The Capacity of Trading Strategies *

Augustin Landier† Guillaume Simon‡ David Thesmar§

March 11, 2015

Abstract

Due to non-linear transaction costs, the financial performance of a trading strategy decreases with portfolio size. Using a dynamic trading model a la Garleanu and Pedersen (2013), we derive closed-form formulas for the performance-to-scale frontier reached by a trader endowed with a signal predicting stock returns. The decay with scale of the realized Sharpe ratio is slower for strategies that (1) trade more liquid stocks (2) are based on signals that do not fade away quickly and (3) have strong frictionless performance. For an investor ready to accept a Sharpe reduction by 30%, portfolio scale (measured in dollar volatility) is given by $\frac{SR^*}{10\lambda\phi^2}$, where SR^* is the frictionless Sharpe, λ a measure of price impact, and ϕ a measure of the speed at which the signal fades away. We apply the framework to four well-known strategies. Because stocks have become more liquid, the capacity of strategies has *increased* in the 2000s compared to the 1990s. Due to high signal persistence, the capacity of a "quality" strategy is an order of magnitude larger than the others and is the only one highly scalable in the mid-cap range.

*We thank Pierre Collin-Dufresne, Gulten Mero, Mark Potters, Emmanuel Sérié, David Sraer and participants at the QuantInvest 2014 and 7th Annual Hedge Fund Research Conferences for insightful discussions. Bruno Durin's input, at a very early stage of this project, was critical.

†Toulouse School of Economics and CFM (e-mail: augustin.landier@tse-eu.fr)

‡CFM (e-mail: guillaume.simon@cfm.fr)

§HEC, CEPR and CFM (e-mail: thesmar@hec.fr)

1 Introduction

The empirical asset-pricing literature documents a wide array of variables that predict stock returns (Harvey et al., 2014). In various settings, these "signals" are shown to correlate with future returns in a statistically significant way: This is typically done either by using Fama-MacBeth stock-level regressions or by computing Jensen's alphas of sorted long-short portfolios, controlling for known risk factors. The interpretation of such findings is either that the Efficient Market Hypothesis fails, or that the benchmark asset pricing model is missing an important risk factor. Under both interpretations, statistical significance implicitly measures the size of the economic "anomaly". The asset management industry can exploit these signals by offering investment strategies that can be viewed either as profitable arbitrage portfolios or as diversifying exposure to a new risk factor. However, even when the predictive power of the underlying signal is statistically significant, some of these investment strategies are difficult to implement at a reasonably large scale. The reason is that transaction costs increase more than linearly with the size of trading portfolios, which makes profitability shrink (Korajczyk and Sadka (2004), Frazzini et al. (2012)). This concern is especially acute for signals that work mostly with small caps – or equivalently for signals that work better in equal-weighted than value-weighted terms. It is also present when signals are not persistent enough, since they force investors to trade more often. Illiquidity and signal dynamics limit the extent to which asset managers can scale their investments up. Overall, in order to assess the economic significance of an asset-pricing anomaly, it is crucial to determine the amount that can be effectively invested on it, i.e. its capacity.

This paper offers a framework to compute and estimate the capacity of a trading strategy. The three key inputs are: (1) gross performance (the Sharpe ratio of the arbitrage portfolio absent transaction costs), (2) liquidity of underlying securities and (3) dynamics of the signal on which the strategy is based. Intuitively, strategies based on a highly predictive signal, trading more liquid securities, and on slow-moving signals, have greater capacity. To evaluate the relative importance of these three ingredients, we use the dynamic trading model proposed by Garleanu and Pedersen (2013). In this model, the investor's flow utility is equal to trading profits, minus a penalty for portfolio risk (scaled by risk aversion, as in the classical Markowitz model) and another penalty reflecting the cost of trading. The main advantage of this model is that it rests on intertemporal optimization, which turns out to be critical to study the effect of signal dynamics on capacity. The

intuition is that the model optimally "slows-down" trading to mitigate the impact of transaction costs. When the signal is relatively slow-moving, it may be optimal to trade more aggressively that in a purely static set-up. This comes from the fact that the position being built now will pay off during a longer period. A dynamically trading investor takes this effect into account, while a static optimizer does not. A constraint of the intertemporal framework is that, in order to solve explicitly the optimization problem, we are restricted to quadratic transaction costs; Its advantage is that we obtain simple tractable formulas.

We use this model to obtain a closed-form relationship between the trading scale of a strategy and its effective risk-adjusted performance. We define trading scale as the dollar volatility targeted by the investor, which measures the amount of money effectively put at work in the strategy. For a given trading scale, the model pins down the optimal portfolio at each point in time, and therefore its Sharpe ratio net of transaction costs. We call this curve the "performance-to-scale" frontier: It tells the net Sharpe that an optimizing trader can expect when targeting a given P&L volatility.

For a large class of "fundamental" signals and assuming investments at the scale of the asset-management industry, we find closed forms for the performance-to-scale frontier. Consider a strategy which, neglecting transaction costs, has a Sharpe ratio SR^* in a given pool of stocks. We find the following simple expression relating the scale of the portfolio (measured by the dollar volatility of profits, Vol) and the realized Sharpe ratio:

$$Vol = \frac{SR}{\lambda\phi^2}\left[\left(\frac{SR^*}{SR}\right)^{2/3} - 1\right]^2. \tag{1}$$

In this formula, λ measures the illiquidity of the stocks that are traded, and ϕ measures the speed at which the signal underlying the trading strategy fades away. In line with economic intuition, the realized Sharpe SR is equal to SR^* at very small portfolio scales (Vol) and declines as Vol increases; This decline is sharper when underlying stocks are more illiquid (i.e. λ higher) or when the signal fades away quickly (i.e. ϕ higher). For instance, for an investor ready to accept a performance degradation by 30% ($SR = 0.7SR^*$), the portfolio scale that can be reached is given by $Vol = \frac{SR^*}{10\lambda\phi^2}$.

Constructing the "scale-to-performance frontier" (1), which links the realized Sharpe and the portfolio scale, is the central result of our paper. This relationship shows that the capacity has an elasticity of 1 w.r.t. to the price impact factor, and an elasticity of 2 w.r.t. to the speed of signal mean-reversion. This result implies that high-frequency signals have capacities that are very small

compared to "fundamental" signals based on corporate accounts (who have lower ϕ's). We also provide a simple formula that shows how much of the gross P&L should be "burnt" in transaction costs by an optimizing trader aiming for a given Sharpe degradation. These amounts are large: Typically, an investor ready to accept a performance degradation by 30% would lose as much as 20% of his P&L in transaction costs.

We then *calibrate* this model on four classical "fundamental" strategies and two different pools of US stocks (large caps, and mid caps. The four strategies are: book-to-market value of equity or "value" (Fama and French, 2006), minus rolling volatility or "lowvol" (Ang et al., 2009), minus growth in shares outstanding or "repurchasers" (Pontiff and Woodgate, 2008) and operating cash flows to assets, or "quality" (Novy-Marx (2013), Asness et al. (2014)). We use US data to estimate, for each strategy, the persistence of the signal and the gross performance. For each pool of stocks, we estimate the price impact coefficient. We then inject these parameters in the model, implicitly assuming that the data-generating process (DGP) and the price impact function used the model are correct. We find that under this assumption, the capacity of "quality" is an order of magnitude larger than the others, with a reachable volatility about $80bn in the large cap pool, and 10bn in the mid-cap range. Also, the calibration exercise shows that the capacity of strategies has *increased* in the 2000s compared to the 1990s, mostly because all pools (including mid caps) have become significantly more liquid. Thus, if anything, these anomalies continue to be large, in spite of the inflow of arbitrage capital.

We also *backtest* the model on historical returns. This alternative investigation allows to look at the robustness of Garleanu and Pedersen (2013)'s assumptions about the DGP. Using historical returns, we compute what would have been the P&L of a trader aiming for a given $ volatility and optimizing dynamically under the model's assumption. We can thus estimate the "realized" performance-to-scale frontier that an investor following the model's trading rule would have realized on historical returns. We find that this realized frontier does not differ very much from the calibrated one. Hence, the model seems reasonably robust: In some cases, the realized performance is even higher in the backtesting than in our calibration exercise. To further investigate the significance of our simulations, we then simulate counterfactual histories of stock returns where the signal does not predict returns. We find that Sharpe ratios of 0.5 are frequently reached at all scales of trading, even when the signal does not actually predict returns. Given this falsification test, the only strategy that has "significant" capacity is again quality. All in all, our investigation suggests that quality is

the largest anomaly of the four, by a large margin.

We finally investigate the robustness of the model to various assumptions. First, we calibrate the loss made by an investor who would make a mistake on the mean-reversion or the price impact parameters. We find that estimating price impact properly is more important than estimating signal dynamics. Second, we investigate the impact on the performance-to-scale frontier of an error in the DGP. While Garleanu and Pedersen (2013) assume that the signal is an AR(1), we simulate signals that have predictive power but a different time-series structure, and estimate the frontier resulting from an investor trading under the AR(1) hypothesis. We find that for a class of "long memory signals" (AR(k) where $k > 1$, like "lowvol") realized capacity is actually larger than expected because the investor systematically underestimates the persistence of the signal. For signals that are natural moving averages (like "momentum" or "repurchasers"), the framework has in contrast a very low realized capacity. The take-away of these simulations is that understanding the exact dynamics of signal may be a critical ingredient. Finally, in this robustness section, we estimate the loss that a static trader would make compared to an investor that trades dynamically. We find this loss to be significant.

This paper relates to the recent literature that studies the impact of trading costs on the performance of various strategies. This literature departs from the traditional market microstructure literature in that it focuses on price impact, rather than fees and bid-ask spreads. Garleanu and Pedersen (2013) develop a framework that produces an optimal trading rule given a strategy and costs of execution. We use their results in order to compute the capacity of an optimally traded signal. Novy-Marx and Velikov (2014) propose a new rule to avoid trading "too much" when implementing a series of well-known strategies. Their insight is that investors should not react too quickly to changes in signals, in order to avoid useless round-trips. They do not focus on the persistence of strategies, as we do and do not investigate price-impact increasing in trade size (e.g. quadratic transaction costs) but rather focus on linear transaction costs. Korajczyk and Sadka (2004) study how price impact deteriorates returns in trading momentum; Using a calibration, they find that capacity of value-weighted momentum ($ 2 Bn) is much higher than its equal-weighted capacity ($200 Mil.). But their optimization is static which leads to significant underestimation of capacity, as we show in our last Section. Frazzini et al. (2012) analyze actual trade data with both executed and intended trades to infer trading costs, and then analyze the performance of well-known strategies once these trading costs are taken into account. Using these data, they find larger capacity for these strategies

than existing papers. Compared to these studies, our paper uses the dynamic model of Garleanu and Pedersen (2013), which features an investor that takes into account the future implications of its current trading. This turns out to be particularly important for fundamental signals. The dynamic optimization allows us to trade efficiently even for large capacities, so that net-of-transaction cost performance does not become negative in the large investment range, in contrast to Korajczyk and Sadka (2004) and Frazzini et al. (2012).

This paper also indirectly relates to the "skeptical" literature that asks whether published market anomalies are really present in the data. Lean and Pontiff (2014) show that many anomalies disappear as soon as they are published on SSRN, thereby suggesting that arbitrage capital moves in quickly and makes anomalies go away. Harvey et al. (2014) provide a recent discussion of why most trading strategies uncovered by the literature are actually not statistically significant, because of data-snooping bias. Our paper focuses on another line of criticism, which is that anomalies may exist statistically, but that no investor with significant capacity can take advantage of them. We find that this is not the case for a class of slow moving signals. We also find that the increase in stock liquidity between the 1990s and the 2000s has been large enough to compensate for the reduction in the performance of these strategies. So, even if the alpha generated by some anomalies is lower today that in the 1990s, the "size" of these anomalies is as large, sometimes even larger than before. Finally, our analysis strongly underscores the fact that "quality" (Novy-Marx (2013), Asness et al. (2014)) is a very large anomaly, by far the largest of the four that we document here. Such a large deviation from the standard asset-pricing model begs for an explanation.

The next section lays out the dynamic trading framework which is an application of Garleanu and Pedersen (2013). We derive in this section an explicit formula for the link between $ volatility of a strategy (the amount invested) and its Sharpe ratio. We define the notion of capacity as the level of $ volatility that is consistent with a given target Sharpe ratio. We then investigate the effect of signal persistence on volatility, and how this interact with liquidity. We use approximations and make approximations to build intuition and get functional forms. Section 3 describes the data. Section 4 calibrates and backtests the model. Section 5 investigates the robustness of the model. Section 6 concludes.

2 Framework

2.1 Set-Up

2.1.1 Data Generating Process

In this section, we discuss the return generating process, which is essentially the same as Garleanu and Pedersen (2013) (henceforth, GP). Let s_t be a signal that is used to forecast returns. s_t is assumed to be a vector whose dimension is equal to the number of securities traded. We assume that this signal can be described as an AR(1) process with persistence parameter ϕ:

$$\Delta s_{t+1} = -\phi s_t + \epsilon_{t+1} \tag{2}$$

where ϕ is a scalar and $\Delta s_{t+1} = s_{t+1} - s_t$. This assumption on the signal's behavior is important to solve the dynamic programming problem. We explore its consequences numerically in Section 5.2.

The signal has some forecasting power over returns. To simplify exposition, we assume that the signal at t is the only variable forecasting returns at $t+1$:

$$r_{t+1} = p_{t+1} - (1 + r_f)p_t = Bs_t + u_{t+1}. \tag{3}$$

In equation (3) we omit exposure to risk factors that price returns. This omission does not affect our results but keeps exposition simple. r_f is the risk-free rate. B is a scaling factor. Another consequence of equation (3) is that past values of the signal (i.e. s_{t-k} for $k > 0$) do not add useful information to predict r_{t+1}, conditionally on s_t. Note also that we follow GP in defining r_t as share price change (adjusted for dividends and splits) rather than returns for tractability of the dynamic optimization problem.

This assumption about the data-generating process fits the GP framework. Their model is written in the spirit of APT: there is a set of K factors, and each security may have a different –but constant – loading on each of the factors. In our paper, we assume that there is one factor per security, and that each security has a loading of B on this factor. So our model corresponds to a special case of theirs, and their formulae apply.

Last, a critical assumption behind equation (3) is that the forecasting power of the signal does not change over time. One possible alternative model would be that the regression coefficient B is itself

noisy, i.e. that $r_{t+1} = (B + \eta_t)s_t + u_{t+1}$. Such a model accommodates better the fluctuations in the performance of the strategy based on the signal s_t. Since the focus of this paper is on transaction costs, it is natural to start with the simple model in equation (3). Also, the GP model needs a constant variance-covariance matrix, so our analytical results do not hold under this alternative assumption.

2.1.2 Portfolio Optimization

We use the results derived in Garleanu and Pedersen (2013). They assume quadratic trading costs (e.g. linear and temporary price impact) defined by a mean variance criterion with risk-aversion of γ and trading costs determined by a liquidity matrix $\lambda\Sigma_u$, where Σ_u is the variance-covariance matrix of price changes (conditional on the signal vector s_t) and λ is a scalar reflecting the illiquidity of the pool of stocks being traded.

The dynamic problem the trader faces is to optimize dynamically over (x_t), while taking into account expected returns, risk and trading costs and the discount rate δ:

$$\max_{(x_{t+s})} E_t\Big\{ \sum_{s\geq 0} \frac{1}{(1+\delta)^s}\Big[-\frac{\lambda}{2}(\Delta x'_{t+s})\Sigma_u(\Delta x_{t+s}) + \frac{1}{1+\delta}\Big(x'_{t+s}r_{t+s+1} - \frac{\gamma}{2}x'_{t+s}\Sigma_u x_{t+s}\Big)\Big]\Big\} \tag{4}$$

where $\Delta x_t = x_t - x_{t-1}$.

Solving the dynamic problem, they obtain the following formula for the optimal portfolio, in number of shares of each stock:

$$x_t = (1-\tau)x_{t-1} + \tau x_t^\star \tag{5}$$

where τ, which Garleanu and Pedersen (2013) label the "trading rate" is the solution of the second order equation:

$$\tau^2 + (\frac{\gamma}{\lambda} + \delta)\tau - \frac{\gamma}{\lambda} = 0 \tag{6}$$

The trading rate τ has nice properties. First, it is smaller than 1. Second, it is a decreasing function of δ. This is in part due to the fact that execution costs are paid one period before returns are obtained: When investors become more impatient (δ goes up), trading costs become more important in present value terms, and investors trade less. Finally, the trading rate is an increasing

function of γ/λ: Less risk-averse (i.e. bigger) investors care relatively more about execution costs, and therefore prefer to trade slowly. At the same time, investors operating on relatively less liquid markets (higher λ) trade more slowly.

Garleanu and Pedersen (2013) call $x_t^\star$ the "aimed portfolio", and is given by:

$$x_t^\star = \frac{1}{\gamma + \phi\lambda\tau}\Sigma_u^{-1} B s_t = \frac{\gamma}{\gamma + \phi\lambda\tau} x_t^M \tag{7}$$

where x_t^M is the Markowitz portfolio $(\gamma\Sigma_u)^{-1} B s_t$.

The aim does not fully respond to changes in the Markowitz portfolio because $\phi\tau > 0$. This comes from the fact that the trader expects the signal to mean-revert (more likely if ϕ is larger). Because trading is costly, the traders knows she is likely to have to wind down the position in the future (at a quicker pace if τ is larger). Thus, the aim is simply a less levered version of the Markowitz portfolio to account for the cost of potential unwinding.

When bringing the Garleanu-Pedersen (henceforth GP) model to the data, four caveats arise, that we discuss here. First, the model omits costs due to broker fees and financing fees. These costs are linear in the amount of $ traded, and hence fees per $ traded do not depend on the pool of stocks traded. They thus shift all of our result on returns by the same amount, but do not affect the comparisons across pools of stocks (for instance, large and mid caps).

Second, the GP model omits shorting fees: While these can vary across stocks, they do not substantially differ across the pools of stocks that we study. True, as mid-caps are more likely to be on special, shorting fees tend to be higher for smaller stocks (see for instance Stambaugh et al. (2012) and Dreschler and Dreschler (2014)). Using proprietary data from a large asset manager, we show that for sufficiently large amounts traded this effect is second order compared to price-impact concerns. Brokers typically split stocks between "General Collateral" (GC) and "Hard to Borrow" (HTB or "on special"). Most stocks belong to the GC category, and for these stocks, the shorting fee is the same (the "GC rate") and hovers around 10bp annually. HTB stocks are those for which the shorting demand is unusually high, or for which stock lenders are difficult to find. For these stocks, the shorting fee can be several percent in annualized terms. In Figure 2, we use our proprietary data on quotes, and show the difference in cost between GC and HTB stocks during the 2012-2014 period. The Panel A of the Figure reports the average excess rate charged for all hard-to-borrow stocks, removing all stocks for which the rate is above 2%. Even for the pool of relatively illiquid

stocks (between the 1000^{th} and the 1500^{th} rank of stock market capitalization, the average cost of shorting for hard-to-borrow stocks is around 10bp. The right panel (Panel B) plots the fraction of *extremely* HTB stocks, i.e. stock for which the extra lending fee is above 2%. This fraction lies between 2% for the least liquid pool, and almost 0% for the most liquid pool. Thus, even extremely HTB stocks are not that expansive to short, compared to the trading costs typically incurred by large asset managers. They are also a very small fraction of the stock universe. On average, the data reveals that the gap in shorting costs between top 500 US stocks (ranked by market cap) vs. stocks that are in the (1000-2000) size range is less than 5bps annualized, which for a strategy with unlevered returns above 1% implies a Sharpe reduction of less than 5 percent.

Third, the GP model omits some rebalancing trades in the objective function by assuming that price changes as opposed to returns are normally distributed. When implementing the model, such rebalancing trades will however show-up as the variance-covariance matrix is evaluated using rolling windows. This will bias simulation results regarding the Sharpe deterioration in a conservative direction as our trades won't be optimal vis-a-vis a more realistic dynamic structure of variance (see Collin-Dufresne et al. (2012)).

Fourth, our trading-cost model assumes no permanent price-impact (i.e. prices revert instantaneously after trading). It is possible to introduce slow reversal of price-impact in GP, but this comes at the cost of not having closed form solutions any more (a very attractive feature of the GP framework). Introducing such costs would lead us to be somewhat more conservative on capacity estimates as it would force traders to slow down their trading further. Brokmann et al. (2014) show that full reversal of prices post trading typically takes a few days. By contrast Frazzini et al. (2012) find in their trading data that 70% of price-impact is permanent.

2.2 Defining Capacity

2.2.1 The Sharpe-to-volatility Frontier

We assume that stock-return variance is largely driven by idiosyncratic noise rather than differences in signals. Under this assumption, we can use equations (5)-(6)-(7) to derive an explicit formulation for the Sharpe-to-volatility frontier:

Proposition 1. *Assume that* $B^2 E s_t' s_t << \Sigma_u$. *Then, for each trading rate* τ *used, we can explicitly*

compute the Sharpe Ratio and the $ volatility that are reached:

$$SR(\gamma) = \Big[1 - \frac{2\phi\tau}{\gamma/\lambda + \phi\tau}\frac{1}{2-\tau}\Big]\Big[\frac{1-(1-\tau)^2}{1-(1-\tau)^2(1-\phi)^2)}\Big]^{1/2} SR^\star$$
$$Vol(\gamma) = \frac{1}{\lambda}\frac{1}{\gamma/\lambda + \phi\tau}\Big[\frac{\tau}{2-\tau}\frac{1+(1-\phi)(1-\tau)}{1-(1-\phi)(1-\tau)}\Big]^{1/2} SR^\star$$

where $SR^\star$ is the Markovitz Sharpe ($SR^\star = B.E(s_t'\Sigma_u^{-1}s_t)^{1/2}$) and τ is function of γ given by equation (6).

Proof. See calculations in Appendix A □

The above system of equations thus describes the Sharpe-to-volatility frontier: An investor of risk aversion γ while optimally reach a risk-adjusted performance $SR(\gamma)$, and a volatility $Vol(\gamma)$. The formulae include another parameter, the trading rate τ which is in fact a function of γ as shown in equation (6). τ increases from 0 when $\gamma = 0$ to 1 when $\gamma \to \infty$. An infinitely risk-averse traders thus aims for a zero volatility and a 100% trading rate since the relative weight of trading costs is negligible (γ much larger than λ). $SR(\gamma)$ goes to $SR^\star$ because the portfolio is infinitely small.

As risk-aversion decreases, it is possible to show that there is a level of risk-aversion below which the target $ volatility increases. The main reason for this is that increasing risk aversion reduces the size of the "aimed" portfolio via two channels: (a) even without dynamic trading, more risk-averse investors take smaller positions (this is the standard static Markowitz effect – the γ in $(\gamma/\lambda + \phi\tau)$) and (b) risk-aversion makes investors more sensitive to potential reversal in the signal that will occur in the long run, so they want to reduce the scale of their aimed portfolio (the τ term in $(\gamma/\lambda + \phi\tau)$, since a is an increasing function of γ). This effect is however counteracted by the fact that the trading speed τ is also an increasing function of γ (risk-averse investors care relatively less about transaction costs). This tends to increase the volatility of the portfolio. Quite reasonably, this force is dominated for scales of capacity that correspond to industry numbers (γ low enough to reach at least million dollar of investments in the strategy); we make the corresponding approximations explicit in the next section and make the link with orders of magnitude coming from the data.

Simultaneously, in this parameter range (γ is low enough; see next Section) the Sharpe ratio decreases as γ increases. The intuition is that very risk-averse investors tend to take smaller positions (to reduce $ volatility). Since their trading is small, performance is not very much impaired by price impact. The formula also receives a simple economic interpretation. The "pure" Markowitz Sharpe

ratio, SR^*, is reduced by optimal trading for two reasons, which correspond to the two terms in the product. The first term corresponds to the impact of trading costs. Trading costs are bigger when the signal is less persistent (ϕ larger) which tends to reduce the effective Sharpe of the strategy. This reduction is there even if the trader trades infinitely fast ($\tau = 1$). The second term corresponds to the loss of Sharpe coming from the fact that the trader is never exactly on a multiple of the Markowitz portfolio (here, the aim). This reduces the Sharpe as portfolio composition is suboptimal from a Sharpe viewpoint. If the signal is perfectly persistent ($\phi = 0$), or when the trading rate is 1, there is no such gap and the term is equal to 1.

Asymptotically, it is easy to see that when γ goes to zero, which corresponds to risk-neutrality, the trading rate goes to zero, $ volatility goes to infinity and SR goes to zero. This is interesting because it shows that in a dynamic trading model, the capacity of a strategy cannot be defined by the break-even constraint $SR = 0$ (as for instance in Frazzini et al. (2012)): When the portfolio becomes large, the trader slows down trading, which makes the Sharpe go to zero, but never become negative.

Together, the two equations of Proposition 1 determine the Sharpe-to-volatility frontier. Higher scale ($ volatility) is reached by less risk-averse investor; Less risk-averse investors invest more, and therefore face bigger trading costs, which reduces the effective performance of the investment strategy. The formulae are however a bit opaque and the dependence of both SR and Vol in γ is ambiguous. To clarify intuition and generate additional comparative statics, we thus study an approximation in the next Section.

2.2.2 Large Investments Approximation

In this Section, we use two approximations to rewrite Sharpe-to-volatility frontier in a more easily interpretable way. The approximation focuses on very large scales of investment for which the optimal trading rate τ is a few percentage points. The second approximation focuses on slow-moving signals. Both approximations will hold in the data.

Assumption 1. ***Large Investments***

1. $\left(\frac{\gamma}{\lambda}\right)^{1/2} \ll 1.$
2. $\delta \ll \left(\frac{\gamma}{\lambda}\right)^{1/2}$

These two assumptions are easy to interpret. When they both hold, it is easy to show that $\tau \approx \sqrt{\gamma/\lambda}$. Hence, the first approximation means that the effective trading rate has to be small: At most a few percents of the portfolio can be traded every day. The second interpretation means that discounting issues (at the daily horizon, again) are negligible compared to the trading rate. In our applications, both assumptions are easily satisfied.

Under the large investment approximation, we can then derive the Sharpe-to-volatility frontier:

Proposition 2. *Assume that the "large investment approximation" holds. Then:*

- *The trading speed is given by*

$$\tau = (\gamma/\lambda)^{1/2}$$

- *The large investment approximation therefore rewrites as $\tau \ll 1$ and $\tau \gg \delta$. Daily portfolio churn needs to be small, but larger than the daily discount rate.*

- *The Sharpe-to-volatility frontier writes as:*

$$SR \approx \Big(\frac{1}{\phi+\tau}\Big)\frac{\tau^{3/2}}{1-\phi}\Big(\frac{1}{\frac{(1-(1-\phi)^2)}{2(1-\phi)^2}+\tau)}\Big)^{1/2} SR^*$$
$$Vol \approx \frac{1}{\lambda\tau^{1/2}}\frac{1}{\tau+\phi}\Big[\frac{1-\phi/2}{\phi+\tau(1-\phi)}\Big]^{1/2} SR^\star$$

Proof. See Appendix B. □

This simply rewrites equations of Proposition 1 under the large investment approximation. As previously discussed the trading rate τ takes a simple form. As is apparent from the equations, γ does not appear explicitly any more but only through the trading rate τ.

The above formula is still hard to interpret. We now add a second approximation:

Assumption 2. ***Slow Signal*** $\phi \ll 1$.

This second approximation is valid for most strategies except the most high frequency ones. Note that ϕ measures the speed of mean-reversion at the daily frequency. For instance, for a signal with a half-life of 10 days, we have that $\phi = .066$. Hence, the "slow signal" approximation is not valid for the daily mean-reversion, but it is for instance valid for the "leader-laggards" strategy which buys small stocks in industries where large firms have just announced favorable earnings (Hou, 2007). For the fundamental signal that we study in this paper, ϕ is in the 10^{-3} range.

Using the formulae of Proposition 2 and the slow signal approximation, we obtain the following set of results:

Proposition 3. *Assume the "large investment" and the "slow signal" approximations hold:*

- *The large investment approximation rewrites:* $\sqrt{\frac{SR}{\lambda Vol}} \ll 1$ *and* $\sqrt{\frac{SR}{\lambda Vol}} \gg \delta$.
- *The targeted \$ volatility is an explicit function of the pure and targeted Sharpes, the signal persistence and the pool's liquidity:*

$$Vol = \frac{SR}{\lambda \phi^2} \left[\left(\frac{SR^*}{SR} \right)^{2/3} - 1 \right]^2$$

  For a given targeted Sharpe, the \$ volatility has an elasticity of 2 w.r.t. signal persistence ϕ*, and an elasticity of 1 w.r.t. to price impact* λ.

- *For a given \$ volatility, performance decreases faster with liquidity when the signal is less persistent:*

$$\frac{\partial^2 SR}{\partial \lambda \partial \phi} < 0$$

- *Transactions costs are a simple fraction of the gross profit of the strategy:*

$$\frac{TC}{ER} = \frac{\phi}{\phi + \tau} = 1 - \left(\frac{SR}{SR^*} \right)^{2/3}$$

  where ER *is the steady state expected \$ gross profit (i.e. net profit plus transaction costs) from the strategy.*

Proof. See Appendix C. □

The above Proposition summarizes the core message of this paper, i.e. that trading costs matter less when the signal is more persistent. The first bullet point defines the range for which the approximations are valid. In the following, we will assume $\delta \approx 8.10^{-5}$, which corresponds to an annualized discount rate of 2%. As we will see later in our data, the price impact parameter $\lambda \approx 10^{-5}$ in the mid-cap range (we describe the data and the calibrated parameters in greater detail in Section 3). Assume to simplify that we are targeting a Sharpe ratio of 0.5. The first bullet point of Proposition 3 simply states that, as long as Vol is at least 100m\$, and below 40bn\$, the two parts

of large investment approximation are satisfied.[1]

The second bullet point combines the two equations of Proposition 2 to compute the maximum $ volatility compatible with a given Sharpe ratio. This value is the "capacity" of a strategy. It is obviously an increasing function of the "pure" Markowitz Sharpe SR^*. Assume for instance that the investor aims for an effective Sharpe of .5. Then, the term in $SR\left[(SR^*/SR)^{2/3}-1\right]^2$ will be worth .008 if the pure Sharpe is .6 (as for instance for the low vol strategy), and .11 if the pure Sharpe is .9 (as for the cash-flow strategy). Thus, the model suggests that, even if one assumes equal persistence and liquidity, the cash-flow strategy will have 10 times as much capacity as the low vol strategy, simply because its pure performance is better.

A simple rule of thumb about Sharpe decay can be derived as follows from the second bullet point: SR is 30% lower than the frictionless Sharpe ratio SR^* when the portfolio scale (measured in dollar volatility) is

$$Vol = \frac{SR^*}{10\lambda\phi^2}.$$

Thus, Sharpe decay with scale is sharper when underlying stocks are more illiquid (i.e. λ higher) or when the signal fades away quickly (i.e. ϕ higher).

We show these comparative statics graphically in Figures 4 and 5, where we investigate the effects of ϕ and λ on the Sharpe-to-volatility frontier defined by the formula in Proposition 2. Figure 4 investigates the effect of liquidity. We consider a fictitious strategy whose frictionless Sharpe ratio is equal to 1, and whose persistence $\phi = 2.10^{-3}$, which roughly corresponds to the average persistence of our fundamental signals. We use 4 different values of λ, which correspond to the median λ in the mid pool in 1991-1995 (1.310^{-4}), 1996-2000 ($1.5.10^{-4}$), 2001-2005 ($7.6.10^{-4}$) and 2006-2013 ($1.8.10^{-4}$). We see there that liquidity has a very large impact on the capacity of strategies. For instance, the Sharpe loss due to trading $ 5bn could be as large as .8 with the liquidity level of the early 1990s, it is not more than .3 with the liquidity level of the early 2000s. This suggest that the increase in liquidity witnessed in the 2000s –in particular in the mid-cap range– led to a considerable increase in the capacity of the fundamental strategies we are studying here.

[1] If Vol = 100m$, then

$$\sqrt{\frac{SR}{\lambda Vol}} = .02 \ll 1$$

while if Vol = 40bn$, then

$$\sqrt{\frac{SR}{\lambda Vol}} = .001 \gg 810^{-5}$$

Figure 5 highlights the effect of the speed of signal mean-reversion ϕ on the scale-performance frontier. Again, we consider a fictitious strategy whose frictionless Sharpe ratio is equal to 1, and a pool whose price impact $\lambda = 1.8e-6$, which corresponds to the liquidity that prevails in the mid pool in 2012. We use 4 different values of ϕ, which correspond to the ϕ of book-to-market in the large (0.610^{-3}, the most persistent strategy), book-to-market in the mid (0.810^{-3}), low vol (1.310^{-3}) and net shares growth (1.810^{-3}, the least persistent strategy). ϕ has a discernable impact on capacity, although less pronounced than liquidity. Assuming the target is a volatility of \$ 5bn, going from the least to the most persistent strategy allows to reduce the Sharpe degradation by about .1. This shows that, in the range of persistence for the fundamental strategies we are looking at, capacity does not differ much.

The relative insensitivity of capacity w.r.t. ϕ comes from the fact that we focus on a group of similarly slow strategies. More "transient signals" have little capacity, and this is why we do not include them in our study. Take for instance the standard daily mean-reversion, which uses minus last day's return as the signal. This high-speed strategy has a $\phi \approx 1$ –since returns are almost i.i.d (see for instance Lo and Khandani (2008) for a description). For such values of ϕ, the "large investment approximation" is still valid, but the "slow signal" approximation is not any more

If one however assumes that $\phi \gg \tau$, which is typically the case when $\phi \approx 1$, it is then easy to show that the capacity frontier is given by:

$$\text{Vol} = \frac{SR\left(1-\frac{\phi}{2}\right)^{1/3}}{\lambda\phi^2}\left(\frac{SR^*}{SR}\right)^{4/3}$$

which shows quite clearly that the capacity becomes tiny when ϕ is near unity. This comes from the fact that the Sharpe decreases with ϕ^2 both for high and low values of mean-reversion, so going from $\phi \approx 1$ (daily mean-reversion) to $\phi \approx 10^{-3}$ (our fundamental strategies) essentially multiplies capacity by a factor of about 10^6. Given the above formula, assuming for instance that $SR^* = 5$, $SR = .5$ and $\phi = 1$, one obtains a capacity of about \$ 760k, a tiny fraction of what can be obtained with slower strategies. So in general, signal persistence has a big effect, just not so much in the range ($\phi \ll 1$) we focus on in the empirical application of this paper.

Going back to Proposition 2, the third bullet point shows that, for given \$ volatilities, the Sharpe reduction due to trading costs $\partial SR/\partial\lambda$ is bigger (more negative) when the signal is less persistent. This equation embodies an important effect, i.e. that "slow-moving" strategies are relatively more

scalable when stocks are more liquid.

The fourth bullet point gives a useful intuition about the levels of transaction costs. The first part of the equality shows that how much transaction costs are "burnt" as the optimal trading increases with ϕ/τ. When the signal reverts more quickly, effective transaction costs are bigger. When the trader trades more quickly, transaction costs are lower. This second effect comes from the fact that fast traders in the model are the ones that care the about transaction costs (compared to risk). They prefer low capacity, and low trading costs. Note in passing that the first equality is valid even outside of the "slow signal" approximation (see proof). Going back to the example of daily mean-reversion discussed above, assuming ϕ 1 $\gg \tau$ we obtain that $TC/ER \approx 1$: Even if they are traded optimally, the profits of high-frequency strategies are almost entirely wiped out by transaction costs.

The second part of the equality provides a simple rule that ties the Sharpe degradation due to transaction costs with the fraction of the PNL that is lost in trading. For instance, assume that a strategy has a Markowitz Sharpe of .7 but is traded with an effective Sharpe of .5, the formula suggests that –provided the DGP that is assumed is correct– effective trading costs should be around 20% of the realized PNL. If one starts from a pure Sharpe of 1, trading costs would be as high as 37% of the effective PNL. This formula can be used to estimate the minimal trading costs that a trader aiming for a particular risk-adjusted performance should expect to incur.

3 Data & Definitions

3.1 Data

Our analysis of raw returns (gross of transaction costs) is done using monthly returns from CRSP and annual accounting variables from COMPUSTAT, as it is done in most of the asset-pricing literature. Our period range is 1990-2013. When we move to the optimal trading analysis (the core of the paper), we use daily split- and dividend-adjusted returns from CRSP. We believe that it is important to allow investors to trade at the daily frequency in order to account for the potential fast decay of a signal's predictive power. At the end of the paper, we implement some simulations on monthly data for pure computational convenience.

We start with monthly data. From the CRSP universe of stocks, we extract two pools: "Large" and "Mid". Every month, we sort stocks by market capitalization computed at the end of the

previous month. "Large" is the set of the largest 500 stocks. "Mid" is the set of stocks ranking from 501 to 1500 in terms of this measure of size. Overall, the two pools change composition every month, but turnover is low. Every month, on average 1.2% of the stocks leave and enter the "large" pool. Monthly turnover is 2.4% in the "mid" pool, consistent with the intuition that stocks can move both up and down in and out of the mid-cap range.

Table 1, panel A, gathers the main descriptive statistics of our data. The average turnover is similar in the large and mid pool (24% among large stocks, versus 29% among mid caps). The total volume traded in mid caps is about $ 3.5tn annual versus $ 15tn among large caps. So, just looking at volume and without discussing price impact issues at this stage, it looks like it is possible to increase the capacity of a strategy by about 20% by moving into the mid cap range.

3.2 Calibrating the Price Impact Parameter λ

We now describe the parameter λ which measures the illiquidity of a given pool of stocks. Garleanu and Pedersen (2013) propose the following calibration, based on Engle et al. (2008) : trades amounting to 1.59% of the daily volume in a stock have a price impact of about 0.10%. Using this approximation, for each stock i, we can compute a liquidity parameter λ_i as the solution of:

$$1.59\% \times volume_i \times \frac{\lambda_i}{2} \times \sigma_i^2 = 0.1\%,$$

where σ_i is the daily volatility of stock i and $volume_i$ is its average dollar daily volume. This leaves us for each stock with the following formula:

$$\lambda_i = \frac{1}{8 \times volume_i \times \sigma_i^2}.$$

Using CRSP daily data, we compute for each year a stock-level λ_i. Separately for the "large" and the "mid" pools, we then define for each year the pool's λ as the median of the liquidity parameters (λ_i) of stocks belonging to the pool.

Time-series changes in each pool's λ are reported in Figure 3. We observe a strong increase in liquidity during 1990-2000 for both pools. Given the way we measure liquidity (proportional to the inverse of Volume $\times \sigma^2$), the sharp decrease in λ mostly comes from the increase in trading volume already documented for instance by Chordi et al. (2011) and Novy-Marx and Velikov (2014). Chordi et al. (2011) for instance document a fivefold increase in turnover between 1995 and 2009.

In our calibrations, we use the average values of λ in the last five years of our sample, i.e. values from 2009-2013, in order to examine the scalability of strategies in the context of contemporary liquidity conditions. This leads to $\lambda = 4\ 10^{-6}$ in the large pool and $\lambda = 2.4\ 10^{-5}$ in the mid pool. We report these numbers in Table 1.

3.3 Definition of Anomalies Traded

We focus on four well-known "fundamental" anomalies. Signals are updated monthly. Let us denote t the month of trading.

- First, we compute a standard " *Value* " signal, which is the ratio of book value of equity to market value of equity (Fama and French, 2006). Book value of equity (item CEQ in COMPUSTAT) is taken from the most recent annual accounts corresponding to the fiscal year ended in month $t-7$.[2] Market value is computed as the end of December of the last calendar year.

- Second, we compute a "*Low Vol*" signal, which is equal to *minus* the volatility of daily returns computed using returns from month $t-4$ to month $t-1$. This signal takes inspiration from papers documenting the fact that low volatility stocks tend to perform well in the long run. Frazzini and Pedersen (2013) use the stock's β as a measure of its riskiness. Ang et al. (2009) use a measure of idiosyncratic volatility closer in spirit to the one we use here. One interpretation for this anomaly is that these stocks provide "embedded leverage" to investors who are not allowed to borrow.

- Third, we compute a "*Net Repurchaser*" signal with is equal to minus the growth rate in (split adjusted) shares outstanding between $t-24$ and $t-1$, where $t-1$ denotes the last available calendar month (Pontiff and Woodgate, 2008). The economic intuition for this signal is that firms trade their own stock with superior information, so that their trading predicts returns.

- Finally, the fourth signal we look at is the "*Cash-Flows*" signal, which is equal to net operating cash flows (item OANCF in COMPUSTAT) normalized by total assets (item AT). These accounting items are taken from the last available annual accounts available 7 months before the current month. The economic intuition as to why this signal predicts returns is not

[2] Our assumption is therefore that the information available today from COMPUSTAT was available 6 full months after the end of the fiscal year.

fully understood: Novy-Marx (2013) hypothesizes that return on assets –which our cash-flow measure approximates– captures some measure of risk exposure of the firm; and therefore predicts returns. Asness et al. (2014) provide evidence consistent with the idea that investors persistently underestimate good quality stocks. Another possibility is that firm earnings attract too much attention compared to cash-flows, which are a better predictor of value (Sloan, 1996).

We then normalize each of these signals in the following manner. First, we compute for each stock the rank of the stock according to the signal using information available at the beginning of month t, in the pool that is considered (mid or large). We then normalize these ranks so that they lie -0.5 and +0.5.

3.4 Persistence Parameters ϕ

To estimate ϕ, for each signal $s_{i,t}$, we estimate the following regression on monthly data:

$$s_{i,t+1} = a + b.s_{i,t} + \epsilon_{i,t}$$

via plain OLS (we are not interested in the standard error). We estimate the above equation separately for each strategy, and for each pool of stocks. This allows us to retrieve the daily persistence parameter $\phi = 1 - b^{1/20}$. In Table 1, we report the values of ϕ for each pool and each strategy. Looking at Table 1, several noteworthy features emerge. First, all these strategies are very persistent. The plain book-to-market signal is the most persistent of the four. For mid caps, the half-life of this signal $(-\log 2/(250 * \log(1-\phi)))$ is about 4.6 years in the "mid" pool, versus 2 years for "low vol". "repurchasers" is the least persistent strategy, with a half-life of about 1.5 years. All these signals are therefore very slow-moving.

3.5 Scaling Parameter B

To estimate the signal scaling parameter B, separately for each strategy, we run the following regressions:

$$r_{i,t+1} = A + B.s_{i,t} + u_{i,t}$$

using monthly data on returns and signal described above. The coefficient B is estimated through

OLS –we are not interested in the standard errors in this paper, though we will provide evidence on the risk-adjusted performance of all strategies in Section 4. To run the estimation, we use the entire sample period 1990-2013. The simulation will therefore have some element of look ahead bias, but this is not critical here since, again, the focus of the study is the analysis of trading costs rather than the actual risk-adjusted performance of stock market anomalies.

4 Back-Testing

To validate our approach, back-testing our trading rule on real-life signals is a crucial step. A large gap between our theoretical predictions and effective trading performance could arise if our initial model is too largely mis-specified. Remember the model assumes that the signal is AR(1) and that returns only depend on the most recent value of the signal. But these assumption may not hold in the data. For example, consider a fundamental signal based on yearly accounting data. By definition, such signal persists at least one year and will therefore be categorized as a slow signal. However, it might for instance be possible that arbitrageurs act massively when accounting data become public, such that the signal's predictive power on returns quickly fades away after data publication: in such case, the signal, while persistent, would not be one that could be traded slowly and our model which assumes that predictive power is constant as long a a signal does not change would be highly flawed. Other structural assumptions of our model will also fail to hold exactly in the data: (1) the variance-covariance matrix Σ_u is probably not stationary, (2) the predictive power of the the signal may vary over time, and (3) the stationary DGP probably involves stock returns, rather than price changes. All of these assumptions (except the second one) are needed to find a closed form solution to the dynamic problem and are susceptible to create slippage between theoretical transaction costs and simulation results.

To alleviate concerns about model mis-specification, it is therefore needed to compute the realized performance of anomalies, using the trading rules derived by theory, on actual returns data. Throughout the analysis, a crucial assumption is that the price impact does not affect the dynamics of returns. Another important feature is that such an analysis is contingent on a given history of realized returns. We will explore the sensitivity of our results to this last assumption at the end of this Section.

4.1 Back-testing Procedure

This Section describes in detail how we run the back-testing procedure. We construct our trading portfolios on each pool of stocks separately.

At the beginning of each round, we first fix γ, the level of risk aversion. We use the dynamic trading rule summarized in equations (5) and (7). The time-unit for portfolio trading in this equation is the day. We use the estimates of λ for each pool, and ϕ for each strategy, from the data section.

We assume that in January 1991, the investor holds the null portfolio. Equation (5) then tells us how to update the portfolio holdings x_t, but it requires the computation of the "aimed" portfolio $\gamma/(\gamma + a\phi)x_t^M$, where x_t^M is the Markowitz portfolio. In order to compute it, we need to compute the inverse variance-covariance matrix. To do so in closed-form, we assume a one-factor structure for the error term in the returns generating process:

$$r_{i,t+1} = B.s_t + \beta_i r_{M,t+1} + \epsilon_{i,t+1}$$

$r_{M,t}$ is the market return and we assume that all shocks $\epsilon_{i,t}$ have the same variance σ_ϵ^2. We note σ_M^2 the variance of the common factor $r_{M,t}$ and β the vector of stock betas. Under this simple risk structure, it is easy to show (see Appendix F) that the Markowitz portfolio is given by:

$$x_t^M = \frac{B}{\gamma\sigma_\epsilon^2}\{s_t - \frac{\sigma_M^2}{\sigma_\epsilon^2}\frac{\beta' s_t}{1 + \frac{\sigma_M^2}{\sigma_\epsilon^2}(\beta'\beta)}\beta\} \tag{8}$$

which can be easily interpreted. First, if stocks are volatile or the investor risk averse, the portfolio is less levered (the term in $\gamma\sigma_\epsilon^2$). Second, other things equal, the Markowitz portfolio underweights high beta stocks, in particular if the market is more volatile. This is to reduce exposure to risk factor and hence volatility. Third, the correlation between the betas and signals plays a critical role in the size of the beta and net dollar exposure of the portfolio. If signal and beta or uncorrelated, the Markovitz portfolio is beta-neutral.[3] For instance, if beta and signal are negatively correlated (as is the case with "low vol"), then the Markowitz portfolio has a long bias.

The parameters required to compute the Markowitz portfolio x^M are estimated on a rolling

[3]It is easy to see that the beta of the Markowitz portfolio is given by:

$$B\frac{\beta' s_t}{\gamma\sigma_\epsilon^2}\Big\{1 - \frac{\sigma_M^2\beta' s_{t-1}}{\sigma_\epsilon^2 + \sigma_M^2\sigma_\epsilon^2(\beta'\beta)}\Big\}$$

. It is equal to zero if β is uncorrelated with s in the cross-section, since $Es = 0$ by construction.

basis –thus we allow the variance-covariance matrix to move slowly over time, in contrast with the assumptions of the GP model. Using monthly returns data, we estimate all the covariance parameters (σ_M , σ_ϵ, β) based on 24 months rolling windows: Every month t, for every stock i, we compute the volatility of returns over the past 24 months. σ_ϵ^2 is then estimated as the cross-sectional average of idiosyncratic variances over the pool. We also compute the univariate $\beta_{i,t}$ as the univariate beta of the stock's return with the market returns –net of the risk-free rate– over the past 24 months. Finally, we compute $\sigma_{M,t}^2$ as the volatility of monthly market returns (adjusted for the risk-free rate), over the past 24 months. Both signals and inverse variance matrix are thus estimated with monthly returns, so that the portfolio weights change only once a month. x_t^M is thus updated on the first day of each month.

Using equation (5), we thus compute portfolio holdings x_t every month, and the realized returns $x_t' r_{t+1}$. The resulting time series of portfolio returns are then used to compute the realized Sharpe ratio and $ volatility. We then iterate the entire process with a different level of risk-aversion. We span the Sharpe-to-volatility with values of γ going from 10^{-25} (low Sharpe, high volatiltiy) to 10^{-7} (high Sharpe, low volatility).

4.2 Performance under Zero Transaction Costs

As a benchmark, we use our back-testing process to analyze performance in the absence of transaction costs, which is a particular case of the back-testing described above. Making transaction costs go to zero leads the trader to invest instantaneously in the Markowitz portfolio. When the trader invests in the Markowitz portfolio, the expected Sharpe is given by $BE\sqrt{s_t'\Sigma_u^{-1}s_t}$ which does not depend on γ. The Sharpe-to-volatility frontier is a flat line, where the same Sharpe –the maximum possible one– is reached whatever the $ volatility aimed.

Note that backtesting the zero transaction cost strategy is an essentially monthly exercise in our setting. Because of the absence of transaction costs here, for given risk-aversion, the investor immediately reaches at the beginning of each month the Markowitz portfolio described in equation (8), and does not move until the end of the month, since no new information comes in (we update our signals on a monthly basis).

In Figure 1, we show for each strategy and in each size pool, the performance of the Markowitz portfolio (also referred to as the "pure" performance). As discussed in introduction, our 4 strategies tend to perform better on smaller capitalizations. But even on the "mid" pool, the realized Sharpe

ratios of book-to-market and low vol are below 0.5 - a level that we will later consider as critical.

We also report the annualized Sharpe ratios of these strategies in Table 1, assuming no trading friction at this stage. We directly report the Sharpe of the Markowitz portfolio. Looking at the Table, we observe three salient features. First, the raw Sharpe of book-to-market and low vol are extremely low, even slightly negative for large caps. The raw performances of repurchasers and cash-flows are higher. Secondly, for all strategies, the performance is higher in mid caps than in large caps. Thirdly, the hedging procedure improves risk-adjusted performance a lot, from .20 to .55 in the case of low vol. This is all the more striking because our method puts a lot of structure on the variance-covariance matrix of returns –in particular, it assumes only one factor, as well as homoskedastic returns. This confirms the intuition of Barroso and Santa-Clara (Forthcoming) that even a simple hedging procedure can significantly improve the risk-adjusted performance of some strategies.[4]

4.3 Calibrating the Sharpe-to-volatility frontier

At this stage, a very natural first step consists of "taking the model seriously", by using the closed form of the Sharpe-to-volatility frontier equation in Proposition (3). Provided that the DGP of the model is correct, using this formula directly would give us a good idea of the capacity of each strategy. This is obviously a strong assumption, and our next step will be to confront the trading model with the real returns data using the procedure described in Section 4.1.

In order to use this formula, we need to use the following three parameters for each strategy: Markowitz Sharpe SR^*, persistence ϕ and liquidity λ. We then assume that the investor aims for a net-of-cost Sharpe ratio of .3. Proposition 3 tells us that the \$ volatility is given by:

$$\text{Vol} = \frac{.3}{\lambda\phi^2}\Big(\Big(\frac{SR^*}{.3}\Big)^{2/3} - 1\Big)^2 \tag{9}$$

We compute the \$ volatility for each strategy, each pool of stocks, and for two separate decades: 1991-2000 and 2001-2010.[5] We do this because, as we saw in Section 3.2, liquidity increased markedly in the 2000s, in particular for midcaps. Of course, when the hedged Markowitz Sharpe is below .3,

[4] Barroso and Santa-Clara (Forthcoming) focus on momentum, and use a different hedging procedure from ours. They hedge for changes in volatility by essentially scaling the momentum factor by the inverse of rolling momentum volatility. While the philosophy is different from our paper –in the GP model a key assumption is that Σ_u is very slow moving– the main point remain that simple hedging procedures have a big impact on risk-adjusted performance.

[5] We exclude the last 3 years (2011-2013) of our sample in this Section for the sake of symmetry. We will add them back in the next Section.

we set the capacity to 0, since it is not feasible to reach the minimum Sharpe of .3.

We report the results of this simple calibration in Table 2. Several salient facts emerge. First, the increase in liquidity appears to have a strongly positive impact on the predicted capacity of strategies in the mid-cap range. For both pools, the λ has been divided by 10, which, in formula (9) automatically leads to a tenfold increase in capacity. For instance, the capacity of "cash-flows" in the mid cap range has been multiplied by 10 between the 1990s and the 2000s, an effect entirely attributable to the increase in liquidity. Thus, even though the "pure" performance of cash-flows in the mid-cap range has, if anything, decreased, the effective size of the anomaly has been multiplied by a factor of 10, simply because mid-sized stocks have become more liquid. Given that λ is defined at the pool level, this effect is of course visible for all strategies.

The second feature of Table 2 is that the "pure" Sharpe ratio is a critical determinant of capacity. Many strategies have zero capacity in the large cap pool because their Sharpe ratio, even in the absence of adjustment costs, does not even reach 0.3. For instance, compare "low vol" and "repurchasers" in the large cap range. In the 2000s, both strategies have an annualized ϕ of about 0.3 and face the same liquidity parameter, but repurchasers have more than six times as much capacity as low vol because its "pure" Sharpe is .45 compared to 0.34. Capacity is therefore quite sensitive to the "pure" performance.

Finally, ϕ does not play much of a role in the restricted set of very slow strategies that we explore. But this does not mean that persistence plays no role in general. As we discussed earlier, it is the case that, *within the class of strategies that we explore*, differences in persistence do not matter very much, but strategies whose signal moves a bit faster can have their capacity much reduced. To fix ideas about the effect of ϕ, let us discuss the standard momentum signal, which we define here as the cumulative return of the stock between month $t-12$ and $t-2$. This signal has a daily $\phi = .008$ (obtained through OLS over 1990-2013), which is thus 8 times larger than the fundamental strategies that we focus on here. Assuming momentum had the same Sharpe ratio as, say quality, in the mid-cap range (it does not), this would lead to a predicted capacity 64 times smaller for momentum than for cash flows. So ϕ does play a big role in determining capacity, but as we have already noticed, not within the restricted set of strategies that we focus on here in this paper.

4.4 Back-testing the Sharpe-to-volatility frontier

In this Section, we analyze fully back-tested results based on past returns, in order to obtain an empirical view of the capacity of each strategy. This approach is the most conservative as it adjusts for the fact that the DGP used in the model may differ from the true process underlying returns and signal data. For instance, our very simplified representation of the covariance structure might miss various sources of correlation. Therefore, portfolio trading that is optimal in the model might not be optimal in reality. Moreover, contrary to the model, real covariance evolves over time, which we take into account by updating the covariance matrix. However, this means that trading costs might be higher than expected in the model. If we can show that, despite the trading rule being based on a quite simplified covariance structure, back-testing yields significantly high Sharpes even at high capacities, we could be quite confident about our message that slow signals have capacity in the mid-cap range.

We report in Figure 6 our simulation results. The thick line represents the Sharpe ratio (net of trading costs) that is realized at various levels of volatility. To span these different levels of volatility, we vary γ. For a very high γ, the trader trades arbitrarily small amounts, leading to vanishing transaction costs, thus in these graphs for a near-zero dollar volatility the Sharpe is equal to the Markovitz Sharpe.

We note that when signals have a relatively large pure Sharpe ratio, such as "cash-flow" or "low vol" in the mid pool, the back-tested Sharpe decays as expected in theory. The most striking feature is the large capacity of the cash-flows strategy in the mic-cap range: At an annual volatility of 15 Bn. dollar, the realized Sharpe of the Cash-flows strategy in the mid remains above 1. So, even in the mid-cap range, a strategy based on persistent stock characteristics with a high Markowitz Sharpe such as quality has high capacity, well above 10 Bn dollar of annual volatility. Interestingly, back-testing the trading rule on actual data gives even more aggressive estimates of capacity for "cash-flows" in the mid-cap range than the calibration in Table 2.

Figure 6 also provides information on what the Sharpe-to-volatility frontier could be, under the null hypothesis that signals had no predictive power on returns. We obtain these intervals by running Monte-Carlo simulations where we assume that the DGP such that (1) the signal has the same persistence as in the data (and thus differs across strategies and pools), while (2) the signal has no predictive power, i.e. $r_{i,t+1} = R + u_{i,t+1}$ where $u_{i,t+1}$ is i.i.d and independent of s_{it}. In

each of these simulations, all data, including stock-returns are drawn from the calibrated generating process. For a given draw of the data, we vary γ such as to span all levels of volatility and obtain the Sharpe-to-volatility frontier for this draw.

We then repeat this procedure 100 times with 100 new simulated data. The confidence interval for a given volatility corresponds to twice the standard deviation each side around the mean of the distribution of Sharpe ratios. This mean is a priori non-zero because some signals have correlation with beta, which leads to portfolios that have a non-zero beta and thus some exposure to the market risk-premium. The set of possible Sharpe-to-volatility frontiers also vary across signals because they have different persistence parameters ϕ. Interestingly, we see that a realized Sharpe of say .4 is typically not outside these simulated 95% confidence intervals, meaning that realizing a track-record with such Sharpe ratios does not automatically mean that the underlying signal has real predictive power. In the large cap range, "cash-flows" is the only strategy for which we can reject the hypothesis that the Sharpe-to-volatility frontier comes from a non-predictive signal. In the mid-cap pool, the same is true for both "cash-flows" and "low-vol". The "repurchasers" strategy does not, however, manage to emerge from the noise present in the data.

4.5 Anatomy of slow trading: example of Cash-Flows

In this last section of our back-testing analysis, we look in detail at how increasing portfolio volatility impacts optimal trading. We build on our back-testing data and extract from them some easily interpretable metrics.

In Figure 7, we use back-testing results from the Cash-Flows strategy in the mid-pool, following the back-testing technique described in Section 4.1. In all graphs from panels A,B,C, the horizontal axis is the portfolio volatility targeted by the trader. Panel A shows together the gross (of transaction costs) Sharpe and the realized Sharpe (i.e. net of transaction costs). The striking fact is that the Sharpe of the gross PNL (i.e. "gross of transaction costs") does not decrease much: this means that the "slowing down" in trading and the deformation it induces vis-a-vis the target portfolio has only a small impact on the Sharpe. Most of the Sharpe deterioration actually comes from trading costs. Panel B shows average monthly dollar turnover divided by the gross market value of portfolio (i.e. the dollar value of the long positions plus the dollar value of the short positions). We see that for small portfolios, about 12% of all positions are traded monthly: this is the rebalancing dictated by the frictionless Markowitz portfolio. Due to transaction costs, the trader slows down trading and

we see that turnover is about only half of its original level when annual volatility is above $ 10bn. Panel C shows average transaction costs per $ of PNL. We see that this rate goes up in a concave manner with the portfolio's scale, reaching about one third at $10bn, when the realized Sharpe is halved, in line with predictions of theory.

In Panel D, we show how effective holdings respond to a one-off unit shock to "aimed" holdings. The methodology takes inspiration from impulse responses obtained in VAR analyses. We first run our back testing procedure fr two separate portfolios: a "small portfolio" corresponding to a target volatility of $ 10m, and a "large portfolio" corresponding to a target volatility of $ 15bn. For each of these portfolios, we obtain a time series of vectors of stock holdings $x_i(t)$, as well as a time-series of aimed portfolios $aim_{i,t}$. We then run the following OLS regression in the back-testing portfolio panel corresponding to each targeted volatility:

$$x_{i,t} = \sum_{k=0}^{25} c_k aim_{i,t-k} + \epsilon_{i,t}.$$

and retrieve the coefficients c_k. Then, we define the cumulative response at $t+k$ in the holdings of a generic portfolio stock i to a unit shock to $aim_{i,t}$ as $C_{t+k} = \sum_{k'=0}^{s} c_{k'}$. We report these cumulative responses C_{t+k}, as a function of k in panel D. k is in month since our data here are monthly.

We observe that for small portfolios, a quarter after the shock, the holdings have reached roughly 80% of the aimed level whereas in the large portfolio, after one year only 40% of the aim has been reached. This means that after a signal is "on" for a stock, for large portfolios, the trader keeps buying at a constant rate more than two years later –notice that, as long as signals do not change, the aimed portfolio remains fixed. This "ramp-up" period is much shorter for the small portfolio. This is the optimal behavior only because the signal is highly persistent. These patterns come from back-tested data: They reflect the real dynamics of a portfolio trading the "Cash-Flows" signal according to the trading rule derived from the model, taking into account the effective realizations of signals and returns.

Panel E shows the times series of the gross exposure in $ bn (i.e. long market value plus short market value) of the large portfolio (the small portfolio has very similar features, on a much smaller scale). This illustrates how Markowitz optimization (here with $\gamma = .5E-10$) picks a portfolio scale that varies negatively with the volatility predicted by the variance-covariance matrix: in periods of turmoil, such as the financial crisis, the portfolio is endogenously descaled. This is because

our variance-covariance matrix is updated monthly in back-testing; In this application, we assume persistence ϕ, liquidity λ and forecasting power B to be constant since we estimate them on the entire sample. If assets under management were constant, this would correspond to an endogenous choice of volatility / leverage. The order of magnitude of the gross market value at the end of the time series, $ 500bn could for instance be generated by a fund with $ 100bn of AUM, with a long and a short of equal size ($ 250bn) and a $\times 2.5$ leverage.

Last, Panel F shows cumulative effective returns on gross exposure (i.e. PNL divided by gross exposure) for the large and small portfolios. It shows in a metric that is quite familiar, that the impact of trading costs is large. Returns on gross exposure are scale-free, thus the difference between the two curves can be attributed fully to transaction costs incurred when trading at large scale. The cumulative returns of the small portfolio are very close to those that would be found if we were ignoring transaction costs altogether and simply analyzing the "pure" Markowitz portfolio. Interestingly, we see that the large portfolio loses money in the very first years of the time-series: this corresponds to a period of fast scaling-up of the strategy toward the aimed portfolio, where transaction costs are higher than in steady state (we initialize holdings at zero). This last observation suggests a crucial role for the discount rate used in the model in the transition period. While the investor reaches his desired size, he can choose to trade slowly to preserve performance on the transition path, or to reach the transition path more quickly. Investors with a higher discount rates are more impatient: They value future PNLs less, and trade more slowly. In our application, we have set $\delta = 2\%$ on an annualized basis, which is low and may explain why the trader in our backtesting is so "eager" to reach the aim that he incurs losses on the transition path.

5 Robustness

To explore the practical relevance of our framework, we explore how robust it is to various types of noise sources. First, we explore the effect of "parameter noise": We assume that the model holds but that the trader only has noisy measures of the underlying parameters (λ, ϕ) and we ask how this impacts the performance of his trading. Second, we consider the impact of "model noise": We assume that the true data generating process is different from that assumed in the model, and ask how a trader using the model to trade would perform. Third, we examine how needed the dynamic optimization approach chosen in the paper is, by comparing its outcomes to those of a myopic trader

that would just use static optimization.

5.1 Sensitivity to Noise in Parameter Estimates

In this subsection, we check whether a mistake in estimating parameters strongly affects the performance of the strategy. We focus on two parameters: signal persistence ϕ and liquidity λ. We run the following thought experiment: The trader believes that the true value of these parameters differs from their actual values. The result of these beliefs is that the trader does not trade optimally. For instance, if the trader overestimates the persistence of the signal, he will "aim too high", i.e. he will trade too aggressively on the signal.

We can easily derive expressions for the Sharpe-to-volatility frontier, assuming that the trader makes such mistakes. To make things simple and comparable to our previous results, we make the "large investment approximation". We summarize our results in the following Proposition:

Proposition 4. *Assume the "large investment" and the "slow signal" approximations hold. Let* $\tau = (\gamma/\lambda)^{1/2}$. *Then:*

- *If the trader wrongly believes that the price impact coefficient is equal to* λ, *but the true liquidity is* λ^*, *then the Sharpe-to-volatility frontier is given by:*

$$SR = \Big(1 - \frac{\lambda^*}{\lambda}\frac{\phi}{\phi+\tau}\Big)\Big(\frac{\phi}{\phi+\tau}\Big)^{1/2} SR^*$$
$$Vol = \frac{1}{\lambda\tau}\Big(\frac{\phi}{\phi+\tau}\Big)^{1/2} SR^*$$

- *If the trader wrongly believes that the persistence coefficient is equal to* ϕ, *but the true persistence is* ϕ^*, *then the Sharpe-to-volatility frontier is given by:*

$$SR = \Big(1 - \frac{\phi}{\phi^*+\tau}\Big)\Big(\frac{\tau}{\phi^*+\tau}\Big)^{1/2} SR^*$$
$$Vol = \frac{1}{\lambda\tau}\Big(\frac{1}{\phi^*+\tau}\Big)\Big(\frac{\tau}{\phi^*+\tau}\Big)^{1/2} SR^*$$

Proof. See Appendix D □

We use the two sets of formulae above to investigate quantitatively the effect of an error on the models parameters. This approach assumes that the DGP used by the Garleanu-Pedersen model is

correct, and that the trader only makes a mistake about the parameter values of either ϕ or λ.

We show the results of our investigations in Figures 8 and 9. Figure 8 explores the effect of a mistake about liquidity. In this first Figure, we assume that $\phi = .001$ and the actual $\lambda = 2.10^{-5}$ (the price impact of midcaps). We then plot 4 different curves where the trader wrongly believes that λ is 10^{-6}, 10^{-5}, 2.10^{-5} and 5.10^{-5}. The main lesson of this Figure is that small mistakes about λ do not affect the capacity of the strategy very much: Taking 10^{-5} or 5.10^{-5} instead of 2.10^{-5} does not change the capacity very much, compared to taking the "true" value of λ. If however, the trader believes that the price impact is 10^{-6} (instead of 2.10^{-5}), i.e. if the trader thinks mid-caps are as liquid as large caps, then capacity is a lot smaller. Overall, believing that midcaps are as liquid as large caps will have a huge impact on performance, but smaller (and more reasonable) mistakes will be more forgiving.

Figure 9 explores the effect of a mistake about the persistence parameter. We also assume that $\phi = .001$ and $\lambda = 2.10^{-5}$, but now we assume that, in each of the 4 curves, the trader makes a counterfactual assumption on ϕ: .001, .0011, .0015, .002 and .005. Here, the mistake has very little impact on the actual capacity of the strategy, as long as we remain in the range of slow-moving strategies. Overall, our analysis suggests that the Sharpe-to-volatility frontier is not too sensitive to reasonable "mistakes" in the persistence or liquidity coefficient. As mentioned earlier, this relative insensitivity to estimation error comes from the fact that we focus on a relatively narrow range of slow-moving strategies. A similar mistake on faster strategies would obviously have a more dramatic impact.

5.2 Model robustness to alternative signal process

In this section we explore the model's robustness (beyond the issue on sensitivity to ϕ and λ that has already been studied in Section 5.1). We want to evaluate the extent to which our assumption that the signal follows an AR(1) is important or not.

We thus run the following thought experiment: Imagine that the signal were to follow another process and let the trader evaluate an AR(1) in the data and trade according to the model. We then ask whether the realized Sharpe massively differs from that predicted by the GP trading model. We answer this question by performing Monte-Carlo simulations. We simulate data where the signal allows to forecast returns as in equation (3), but the signal comes from a non-AR(1) generating process. We then assume that the trader fits an AR(1) model on the signal data, and trades

according to the GP model. We compute realized Sharpe and $ volatility for this process, and compare them with the Sharpe and $ volatility that would emerge if the model was correct (i.e. if the signal really was an AR(1)). These values are directly taken from the closed-form equation (9).

More precisely, we consider two families of alternative signal-generating processes. First, we assume that:

$$s_{t+1} = \sum_{i=1}^{N} \frac{(1-\phi)}{N} s_{t-i} + \epsilon_{t+1}$$

.

Our model assumes $N = 1$. We simulate such process for $N = 2, 4, 6$. We do so on a panel of 1000 stocks with same liquidity and time length as our mid-cap pool. We generate return data using the signal above and assuming that the predictive power of the signal is that of the cash-flows signal in real data and idiosyncratic volatility matches real data. For each dataset of signals and returns that is generated, we compute the parameter $\hat{\phi}$ that results from estimating in those synthetic data an AR(1) : $s_{t+1} = (1-\phi)s_{t-i} + \epsilon_{t+1}$. Then, we trade according to the model's optimal rule and compute performance net of transaction costs for various target volatilities. We perform this for a large number of synthetic (signal-return) draws. Finally, we report in Table 3 the average Sharpe ratio decay (i.e. SR/SR^*) that is reached at various volatility scales, and compare them to the prediction of the model given the persistence parameter $\hat{\phi}$ estimated from the data.

We find that the realized Sharpe ratios are quite similar to prediction, even for high $ volatility. In fact, the realized Sharpe is even higher than anticipated. The reason is that at low frequency, the signal is actually more persistent than what the estimation of a monthly AR(1) suggests, therefore trading costs are smaller than expected. This result does not depend on all coefficients in the AR(n) being equal.

We consider a second family of alternative processes, closer to a moving average than an auto-regressive process. We start from a random walk variable $x_{t+1} = x_t + \eta_{t+1}$. Then, we construct s_t as the normalized rank at time t of $x(t)/x(t-N)$. "repurchasers" or "momentum" which is defined as cumulative returns over a rolling window) are signals generated by such process –and are thus likely to differ significantly from the AR(1) we use in our model. To generate the data, we use parameters from the "repurchasers" strategy, because this signal is constructed precisely in this manner using outstanding shares as x_t: We thus use for σ_η the estimate that we get from the volatility in stock-

level change of shares outstanding from the mid-cap pool data. We also assume the same return predictability (B coefficient) as that of the "repurchasers" signal. Then, as previously, we generate a large number of synthetic datasets based on independent (signal,return) draws. We show results in Table 3 for $N = 24$ and $N = 12$. Here, we observe that the realized performance is very poor compared to realizations. For this kind of process, the AR(1) representation is very misleading for the trader, which may explain why, in spite of a reasonable pure performance, the "repurchasers" signal does not do so well in our backtesting exercise. This suggest that the GP model that we using is going to be ill-suited to trade a momentum signal.

All in all, this robustness exercise suggests that our modeling approach remains valid for signals that are AR(n) with positive coefficients.For such signals, the AR(1) estimation does not lead to trading realizations that are far from expectations. Sorting on a persistent fundamental characteristic of a firm is therefore well represented by our model. By contrast, our approach seems less robust for signals that are constructed by sorting on the growth rate of a persistent variable, such as "repurchasers".

5.3 Myopic Trading

This final section shows that the dynamic setting of this paper cannot be adequately approximated by a myopic optimization. The dynamic apparatus of GP is critical for the slow-moving signals that we focus on here; As we will see, a static approximation is very suboptimal.

To show this, we consider a "myopic" alternative to the model, where the investor does not take into account the future trading costs that his current trading decision would entail. This is, in some sense, the framework used in most existing studies of capacity (Frazzini et al. (2012), Korajczyk and Sadka (2004)): Traders optimize the expected risk-adjusted performance, taking into account the instantaneous impact of their trades. They do not, however, taking into account the effect of current trading on future profits (if the signal persists) or future trading costs (if the signal mean-reverts). To remain consistent with the paper's framework, we model the optimization problem as:

$$\max_{x_t} E_t\Big[-\frac{\lambda}{2}(\Delta x_t')\Sigma_u(\Delta x_t) + \frac{1}{1+\delta}\Big(x_t' r_{t+1} - \frac{\gamma}{2} x_t'\Sigma_u x_t\Big)\Big] \tag{10}$$

for each period t and taking x_{t-1} as given. This linear-quadratic static optimization is easy to solve. The FOC yields:

$$x_t = (1-\pi)x_{t-1} + \pi x_t^M \tag{11}$$

where $x_t^M = (\gamma\Sigma_u)^{-1}s_t$ is the Markowitz portfolio and $\pi = \frac{\gamma}{\gamma+\lambda(1+\delta)}$ is the trading speed of the myopic agent.

To fix ideas, we make a large investment approximation similar (less restrictive) to the one we explored in Section 2.2: $(\gamma/\lambda) \ll 1$ and $(\gamma/\lambda) \gg \delta$. Under this approximation, the trading speed becomes $\pi \approx \frac{\gamma}{\lambda} \ll 1$. Conditionally on this approximation, it is easy to see that the difference between the myopic and the dynamic model is twofold. On the one hand, the myopic investor trades too slowly. The dynamically optimal trading rate is $\tau = (\gamma/\lambda)^{1/2}$ while the myopic one is $\pi = \gamma/\lambda$. Second, the myopic investor aims "too high", since he goes towards the Markowitz portfolio, which is more levered than the "aimed" portfolio. The difference between the two targets is larger when mean-reversion ϕ is bigger, and when risk-aversion γ is bigger: In both cases, the myopic trader makes bigger mistakes because he does not take into account the future cost of exiting the position.

In order to see more clearly the effect on capacity and implement some simple quantification, we explictly formulate the Sharpe-to-volatility frontier of the myopic trader in the following Proposition:

Proposition 5. *Under the "large investment" approximation, the Sharpe-to-volatility frontier realized by a myopic trader writes as:*

$$Vol = \frac{SR}{\lambda\phi}\Big[\Big(\frac{SR^*}{SR}\Big)^2 - 1\Big]$$

Proof. See Appendix E. □

Let us consider trader aiming for a Sharpe of SR and trading a signal of mean-reversion ϕ. Let Vol_{Dyn} be the \$ volatility reached if he trades dynamically, and Vol_{Myo} the \$ volatility if he trades myopically. In the large investment approximation, the ratio of the two volatilities is thus given by:

$$\frac{\text{Vol}_{Dyn}}{\text{Vol}_{Myo}} = \frac{1}{\phi}.\frac{[(SR^*/SR)^{2/3} - 1]^2}{(SR^*/SR)^2 - 1}$$

The myopic capacity frontier is much smaller than the dynamically traded frontier. This comes from the fact that, for our strategies $\phi \approx .001$ (see data Section). When $\frac{SR^*}{SR} \approx 1.1$ (the trader aims for a small reduction in performance), then $\frac{\text{Vol}_{Dyn}}{\text{Vol}_{Myo}} \approx 20$. When the trader is more ambitious,

say, $\frac{SR^*}{SR} \approx 2$, then $\frac{\text{Vol}_{Dyn}}{\text{Vol}_{Myo}} \approx 100$. So in the range of parameters that we explore, not optimizing dynamically reduces capacity by a factor of 20 to 100. Of course, for shorter-lived signals the difference is much smaller.

6 Conclusion

Anomalies are "abnormal" only if substantial dollar amounts can be profitably put at work to arbitrage them. This paper explores the deterioration of the Sharpe ratio of a trading signal when the dollar scale of the traded portfolio increases. We show that, using dynamic optimization under quadratic transaction costs, yields closed form formula for the Sharpe-to-scale frontier. When signals are persistent enough, arbitrageurs can put large amounts of money at work by trading more slowly. Back-testing optimal trading rules shows that even in the mid-cap range, strategies based on persistent stock characteristics such as quality have high capacity, well above 10 Bn dollar of annual volatility.

References

Ang, Andrew, Robert Hodrick, Yuhang Xing, and Xiaoyan Zhang, "High Idiosyncratic Volatility and Low Returns: International and Further U.S. Evidence," *Journal of Financial Economics*, 2009, *91* (1), 1–23.

Asness, Cliff, Andrea Frazzini, and Lasse Pedersen, "Quality Minus Junk," 2014.

Barroso, Pedro and Pedro Santa-Clara, "Momentum has its Moments," *Journal of Financial Economics*, Forthcoming.

Brokmann, Xavier, Emmanuel Sérié, Julien Kockelkoren, and Jean-Philippe Bouchaud, "Slow Decay of Impact in Equity Markets," *CFM working paper*, 2014.

Chordi, Tarun, Richard Roll, and Avanidhar Subrahmanyam, "Recent trends in trading activity and market quality," *Journal of Financial Economics*, 2011, *101*, 243–315.

Collin-Dufresne, Pierre, Kenneth Daniel, Ciamac Moallemi, and Mehmet Saglam., "Strategic asset allocation with predictable returns and transaction costs," 2012.

Dreschler, Itamar and Qingyi Dreschler, "The Shorting Premium and Asset Pricing Anomalies," 2014.

Engle, Robert, Robert Ferstenberg, and Jeffrey Russel, "Measuring and Modelling Execution Risk," 2008.

Fama, Eugene and Kenneth French, "The Value Premium and the CAPM," *Journal of Finance*, 2006.

Frazzini, Andrea and Lasse Pedersen, "Betting Against Beta," *Journal of Financial Economics*, 2013, *111*, 1–25.

_ , Ronen Israel, and Tobias Moskowitz, "Trading Costs of Asset Pricing Anomalies," 2012.

Garleanu, Nicolae and Lasse Pedersen, "Dynamic Trading with Predictable Returns and Transaction Costs," *Journal of Finance*, 2013, *68* (6), 2309–2340.

Harvey, Campbell, Yan Liu, and Heqing Zhu, "... and the Cross-Section of Expected Returns," 2014.

Hou, Kewei, "Industry Information Diffusion and the Lead-Lag Effect in Stock Returns," *Review of Financial Studies*, 2007, *20* (4), 1113–1138.

Korajczyk, Ron and Roni Sadka, "Are momentum profits robust to trading costs?," *Journal of Finance*, 2004, *59*, 1039–1082.

Lean, David Mc and Jeffrey Pontiff, "Does academic research destroy stock return predictability," 2014.

Lo, Andrew and Amir Khandani, "What Happened to the Quants in August 2007?: Evidence from Factors and Transactions Data," 2008.

Novy-Marx, Robert, "The Other Side of Value: The Gross Profitability Premium,," *Journal of Financial Economics*, 2013, *108* (1), 1–28.

_ **and Mihail Velikov**, "A Taxonomy of Anomalies and their Trading Costs," 2014.

Pontiff, Jeffrey and Artemiza Woodgate, "Share Issuance and Cross-sectional Returns," *Journal of Finance*, 2008, *63* (2), 921–945.

Sloan, Richard, "Do stock prices fully reflect information in accruals and cash flows about future earnings?," *The Accounting Review*, 1996, *71*, 289–315.

Stambaugh, Robert, Jianfeng Yu, and Yu Yuan, "The Short of It: Investor Sentiment and Anomalies," *Journal of Financial Economics*, 2012.

Table 1: Large vs Mid Caps

	Large	Mid
Panel A: Sample Statistics		
Total Capitalization (bn)	8389	1258
Total Volume (bn)	14771	3411
Average Turnover (%)	24	29
λ ($\times 10^{-6}$)	4	24
Panel B: Strategy-level Statistics		
Book-to-market		
Raw (unhedged) Sharpe	-.05	.06
Markowitz Sharpe	-.04	.19
ϕ (persistence $\times 10^{-3}$)	.6	.8
Low vol		
Raw (unhedged) Sharpe	-.05	.06
Markowitz Sharpe	.5	.71
ϕ (persistence $\times 10^{-3}$)	1.3	1.3
Repurchasers		
Raw (unhedged) Sharpe	.19	.37
Markowitz Sharpe	.41	.45
ϕ (persistence $\times 10^{-3}$)	1.8	1.8
Cash-Flows		
Raw (unhedged) Sharpe	.5	.88
Markowitz Sharpe	.56	1.1
ϕ (persistence $\times 10^{-3}$)	.9	1.3

Note: This table reports summary statistics on the pools of "Large" and "Mid" caps. Every month, stocks are sorted by stock market capitalization. The largest 500 ones belong to the "Large" pool. Stocks ranking between 501 and 1500 belong to the "mid" pool. Panel A reports summary statistics for each pool: Aggregate volume in $bn; aggregate market capitalization; average monthly turnover (annualized) and the illiquidity parameter λ. Panel B reports statistics for each of the 4 strategies we cover in this paper. λ has no unit. Volume, turnover, persistence and Sharpe ratios are annualized.

Table 2: Capacity: Results from the Calibrated Model

	Mid				Large			
	Frictionless Sharpe	ϕ ($\times 10^{-3}$)	λ ($\times 10^{-6}$)	Volatility ($ bn)	Frictionless Sharpe	ϕ ($\times 10^{-3}$)	λ ($\times 10^{-6}$)	Capacity ($ bn)
Book-to-market								
1990-2001	.33	.9	201	.0086	.092	.6	40	0
2002-2013	.01	.8	28	0	-.36	.5	5	0
Low vol								
1990-2001	.47	1	201	.15	.25	1.3	40	0
2002-2013	.96	1.6	28	5.1	.62	1.3	5	13
Repurchasers								
1990-2001	.56	1.8	201	.12	.34	2	40	.015
2002-2013	.7	1.8	28	1.9	.44	1.6	5	1.8
Cash-Flows								
1990-2001	1.3	1.4	201	2	.82	1	40	6.6
2002-2013	1	1.3	28	9.7	.82	.8	5	79

Note: This Table reports the theoretical capacity of each strategy by sub-period and by pool. We use the formula derived in the main text:

$$Vol = \frac{SR}{\lambda \phi^2} \left[\left(\frac{SR^*}{SR} \right)^{2/3} - 1 \right]^2$$

and define capacity as the annualized dollar volatility at which the realized Sharpe (SR) becomes equal to For each sub-period, and each strategy, we compute the persistence parameter ϕ by regressing the signal on its lagged value, using monthly data. We compute the price impact coefficient λ by taking the median of (2/15.9) $\times$ daily shares traded/daily vol^2 over all months-stocks in the pool and the period. λ is thus period and pool specific, but it the same across strategies. Finally, we estimate the Sharpe ratio of the strategy using the simplified Markowitz procedure described in Section 4.1: This procedure assumes a simplified correlation structure of returns, and no transaction costs. We use the estimated SR^*, ϕ and λ to compute the $ volatility reached with a Sharpe $SR = .3$. The volatility is set to 0 when the pure Sharpe is already lower than .3. Columns 1-4 are for the pool of "mid caps" while columns 5-8 are for the pool of large caps. In columns 1 and 5, we report the value of the frictionless Sharpe; In columns 2 and 6, the annualized value of ϕ (the formula uses the daily version); In columns 3 and 7 we report the λ and in columns 5 and 8 the $ volatility reached. Sharpes, volatility and ϕ are all annualized.

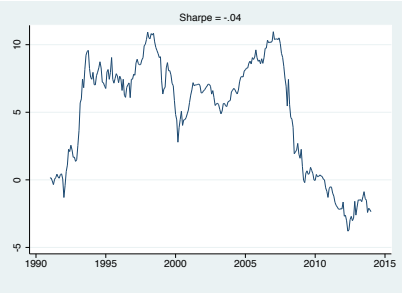
Table 3: Expected vs. Simulated Performance-Capacity Frontier for non-AR(1) Signals

		SR/SR^*	
Real signal	$Volatility	Expected	Realized
AR(6)	vol=5	0.75	0.76
	vol=10	0.68	0.75
	vol=20	0.54	0.7
AR(4)	vol=5	0.78	0.81
	vol=10	0.69	0.72
	vol=20	0.58	0.65
AR(2)	vol=5	0.81	0.77
	vol=10	0.74	0.75
	vol=20	0.63	0.64
$x(t)/x(t-24)$	vol=5	0.68	0.1
	vol=10	0.69	-0.11
	vol=20	0.35	-0.15
$x(t)/x(t-12)$	vol=5	0.64	0.29
	vol=10	0.69	-0.32
	vol=20	0.35	-0.34

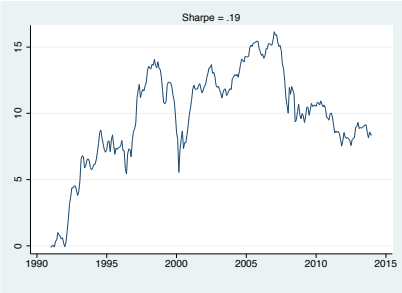
Note: This Table reports the Sharpe decay, $\frac{SR}{SR^*}$ as a function of dollar volatility (in annualized Bn dollar) for several synthetic signals that do not follow an AR(1) process. We perform monte-carlo simulation, where an AR(1) estimation is performed on the signal, and trading occurs according to those estimates. For each alternative process, we draw a large number of synthetic (signal,returns) datasets, on which we perform simulations. The number of stocks, liquidity, time length and volatility parameters are based on our mid-cap pool. For each draw of (signal,return), we estimate an AR(1) and simulate the performance of the trading rule recommended by the model. We vary γ to span a large spectrum of volatilities. The first three alternative signals correspond to an AR(n) process of the form: $s_{t+1} = \sum_{i=1}^{N} \frac{(1-\phi)}{N} s_{t-i} + \epsilon_{t+1}$. The ϕ parameter and signal noise are taken from the Cash-Flows signal. The last two signal processes are constructed as follows: we generate a variable $x_{t+1} = x_t + \eta_{t+1}$. Then, we construct s_t as the normalized rank at time t of $x(t)/x(t-N)$. We show results for $N = 24$, $N = 12$. Data are generated using parameters (notably return predictability and σ_η) from the "Repurchasers" strategy.

Figure 1: Performance of Four Strategies: Large vs Midcaps

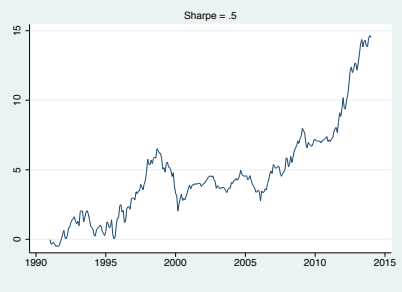
Panel A: Book-to-Market Big



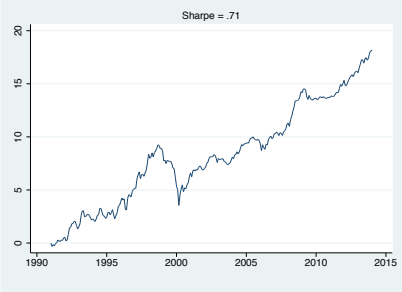
Book-to-Market Mid



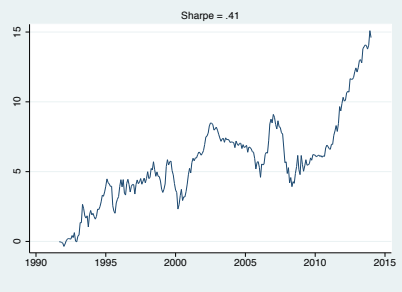
Panel B: Low Vol Big



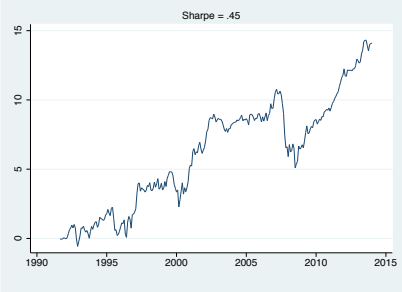
Low Vol Mid



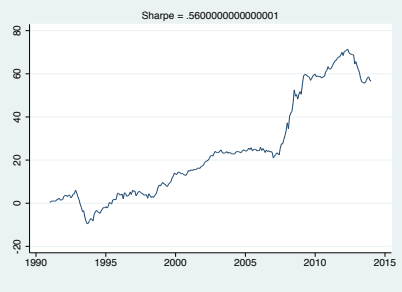
Panel C: Net repurchasers Big



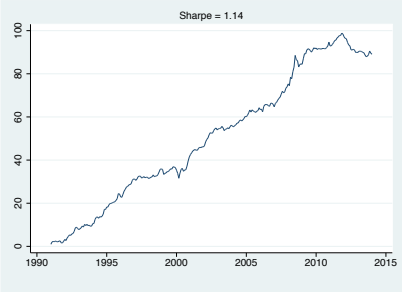
Net repurchasers Mid



Panel D: Cash-Flows Big

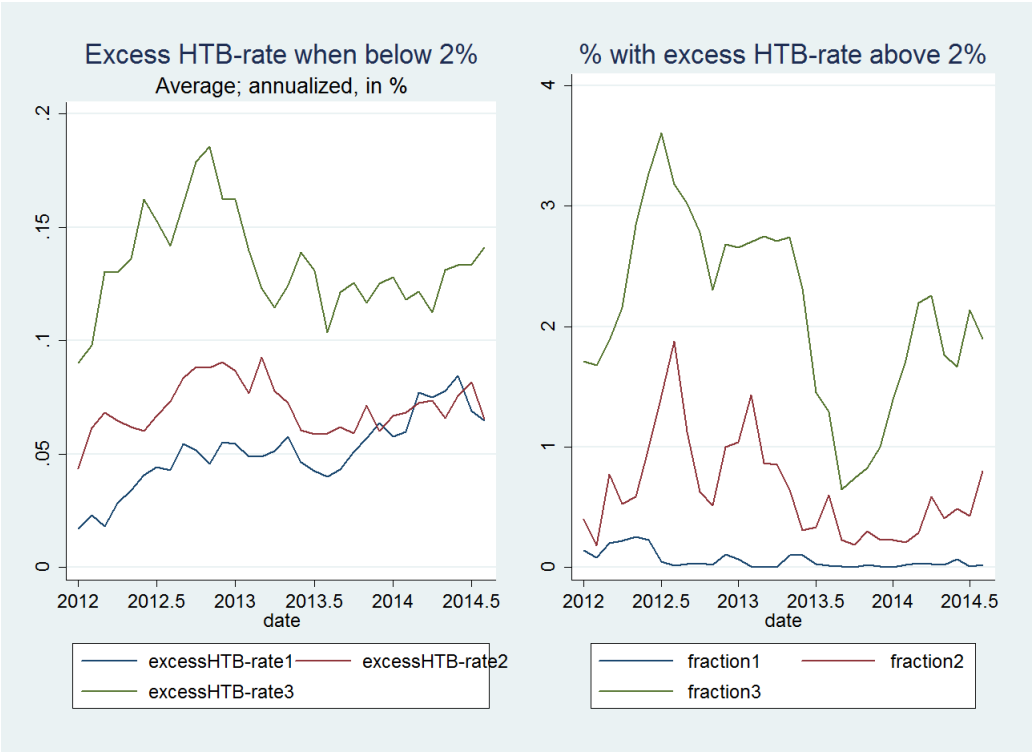


Cash-Flows Mid



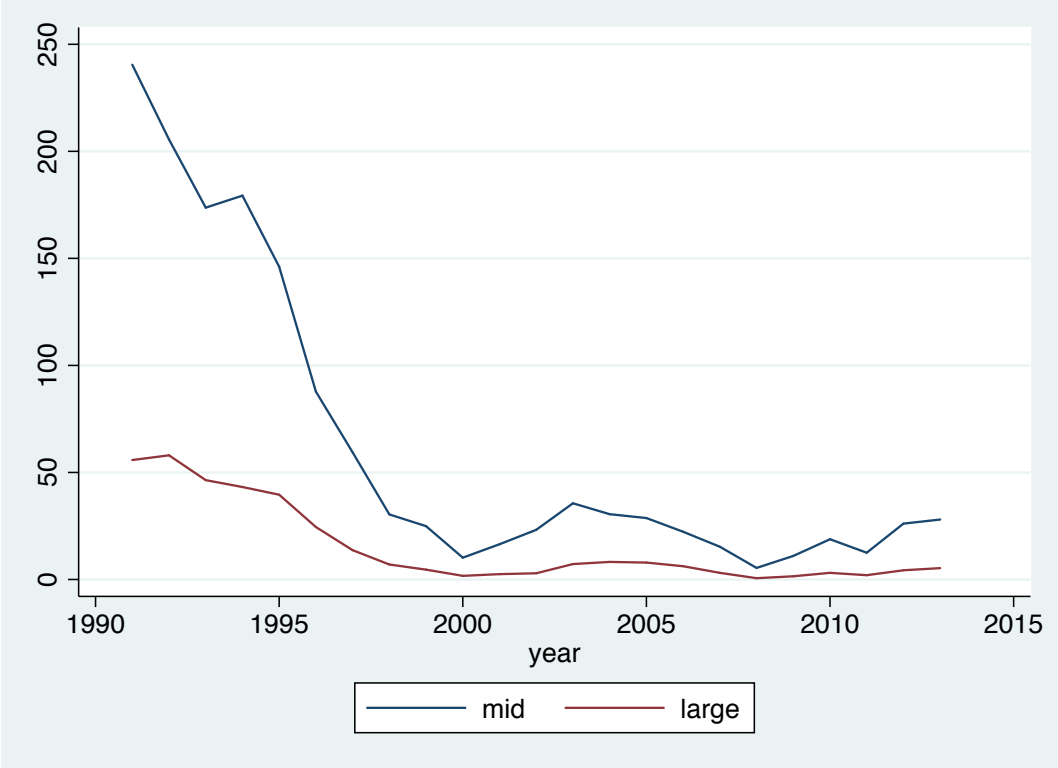
Note: These four panels correspond to four classical signals: Book-to-market, 1/rolling volatility, decrease in shares outstanding and cash-flows. Portfolio weights are computed as the rank of each stock, normalized so as to lie between -.5 and .5. We then hedge the portfolio using the simplified Markowitz procedure described in Section 4.1: This procedure assumes a simplified correlation structure of returns, and no transaction costs.

Figure 2: Cost of shorting: Hard to borrow rates by size pools



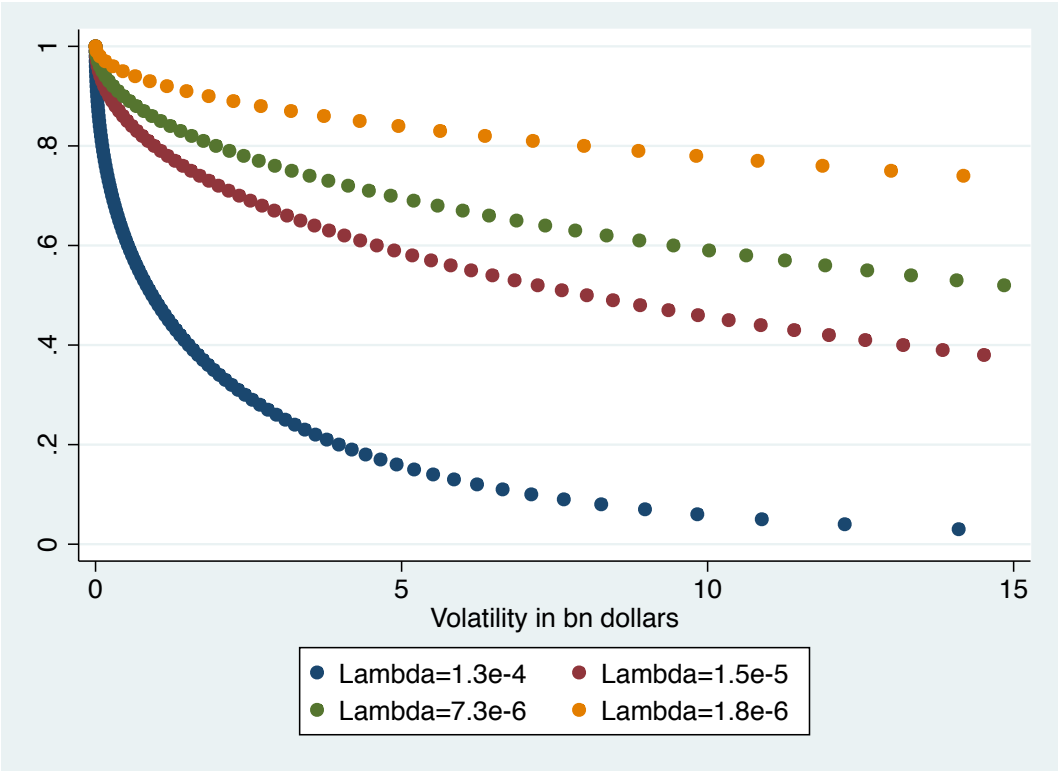
Note: Hard-to-borrow rates minus the General Collateral rate (or "easy-to-borrow rate") across size pools, using prop data. Rates are annualized, in %. The "General Collateral rate" corresponds to stocks which are not "'hard-to-borrow"' at a given point in time. We use the top 1500 US stocks sorted by size at each point in time. Pool 1 denotes the top 500 stocks, pool 2 is (500,1000) and pool 3 is (1000,1500).

Figure 3: Illiquidity lambdas by size pool (X 10-6)



Note: Using CRSP daily data, we compute for each year a stock-level λ using the formula $\lambda_i = \frac{1}{8 \times volume_i \times \sigma_i^2}$. We divide stocks into two time-varying pools: the "large" pool is composed of the top 500 stocks by market capitalization and the "mid" pool is that of stocks ranked between 500 and 1500. For each pool of stocks, we then define for each year the pool's λ as the median of the liquidity parameters (λ_i) of stocks belonging to the pool. Scaling of lambdas is reported in (X 10-6).

Figure 4: Pool Liquidity and Capacity

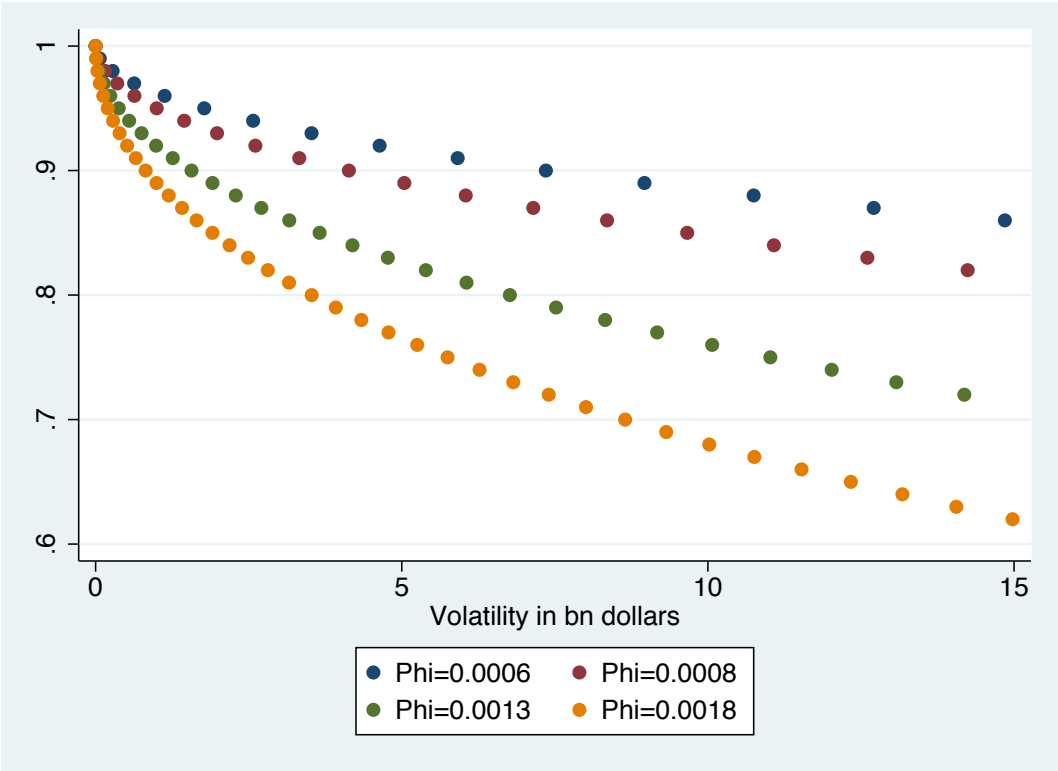


Note: This chart illustrates the impact of liquidity on strategy capacity. We consider a fictitious strategy whose frictionless Sharpe ratio is equal to 1, and whose persistence $\phi = 2.10^{-3}$ (25% in annualized terms), which roughly corresponds to the average persistence of our fundamental signals (see Table 1). Then, for each value of the Sharpe between 0 and 1, we compute the $ volatility reached using the formula drived in the main text:

$$Vol = \frac{SR}{\lambda \phi^2} \left[\left(\frac{SR^*}{SR} \right)^{2/3} - 1 \right]^2$$

We use 4 different values of λ, which correspond to the median λ in the mid pool in 1991-1995 (1.3e-4), 1996-2000 (1.5e-5), 2001-2005 (7.6e-6) and 2006-2013 (1.8e-6). We then draw the 4 frontiers with $ volatility on the x axis. Hence the blue line corresponds to the Sharpe-to-volatility frontier in the early 1990s, while the yellow line corresponds to the frontier in the later 2000s.

Figure 5: Signal Persistence and Capacity



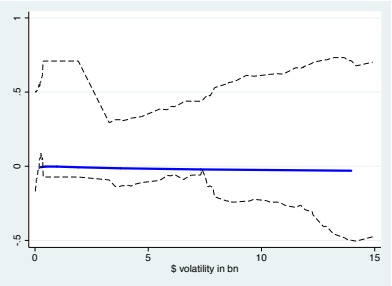
Note: This chart illustrates the impact of signal persistence on strategy capacity. We consider a fictitious strategy whose frictionless Sharpe ratio is equal to 1, and a pool whose price impact $\lambda = 1.8e-6$, which corresponds to the liquidity that prevails in the mid pool in 2012 (see Table 1). Then, for each value of the Sharpe between 0 and 1, we compute the $ volatility reached using the formula drived in the main text:

$$Vol = \frac{SR}{\lambda \phi^2} \left[\left(\frac{SR^*}{SR} \right)^{2/3} - 1 \right]^2$$

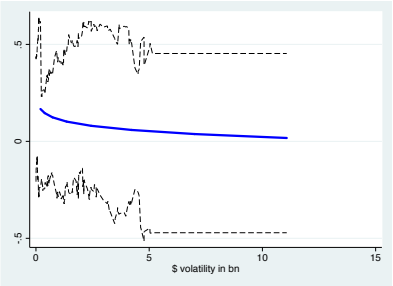
We use 4 different values of ϕ, which correspond to the ϕ in the 2000s for book-to-market (.2), cash-flows (.27), low vol (.31) and net shares outstanding growth (.36). We then draw the 4 frontiers with $ volatility on the x axis. Hence the blue line would correspond to the Sharpe-to-volatility frontier of book-to-market in the "mid" pool if it had a Sharpe of 1, while the yellow line would be the frontier of "net repurchasers" if it had a Sharpe of 1.

Figure 6: Performance-Capacity Frontiers: Backtesting Results

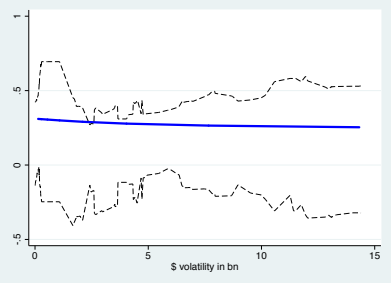
Panel A: Book-to-Market Big



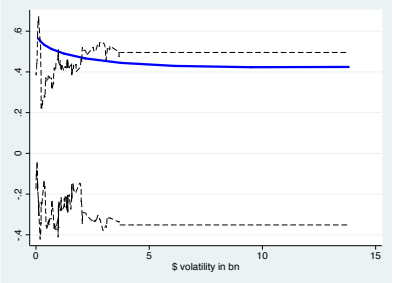
Book-to-Market Mid



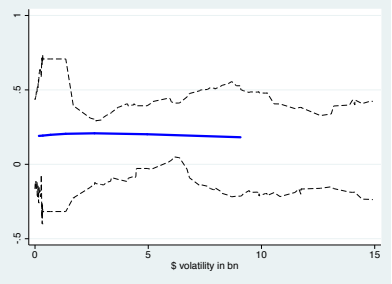
Panel B: Low Vol Big



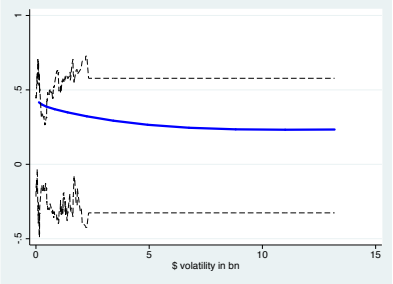
Low Vol Mid



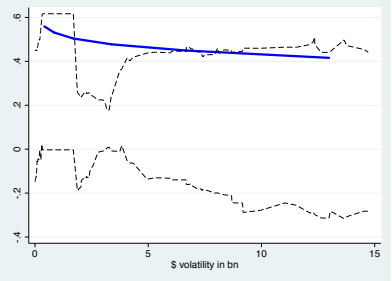
Panel C: Net repurchasers Big



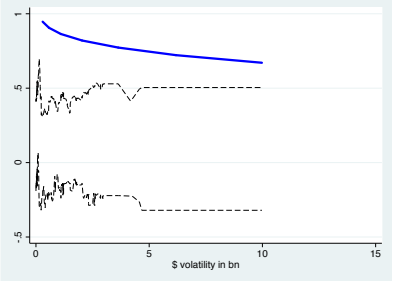
Net repurchasers Mid



Panel D: Cash-Flows Big



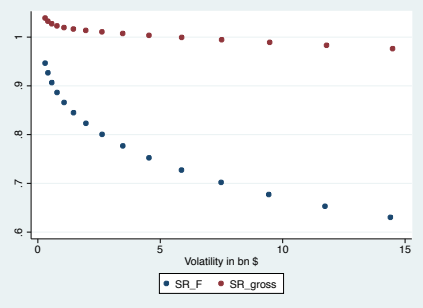
Cash-Flows Mid



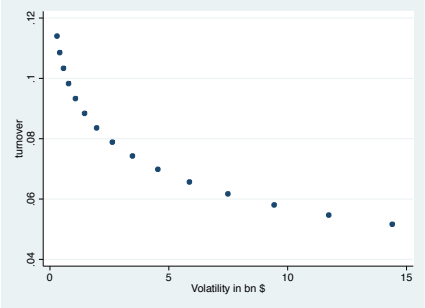
Note: The thick line represents the Sharpe-to-volatility frontier obtained by optimally trading on each of the 4 signals. The procedure is detailed in Section 4.1. The dashed lines represent the top and bottom 5% Sharpe-to-volatility frontiers obtained in simulated data where the signal does not predict returns.

Figure 7: Slowing-down Trading: Example of Cash-Flows in the Mid Pool

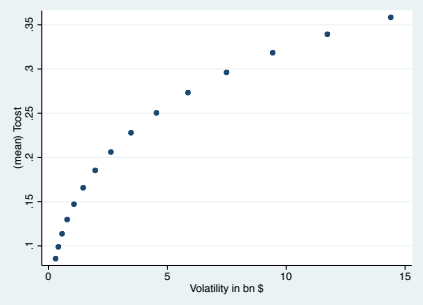
Panel A: Gross and Net Sharpe Deterioration



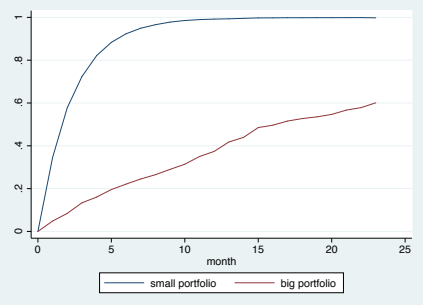
Panel B: Dollar turnover of gross portfolio (monthly)



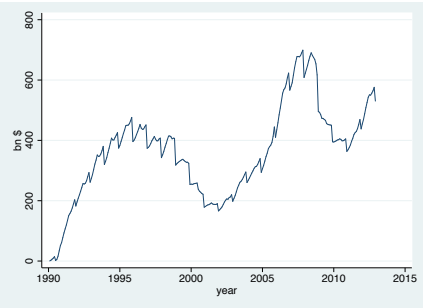
Panel C: Transactions costs per unit of gross pnl



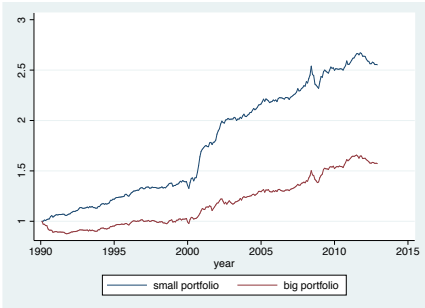
Panel D: Impulse Response of $x_i(t)$ to $aim_i(0) = 1$



Panel E: Time-series of Gross Exposure (large portfolio)

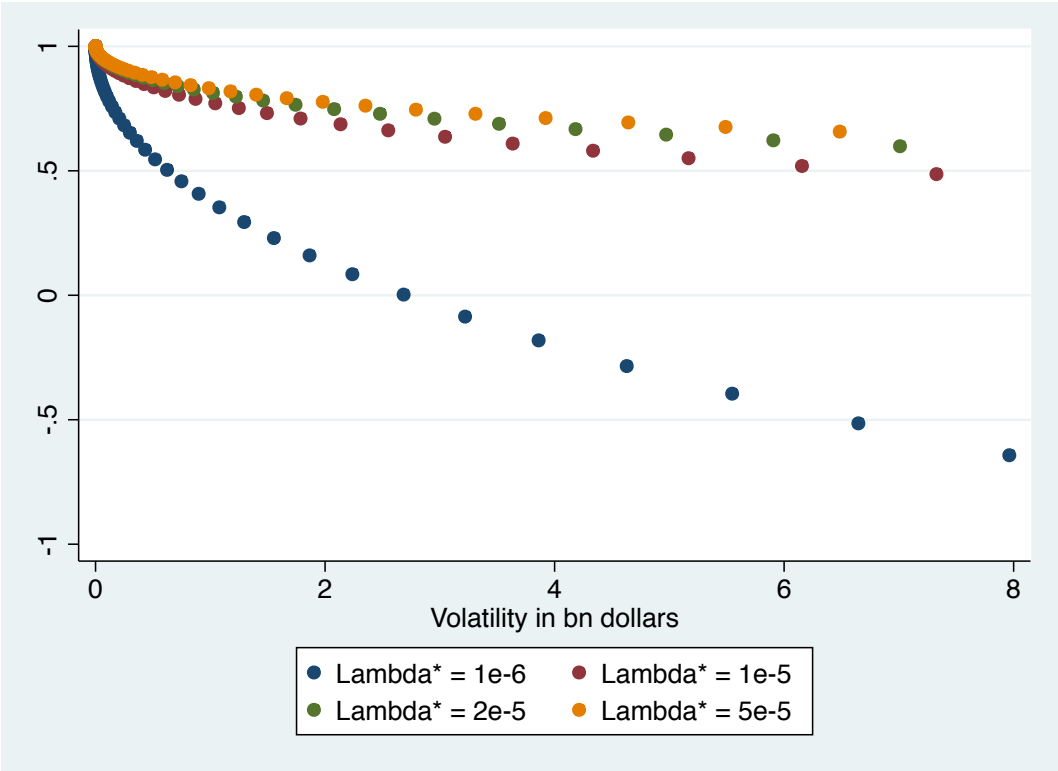


Panel F: Cumulative Returns on Gross Exposure



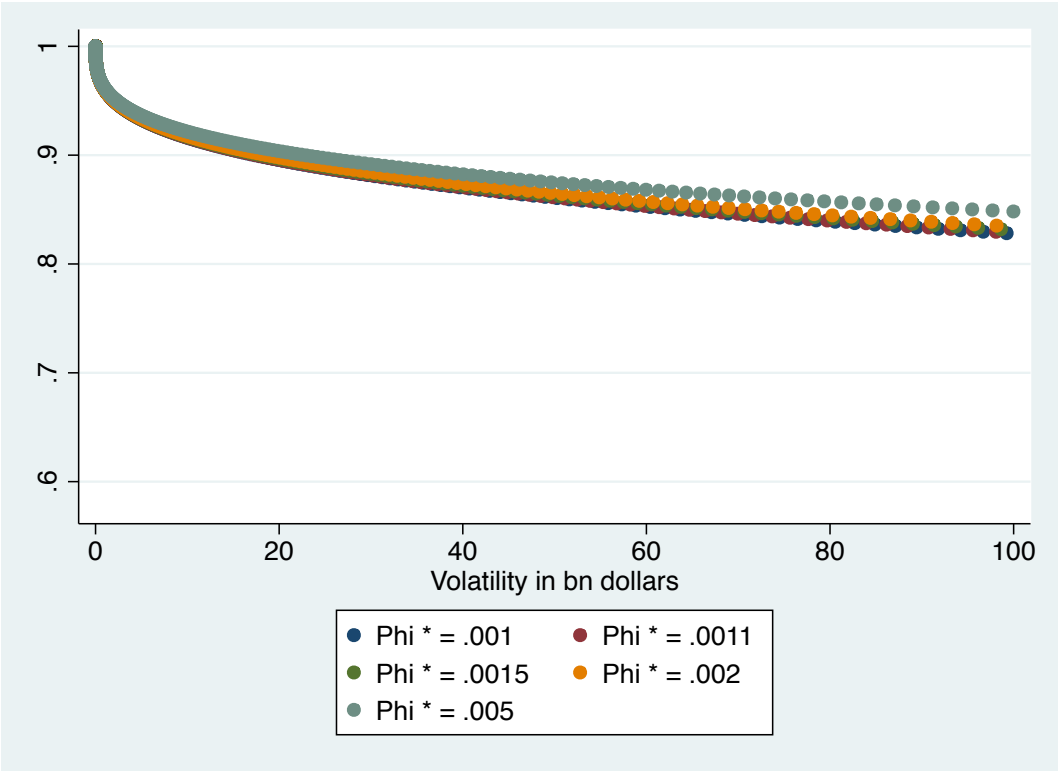
Note: We use back-testing results from the Cash-Flows strategy in the mid-pool. In panels A,B,C, the horizontal axis is the portfolio volatility targeted by the trader. Panel A shows together the Sharpe gross and net (of transaction costs). Panel B shows average monthly dollar turnover divided by the gross market value of portfolio (i.e. the dollar value of the long positions plus the dollar value of the short positions). Panel C shows average transaction costs per dollar of pnl. In Panel D, we show the cumulative response of $x_i(t)$ to $aim_i(0) = 1$, estimated in the back-tested portfolio panel using regression of $x_{i,t}$ on 25 lags of $aim_{i,t}$. "Small portfolio" corresponds to a target volatility of $ 10 mil., and "large portfolio" corresponds to a target volatility of $15 Bil. The horizontal axis is months since the shock. Panel E shows the times series of the gross exposure (i.e. long market value plus short market value) of the large portfolio. Panel F shows cumulative returns on gross exposure (i.e. pnl divided by gross exposure) for the large and small portfolios. The back-testing technique used to produce these results is detailed in Section 4.1. The time period is 1990-2013.

Figure 8: The Effect of a Mistake on Liquidity



Note: In this Figure, we assume that $\phi = .001$ and the actual $\lambda = 2.10^{-5}$. To construct each one of the curves, we then assume that the trader believes that λ is 10^{-6}, 10^{-5}, 2.10^{-5} (the correct assumption) and 5.10^{-5}. For each of these expected values of λ, we use the results in Proposition 4 to compute the hypothetical capacity frontier of a trader expecting liquidity λ while its actual value is 2.10^{-5}. Thus, the green curve corresponds to a trader using the correct parameters to calibrate his trading.

Figure 9: The Effect of a Mistake in Signal Persistence



Note: This chart illustrates the Sharpe-to-volatility frontiers of traders who hold "wrong beliefs" in the actual persistence of the signal. In all simulations, the price impact is given by $\lambda = 1.5e-5$ and the actual $\phi = .001$. To construct each one of the curves, we then assume that the trader believes that ϕ is 10^{-3} (the correct assumption), 1.110^{-3}, $1.5.10^{-3}$, 2.10^{-3} and 5.10^{-3}. For each of these actual values of ϕ, we use the results in Proposition 4 to compute the hypothetical capacity frontier of a traders believing liquidity is ϕ while the true parameter is 10^{-3}. Thus, the blue curve corresponds to a trader using the correct parameters to calibrate his trading.

APPENDIX

A Proof of Proposition 1

We note E is the unconditional expectation operator, which can be applied to any stationary process. We will use repeatedly two simple time-series properties: First if $E_t(y_{t+1}) = 0$ then for any stationary variable x_t , $E(x_t(y_{t+1})) = 0$; second, $E(x_t) = E(x_{t-1})$.

Step 1: Expected PNL.

We first compute the expected return of the optimal portfolio, without transaction costs:

$$\begin{aligned} ER_t &= E\big[x_t' r_{t+1}\big] \\ &= E\big[x_t' s_t\big] \\ &= E\Big\{\big[(1-\frac{a}{\lambda})x_{t-1} + \frac{a}{\lambda}x_t^\star\big]'\big[(1-\phi)s_{t-1}+\epsilon_t\big]\Big\} \\ &= (1-\frac{a}{\lambda})(1-\phi)ER_{t-1} + \frac{a}{\lambda}E(x_t^{\star\prime}s_t) \\ &= (1-\frac{a}{\lambda})(1-\phi)ER_t + \frac{a}{\lambda}\theta E(s_t'\Sigma_u^{-1}s_t) \\ &= \frac{a/\lambda}{1-(1-\phi)(1-a/\lambda)}\theta E(s_t'\Sigma_u^{-1}s_t) \end{aligned}$$

where

$$\theta = \frac{1}{\gamma + a\phi}$$

Step 2: Volatility of PNL.

Next, we compute the unconditional variance of R_t. Using the assumption $Es_t's_t << \Sigma_u$, we get:

$$\begin{aligned} (Vol(R_t))^2 &= E\big[x_t' r_{t+1} r_{t+1}' x_t\big] \\ &= E\big[x_t'\Sigma_u x_t\big] \\ &= (1-\frac{a}{\lambda})^2 E(x_{t-1}'\Sigma_u x_{t-1}) + 2\frac{a}{\lambda}(1-\frac{a}{\lambda})E(x_{t-1}'\Sigma_u x_t^{\star\prime}) + (\frac{a}{\lambda})^2 E(x_t^{\star\prime}\Sigma_u x_t^\star) \\ &= \frac{1}{1-(1-a/\lambda)^2}\big[2\frac{a}{\lambda}(1-\frac{a}{\lambda})E(x_{t-1}'\Sigma_u x_t^{\star\prime}) + (\frac{a}{\lambda})^2 E(x_t^{\star\prime}\Sigma_u x_t^\star)\big] \end{aligned}$$

Now, we use $x_t^\star = \frac{1}{\gamma+a\phi}\Sigma_u^{-1}s_t = \theta\Sigma_u^{-1}s_t$:

$$\begin{aligned} (Vol(R_t))^2 &= \frac{1}{1-(1-a/\lambda)^2}\big[2\frac{a}{\lambda}(1-\frac{a}{\lambda})\theta E(x_{t-1}'((1-\phi)s_{t-1}+\epsilon_t)) + (\frac{a}{\lambda}\theta)^2 E(s_t'\Sigma_u^{-1}s_t)\big] \\ &= \frac{1}{1-(1-a/\lambda)^2}\big[2\frac{a}{\lambda}(1-\frac{a}{\lambda})\theta(1-\phi)E(x_{t-1}'(s_{t-1})) + (\frac{a}{\lambda}\theta)^2 E(s_t'\Sigma_u^{-1}s_t)\big] \\ &= \frac{1}{(a/\lambda)(2-a/\lambda)}\big[2\frac{a}{\lambda}(1-\frac{a}{\lambda})\theta(1-\phi) + (\frac{a}{\lambda}\theta)(1-(1-\phi)(1-a/\lambda))\big]ER_t \\ &= \frac{\theta}{(2-a/\lambda)}\big[2(1-\frac{a}{\lambda})(1-\phi) + (1-(1-\phi)(1-a/\lambda))\big]ER_t \\ &= \frac{\theta}{(2-a/\lambda)}\big[1+(1-\phi)(2(1-a/\lambda)-(1-a/\lambda)\big]ER_t \\ &= \frac{\theta}{(2-a/\lambda)}\big[1+(1-\phi)(1-a/\lambda)\big]ER_t \end{aligned}$$

or:

$$Vol(R_t) = \frac{\theta\big[1+(1-\phi)(1-a/\lambda)\big]}{2-a/\lambda}\frac{E(R_t)}{Vol(R_t)}$$

Step 3: Expected transaction costs.

$$
\begin{aligned}
E(TC_t) &= \lambda E \Delta x_t' \Sigma_u \Delta x_t \\
&= 2\lambda \big[E(x_t' \Sigma_u x_t) - E(x_{t-1}' \Sigma_u x_t) \big] \\
&= 2\lambda \big[E(x_t' \Sigma_u x_t) - E\big(x_{t-1}' \Sigma_u ((a/\lambda) x_t^\star + (1 - a/\lambda) x_{t-1}) \big) \big] \\
&= 2\lambda (a/\lambda) \big[E(x_t' \Sigma_u x_t) - E\big(x_{t-1}' \Sigma_u x_t^\star \big) \big] \\
&= 2a \big[E(x_t' \Sigma_u x_t) - \theta(1-\phi) E(R_t) \big] \\
&= 2a \big[(Vol(R_t))^2 - \theta(1-\phi) E(R_t) \big]
\end{aligned}
$$

where we use:
$E\big(x_{t-1}' \Sigma_u x_t^\star\big) = E\big(x_{t-1}' \Sigma_u \theta \Sigma_u^{-1} s_t\big) = \theta(1-\phi) E\big(x_{t-1}' s_{t-1}\big) = \theta(1-\phi) E R_{t-1} = \theta(1-\phi) E R_t$
This can be conveniently expressed by unit of volatility:

$$
\begin{aligned}
\frac{E(TC_t)}{Vol(R_t)} &= 2a \big[(Vol(R_t)) - \theta(1-\phi) \frac{E(R_t)}{Vol(R_t)} \big] \\
&= 2a \big[\frac{\theta \big[1 + (1-\phi)(1 - a/\lambda) \big]}{2 - a/\lambda} - \theta(1-\phi) \big] \frac{E(R_t)}{Vol(R_t)} \\
&= 2a\theta \big[\frac{1 + (1-\phi)(1-a/\lambda) - (1-\phi)(2 - a/\lambda)}{2 - a/\lambda} \big] \frac{E(R_t)}{Vol(R_t)} \\
&= \frac{2a\theta\phi}{2 - a/\lambda} \frac{E(R_t)}{Vol(R_t)}
\end{aligned}
$$

Step 4: PNL Sharpe Ratio.

$$
\begin{aligned}
SR &= \frac{E(R_t - TC_t)}{Vol(R_t)} \\
&= [1 - \frac{2a\theta\phi}{2 - a/\lambda}] \frac{E R_t}{Vol(R_t)}
\end{aligned}
$$

$$
\begin{aligned}
\Big(\frac{E(R_t)}{Vol(R_t)} \Big)^2 &= \frac{2 - a/\lambda}{\theta \big[1 + (1-\phi)(1-a/\lambda) \big]} E R_t \\
&= \frac{2 - a/\lambda}{\theta \big[1 + (1-\phi)(1-a/\lambda) \big]} \frac{\theta a/\lambda}{1 - (1-\phi)(1-a/\lambda)} E(s_t' \Sigma_u^{-1} s_t) \\
&= \frac{2 - a/\lambda}{\big[1 + (1-\phi)(1-a/\lambda) \big]} \frac{a/\lambda}{1 - (1-\phi)(1-a/\lambda)} E(s_t' \Sigma_u^{-1} s_t) \\
&= \frac{1 - (1 - a/\lambda)^2}{1 - ((1-\phi)(1-a/\lambda))^2} E(s_t' \Sigma_u^{-1} s_t)
\end{aligned}
$$

or

$$
\frac{E(R_t)}{Vol(R_t)} = \Big(\frac{1 - (1-a/\lambda)^2}{1 - ((1-\phi)(1-a/\lambda))^2} \Big)^{1/2} SR_t^\star
$$

The formula for the Sharpe Ratio follows directly.
The last step consists of computing the portfolio volatility:

$$
\begin{aligned}
Vol(R_t) &= \theta \frac{1 + (1-\phi)(1-a/\lambda)}{2 - a/\lambda} \frac{E(R_t)}{Vol(R_t)} \\
&= \frac{1}{\gamma + a\phi} \frac{1 + (1-\phi)(1-a/\lambda)}{2 - a/\lambda} \Big(\frac{1 - (1-a/\lambda)^2}{1 - ((1-\phi)(1-a/\lambda))^2} \Big)^{1/2} SR^\star \\
&= \frac{1}{\gamma + a\phi} \frac{1 + (1-\phi)(1-a/\lambda)}{2 - a/\lambda} \Big(\frac{(2 - a/\lambda)(a/\lambda)}{(1 + (1-\phi)(1-a/\lambda))^2} \Big)^{1/2} SR^\star \\
&= \frac{1}{\gamma + a\phi} \Big[\frac{a/\lambda}{2 - a/\lambda} \frac{1 + (1-\phi)(1-a/\lambda)}{1 - (1-\phi)(1-a/\lambda)} \Big]^{1/2} SR^\star
\end{aligned}
$$

QED.

B Proof of Proposition 2

First, we compute the approximation for the trading rate $\tau = a/\lambda$. It is given by:

$$\begin{aligned}\tau &= \frac{1}{2}.\Big(-(\frac{\gamma}{\lambda}+\delta)+\sqrt{(\frac{\gamma}{\lambda}+\delta)^2+4\frac{\gamma}{\lambda}}\Big)\\ &= \frac{1}{2}.\Big(-(\frac{\gamma}{\lambda}+\delta)+\sqrt{(\frac{\gamma}{\lambda})^2+2\frac{\gamma}{\lambda}\delta+\delta^2+4\frac{\gamma}{\lambda}}\Big)\\ &= \frac{1}{2}.\Big(-(\frac{\gamma}{\lambda}+\delta)+\sqrt{(\frac{\gamma}{\lambda})^2+2\frac{\gamma}{\lambda}(\delta+2)+\delta^2}\Big)\end{aligned}$$

Under the large investment approximation, $\delta \ll (\gamma/\lambda)^{1/2} \ll 1$, so the trading rate simplifies into:

$$\tau \approx \frac{1}{2}.\Big(-(\frac{\gamma}{\lambda}+\delta)+\sqrt{(\frac{\gamma}{\lambda})^2+\delta^2+4\frac{\gamma}{\lambda}}\Big)$$

Also, $\frac{\gamma}{\lambda} \ll 1$ thus:

$$\tau \approx \frac{1}{2}.\Big(-(\frac{\gamma}{\lambda}+\delta)+\sqrt{\delta^2+4\frac{\gamma}{\lambda}}\Big)$$

Finally, $\frac{\gamma}{\lambda} \gg \delta^2$ which leads to the simple form:

$$\tau \approx \frac{1}{2}.\Big(-(\frac{\gamma}{\lambda}+\delta)+\sqrt{4\frac{\gamma}{\lambda}}\Big)$$

Thanks to the assumptions that $\frac{\gamma}{\lambda} \ll 1$ and $\frac{\gamma}{\lambda} \gg \delta^2$, the term in square root dominates so that:

$$\tau \approx \sqrt{\frac{\gamma}{\lambda}}$$

Given the above formula for the trading speed τ, the large investment approximation implies that: $t \ll 1$ and $t \gg \delta$. Only a few percent of the portfolio are traded every day, but this churning rate is much larger than the daily discount rate.

We then compute the Sharpe ratio. The first term of the Sharpe Ratio formula is given by:

$$\begin{aligned}1-\frac{2\tau\phi}{\gamma/\lambda+\tau\phi}\frac{1}{1+(1-\tau)} &= 1-\frac{2\phi}{\phi+\tau}\frac{1}{2-\tau}\\ &\approx \frac{\tau}{\phi+\tau}\end{aligned}$$

since $\tau \ll 1$.

The second term of the Sharpe ratio is given by

$$\begin{aligned}\Big[\frac{1-(1-\tau)^2}{1-(1-\tau)^2(1-\phi)^2)}\Big]^{1/2} &= \tau^{1/2}\Big[\frac{2-\tau}{1-(1-\tau)^2(1-\phi)^2}\Big]^{1/2}\\ &\approx \frac{\tau^{1/2}}{1-\phi}.\Big[\frac{1}{\frac{(1-(1-\phi)^2)}{2(1-\phi)^2}+\tau)}\Big]^{1/2}\end{aligned}$$

where we also use the fact that $t \ll 1$.

Combining the two parts, we obtain that:

$$SR \approx \Big(\frac{1}{\phi+\tau}\Big)\frac{\tau^{3/2}}{1-\phi}\Big(\frac{1}{\frac{(1-(1-\phi)^2)}{2(1-\phi)^2}+\tau)}\Big)^{1/2}SR^*$$

which is the expression for the Sharpe ratio in the proposition.

We now compute volatility:

$$Vol = \frac{1}{\lambda}\frac{1}{\tau^2+\tau\phi}\Big[\frac{\tau}{2-\tau}\frac{1+(1-\phi)(1-\tau)}{1-(1-\phi)(1-\tau)}\Big]^{1/2}SR^\star$$

Using the assumption that $\tau \ll 1$ we get that:

$$Vol \approx \frac{1}{\lambda\tau^{1/2}} \frac{1}{\tau+\phi} \Big[\frac{1-\phi/2}{\phi+\tau(1-\phi)}\Big]^{1/2} SR^{\star}$$

This proves the second equation. QED

C Proof of Proposition 3

First, we take the results from Proposition 2, and let $\phi \ll 1$, which leads to:

$$SR \approx \Big(\frac{\tau}{\phi+\tau}\Big)^{3/2} SR^{*}$$
$$Vol \approx \frac{1}{\lambda\tau^2} \Big[\frac{\tau}{\tau+\phi}\Big]^{3/2} SR^{\star}$$

Combining the two equations, it is easy to see that: $Vol = SR/(\lambda\tau^2)$ which leads to the following expression that combines τ with the SR and the \$ volatility:

$$\tau = \sqrt{\frac{SR}{\lambda Vol}}$$

The validity conditions in the Proposition immediately follow from $\delta \ll \tau \ll 1$.

To obtain the second result, we just need to reverse the expression for the Sharpe ratio obtained previously:

$$\tau = \phi \frac{1}{(SR/SR^{*})^{2/3} - 1}$$

We then combine the two definitions of τ and obtain the formula in the proposition.

To show the third bullet point in the proposition, we first note $F(x) = x((SR^{*}/x)^{2/3} - 1)^2$. It is easy to see that F is convex (it is the square of a convex function) and decreasing. Hence, its inverse is concave and decreasing. Let us call it G, then:

$$SR = G(\lambda\phi^2 \mathrm{Vol})$$

thus:

$$\begin{aligned}\frac{\partial^2 SR}{\partial\lambda\partial\phi} &= 2\phi.G'.\mathrm{Vol} + 2\lambda\phi^3.G''.\mathrm{Vol} \\ &< 0\end{aligned}$$

since $G'' < 0$ and $G' < 0$. QED

For the last bullet point of the proposition, we go back to expressions of TC in the proof of Proposition 1 and combine them with the approximation:

$$\begin{aligned}\frac{TC}{ER} &= \frac{2\tau\phi}{2-\tau} \frac{1}{\gamma/\lambda + \tau\phi} \\ &\approx \frac{2\tau\phi}{2-\tau} \frac{1}{\tau^2 + \tau\phi} \\ &\approx \frac{\phi}{\phi+\tau}\end{aligned}$$

which is valid even when the "slow signal" approximation is not valid.

We combine the above expression with the formula of SR at the beginning of this proof, and obtain the stated result. QED

D Proof of Proposition 4

Let us first show the first bullet point. The fact that the true transaction cost parameter is λ^{*} only changes the transaction costs:

$$\frac{TC}{Vol(R_t)} = \frac{2\lambda^{*}\tau\phi}{2-\tau} \frac{1}{\lambda((\gamma/\lambda) + \phi\tau)} \frac{E(R_t)}{Vol(R_t)}$$

We inject this new formula into the definition of the net Sharpe ratio, and obtain:

$$\frac{ER-TC}{Vol(R_t)} = \Big(1 - \frac{2\lambda^*\phi}{\lambda(2-\tau)}\frac{\tau}{(\gamma/\lambda)+\phi\tau}\Big)\Big(\frac{1-(1-\tau)^2}{1-((1-\phi)(1-\tau))^2}\Big)^{1/2} SR_t^\star$$

We then make the large investment approximation, so that $\tau \approx \sqrt{\gamma/\lambda} \ll 1$. This leads to:

$$SR = \Big(1 - \frac{\lambda^*}{\lambda}\frac{\phi}{\tau+\phi}\Big)\Big(\frac{2\tau}{1-(1-\phi)^2(1-2\tau)}\Big)^{1/2} SR_t^\star$$

Then, we make the slow signal approximation: $\phi \to 0$, so that:

$$SR = \Big(1 - \frac{\lambda^*}{\lambda}\frac{\phi}{\tau+\phi}\Big)\Big(\frac{\tau}{\phi+\tau}\Big)^{1/2} SR_t^\star$$

The algebra for volatility follows similar steps.

The second result is shown using a similar method, except that, if the investor makes a mistake on ϕ, it does not affect his trading speed but only the aimed portfolio. QED.

E Proof of Proposition 5

Given the dynamics in equation (11), the formulae developed in Appendix A are still valid, except that we need to replace a/λ by π and θ by $1/\gamma$. The formula for the Sharpe ratio becomes:

$$\begin{aligned} SR(\gamma) &= \Big[1 - \frac{2\pi\lambda\phi}{\gamma}\frac{1}{1+(1-\pi)}\Big]\Big[\frac{1-(1-\pi)^2}{1-(1-\pi)^2(1-\phi)^2}\Big]^{1/2} SR^* \\ &\approx \Big[\frac{\pi}{\pi+\phi}\Big]^{1/2} SR^* \end{aligned}$$

while the formula for the volatility becomes:

$$\begin{aligned} \mathrm{Vol}(\gamma) &= \frac{1}{\gamma}\Big[\frac{\pi}{2-\pi}\frac{1+(1-\pi)(1-\phi)}{1-(1-\pi)(1-\phi)}\Big]^{1/2} SR^* \\ &\approx \frac{1}{\gamma}\Big[\frac{\pi}{\pi+\phi}\Big]^{1/2} SR^* \end{aligned}$$

We combine the two equations and obtain:

$$\mathrm{Vol} = \frac{SR}{\lambda\phi}\Big[\Big(\frac{SR^*}{SR}\Big)^2 - 1\Big]$$

QED.

F A simple implementation of Markowitz hedging

Finding the Markowitz portfolio requires inverting the variance-covariance matrix of stocks. A simple way to do this is to assume a specific albeit quite general form to this matrix. In this section, we assume that within a given liquidity pool, we can describe returns as:

$$r_{i,t} = S_t + \beta r_{M,t} + \epsilon_{i,t}$$

where $r_{M,t}$ is the market return (of variance σ_M^2), β the vector of betas, S_t the vector of signals and $\epsilon_{i,t}$ a common idiosyncratic shock of variance σ_ϵ^2. Therefore, the variance-covariance matrix is (omitting the time subscript for simplicity):

$$\Sigma = \sigma_M^2 \beta\beta' + \sigma_\epsilon^2 Id$$

Proposition 6.

$$\Sigma^{-1} = \frac{1}{\sigma_\epsilon^2}\{Id - \frac{\sigma_M^2}{\sigma_\epsilon^2}\frac{1}{1+\frac{\sigma_M^2}{\sigma_\epsilon^2}(\sum_i \beta_i^2)}\beta\beta'\}$$

and the Markowitz portfolio is

$$\frac{1}{\gamma\sigma_\epsilon^2}\{s_t - \frac{\sigma_M^2}{\sigma_\epsilon^2}\frac{\sum_i \beta_i s_i}{1+\frac{\sigma_M^2}{\sigma_\epsilon^2}(\sum_i \beta_i^2)}\beta\}$$

Proof. Note that $(\beta\beta')^k = (\sum_i \beta_i^2)^{k-1}(\beta\beta')$ and

$$\Sigma^{-1} = \frac{1}{\sigma_\epsilon^2}\big(\sum_i (-\frac{\sigma_M^2}{\sigma_\epsilon^2})^k (\beta\beta')^k\big)$$

□

It is then easy to compute the Sharpe ratio of the hedged portfolio:

$$(SR^*)^2 = s'\Sigma^{-1}s = \frac{1}{\sigma_\epsilon^2}\cdot\Big(\sum s_i^2 - \sigma_M^2 \frac{\sum s_i\beta_i}{\sigma_\epsilon^2 + \sigma_M^2 \sum \beta_i^2}\Big)$$

which is the formula we show in the text. QED