \begin{bmatrix} \pi_1^{(1)} \\ \pi_2^{(1)} \\ \pi_1^{(2)} \\ \pi_2^{(2)} \end{bmatrix} = \begin{bmatrix} \theta_2^{(1)} X_1 \\ \left(\theta_2^{(1)} - \Delta\right) X_2 \\ S_1 \\ S_2 \end{bmatrix}$$

There are infinitely many solutions as matrix A is singular. We choose $\pi_1^{(2)}$ to parametrize the solutions. A similar procedure follows in $\mathcal{N}_2$ and $\mathcal{S}_2$, yet solutions are unique. Borrowing is determined by the market clearing condition $\pi_0^{(2)} = -\pi_0^{(1)} = X_2 - \pi_1^{(2)} - \pi_2^{(2)}$. $\blacksquare$

Proof of Proposition 5. The results in (17) and (29) in the volatility representation (31) give an explicit form for σ_1 in $\{\mathcal{L}_2, \mathcal{N}_2\}$. Since $1 + \ell - 2s \geq 0$ and given $a \geq 0$, and

$$(1+r)b_1(a)(1-s)^r + (1-r)b_2(a) \leq 0$$

holds, the derivative

$$q'(s) = \begin{cases} \frac{1}{2}a(1+\ell-2s)(1-s)^{-\frac{1}{2}(1+\ell)}s^{-\frac{1}{2}(1-\ell)}, & \mathcal{L}_2, \\ -\frac{1}{2}\left((1+r)b_1(a)(1-s)^r + (1-r)b_2(a)\right)(1-s)^{-\frac{1}{2}(1+r)}, & \mathcal{N}_2, \end{cases}$$

is nonnegative and it follows that $\sigma_1(s) > \sigma_e$ in $\{\mathcal{L}_2, \mathcal{N}_2\}$. q is constructed under the conjecture that $\sigma_1 > 0$ and $\lim_{s->1} \sigma_1 = \sigma_e$. From the clearing condition of the short market and individual optimality conditions, we have

$$\varphi(s)\left(\sigma_1(s)q(s) - \left(\theta_1^{(1)}(s) - \Delta + \phi\right)(1-s)\right) + \phi\left(\theta_1^{(1)}(s) - \Delta + \phi\right)(1-s) = 0 \tag{41}$$

In addition, from (6) and (17)

$$v(s) = (1-s)\left(\Delta - \phi + \varphi(s)\right), \tag{42}$$
$$\theta_1^{(1)}(s) = v(s) + \sigma_e - \varphi(s), \tag{43}$$

so that

$$(\sigma_1(s) - \sigma_e)\frac{q(s)}{q'(s)} = (1-s)s\left(\Delta - \phi + \varphi(s)\right) \tag{44}$$

Using (42) and (43) in (41), a linear equation for $\varphi(s)$ obtains so that

$$\varphi(s) = \frac{\phi x(s)}{\sigma_1(s)q(s) + x(s)} \tag{45}$$

where

$$x(s) \equiv (\sigma_1(s) - \sigma_e)\frac{q(s)}{q'(s)} - (1-s)\sigma_e = (1-s)\left(s\left(\Delta - \phi + \varphi(s)\right) - \sigma_e\right). \tag{46}$$

Note that $x(s) \geq 0$ in $\mathcal{S}_2$. Using (44) and (45), gives the following quadratic equation for x

$$(1 + q'(s))x^2 + c_1(s)x + c_2(s) = 0$$

with

$$\begin{aligned}
c_1(s) &\equiv \sigma_e(q(s) + (1 - s)q'(s)) - (1 - s)(\Delta s - \sigma_e)(1 + q'(s)) + \phi(1 - s)sq'(s), \\
c_2(s) &\equiv \sigma_e(1 - s)(q(s) + (1 - s)q'(s))(\sigma_e - s(\Delta - \phi)).
\end{aligned}$$

The solution is given by

$$x(s) = \frac{-c_1(s) \pm \sqrt{c_1(s)^2 - 4(1 + q'(s))c_2(s)}}{2(1 + q'(s))},$$

and thus there are two real roots (assuming the square root is well defined - note that it is indeed the case at the limits $s^{**}, 1$). The expression in (32) follows by using the upper branch in the definition of x in (46) and solving for σ_1. The upper branch is selected as

$$\begin{aligned}
\lim_{s \downarrow s^{**}} \sigma_1(s) &= 1_{\{-\}} \left(\frac{\sigma_e \left(q(s^{**})(\Delta - \phi)^2 + \Delta q'(s^{**})(\Delta - \sigma_e - \phi) \right)}{q(s^{**})(1 + q'(s^{**}))(\Delta - \phi)^2} \right) \\
&\quad + 1_{\{+\}} \left(1 + \frac{q'(s^{**})}{q(s^{**})}(1 - s^{**}) \right) \sigma_e. \\
\lim_{s \uparrow 1} \sigma_1(s) &= 1_{\{-\}} \left(\frac{\sigma_e}{1 + q'(s)} \right) + 1_{\{+\}} \sigma_e.
\end{aligned}$$

$\blacksquare$

Table 1: Summary statistics for decile portfolios. Table reports mean (*median*) quantities associated with fee deciles. The data cover the sample period from January 2004 to December 2014. As done in previous literature, we filter the data to exclude assets with a stock price below $5 per share. There are 270 firms per decile.

Decile	Fee (bps)	Vol (%)	IO (%)	DISP (%)	SIR (%)	Util (%)	MktCap ($bil)	$r_{it} - r_{ft}$ (%)
1 (Cheap)	4.6	2.2	65.6	8.9	3.5	6.86	14.3	0.5
	4.6	*1.8*	*72.5*	*8.9*	*1.8*	*2.95*	*3.0*	*0.4*
2	8.7	2.3	73.6	10.0	4.3	12.14	7.1	0.9
	8.0	*1.9*	*78.8*	*10.0*	*3.2*	*7.75*	*2.2*	*0.8*
3	10.1	2.4	73.6	10.0	4.9	14.17	4.0	1.1
	9.1	*2.0*	*78.8*	*10.0*	*3.8*	*10.13*	*1.3*	*0.8*
4	10.9	2.5	70.7	12.6	4.9	14.19	2.5	1.1
	9.8	*2.1*	*76.2*	*12.6*	*3.8*	*10.24*	*0.9*	*0.9*
5	11.6	2.6	67.1	13.9	4.5	13.6	1.9	1.0
	10.2	*2.2*	*71.7*	*13.9*	*3.5*	*9.44*	*0.6*	*0.6*
6	12.4	2.7	62.3	15.9	4.0	12.7	1.8	1.0
	10.7	*2.3*	*65.2*	*15.9*	*2.9*	*8.00*	*0.4*	*0.6*
7	14.3	2.8	59.0	18.2	4.2	13.97	2.5	1.2
	12.6	*2.3*	*60.3*	*18.2*	*2.6*	*8.06*	*0.3*	*0.6*
8	21.3	2.9	58.3	21.6	5.4	19.95	3.1	1.3
	17.0	*2.4*	*59.5*	*21.6*	*3.3*	*13.9*	*0.4*	*0.5*
9	52.1	3.2	52.3	22.7	6.3	26.76	3.6	1.1
	36.6	*2.6*	*51.4*	*22.7*	*3.4*	*22.06*	*0.3*	*0.2*
10 (Expensive)	642.7	3.9	41.7	31.1	9.7	50.05	1.8	0.6
	284.3	*3.2*	*35.3*	*31.1*	*6.2*	*55.8*	*0.2*	*-0.3*

Notes:
Fee (bps/year): MSF reports the value-weighted average lending fee for each security over the past 1, 3, 7, and 30 days, where the value weight assigned to a loan fee is the dollar value of the outstanding balance of the loan for that transaction divided by the total dollar value of outstanding balances for that time period. In keeping with the literature, we analyze trading strategies that are rebalanced monthly, and therefore use the 30-day value-weighted average fee as our measure of a stock's shorting fee. If an observation is missing the 30-day value-weighted average fee, we drop it from the sample.
Vol (%): Monthly volatility of the stock return is the sum of the squared daily stock returns over a month (see French et al. (1987) and Schwert (1989)).
IO(%): Institutional ownership. Data on institutional holdings is from Thompson Reuters Financial.
DISP(%): Analyst earnings forecast dispersion (DISP) is the standard deviation of annual earnings-per-share forecasts scaled by the absolute value of the average outstanding forecast (see Diether et al. (2002b)). Analyst forecast data is from I/B/E/S.
SIR(%): Short interest ratio. The raw short interest numbers from Compustat are reported as the number of shares that are held short in a given firm are divided by shares outstanding from CRSP.
Util(%): Utilization is a measure provided by Markit defined as the percentage of value of assets on loan (from Beneficial Owners) divided by the total lendable assets (Inventory Value).
$r_{it} - r_{ft}$: Monthly excess returns. r_{it} is the return of portfolio i computed using equal weights. r_f is the risk free rate from Kenneth French's website.

Table 2: Alphas for adjusted returns per decile portfolio and cheap-minus-expensive (CME) portfolios. CME portfolios are constructed using long-short legs with deciles 1-10 and 1-10b. Decile 10b corresponds to the more expensive half of decile 10. t-stats are computed using Newey-West standard errors with 12 lags. The No Adj column runs (34) with no adjustment term. Model (1) and (2) columns apply (35) and (36) to (34) respectively. *,**,*** indicate significance at the 10%, 5% and 1% levels.

Factors	decile	α	*t-stat*	α	*t-stat*	α	*t-stat*
			(a)				
		No Adj		Model (1)		Model (2)	
FF3	1	0.01	(0.14)	0.01	(0.14)	0.01	(0.21)
	2	0.32***	(3.35)	0.32***	(3.36)	0.32***	(3.42)
	3	0.34***	(6.31)	0.34***	(6.31)	0.34***	(6.43)
	4	0.32***	(3.47)	0.32***	(3.47)	0.33***	(3.56)
	5	0.20	(1.52)	0.20	(1.52)	0.21	(1.59)
	6	0.21*	(1.75)	0.21*	(1.75)	0.22*	(1.83)
	7	0.53***	(3.25)	0.53***	(3.25)	0.54***	(3.32)
	8	0.46**	(2.37)	0.46**	(2.37)	0.48**	(2.46)
	9	0.37	(1.47)	0.37	(1.48)	0.41*	(1.65)
	10	-0.32**	(-1.96)	-0.26	(-0.77)	0.23	(0.68)
Long/Short	1-10	0.51	(1.35)	0.46	(1.22)	-0.21	(-0.64)
	1-10b	0.84**	(2.05)	0.75*	(1.89)	-0.23	(-0.62)
			(b)				
		No Adj		Model (1)		Model (2)	
FF3+mom	1	0.02	(0.40)	0.02	(0.41)	0.02	(0.47)
	2	0.35***	(4.15)	0.35***	(4.16)	0.35***	(4.22)
	3	0.36***	(6.89)	0.36***	(6.89)	0.36***	(6.98)
	4	0.35***	(5.21)	0.35***	(5.22)	0.36***	(5.32)
	5	0.24**	(2.34)	0.24**	(2.35)	0.24**	(2.43)
	6	0.24**	(2.08)	0.24**	(2.08)	0.25**	(2.16)
	7	0.57***	(4.51)	0.57***	(4.52)	0.58***	(4.61)
	8	0.52***	(4.12)	0.52***	(4.13)	0.54***	(4.28)
	9	0.44**	(2.40)	0.44**	(2.41)	0.48**	(2.68)
	10	-0.20*	(-1.94)	-0.14	(-0.61)	0.35	(1.59)
Long/Short	1-10	0.42	(1.40)	0.37	(1.23)	-0.31	(-1.21)
	1-10b	0.77**	(2.32)	0.68*	(1.97)	-0.30	(-0.95)

Table 3: Fama-MacBeth regressions, t-stats are reported in parenthesis. Panel (a) applies (35). Panel (b) applies (36). Factors are drawn from Kenneth French's data library.

Model	$\bar{\alpha}$	$\bar{\gamma}$	$\bar{\lambda}_{\text{mkt}}$	$\bar{\lambda}_{\text{hml}}$	$\bar{\lambda}_{\text{smb}}$	$\bar{\lambda}_{\text{mom}}$
			(a)			
FF3	-0.42	-19.49**	1.74	2.63	0.42	
	(-0.27)	(-1.97)	(0.49)	(0.62)	(0.19)	
FF3+mom	-0.44	-19.60**	1.77	2.62	0.43	-2.41
	(-0.28)	(-2.05)	(0.49)	(0.66)	(0.22)	(-0.65)
			(b)			
FF3	-0.51	-1.59**	1.82	2.51	0.46	
	(-0.30)	(-1.98)	(0.51)	(0.60)	(0.21)	
FF3+mom	-0.49	-1.61**	1.79	2.52	0.46	-2.45
	(-0.30)	(-2.06)	(0.50)	(0.64)	(0.23)	(-0.66)

Notes: mkt = CRSP value-weighted market excess return; smb (hml) = small minus big size (high minus low value) factor; mom = up minus down momentum factor, respectively. *,**,*** indicate significance at the 10%, 5% and 1% levels, respectively.

Table 4: Summary statistics and OLS estimates predictive regression.

Variable	Mean	Median	Std dev	b	*t-stat*	R^2
DP	-3.92	-3.95	0.16	0.16	(0.24)	0.16
DY	-3.92	-3.94	0.16	0.38	(0.63)	0.85
EP	-3.04	-2.90	0.47	-0.12	(-0.21)	0.08
DE	-0.88	-1.09	0.59	0.14	(0.23)	0.11
RVOL (ann)	0.14	0.12	0.06	0.42	(1.13)	1.05
BM	0.31	0.32	0.05	0.38	(0.95)	0.86
NTIS	-0.00	0.00	0.02	-0.82	(-1.37)	3.90
TBL (ann)	0.01	0.00	0.02	0.37	(1.30)	0.78
LTY (ann)	0.04	0.04	0.01	0.74^{**}	(2.47)	3.13
LTR	0.01	0.01	0.03	0.33	(0.84)	0.64
TMS (ann)	0.03	0.03	0.01	-0.01	(-0.03)	0.00
DFY (ann)	0.01	0.01	0.01	-0.33	(-0.48)	0.65
DFR	0.00	0.00	0.02	0.64	(0.87)	2.38
INFL	0.00	0.00	0.00	-0.49	(-0.99)	1.39
SII (-)	0.00	-0.19	1.00	0.99^{*}	(1.88)	2.84
VARPHI (-)	0.86	0.70	0.58	1.03^{*}	(1.76)	2.43
VARPHI(-)–PC	–	–	–	3.09^{**}	(1.98)	4.57

Notes: The database contains 132 monthly observations from January 2004 to December 2014. Table 4 displays summary statistics for 14 predictor variables from Welch and Goyal (2008), the aggregate short interest in Rapach et al. (2016), labeled SII, and $\hat{\varphi}q$, labeled VARPHI, from (37). DP is the log dividend-price ratio, DY is the log dividend yield, EP is the log earnings-price ratio, DE is the log dividend-payout ratio, RVOL is the volatility of excess stock returns, BM is the book-to-market value ratio for the DJIA, NTIS is net equity expansion, TBL is the interest rate on a three-month Treasury bill, LTY is the long-term government bond yield, LTR is the return on long-term government bonds, TMS is the long-term government bond yield minus the Treasury bill rate, DFY is the difference between Moody's BAA- and AAA-rated corporate bond yields, DFR is the long-term corporate bond return minus the long-term government bond return, and INFL is inflation calculated from the CPI for all urban consumers.

Each predictor variable is standardized to have a standard deviation of one. Brackets below the b estimates report heteroskedasticity- and autocorrelation-robust t-statistics for testing $H_0 : b = 0$ against $H_A : b > 0$. *,**,*** indicate significance at the 10%, 5% and 1% levels, respectively.

Figure 2: Comparative statics sharing rule $q(s)$. Base parameters $(\rho, \bar{\mu}, \eta, \mu_e, \sigma_e) = (.01, .008, .5, .02, .03)$

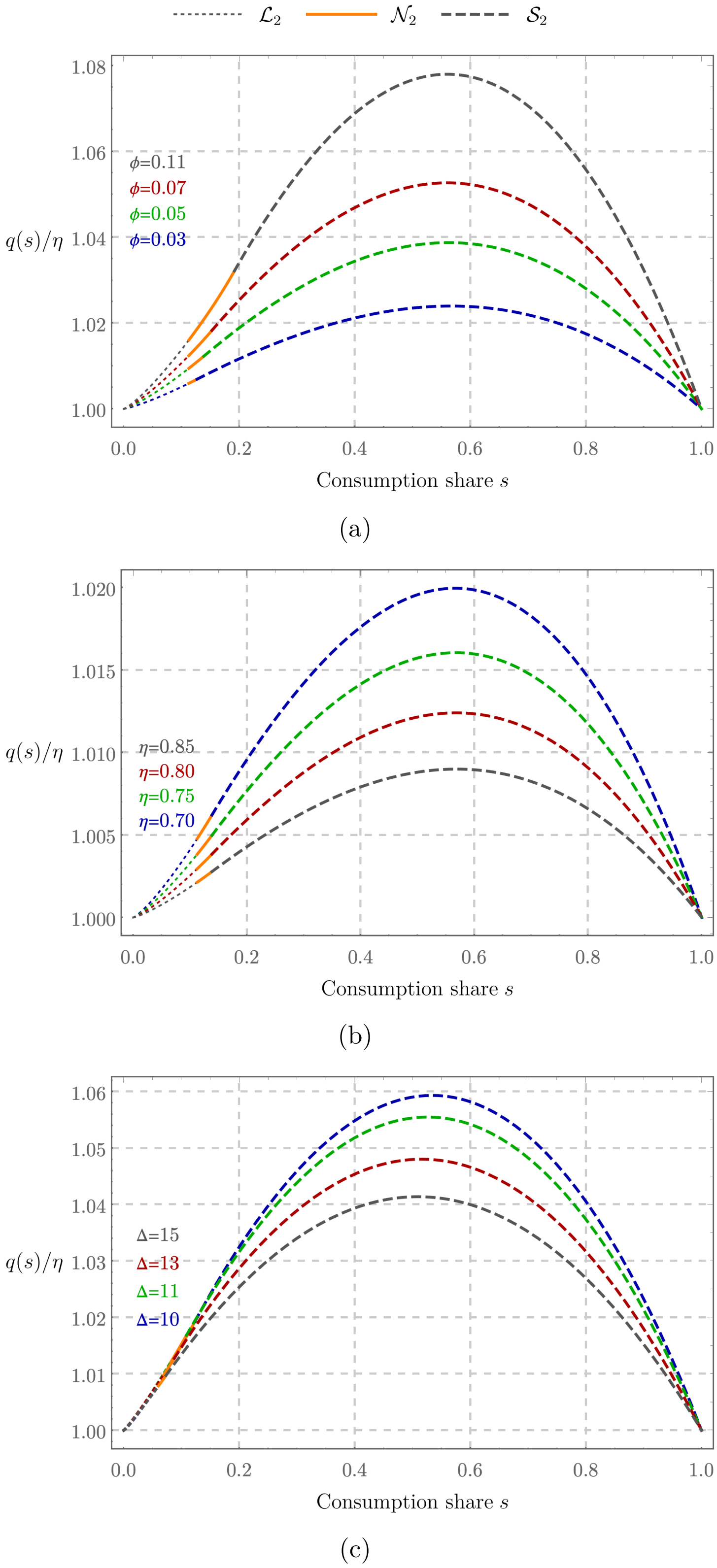
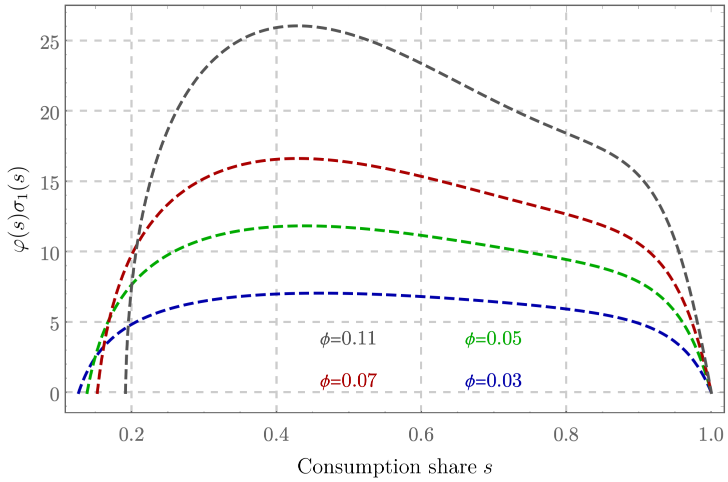
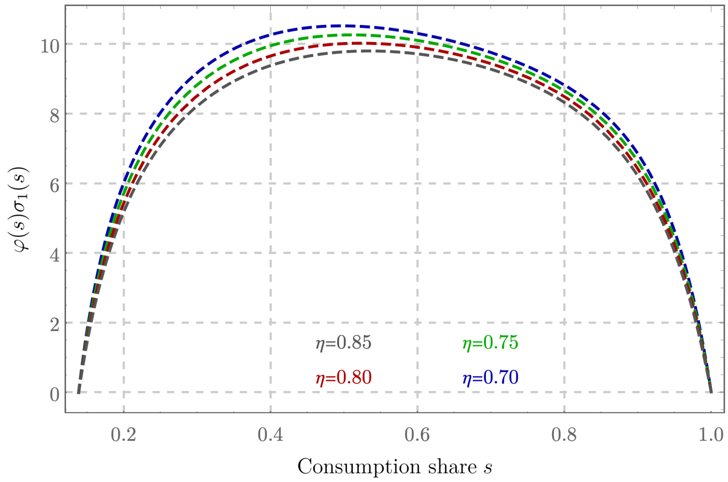


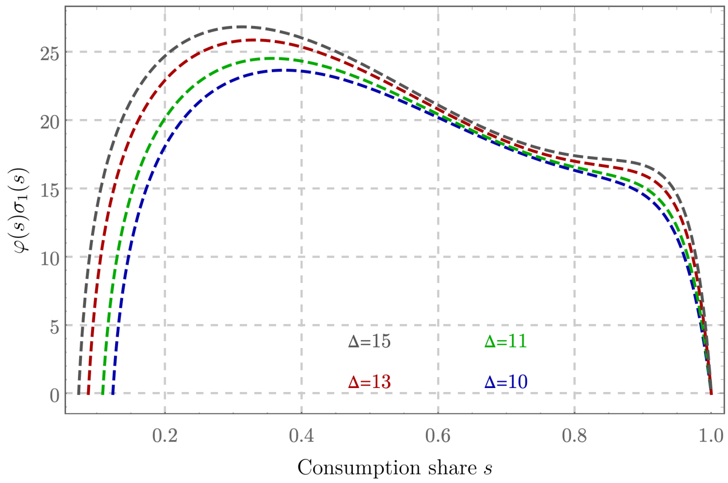
Figure 3: Comparative statics convenience yield $\varphi(s)\sigma_1(s)$. Base parameters $(\rho, \bar{\mu}, \phi, \mu_e, \sigma_e, \eta) = (.01, .008, .005, .02, .04, 0.5)$



(a)

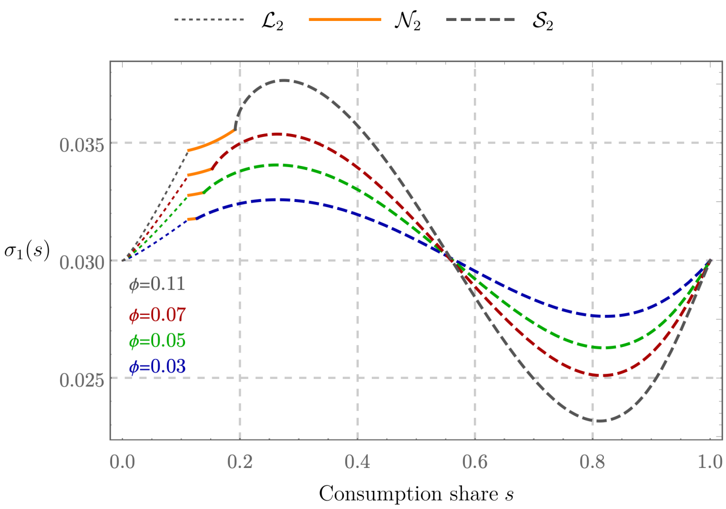


(b)

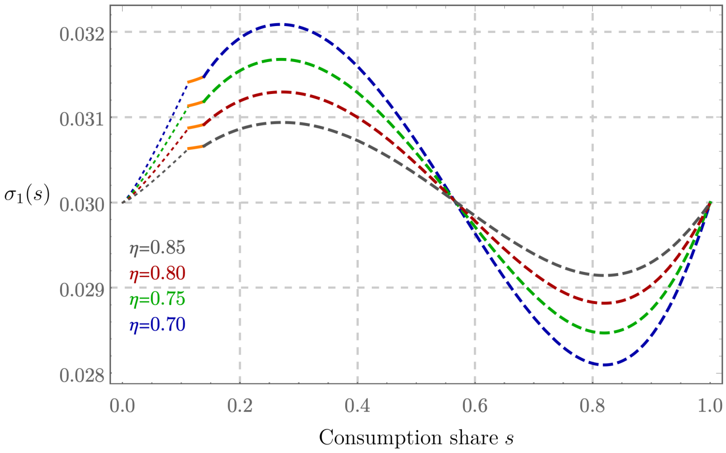


(c)

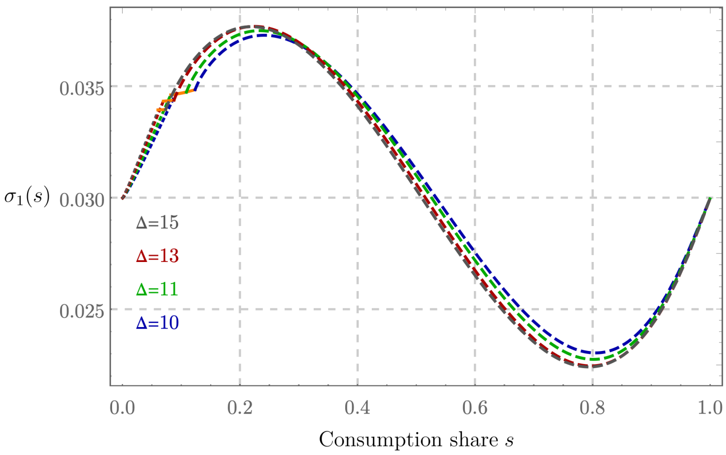
Figure 4: Comparative statics for volatility of Stock 1. Base parameters $(\rho, \bar{\mu}, \phi, \mu_e, \sigma_e, \eta) = (.01, .008, .005, .02, .04, 0.5)$



(a)



(b)



(c)

References

Atmaz, A., Basak, S., 2019. Option prices and costly short-selling. forthcoming, Journal of Financial Economics.

Banerjee, S., Graveline, J. J., 2014. Trading in derivatives when the underlying is scarce. Journal of Financial Economics 111 (3), 589–608.

Basak, S., Croitoru, B., 2000. Equilibrium mispricing in a capital market with portfolio constraints. Review of Financial Studies 13 (3), 715–748.

Beneish, M. D., Lee, C. M., Nichols, D., 2015. In short supply: Short-sellers and stock returns. Journal of Accounting and Economics 60 (2), 33–57.

Blocher, J., Reed, A. V., Van Wesep, E. D., 2012. Connecting two markets: An equilibrium framework for shorts, longs, and stock loans. Journal of Financial Economics.

Blocher, J., Whaley, R. E., 2015. Passive investing: The role of securities lending. Vanderbilt Owen Graduate School of Management Research Paper.

Brennan, M. J., 1991. The price of convenience and the valuation of commodity contingent claims. Stochastic models and option values 200, 33–71.

Buss, A., Dumas, B., 2017. Trading fees and slow-moving capital. Working paper, Insead.

Cherian, J. A., Jacquier, E., Jarrow, R. A., 2004. A model of the convenience yields in on-the-run treasuries. Review of Derivatives Research 7 (2), 79–97.

Cherkes, M., Jones, C. M., Spatt, C. S., 2013. A solution to the palm-3com spin-off puzzles. Columbia Business School Research Paper (12/52).

Cochrane, J. H., 2002. Stocks as money: convenience yield and the tech-stock bubble. Tech. rep., National Bureau of Economic Research.

Cohen, L., Diether, K., Malloy, C., 2007. Supply and demand shifts in the shorting market. Journal of Finance 62, 2061–2096.

Cuoco, D., Cvitanić, J., 1998. Optimal consumption choices for a large investor. Journal of Economic Dynamics and Control 22 (3), 401–436.

Cuoco, D., Liu, H., 2000. A martingale characterization of consumption choices and hedging costs with margin requirements. Mathematical Finance 10 (3), 355–385.

D'Avolio, G., 2002. The market for borrowing stock. Journal of Financial Economics 66, 271–306.

Diether, K., Malloy, C., Scherbina, A., 2002a. Differences of opinion and the cross section of stock returns. Journal of Finance 57 (5), 2113–2141.

Diether, K. B., Malloy, C. J., Scherbina, A., 2002b. Differences of opinion and the cross section of stock returns. The Journal of Finance 57 (5), 2113–2141.

Drechsler, I., Dreschler, Q. F., 2016. The shorting premium and asset pricing anomalies. Working paper, NBER and Wharton UPenn.

Duffie, D., 1996. Special repo rates. Journal of Finance 51 (2), 493–526.

Duffie, D., Gârleanu, N., Pedersen, L. H., 2002. Securities lending, shorting, and pricing. Journal of Financial Economics (66), 307–339.

French, K. R., Schwert, G. W., Stambaugh, R. F., 1987. Expected stock returns and volatility. Journal of financial Economics 19 (1), 3–29.

Gallant, A., Rossi, P., Tauchen, G., 1992. Stock prices and volume. Review of Financial studies 5 (2), 199–242.

Gallmeyer, M., Hollifield, B., 2008. An examination of heterogeneous beliefs with a short sale constraint in a dynamic economy. Review of Finance 12 (2), 323–364.

Hajek, B., 1985. Mean stochastic comparison of diffusions. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 68 (3), 315–329.

Harrison, J. M., Kreps, D. M., 1978. Speculative investor behavior in a stock market with heterogeneous expectations. The Quarterly Journal of Economics 92 (2), 323–336.

Jarrow, R., 1980. Heterogeneous expectations, restrictions on short sales, and equilibrium asset prices. Journal of Finance 35, 1105–1113.

Krishnamurthy, A., 2002. The bond/old-bond spread. Journal of Financial Economics 66 (2), 463–506.

Lamont, O. A., Thaler, R. H., 2003. Can the market add and subtract? Mispricing in tech stock carve-outs. Journal of Political Economy 111 (2), 227–268.

Miller, E., 1977. Risk, uncertainty, and divergence of opinion. Journal of Finance 32, 1151–1168.

Mitchell, M., Pulvino, T., Stafford, E., 2002. Limited arbitrage in equity markets. Journal of Finance 57, 551–584.

Muravyev, D., Pearson, N. D., Pollet, J. M., 2018. Is there a risk premium in the stock lending market? evidence from equity options. Working paper, Boston College.

Nielsen, L., 1989. Asset market equilibrium with short-selling. Review of Economic Studies 56, 467–473.

Nutz, M., Scheinkman, J. A., 2019. Shorting in speculative markets. Forthcoming, Journal of Finance.

Pindyck, R. S., May 1992. The present value model of rational commodity pricing. Working Paper 4083, National Bureau of Economic Research.

Prado, M. P., 2015. Future lending income and security value. Journal of Financial and Quantitative Analysis 50 (04), 869–902.

Radner, R., 1972. Existence of equilibrium of plans, prices and prices expectations in a sequence of markets. Econometrica 40 (2), 289–303.

Rapach, D. E., Ringgenberg, M. C., Zhou, G., 2016. Short interest and aggregate stock returns. Journal of Financial Economics 121 (1), 46–65.

Reed, A. V., 2013. Short selling. Annual Review of Financial Economics 5 (1).

Schwert, W. G., 1989. Why does stock market volatility change over time. Journal of Finance 44, 1115–1153.

Vayanos, D., Weill, P.-O., 2008. A search-based theory of the on-the-run phenomenon. The Journal of Finance 63 (3), 1361–1398.

Welch, I., Goyal, A., 2008. A comprehensive look at the empirical performance of equity premium prediction. Review of Financial Studies 21 (4), 1455–1508.