Journal of Financial Economics 114 (2014) 405–423

Contents lists available at ScienceDirect

Journal of Financial Economics

journal homepage: www.elsevier.com/locate/jfec

Price pressures ☆

Terrence Hendershott [a], Albert J. Menkveld [b,c,d,*]

[a] *Haas School of Business, University of California at Berkeley, United States*
[b] *VU University Amsterdam, FEWEB, De Boelelaan 1105, 1081 HV Amsterdam, Netherlands*
[c] *Duisenberg school of finance, Netherlands*
[d] *Tinbergen Institute, Netherlands*

ARTICLE INFO

Article history:
Received 8 June 2013
Received in revised form
20 August 2013
Accepted 7 October 2013
Available online 14 August 2014

JEL classification:
G12
G14
D53
D61

Keywords:
Liquidity
Inventory risk
Intermediary
Volatility

ABSTRACT

We study price pressures, i.e., deviations from the efficient price due to risk-averse intermediaries supplying liquidity to asynchronously arriving investors. Empirically, New York Stock Exchange intermediary data reveals economically large price pressures, 0.49% on average with a half life of 0.92 days. Theoretically, a simple dynamic inventory model captures an intermediary's use of price pressure to mean-revert inventory. She trades off revenue loss due to price pressure against price risk associated with staying in a nonzero inventory state. The closed-form solution identifies the intermediary's risk aversion and the investors' private value distribution from the observed time series patterns of prices and inventories. These parameters imply a relative social cost due to price pressure, a deviation from constrained Pareto efficiency, of approximately 10% of the cost of immediacy.

1. Introduction

The cost of trading large quantities quickly, often referred to as liquidity, plays a fundamental role in facilitating risk sharing and allocating resources through securities markets. A market might turn illiquid through two main channels. Informational asymmetry makes counterparties charge a permanent price impact to protect against losses due to information-motivated traders (e.g., Glosten and Milgrom, 1985). In addition, a natural counterparty might not yet be in the market and a risk-averse intermediary charges a price impact for temporarily holding the position (e.g., Grossman and Miller, 1988). This price impact is transitory as the price rebounds once the natural counterparty arrives, allowing the intermediary to liquidate its initial position. These transitory price effects or price pressures are the focus of this study.

This paper generates systematic evidence on the price pressure channel and interprets it based on a structural

☆ We benefit from in-depth discussions with Viral Acharya, Bruno Biais, Sabrina Buti, Darrell Duffie, Robert Engle, Joel Hasbrouck, Charles Jones, Jeremy Large, Sunny Li, Virgiliu Midrigan, Mark Seasholes, Stijn van Nieuwerburgh, Marios Panayides, Matthew Pritsker, Jan van Schuppen, Elvira Solji, Clara Vega, Pierre-Olivier Weill, and audience participation at the Bank of Canada, Cass Business School, Université Paris Dauphine, ESSEC Business School, National Bureau of Economic Research micro-structure, Society for Financial Econometrics (SoFiE) 2011 Conference held in Chicago, Illinois, Stanford University, Swiss Central Bank, and University of Toronto. The project was inspired by the Lars Ljungqvist and Thomas J. Sargent New York University (NYU) PhD course based on their book *Recursive Macroeconomic Theory*. Albert J. Menkveld gratefully acknowledges Robert Engle for sponsoring his NYU visit in August 2010 and VU University Amsterdam for a VU talent grant.

* Corresponding author at: VU University Amsterdam, FEWEB, De Boelelaan 1105, 1081 HV Amsterdam, Netherlands.
Tel.: +31 20 598 6130; fax: +31 20 598 6020.
E-mail address: albertjmenkveld@gmail.com (A.J. Menkveld).

http://dx.doi.org/10.1016/j.jfineco.2014.08.001

model. The empirical challenge is to disentangle price pressure effects associated with an intermediary's trading from informational effects. A state space model decomposes price changes into two components: a permanent change due to information and a transitory change potentially due to price pressure. The empirical model relates the permanent changes to surprise order flow and the transitory changes to the intermediary's inventory. Moreover, the empirical state space model turns out to be the solution of a simple structural model in which a representative intermediary solves a dynamic inventory control problem. Thus, the empirical estimates identify deep parameters such as the intermediary's risk aversion and investors' private value distribution for the assets they trade.

Several reasons can be cited for why the state space approach is preferable to other approaches such as an autoregressive moving average (ARIMA) model (e.g., Hasbrouck, 1991a,b) or the generalized method of moments (GMM). First, maximum likelihood estimation is asymptotically unbiased and efficient. Second, the Kalman filter used in the estimation ensures maximum efficiency in dealing with missing values; i.e., it accounts for level series changes across periods with missing observations. Standard approaches based on the differenced series discard such information. This is particularly beneficial for high-frequency data on inactively traded securities. Third, after estimation, the Kalman smoother (a backward recursion) enables the decomposition of any realized price change into a permanent part and a transitory part conditional on all observations, i.e., past, current, and future observations. Therefore, the decomposition benefits from peering into the future. Durbin and Koopman (2012) provide an in-depth discussion on the use of state space models in economics. Hasbrouck (1999) applies the state-space approach to the intraday dynamics of discrete bid and ask quotes.

The empirical estimates show economically large and persistent price pressures. The state space model is estimated for 697 stocks based on daily prices and inventory positions of NYSE intermediaries from 1994 through 2005. The average size of price pressure is 49 basis points with a half life of 0.92 days. Substantial cross-sectional variation exists. Price pressure for the quintile of largest stocks is 17 basis points with a half life of 0.54 days. For the smallest-stocks quintile, it is 118 basis points with a half life of 2.11 days. These price pressures are roughly the size of the (effective) bid–ask spread.

The state space estimates also yield a novel empirical measure of liquidity: the price pressure an intermediary charges per unit of inventory. This marginal price pressure is the price elasticity of liquidity demand and is an alternative to the standard bid–ask spread measure. Institutional investors in particular often care more about the marginal pressure than the spread. They typically split large orders into small pieces sent to the market sequentially. The cumulative impact of that order i.e., the extent to which their trading drives the price away from the efficient price as the order executes, is typically much larger than the half-spread paid on each individual order. Marginal price pressure as a liquidity measure instead of the spread is in the tradition of earlier work by Stoll (1978), Grossman and Miller (1988), Campbell, Grossman, and Wang (1993), and Pastor and Stambaugh (2003).[1]

Small stocks are much less liquid than large stocks based on marginal pressure than based on the spread. The marginal price pressure is 0.02 basis points per one thousand dollars of idiosyncratic inventory position for the (quintile of) largest stocks and 0.98 basis points for the smallest stocks. The (half-) spread for these two sets of stocks is 8.41 and 46.12 basis points, respectively, implying the largest stocks are 49 times more illiquid when measured by marginal pressure but only five times more illiquid when measured by spread. The analysis is based on idiosyncratic inventory because systematic inventory risk can easily be hedged through offsetting positions in index futures or exchange traded funds.

A simple recursive structural model in the spirit of Ho and Stoll (1981) helps to further understanding of the average size of price pressure, its duration, and the level of marginal pressure. The model studies a risk-averse intermediary who uses price pressure to manage risk though mean-reverting her inventory. She equates the size of the subsidy she pays to speed up mean-reversion, i.e., the price pressure, to the utility cost of one more time unit of price risk on her nonzero position. This pecuniary amount equated to a utility cost facilitates empirical identification of the model's deep parameters. Observed price pressure together with the price risk on the intermediary's inventory identifies her risk aversion. The speed of mean-reversion given the level of price pressure identifies investors' private value distribution as it captures how price pressure impacts investors' buying and selling. If the private distribution is relatively flat (inelastic), then not many buyers are replaced by sellers and the net demand response to price pressure is small. This intuition is formalized in the stochastic inventory control problem and its closed-form solution.

The model's solution along with the time series properties of price pressure generates a rich economic perspective. The intermediary's relative risk aversion is 3.96. The uniform distribution assumed for private values ranges from -185 to $+185$ basis points. The model-implied (half-) spread that makes the intermediary indifferent between participating or not ranges from 7.63 basis points for the largest stocks to 87.41 basis points for the smallest stocks. This range estimate is based entirely on inventory dynamics and price pressure estimates. It is, therefore, comforting that it is the same order of magnitude as the observed half-spread range of 8.41–46.12 basis points.

The model also provides insight into potential investor welfare loss due to asynchronous arrivals and risk-averse intermediation. The intermediary internalizes the price pressure as the cost of risk management (inventory reduction). This cost, however, overestimates the extent to which private value is lost as it is based on the differential valuation of the marginal seller who is substituted by the marginal buyer. A social planner cares about the average

[1] Grossman and Miller (1988, p. 630) criticize not only the bid–ask spread as a liquidity measure but also the liquidity ratio, defined as the ratio of average dollar volume to the average price change. Its reciprocal, the illiquidity ratio, is subject to the same critique.

seller replaced by the average buyer as a planner integrates welfare across all agents. The intermediary considers price pressure to be more costly than the planner and mean-reverts more slowly. Empirically, this results in a 10% higher cost of immediacy relative to constrained Pareto efficiency.

While our empirical analysis is specific to the NYSE structure during our sample period, price pressures capture generic properties of supply and demand and the methodology developed herein is more widely applicable. In the presence of slow-moving capital due to some friction such as inattention, it is natural for some agents to step in to produce immediacy and be compensated through price pressure (see, e.g., Duffie, 2010). Regulators should track the size of marginal price pressure in securities markets to monitor liquidity. They typically are the only ones who have access to the trader identities needed to compute the aggregate position of intermediaries at any time, e.g., Kirilenko, Kyle, Samadi, and Tuzun (2011). Following the Flash Crash, a committee including Nobel laureates proposed pricing incentives for agents with market making strategies that at all times maintain best buy and sell quotations that are "reasonably related to the market" (Born, Brennan, Engle, Ketchum, O'Hara, Philips, Ruder, and Stiglitz, 2011, p. 10). The realized marginal price pressure (a by-product of the model) could indicate whether their quotations were reasonable during market turbulence. Realized pressure could also be useful for issuers who want to evaluate the performance of the designated market makers they hire for their stock (see, e.g., Nimalendran and Petrella, 2003; Venkataraman and Waisburd, 2007; Anand, Tanggaard, and Weaver, 2009; Menkveld and Wang, 2013). In, 2013 the Nasdaq proposed to start such program in the US (see Securities and Exchanges Commission Release No. 34-69195).

The paper's focus on everyday price pressures contributes to a literature largely focusing on instances identifying liquidity demand by investors. Kraus and Stoll (1972) are among the first to show the existence of price pressures by studying large institutional trades. Subsequent studies focusing on other types of trades reveal that price pressures can be large.[2] To more fully characterize the price dynamics arising from investors' trading needs, one needs data on intermediaries who take temporary inventory positions to satisfy liquidity demand. Madhavan and Smidt (1991, 1993) and Hasbrouck and Sofianos (1993) show inventory mean-reversion for NYSE intermediaries (specialists), but not price pressure. Hendershott and Seasholes (2007) use a longer time series of similar data to find evidence consistent with both inventory control and price pressure. They find that a portfolio of stocks for which the intermediary is long outperforms a portfolio of stocks for which she is short. Their results, however, are also consistent with the alternative explanation based on information asymmetry. The state space approach developed in this paper identifies the contribution of both channels.

Section 2 presents estimates of the magnitude of daily price pressures and pricing errors. Section 3 sets up and solves a structural model and shows how the price pressure estimates identify the model's deep parameters. How the unique NYSE market structure and the intermediaries' role in it impact the interpretation of the results is discussed in Section 4. In particular, it discusses an alternative, information-based explanation for the negative correlation between inventory and the pricing error. Section 5 concludes.

2. Empirical identification of price pressure and inventory dynamics

Markets on which investors do not continuously monitor and participate deviate from the standard Walrasian tradition (Townsend, 1978; Grossman and Miller, 1988; Rust and Hall, 2003; and many others). Historically, the NYSE assigned to each stock one intermediary, called a specialist, to act as a market maker. This structure helps identify the amount of liquidity supplied. Data on these intermediaries' inventory positions help identify price pressure both in the cross-section and through time. While the designation of a single intermediary is unique to the NYSE, the fundamental economic forces that generate price pressure and intermediaries' inventory risk exist in all markets on which investor trading needs are not perfectly synchronized. Section 4 further discusses the NYSE structure and how it changed subsequent to our sample period.

2.1. Data and summary statistics

Center for Research in Security Prices (CRSP) data, the NYSE's Trade and Quotes (TAQ) data, and a proprietary NYSE data set provide the end-of-day midquote price (i.e., the average of the bid price and ask price), the end-of-day specialist inventory position, and other variables from 1994 through 2005. A balanced panel is created to make results comparable through time and control for stock fixed effects. All NYSE common stocks matched across TAQ and CRSP with an average share price of more than \$5 and less than \$1,000 are used. The resulting sample contains 697 stocks present throughout the whole sample period.

Stocks are sorted into quintiles based on market capitalization and stay in the assigned quintile throughout the sample. Quintile 1 refers to the large-cap stocks, and Quintile 5 corresponds to the small-cap stocks. The NYSE specialist positions are in shares in the original database. To facilitate comparisons across stocks, we convert inventory into dollars by multiplying the position with each stock's average price each year after adjusting for stock-splits and dividends. This eliminates daily price changes in

[2] Harris and Gurel (1986) and subsequent papers on additions to the Standard and Pool's 500 index find evidence for price pressure. Greenwood (2005) extends this with an examination of transitory price effects upon a weighting change to the Nikkei 225. Coval and Stafford (2007) examine price pressures due to mutual fund redemptions. Less directly, Campbell, Grossman, and Wang (1993) study how price pressures leads to stocks' return autocorrelations declining with trading volume. Gabaix, Gopikrishnan, Plerou, and Stanley (2006) examine how price pressures by institutional investors can theoretically affect stock market volatility. Brunnermeier and Pedersen (2009) propose a theory in which liquidity spirals could lead to excessive price pressures. Degryse, de Jong, van Ravenswaaij, and Wuyts (2005) and Large (2007) show transitory price effects in limit order markets and label it order book resiliency. Nagel (2012) uses a return reversal measure to study the withdrawal of liquidity supply during financial turmoil.

Table 1
Summary statistics.
This table presents summary statistics on the data set that merges data from the Center for Research in Security Prices (CRSP), the NYSE's Trade and Quote (TAQ), and a proprietary NYSE data set. It is a balanced panel that contains daily observations on 697 NYSE common stocks from January 1994 through December 2005. Stocks are sorted into quintiles based on the average market capitalization, where Quintile 1 contains the largest-cap stocks (i.e., a stock is in the same quintile for all years in the sample).

Variable	Description (units)	Source	Mean Q1	Mean Q2	Mean Q3	Mean Q4	Mean Q5	Standard deviation within[a]
$midquote_{it}$	Closing midquote, dividend and split adjusted[b] (dollars)	NYSE	53.76	44.54	36.65	28.58	19.21	22.21
$invent\_share\ s_{it}$	Specialist inventory at the close (thousands of shares)	NYSE	8.19	5.55	4.33	3.23	5.39	34.19
$invent\_dollar_{it}$	Specialist inventory at the close[b] (thousands of dollars)	NYSE, CRSP	412.61	168.98	129.48	75.44	77.90	1,383.32
$shares\_out\ st_{it}$	Shares outstanding (millions)	CRSP	729.94	157.73	70.08	36.27	18.73	283.85
$market\_cap_{it}$	Shares outstanding times price (billions of dollars)	CRSP	34.29	5.34	2.06	0.88	0.29	11.57
$espread_{it}$	Share-volume-weighted effective half-spread (basis points)	TAQ	8.41	12.46	16.50	24.60	46.12	24.20
$dollar\_volume_{it}$	Average daily volume (millions of dollars)	TAQ	88.21	23.44	10.13	3.63	0.99	42.31
$specialist\_particip_{it}$	Specialist participation rate (percent)	NYSE	12.31	12.73	14.10	16.58	20.87	8.30
Number of observations: 697*3,018 (stock*day)								

[a] Denotes that the analysis is based on the deviations from time means, i.e., $x^*_{it} = x_{it} - \bar{x}_i$.
[b] Denotes that all price series were adjusted to account for stock splits and dividends.

the inventory variable allowing its use as an explanatory variable for the transitory price effect in the econometric model. Throughout we use "specialist", "market maker", and "intermediary" interchangeably to refer to the NYSE specialist.

Table 1 presents the mean of various trading variables by size quintile. It includes the within variation (defined as the standard deviation of the data series after removing stock fixed effects) to provide a sense of the variable's variability through time. Several observations emerge from the statistics. First, the within standard deviation in intermediary inventory is \$1.4 million, which is substantial relative to her average position. It suggests that specialists are active intermediaries in matching buyers and sellers at interday horizons. Second, while not the focus of this study, the average position of intermediaries is positive and economically significant.[3] For example, for the large-cap stocks in Q1 they maintain an almost half a million dollar average inventory position. The inventory position for the small-cap stocks in Q5 is considerably smaller at \$77,900. Third, the market capitalization is \$34 billion for Q1 stocks and declines to \$290 million for Q5 stocks. The effective spread (more precisely the half-spread, but for notational convenience we use "spread" throughout), which is defined as the distance between the transaction price and the prevailing midquote price, is 8 basis points for Q1 stocks and 46 basis points for Q5 stocks, demonstrating considerable heterogeneity across stocks.

2.2. State space model to disentangle price pressures and efficient price innovations

One challenge in identifying price pressures is that investors' net order imbalance, the difference between the volume investors buy and sell (which equals the intermediaries' inventory change), can convey information as well as cause pressure that makes prices overshoot. This well-known pattern has been shown in various event studies. For example, Kraus and Stoll (1972) find that prices overshoot in the event of a block trade, i.e., a transaction in which the initiating party wants to trade a large number of shares.

To identify the price pressure effect in the presence of an information effect, we build on the state space approach of Menkveld, Koopman, and Lucas (2007), which models an observed, high-frequency price series as the sum of two unobserved series: a nonstationary efficient price series (information) and a stationary series (pricing error) that captures transitory price effects including price pressures. We use log prices throughout the paper. In its simplest form, the model structure for the log price is

$$p_t = m_t + s_t \quad (1)$$

and

$$m_t = m_{t-1} + \delta_t + w_t, \quad (2)$$

where p_t is the observed price, m_t is the unobserved efficient price, s_t is the unobserved pricing error, δ_t is the required return obtained as the (time-varying) risk-free rate plus β_i times a 6% risk premium, and w_t is the innovation in the efficient price. s_t and w_t are assumed to be mutually uncorrelated and normally distributed. It is immediate from the structure of the model that only draws on w_t affect the security's price permanently as any draw on s_t is temporary because it affects prices only at a particular time. The model can be estimated with maximum likelihood, where the likelihood is constructed using the Kalman filter. Details on the implementation are in Appendix A.

The remainder of this subsection develops the general state space model that is taken to the data.[4]

[3] This average long position is likely driven by a combination of a cost asymmetry between a long and a short position and capital requirements imposed by the exchange that intermediaries choose to invest in stocks. The model in Section 3 interprets the zero position as the deviation from the long-term optimal position of the intermediary.

[4] The state space model developed has since been used in Menkveld (2013), who studies trading by a single large high-frequency trader, by Brogaard, Hendershott, and Riordan (2014), who study trading by

2.2.1. Observed price process

The observation equation in the state space model is

$$p_{it} = m_{it} + s_{it}, \tag{3}$$

where i indexes stocks, t indexes time, p_{it} is the observed log price, m_{it} is the efficient price, and s_{it} is the stationary pricing error. The last two terms each represent a latent state process.

2.2.2. Unobserved efficient price process

We use the model to analyze daily midquote price series by stock-year. The efficient price series is a martingale that consists of two components:

$$m_{it} = m_{i,t-1} + \delta_{it} + \beta_i f_t + w_{it}, \quad w_{it} = \kappa_i \hat{I}_{it} + u_{it}, \tag{4}$$

where the subscript i indexes stocks, t indexes days, δ_t is the required return obtained as the (time-varying) risk-free rate plus beta times a 6% risk premium, f_t is the demeaned market return (CRSP value-weighted return), w_{it} is the idiosyncratic innovation, $\hat{I}_{it}$ is the idiosyncratic inventory innovation obtained as the residual of an AR(p) model ($p=9$ based on the Akaike information criterion) that represents the surprise net order imbalance, which is potentially informative, and u_{it} is the stock-specific innovation orthogonal to this order imbalance innovation and assumed to be a normally distributed white noise process.

The decomposition of an efficient price innovation into a common factor component ($\beta_i f_t$) and an idiosyncratic component (w_{it}) is relevant for our purposes as the latter represents undiversifiable risk for the intermediary. The common-factor risk is easily hedged through highly liquid index products. For the same hedging reason, we remove the common factor from inventory dynamics.

Finally, several econometric issues deserve discussion. First, in the absence of trade information, a canonical microstructure model cannot be estimated as it is not econometrically identified (see, e.g., George and Hwang, 2001). The reason is that the efficient price innovation is likely correlated with the pricing error; i.e., the order imbalance an intermediary absorbs might contain information (affecting the efficient price change) as well as command a price pressure (affecting the pricing error). Incorporating the intermediary's inventory econometrically identifies the permanent and transitory components of price changes in the state space model. Hendershott, Jones, and Menkveld (2013) discuss how price pressures can be identified without the intermediary's inventory if the pricing error follows an AR(2) process. A related point is that the inventory change term in the martingale equation eliminates a potential omitted variable bias, as it is correlated with the inventory term that is assumed to affect the pricing error.

(*footnote continued*)
high-frequency traders as a group at Nasdaq, and by Hendershott, Jones, and Menkveld (2013), who use it to analyze the implementation shortfall on institutional trades.

2.2.3. Unobserved pricing error process

The following process for the stationary pricing error incorporates the intermediary's inventory:

$$s_{it} = \alpha_i I_{it} + \beta_i^0 f_t + \cdots + \beta_i^k f_{t-k} + \varepsilon_{it}, \tag{5}$$

where the error term ε_{it} is normally distributed and uncorrelated with u_{it}. Inventory enters as an explanatory factor to allow price pressure to originate from the intermediary's desire to mean-revert inventory (Section 3 provides a simple dynamic model generating such a linear structure). The f terms enter to capture a documented lagged adjustment to common factor innovation that is particularly prevalent for less actively traded small stocks (see, e.g., Campbell, Lo, and MacKinlay, 1997, p. 76). In the proposed specification, the beta coefficient in the efficient price process captures the long-term impact of a common factor shock on the price of the security and any lagged adjustment shows up through negative beta coefficients in the pricing error equation.

2.3. Preliminary evidence on price pressures

All empirical analyses are done by stock-year to characterize price pressures in the cross-section as well as in the time dimension. Quintile bins aggregate the 697×12 stock-year results according to market capitalization over the full sample. Therefore, each quintile contains the same set of stocks in each year of the sample. Means are calculated for each size-year bin separately and also by pooling all size-bins across years. These quintile-averages are reported in the "all" column of the tables and t-statistics for these are based on stock and year clustered standard errors (cf. Petersen, 2009).

Before turning to the state space model estimates, some straightforward time series statistics are useful to examine for evidence of transitory price effects in the data. Panels A and B of Table 2 report first- and second-order autocorrelation of idiosyncratic midquote returns. The effects of contemporaneous and lagged adjustment to the common factor innovation are removed by regressing the midquote return on the common factor innovation up to four lags. The residuals serve as the idiosyncratic returns. Consistent with the individual stock autocorrelation results in Campbell, Lo, and MacKinlay (1997, p. 72) the average first-order autocorrelation is negative in 65 of the 70 size-year bins. The t-statistics calculated by quintile and across all quantiles are all significant at conventional levels. The negative first-order autocorrelation is consistent with pricing errors as the simple state space model described in Eqs. (1) and (2) implies a negative first-order correlation in midquote returns. The table further shows that the second-order autocorrelation is significantly negative, which indicates that price pressures could carry over days instead of being only an intraday phenomenon. This also implies that the unconditional transitory deviations are potentially much larger relative to fundamental volatility than the simple first-order autocorrelations suggest.[5]

[5] If the stationary term follows an AR(1) process, $s_t = \varphi s_{t-1} + \varepsilon_t$, the first-order autocorrelation in midquote returns is

$$\rho_1 = \frac{-(1-\varphi)\sigma_\varepsilon^2}{(1+\varphi)\sigma_w^2 + 2\sigma_\varepsilon^2},$$

Table 2
Descriptive statistics midquote price and inventory dynamics.

This table presents descriptive statistics on daily log midquote price change and intermediary inventory. It is meant to illustrate these series' dynamics and their interaction building up to the state space model of Table 3. A common factor has been removed from both series. For price changes, the series is regressed on market return and the residuals enter the analysis. For inventory, the series is regressed on a common factor that is defined as the cross-sectional mean each day of the standardized inventory series. The table reports average estimates for each stock each year by market-cap quintile and year. The *t*-statistic pertains to the "All" (column) parameter estimates and is based on a robust standard error obtained by stock and year clustering (see, e.g., Petersen, 2009).

Quintile	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	All	*t*-Statistic
Panel A: First order autocorrelation idiosyncratic log price change														
Q1	−0.03	−0.03	−0.04	−0.08	−0.02	−0.01	0.01	−0.00	−0.05	−0.01	−0.00	−0.02	−0.02	−3.35
Q2	−0.02	−0.02	−0.04	−0.06	−0.01	−0.02	−0.01	0.00	−0.05	−0.02	−0.02	−0.03	−0.02	−4.32
Q3	−0.00	−0.01	−0.03	−0.06	−0.02	−0.04	−0.03	−0.01	−0.06	−0.05	−0.03	−0.04	−0.03	−5.80
Q4	−0.02	−0.02	−0.03	−0.06	−0.01	−0.03	−0.05	−0.03	−0.08	−0.06	−0.08	−0.05	−0.04	−6.19
Q5	−0.05	−0.05	−0.02	−0.00	0.03	0.01	−0.02	0.01	−0.01	−0.02	−0.04	−0.02	−0.02	−2.03
All	−0.02	−0.03	−0.03	−0.05	−0.01	−0.02	−0.02	−0.01	−0.05	−0.03	−0.03	−0.03	−0.03	−6.66
Panel B: Second order autocorrelation idiosyncratic log price change														
Q1	−0.03	−0.05	−0.04	−0.02	−0.02	−0.02	−0.06	−0.05	−0.01	−0.01	−0.02	−0.01	−0.03	−6.35
Q2	−0.03	−0.03	−0.03	−0.01	−0.03	−0.01	−0.02	−0.02	−0.00	−0.02	−0.01	−0.02	−0.02	−5.91
Q3	−0.02	−0.02	−0.02	−0.02	−0.01	−0.00	−0.01	−0.02	0.01	−0.00	−0.02	−0.01	−0.01	−4.71
Q4	−0.01	−0.02	−0.02	−0.03	−0.02	−0.00	−0.01	−0.01	0.01	−0.01	0.01	−0.00	−0.01	−3.20
Q5	−0.01	−0.01	−0.01	−0.01	−0.00	0.00	−0.01	0.00	−0.00	−0.01	0.00	0.00	−0.00	−2.43
All	−0.02	−0.02	−0.02	−0.01	−0.02	−0.01	−0.02	−0.02	−0.00	−0.01	−0.01	−0.01	−0.01	−6.36
Panel C: Standard deviation of idiosyncratic component specialist inventory I_{it} (in thousands of dollars)														
Q1	691	968	813	964	1126	1336	1344	1489	1472	1122	1118	1128	1131	11.06
Q2	473	510	489	524	530	695	820	647	448	391	441	400	531	12.26
Q3	374	429	383	372	430	452	668	437	294	242	266	271	385	10.65
Q4	226	254	255	261	291	315	320	333	229	163	145	147	245	11.95
Q5	167	159	168	235	205	224	211	187	130	109	95	96	166	11.31
All	386	464	421	471	516	605	672	619	514	406	414	408	491	14.68
Panel D: First order autocorrelation idiosyncratic component specialist inventory I_{it}														
Q1	0.28	0.27	0.26	0.22	0.25	0.27	0.28	0.29	0.28	0.34	0.37	0.25	0.28	21.24
Q2	0.47	0.46	0.44	0.38	0.35	0.34	0.36	0.32	0.25	0.28	0.33	0.25	0.35	15.97
Q3	0.59	0.59	0.56	0.51	0.49	0.45	0.41	0.41	0.30	0.31	0.34	0.24	0.43	12.68
Q4	0.74	0.73	0.71	0.66	0.63	0.63	0.59	0.57	0.40	0.38	0.36	0.29	0.56	12.56
Q5	0.82	0.80	0.80	0.77	0.78	0.79	0.77	0.76	0.66	0.61	0.57	0.51	0.72	23.89
All	0.58	0.57	0.56	0.51	0.50	0.50	0.48	0.47	0.38	0.38	0.39	0.31	0.47	19.06
Panel E: Regression coef of log price change on lagged inventory, both idiosyncratic (in basis points per thousand of dollars)														
Q1	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.01	8.71
Q2	0.01	0.01	0.01	0.02	0.01	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.02	10.05
Q3	0.02	0.02	0.02	0.02	0.02	0.03	0.04	0.01	0.05	0.03	0.02	0.03	0.03	7.49
Q4	0.03	0.02	0.04	0.05	0.04	0.06	0.07	0.04	0.11	0.06	0.07	0.06	0.05	7.43
Q5	0.08	0.07	0.03	0.03	0.07	0.06	0.08	0.08	0.15	0.09	0.10	0.13	0.08	6.63
All	0.03	0.03	0.02	0.02	0.03	0.03	0.05	0.03	0.07	0.04	0.04	0.05	0.04	9.30

Panels C, D, and E of Table 2 report the standard deviation of intermediaries' inventory, its autocorrelation, and a cross-correlation with subsequent midquote returns as preliminary evidence on the conjectured relation between pricing errors and inventory. The standard deviation of intermediary end-of-day inventory is $1.131 million for the large-cap stocks and monotonically decreases to $165,000 for the small-cap stocks. It is relatively constant throughout time but tapers off in the last few years in the sample. The cross-sectional variation is undoubtedly due to a higher fundamental volatility and a smaller less active market for small-cap stocks. In this case, the intermediaries shy away from frequent and large nonzero inventory positions, as in the model in Section 3.

The inventory volatility is roughly twice the mean inventory across all quintiles. This is evidence of active intermediation across days as inventory management is not a strictly intraday phenomenon in which intermediaries go home flat. The table further shows a significant first-order autocorrelation in inventories, showing that these positions can last for multiple days. Again, considerable cross-sectional variation exists as the average autocorrelation for the large-cap stocks is 0.28, which monotonically increases to an average autocorrelation of 0.72 for small-cap stocks, corresponding to inventory half-lives of 0.54 days for the largest stocks and 2.11 days for the smallest stocks. The intermediary seems to trade out of most of an end-of-day position in the course of the next day for the large-cap stocks, whereas it takes multiple days for the small-cap stocks. These inventory autocorrelations are lower than the puzzlingly large autocorrelations found in NYSE data from the late 1980s and early 1990s in Madhavan and Smidt (1993) and Hasbrouck and Sofianos (1993).

(*footnote continued*)
where φ is the autoregressive coefficient in the s_t process. The more persistent the price pressure is (the higher φ is), the less negative the return autocorrelation becomes.

Panel E reveals that today's inventory position correlates significantly with tomorrow's midquote return. The positive sign is consistent with models of intermediary risk management. The intermediary benefits from negative price pressure on a long position (relative to the long-term average) as it elicits net public buying (more investor buying than investor selling). The result is that her inventory is expected to become smaller and this, in turn, results in less negative price pressure. This channel creates a positive correlation between today's position and tomorrow's midquote return.

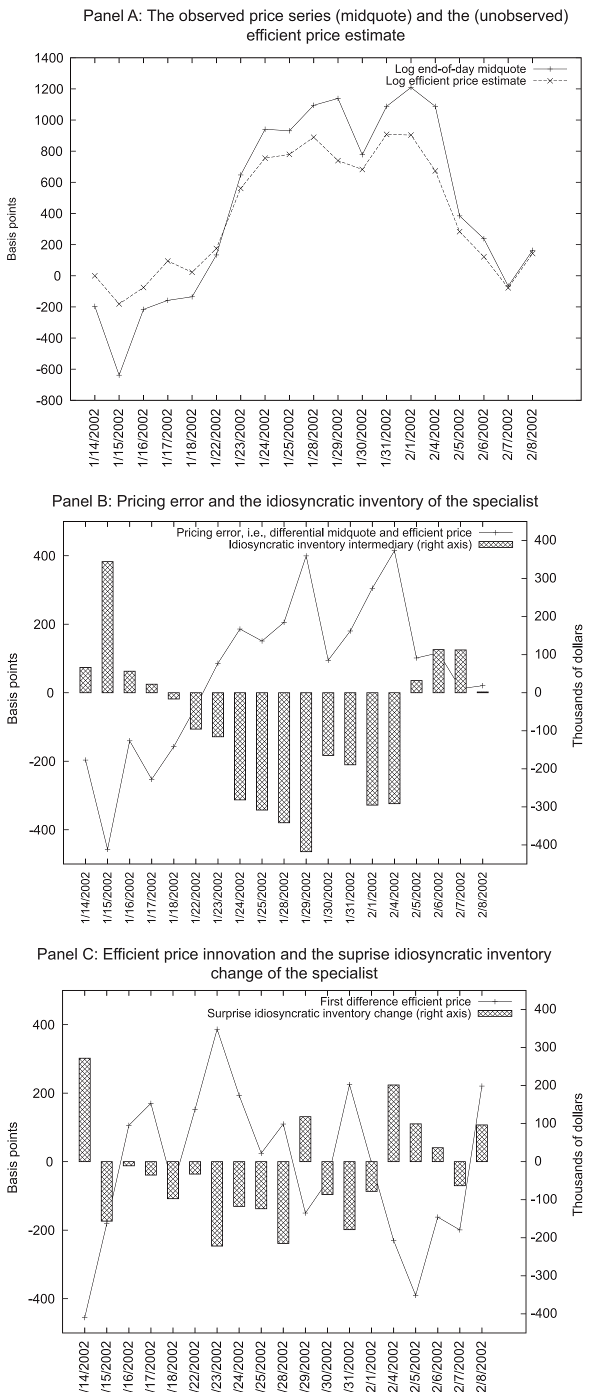

Fig. 1. Price change decomposition conditional on the state space model estimates, not using inventory data. These graphs depict the observed end-of-day log (midquote) price and the unobserved efficient price estimate for a representative stock (Rex Stores Corporation, ticker RSC, Center for Research in Securities Prices PERMNO 68830). The efficient price estimate is calculated with the Kalman smoother (and thus takes all observations into account) after estimating the following state space model:

(observed midquote price) $p_{it} = m_{it} + s_{it}$,

(unobserved efficient price) $m_{it} = m_{i,t-1} + \delta_{i,t} + \beta_i f_t + w_{it}$,

and

(unobserved transitory price deviation)

$$s_{it} = \varphi_i s_{i,t-1} + \beta_i^0 f_t + \cdots + \beta_i^3 f_{t-3} + \varepsilon_{it},$$

where i indexes stocks, t indexes days, m_{it} is the end-of-day unobserved efficient price, s_{it} is the pricing error, δ_t is the required return obtained as the (time-varying) risk-free rate plus beta times a 6% risk premium, f_t is the demeaned market return, p_{it} is the end-of-day observed midquote, β_i^j captures potential lagged adjustment to market returns, and w_{it} and ε_{it} are mutually independent and identically distributed error terms. The model is estimated using maximum likelihood, in which the error terms w_{it} and ε_{it} are assumed to be normally distributed. Panel A presents these series recentered around the first day's estimate of the efficient price. Panel B shows the pricing error, the difference between the observed price and the efficient price, against the idiosyncratic inventory position of the specialist. Panel C depicts the efficient price innovation against the contemporaneous (surprise) idiosyncratic inventory innovation.

2.4. State space model estimates

Before presenting the results of the general state space model defined by Eqs. (3)–(5), we graphically illustrate the idea of the model. The equations relate the permanent price change Δm_t and the pricing error s_t to the intermediary's inventory series. Initially, a reduced version of the model does not utilize inventory data:

$$p_{it} = m_{it} + s_{it}, \quad (6)$$

$$m_{it} = m_{i,t-1} + \beta_i f_t + w_{it}, \quad (7)$$

and

$$s_{it} = \varphi_i s_{i,t-1} + \beta_i^0 f_t + \cdots + \beta_i^3 f_{t-3} + \varepsilon_{it}, \quad (8)$$

where the AR(1) process for s_{it} allows pricing errors to be persistent as suggested by the negative second-order auto correlations in returns (Panel B of Table 2) and the inventory persistence (Panel D of Table 2).

Fig. 1 illustrates the model estimates for 20 trading days in a representative stock (Rex Stores Corporation, ticker RSC, CRSP PERMNO 68830) starting January 14, 2002. It exploits one attractive feature of the state space approach, which is that conditional on the model's parameter estimates the Kalman smoother generates estimates of the unobserved efficient price and the pricing error processes conditional on all observations. In other words, it uses past prices, the current price, and future prices to estimate the efficient price m_{it} and the pricing error s_{it} at any time t in the sample.

The figure leads to a couple of observations. Panel A plots the observed midquote price and the efficient price estimate. The graph illustrates that price deviations from fundamental value can be economically large (hundreds of

Table 3
State space model estimates.

This table presents estimates of the following state space model for a latent efficient price and an observed end-of-day midquote price:

(observed midquote price) $p_{it} = m_{it} + s_{it}$,

(unobserved efficient price) $m_{it} = m_{i,t-1} + \delta_{i,t} + \beta_i f_t + w_{it}, \quad w_{it} = \kappa_i \hat{I}_{it} + u_{it}$,

and (unobserved transitory price deviation) $s_{it} = \alpha_i I_{it} + \beta_i^0 f_t + \cdots + \beta_i^3 f_{t-3} + \varepsilon_{it}$,

where i indexes stocks, t indexes days, m_{it} is the end-of-day unobserved efficient price, s_{it} is the pricing error, δ_t is the required return obtained as the (time-varying) risk-free rate plus beta times a 6% risk premium, f_t is the demeaned market return, p_{it} is the end-of-day observed midquote, I_{it} is the idiosyncratic component of the specialist end-of-day inventory position, $\hat{I}_{it}$ is the (surprise) innovation in this position as obtained from an AR(9) model, β_i^j captures potential lagged adjustment to market returns, and w_{it} and ε_{it} are mutually independent and identically distributed error terms. The model is estimated using maximum likelihood, in which the error terms w_{it} and ε_{it} are assumed to be normally distributed. The table reports average estimates for each stock each year by market-cap quintile and year. The t-statistic pertains to the "All" (column) parameter estimates and is based on a robust standard error obtained by stock and year clustering (see, e.g., Petersen, 2009). Some parameter estimates are omitted from the main text but available in the online appendix.

Quintile	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	All	t-Statistic
Panel A: α_i *conditional price pressure (in basis points per thousand of dollars)*														
Q1	−0.03	−0.02	−0.02	−0.02	−0.02	−0.02	−0.02	−0.02	−0.02	−0.01	−0.01	−0.01	−0.02	−11.03
Q2	−0.06	−0.05	−0.05	−0.04	−0.03	−0.04	−0.04	−0.05	−0.04	−0.03	−0.03	−0.03	−0.04	−8.92
Q3	−0.12	−0.10	−0.09	−0.07	−0.09	−0.11	−0.10	−0.06	−0.09	−0.06	−0.06	−0.06	−0.08	−10.78
Q4	−0.33	−0.26	−0.29	−0.27	−0.26	−0.29	−0.25	−0.23	−0.26	−0.14	−0.18	−0.15	−0.24	−10.49
Q5	−1.05	−1.04	−0.76	−0.76	−1.02	−0.98	−1.06	−1.24	−1.19	−0.86	−0.93	−0.87	−0.98	−9.95
All	−0.32	−0.29	−0.24	−0.23	−0.28	−0.29	−0.30	−0.32	−0.32	−0.22	−0.24	−0.23	−0.27	−11.01
Panel B: $\lvert\alpha_i\rvert\sigma(I)_i$ *explained transitory volatility, i.e., unconditional price pressure (in basis points)*														
Q1	14	13	13	16	19	20	25	30	22	15	14	8	17	
Q2	26	24	21	19	20	30	35	31	24	14	15	13	23	
Q3	39	40	35	31	39	45	44	35	23	16	19	17	32	
Q4	65	64	68	57	63	75	75	65	40	21	23	20	53	
Q5	112	108	113	107	151	170	179	154	107	92	70	54	118	
All	51	50	50	46	58	68	72	63	43	31	28	22	49	
Panel C: κ_i *informativeness order imbalance innovation* $\hat{I}$ *(in basis points per thousands of dollars)*														
Q1	−0.08	−0.07	−0.07	−0.07	−0.08	−0.08	−0.08	−0.07	−0.07	−0.06	−0.05	−0.05	−0.07	−18.01
Q2	−0.16	−0.13	−0.13	−0.14	−0.17	−0.16	−0.17	−0.14	−0.16	−0.15	−0.13	−0.13	−0.15	−20.01
Q3	−0.22	−0.18	−0.17	−0.21	−0.25	−0.24	−0.24	−0.24	−0.25	−0.25	−0.25	−0.22	−0.23	−24.22
Q4	−0.34	−0.27	−0.29	−0.36	−0.43	−0.40	−0.45	−0.42	−0.52	−0.53	−0.49	−0.48	−0.42	−13.81
Q5	−0.58	−0.42	−0.61	−0.61	−0.69	−0.58	−0.76	−0.87	−1.20	−1.37	−1.21	−1.33	−0.85	−8.63
All	−0.27	−0.21	−0.26	−0.28	−0.32	−0.29	−0.34	−0.35	−0.44	−0.47	−0.42	−0.44	−0.34	−12.82
Panel D: $\lvert\kappa_i\rvert\sigma(\hat{I})_i$ *explained permanent volatility (in basis points)*														
Q1	38	36	38	44	57	69	75	55	55	36	29	26	47	
Q2	41	38	40	45	64	75	83	61	56	43	35	34	51	
Q3	43	40	41	54	70	74	79	74	58	47	40	34	54	
Q4	47	39	42	52	70	70	81	73	65	57	46	41	57	
Q5	43	38	49	55	67	65	77	90	79	74	62	59	63	
All	42	38	42	50	66	71	79	70	63	52	42	39	54	

Panel E: $\sigma(w)_i$ *permanent volatility (in basis points)*													
Q1	124	124	129	135	176	202	256	194	188	127	107	103	156
Q2	141	135	138	143	185	217	269	213	206	150	125	123	170
Q3	149	141	145	156	201	216	257	212	204	151	136	133	175
Q4	161	151	153	156	198	215	242	214	199	161	143	148	178
Q5	174	168	179	179	225	235	263	247	243	200	175	186	206
All	150	144	149	154	197	217	257	216	208	158	137	139	177
Panel F: $\frac{\alpha_i^2\sigma^2(I)_i}{\sigma^2(w)_i+\beta_i^2\sigma^2(\hat{f})}$ *ratio of price pressure variance and permanent variance*													
Q1	0.02	0.02	0.02	0.01	0.01	0.02	0.01	0.03	0.02	0.02	0.03	0.01	0.02
Q2	0.08	0.09	0.04	0.03	0.02	0.03	0.03	0.03	0.02	0.01	0.02	0.01	0.03
Q3	0.29	0.25	0.16	0.08	0.11	0.16	0.05	0.08	0.02	0.01	0.03	0.07	0.11
Q4	0.31	0.47	0.48	0.26	0.22	0.37	0.24	0.18	0.05	0.03	0.03	0.05	0.22
Q5	1.09	1.35	1.13	1.11	1.23	2.40	1.82	1.15	0.79	2.08	0.59	0.59	1.28
All	0.36	0.43	0.37	0.30	0.32	0.59	0.43	0.29	0.18	0.43	0.14	0.15	0.33
Panel G: $\frac{\alpha_i^2\sigma^2(I)_i}{\sigma^2(w)_i}$ *ratio of price pressure variance and permanent idiosyncratic variance*													
Q1	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.04	0.03	0.04	0.04	0.01	0.03
Q2	0.10	0.10	0.05	0.04	0.02	0.03	0.04	0.04	0.03	0.02	0.03	0.02	0.04
Q3	0.35	0.28	0.20	0.11	0.17	0.18	0.05	0.10	0.04	0.02	0.05	0.11	0.14
Q4	0.37	0.49	0.54	0.31	0.30	0.40	0.25	0.23	0.08	0.04	0.05	0.05	0.26
Q5	1.25	1.46	1.25	1.28	1.55	2.49	2.04	1.43	1.00	2.32	0.72	0.65	1.45
All	0.42	0.47	0.41	0.35	0.41	0.62	0.48	0.37	0.24	0.48	0.18	0.17	0.38
Panel H: $\sigma(\varepsilon)_i$ *error term transitory volatility (in basis points)*													
Q1	20	17	23	30	24	20	26	19	32	18	12	15	21
Q2	20	20	22	28	24	22	34	18	39	23	17	19	24
Q3	19	18	23	24	28	29	35	26	41	26	20	22	26
Q4	29	26	24	27	18	23	37	29	42	30	30	25	28
Q5	43	42	34	25	22	29	43	35	38	28	29	25	33
All	26	24	25	27	23	24	35	25	39	25	21	21	26

basis points) and persistent as they appear to last for multiple days. Panel B plots the differential between the observed price and the efficient price estimate, i.e., the price pressure, along with the intermediary's inventory deviation from its long-term mean. The clear negative correlation between the two series is consistent with the intermediary using price pressure to mean-revert her inventory toward its desired level. Panel C plots the innovation in efficient price Δm_t against the contemporaneous surprise idiosyncratic inventory change, which is obtained as the residual of an AR(9) model. It indicates that unexpected order flow is informative about the efficient price. A surprise positive inventory change indicates unexpected selling by liquidity demanders, which changes the efficient price downward as the selling could have been driven by information. These observations turn out to be generic features of the data in the general state space model with inventory data.

Table 3 reports the estimates of the general state space model defined by Eqs. (3)–(5). In Panel A, the key parameter α_i that measures the marginal price pressure has the conjectured negative sign and is statistically significant. Prices are temporarily low when the intermediary is on a long position and temporarily high when she is on a short position relative to her long-term mean inventory. Marginal price pressure exhibits substantial cross-sectional variation as α_i is -0.02 for the large-cap stocks and -0.98 for the small-cap stocks. These numbers are economically significant even in the large stocks as a $1.131 million (1 standard deviation) position change in intermediary inventory creates a price pressure of $1131 \times 0.02 = 17$ basis points (cf. effective half-spread of 8 basis points, see Table 1).

Panel B of Table 3 further shows that the average price pressure varies less in the cross-section than the marginal price pressure. The average pressure, which is measured as the marginal pressure times the standard deviation of inventory, is 17 basis points for the large-cap stocks and 118 basis points for the low-cap stocks. The average pressure is roughly seven times higher for the small-cap stocks relative to the large-cap stocks whereas the marginal pressure is 50 times higher with the difference attributable to the intermediary taking smaller positions in the smaller stocks. Panel F also reports the size of these average price pressures relative to permanent variance. The ratio ranges from 0.02 for the large-cap stocks to 1.28 for small-cap stocks. The average ratio is 0.33.

Transitory volatility is often measured using variance ratios of stock returns to capture deviations from a random walk. The 1-to-n variance ratio divides the n-period return variance by n times the one-period return variance. Campbell, Lo, and MacKinlay (1997, Table 2.7) report a 1-to-16 variance ratio for weekly returns for US equities of 0.85 for 1962–1994. The comparable ratio from the state space model's average price pressures relative to permanent variance is $0.88 = 1/[1+(2\times 0.33)/5]$. Campbell, Lo, and MacKinlay (1997) find little cross-sectional difference in the variance ratios, consistent with the return autocorrelations in Table 2. However, Panel F of Table 3 shows significant cross-sectional differences. This highlights the importance of data on intermediaries' inventory positions in identifying and measuring transitory volatility when inventories are persistent.

Consistent with earlier empirical work (for example, Hasbrouck, 1991b), unexpected inventory changes, which equal unexpected investor buying or selling, explain a significant part of efficient price innovations. The κ_i estimates in Panel C are -0.34 on average and statistically significant. In Panel D the average explained permanent volatility (κ_i times the standard deviation of unexpected inventory changes) is 54 basis points. This compares with an average total permanent volatility of 177 basis points in Panel E.

3. A simple dynamic inventory control model to interpret price pressure dynamics

This section models the intermediary's dynamic inventory control as a stochastic optimal linear regulator (SOLR) problem (see Ljungqvist and Sargent, 2004, p. 112). A representative intermediary faces stochastically arriving investors with elastic liquidity demands. The solution characterizes a stationary distribution and the transition probabilities for price pressures, which, along with the empirical results in Section 2, enables identification of the intermediary's relative risk aversion and liquidity demanders' private value distribution. It also allows for an analysis of the costs of trading and social welfare. Given that the model assumes that investors arrive asynchronously, first-best is not achievable. So, the social welfare analysis focuses on deviations from constrained Pareto efficiency. For expositional ease, we often abbreviate constrained Pareto efficiency to Pareto efficiency in the remainder of the paper.

The model is inspired by Ho and Stoll (1981), who frame dynamic inventory control in a standard macro-model of a constant relative risk aversion (CRRA) utility intermediary who controls the public buy rate and sell rate, which are linear in her ask and bid price quotes. To fit the SOLR approach we assume that liquidity demand, which determines the intermediary's buy and sell volume, is exogenous and normally distributed with a mean that is linear in the bid and ask price, respectively; the intermediary maximizes quadratic utility over nonstorable consumption[6]; a security position exposes the intermediary to fundamental value risk, which is modeled as a normally distributed stochastic dividend to avoid a notational burden (cf. Ho and Stoll, 1981, p. 52).

These assumptions have several implications. First, the nonstorable consumption removes the ability of the intermediary to precautionarily save to smooth consumption. Second, quadratic utility leads to risk aversion that increases with wealth. Third, the normality assumption for public buy volume and sell volume implies that they could become negative. To show that the model's predicted price pressure dynamics are robust the supplementary material numerically solves an infinite horizon Ho and Stoll model.

3.1. The environment

Time is discrete and infinite. One durable asset produces a perishable and stochastic (numéraire) consumption good.

[6] Cochrane (2001, p. 155) also considers quadratic utility a natural starting point when he introduces dynamic programming. Madhavan and Smidt (1993) model inventory as having quadratic holding costs. Lagos, Rocheteau, and Weill (2011) model liquidity provision in which intermediaries maximize over nonstorable consumption.

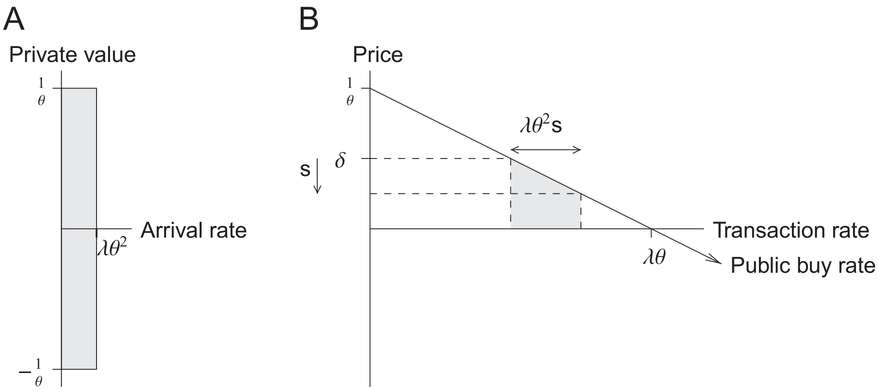

Fig. 2. Illustration of liquidity demand. Panel A: Mean demand for each level of private value. Panel B: Mean public buy volume as function of ask price.

An infinitely lived intermediary supplies liquidity by standing ready to buy or sell the asset to outside investors who demand liquidity. At the start of each period, a single intermediary quotes an ask price at which she commits to sell and a bid price at which she commits to buy.

3.1.1. *Reduced form liquidity demand*

Liquidity demand is modeled to be linear in price consistent with Ho and Stoll (1981). Its shape is characterized by two primitive parameters: λ, which is the aggregate amount of all investors' private values to trade per unit of time, and θ, which captures the dispersion of private values across investors. Investor buying and selling are determined by the intermediary's bid price, $s-\delta$, and ask price, $s+\delta$, which are characterized by the bid–ask spread of 2δ and price pressure of s, analogous to $s_t=p_t-m_t$ in the state space model. The analogy follows from the Ho and Stoll stochastic dividend trick, which makes m_t equal to zero in the model. Specifically, investor buying and selling demand for liquidity are normally distributed variables linear in the bid–ask quotes:

$$q_s(s,\delta)=\lambda\theta(1-\theta(s+\delta))+\varepsilon_s,\quad \varepsilon_s\sim \mathrm{N}\left(0,\frac{1}{2}\sigma_\varepsilon^2\right) \tag{9}$$

and

$$q_b(s,\delta)=\lambda\theta(1+\theta(s-\delta))+\varepsilon_b,\quad \varepsilon_b\sim \mathrm{N}\left(0,\frac{1}{2}\sigma_\varepsilon^2\right), \tag{10}$$

where ε_s and ε_b are independent of each other and identically and independently distributed (IID) each period. The environment is best illustrated by Fig. 2. Panel A illustrates liquidity demand distribution over private values, and Panel B depicts how such demand translates into an expected investor buy rate, which equals the intermediary sell rate, that is linear in the ask price.

3.1.2. *The intermediary's liquidity supply*

The intermediary is risk-averse and, therefore, dislikes the risky dividend associated with a nonzero inventory position. She solves the following infinite-horizon dynamic program:

$$v_{i_0}=\max_{\{s_t(i^t),\delta_t(i^t)\}_{t=0}^{\infty}}\sum_{t=0}^{\infty}\mathrm{E}_{i_0}\left(\beta^t\left((c_t-\bar{c})-\frac{1}{2}\tilde{\gamma}(c_t-\bar{c})^2\right)\right), \tag{11}$$

where i^t represents the history of her inventory position through time t and $\mathrm{E}_{i_0}(\cdot)$ is the expectation operator conditional on starting off with inventory position i_0. The quadratic utility parametrization is such that the Arrow-Pratt coefficient of (absolute) risk aversion (ARA) is $\tilde{\gamma}$ at $\bar{c}$, which is fixed ex ante at the competitive intermediary's average consumption. The intermediary's actual consumption in period t equals net trading revenue plus the stochastic dividend:

$$\begin{aligned}c_t&=(s_t+\delta_t)q_s(s_t,\delta_t)-(s_t-\delta_t)q_b(s_t,\delta_t)+i_t\Delta m_{t+1},\\&=2\lambda\theta(\delta_t-\theta(s_t^2+\delta_t^2))+s(\varepsilon_{st}-\varepsilon_{bt})+\delta(\varepsilon_{st}+\varepsilon_{bt})\\&\quad+i_t\Delta m_{t+1},\quad \Delta m_{t+1}\sim\mathrm{N}(0,\sigma^2),\end{aligned} \tag{12}$$

where Δm_{t+1} is the stochastic dividend at the start of period $t+1$ (which runs from t to $t+1$). The conditional consumption mean and variance, therefore, are equal to

$$\mathrm{E}_t(c_t)=2\lambda\theta(\delta_t-\theta(s_t^2+\delta_t^2)) \tag{13}$$

and

$$\mathrm{var}_t(c_t)=\sigma_\varepsilon^2(s_t^2+\delta_t^2)+\sigma^2 i_t^2. \tag{14}$$

The law of motion for inventory follows from the net trade in the asset:

$$i_{t+1}=i_t-q_s(s_t,\delta_t)+q_b(s_t,\delta_t)=i_t+2\lambda\theta^2 s_t-\varepsilon_{st}+\varepsilon_{bt}. \tag{15}$$

A final step simplifies the problem in two ways. First, the expected utility expression is linearized in mean and variance by a first-order Taylor expansion around the conditional mean. This removes the quadratic difference between the conditional mean and $\bar{c}$ (the unconditional mean), which would make the objective function fourth-order and not possible to solve with standard techniques. Second, the impact of σ_ε^2 in the variance of consumption, Eq. (14), is omitted to focus on how prices reflect the dynamic trade-off between the intermediary's expected cost of a pressured price and its benefit of mean-reverting risky inventory. Omission of the consumption variance effect is innocuous if the utility cost of liquidity demand uncertainty (as captured by $\frac{1}{2}\tilde{\gamma}\sigma_\varepsilon^2$) is small relative to the expected revenue loss due to price pressure (as captured by $2\lambda\theta^2$). The final specification of the dynamic program is, therefore,

$$v_{i_0}=\max_{\{s_t(i^t),\delta_t(i^t)\}_{t=0}^{\infty}}\sum_{t=0}^{\infty}\beta^t\mathrm{E}_{i_0}\left(2\lambda\theta\delta_t-2\lambda\theta^2(s_t^2+\delta_t^2)-\frac{1}{2}\tilde{\gamma}\sigma^2 i_t^2\right) \tag{16}$$

and

$$i_{t+1} = i_t + 2\lambda\theta^2 s_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2). \tag{17}$$

3.2. The recursive form, the closed-form solution, and its characteristics

The IID character of net liquidity demand uncertainty (ε_t) and stochastic dividend (Δm_{t+1}) allows for a recursive formulation of the dynamic program. The intermediary solves the following Bellman equation:

$$v_i = \max_{\{s,\delta\}} E_{i_0}\left(2\lambda\theta\delta - 2\lambda\theta^2(s^2+\delta^2) - \frac{1}{2}\tilde{\gamma}\sigma^2 i^2 + \beta v_{i'}\right) \tag{18}$$

and

$$i' = i + 2\lambda\theta^2 s + \varepsilon. \tag{19}$$

From this point onward, one can apply well-known SOLR results (see Ljungqvist and Sargent, 2004, p. 112) to find the following solution:

$$v_i = \frac{\lambda}{2(1-\beta)} - P\left(\frac{i^2}{\beta} + \frac{1}{1-\beta}\sigma_\varepsilon^2\right), \tag{20}$$

$$s^* = \alpha i, \quad \alpha := \frac{-1}{\frac{1}{P} + Q}, \tag{21}$$

and

$$\delta^* = \frac{1}{2\theta}, \tag{22}$$

where (s^*, δ^*) denote the optimal price controls and

$$P := \frac{-(1-\beta) + \beta RQ + \sqrt{(1-\beta)^2 + 2\beta(1+\beta)QR + \beta^2 Q^2 R^2}}{2Q},$$

$$Q := 2\lambda\theta^2, \quad R := \frac{1}{2}\tilde{\gamma}\sigma^2. \tag{23}$$

The solution has various intuitive features. For example, the marginal price pressure α becomes larger in magnitude when either the intermediary is more risk-averse (higher $\tilde{\gamma}$) or when there is more (dividend) volatility (higher σ). The bid–ask spread (2δ) is wider when the private value distribution is more dispersed (lower θ). And, the state value v_i is highest when the intermediary is at a zero position ($i=0$). It declines monotonically the further away her position is from zero.

3.2.1. Time series properties of price pressure

The solution to the intermediary's control problem implies linear dynamics for price pressure (s) and inventory, consistent with the prior econometric model. More specifically, price pressure is linear in inventory [Eq. (21)] and inventory is characterized by a first-order autoregressive process:

$$i_t = \left(1 - \frac{1}{1+PQ}\right) i_{t-1} + \varepsilon_t = -Q\alpha i_{t-1} + \varepsilon_t. \tag{24}$$

3.2.2. Orthogonality of the spread and optimal inventory control

The intermediary's pricing strategy decomposes into two orthogonal parts. First, she sets a spread that exploits her monopolistic pricing power vis-à-vis the investors. Second, she uses price pressure (the midquote price) to optimally mean-revert her inventory. Mathematically, the control strategy orthogonalizes as the spread does not enter the speed of mean-reversion and the spread is, therefore, constant across inventory states.[7] This result critically depends on the assumption that public buy volume and sell volume are linear in prices in Eqs. (9) and (10).

3.3. Identification of the model's primitive parameters

The remainder of Section 3 uses the empirical results in Section 2 to identify the model's primitive parameters, calculates a model-implied competitive spread, and analyzes Pareto efficiency.

The unobserved primitive parameters to be identified are λ (arrival rate of investors), θ (private value distribution), and $\tilde{\gamma}$ (intermediary risk aversion). The primitive parameters that are observed are σ (the fundamental price risk) and β (the discount factor). The former is set equal to the state space estimate of $\sigma(w)$, which captures the size of efficient price innovations. β is assumed to be equal to the reciprocal of the gross risk-free rate obtained from Kenneth French's website.[8]

The identification strategy for the hidden parameters follows naturally from inspection of the model's solution. The empirical estimates of marginal pressure ($\hat{\alpha}$) along with the rate of inventory mean-reversion [$\hat{\rho}_I := \rho(I_t, I_{t-1})$] enables solving for Q and R, which in turn identify $\lambda\theta^2$ and $\tilde{\gamma}$, respectively. Eqs. (22) and (24) yield

$$\hat{\alpha} = \frac{-P}{1+PQ} \tag{25}$$

and

$$\hat{\rho}_I = \frac{PQ}{1+PQ}, \tag{26}$$

where P is a function of Q and R [see Eq. (23)]. To separate the factors λ and θ from the value of Q, estimates of the intermediated volume (*specialist_particip* × *daily_volume*) and the effective bid–ask spread (*espread*) are needed. An additional moment that is exploited is realized trading activity by the intermediary:

$$intermediated\_volume = \left(\frac{2}{\theta} - \frac{1}{\theta}\right)\lambda\theta^2 = \lambda\theta. \tag{27}$$

This follows from all investor who arrive trading except for those whose private value falls in the region of the bid–ask spread, which is $1/\theta$ wide. Finally, the model's coefficient for absolute risk aversion ($\tilde{\gamma}$) is converted to an estimate of the more standard relative risk aversion by multiplying it

[7] Zabel (1981) and Mildenstein and Schleef (1983) theoretically find that the spread is independent of inventory. Ho and Stoll (1981) finds that inventory has a very small effect on spread. Comerton-Forde, Hendershott, Jones, Seasholes, and Moulton (2010) empirically find that inventory positions affect the spread. A 1 standard deviation shock to market-wide inventories increases bid–ask spreads by roughly 0.20 basis points, which is substantially smaller than the 49 basis point average transitory volatility due to price pressure in Table 3.

[8] The web address is http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/.

Table 4
Model's primitive parameters and an analysis of the spread.

This table presents an economic analysis of the cost to liquidity demanders of a competitive intermediary who uses price pressure to control her inventory position. The model's primitive parameters consist of the liquidity demander's total private value rate λ and its dispersion $1/\theta$, uncertainty about net liquidity demand $\sigma(\varepsilon)$, a discount factor β, and the intermediary's coefficient of relative risk aversion (γ), which is the model's absolute risk aversion ($\tilde{\gamma}$) divided by a competitive intermediary's average consumption. The conditional price pressure estimates along with other trading variables measured in the empirical part of the paper (Panel A contains sample averages taken from Tables 1 and 3) allows for identification of the primitive parameters (Panel B), which in turn allow for a decomposition of the net cost to liquidity demanders (investors) of liquidity supply by a competitive intermediary (Panel C). It expresses this net cost as a fraction of transacted volume (spread) and decomposes it into a spread received by the intermediary and a subsidy enjoyed by the liquidity demanders at times when prices are pressured. It also calculates the Pareto efficient cost if the intermediary were to internalize the subsidy enjoyed by the liquidity demander. Panel B and C report medians and interquartile ranges ($Q_{0.75}-Q_{0.25}$, where Q_i is quantile i) in parentheses. The complete set of results (e.g., disaggregated by year) is added to the online appendix.

	Q1	Q2	Q3	Q4	Q5	All
Panel A: Measured variables that identify model's primitive parameters (cf. Tables 1 and 3)						
Conditional price pressure α_i (in basis points per thousand of dollars)	−0.02	−0.04	−0.09	−0.25	−0.99	−0.28
Standard deviation daily inventory $\sigma(I)$ (in thousands of dollars)	1131	530	385	245	166	491
First order autocorrelation inventory	0.28	0.35	0.43	0.56	0.72	0.47
Price risk inventory $\sigma(w)_i$ (in basis points)	153	170	176	179	207	177
Daily dollar volume[a] (in millions of dollars)	10.86	2.98	1.43	0.60	0.21	3.22
Effective half spread (in basis points)	8.41	12.46	16.50	24.60	46.12	21.62
Daily discount factor[b]	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
Panel B: Identification of model's primitive parameters						
Daily private value rate λ (in thousands of dollars)	58.91 (236.50)	14.43 (36.51)	6.29 (12.80)	2.83 (4.76)	1.45 (2.28)	5.90 (21.91)
Dispersion private value $\frac{1}{\theta}$ (in basis points)	174 (297)	133 (197)	146 (164)	182 (197)	307 (362)	185 (256)
Intermediary's relative risk aversion γ	3.05 (7.90)	3.79 (8.91)	4.02 (9.30)	4.49 (12.61(	4.46 (9.52)	3.96 (9.48)
Panel C: Decomposition of the spread paid by liquidity demander						
(1) Model-implied competitive half spread (in basis points)	7.63 (17.20)	14.32 (30.03)	21.52 (44.83)	42.30 (81.57)	87.41 (178.75)	25.18 (66.49)
(2) Price pressure subsidy to liquidity demander (in basis points)	0.56 (1.11)	1.40 (3.05)	2.54 (5.62)	5.86 (13.32)	15.13 (33.03)	2.82 (9.38)
(3) Net half spread to liquidity demander[c] (1)–(2) (in basis points)	6.97 (15.69)	12.47 (26.13)	19.09 (39.77)	36.34 (69.13)	72.21 (145.93)	22.07 (56.59)
(4) Constrained Pareto efficient half spread (in basis points)	6.20 (14.09)	10.76 (22.14)	16.01 (33.10)	29.28 (54.24)	54.48 (108.47)	18.41 (44.68)
(5) Deadweight loss[c] (3)–(4) (in basis points)	0.61 (1.19)	1.56 (3.41)	2.83 (6.35)	6.58 (15.58)	17.66 (38.62)	3.16 (10.88)

[a] Denotes $dollar\_volume_{it} * specialist\_particip_{it}$ as a proxy for volume intermediated by the specialist.
[b] Denotes that discount factor equals the reciprocal of the gross risk-free rate from Kenneth French's website (http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/).
[c] Denotes that the reported difference is the median of differences, not the difference of (reported) medians.

by a proxy for intermediary wealth consistent with the model. Appendix B provides the expressions for the calibrated parameters.

Table 4 summarizes the results. Panel A repeats the empirical results used for identification of the model's primitive parameters. Panels B and C of Table 4 report median estimates for the overall sample and by quintile. Estimates by quintile and year are available in the online appendix. Panels B and C of Table 4 report interquartile ranges ($Q_{0.75}-Q_{0.25}$) as the model-implied variables are nonlinear functions of the empirical estimates (with estimates in denominators), which creates numerical issues. Medians and interquartile ranges are robust to these issues.

The private value rate λ is highest for the largest-cap stocks: $58,910 per stock per day. It decreases monotonically to $1,450 for the smallest-cap stocks. Private value dispersion $1/\theta$ is 174 basis points for the largest-cap stocks, is 133 for the second largest-cap stocks, and increases monotonically to 307 basis points for the smallest-cap stocks. These private value dispersions are somewhat smaller than the 21% standard deviation that Hollifield, Miller, Sandås, and Slive (2006, p. 2783) report based on structural estimation of a limit order book model. The 21% is the unconditional standard deviation implied by a mixture of two normal distributions, with approximately 85% weight on a distribution with a standard deviation of approximately 5% and a 15% weight on a distribution with a standard deviation of approximately 54%.

The median relative risk aversion is 3.05 for the largest-cap stocks and generally increases to 4.46 for the smallest-cap stocks. The differences across quintiles are small relative to the interquartile ranges within each quintile. The overall median is 3.96. It is somewhat low relative to the relative risk aversion estimated from asset, insurance, and labor markets (e.g., Mehra and Prescott, 1985; Barsky, Juster, Kimball, and Shapiro, 1997; Chetty, 2006; Cohen and Einav, 2007). The lower risk aversion could arise from risk tolerant capital migrating to the intermediation sector or be a sign of an agency conflict between firm and the individual traders it employs to act as intermediaries. If the traders have limited liability, then they could be willing to take more risky bets than the firm's owners would prefer. The structural model also could underestimate true risk aversion as the model ignores increased net demand uncertainty associated with price pressure (see related discussion in Section 3.1). Finally, the risk aversion might

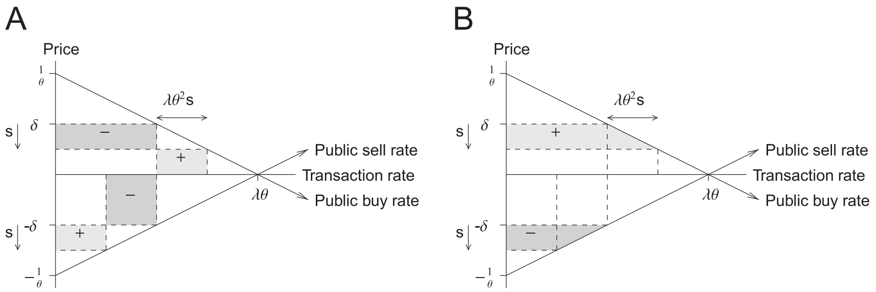

Fig. 3. Deadweight loss. Panel A: Intermediary's cost of price pressure. Panel B: Liquidity demander's benefit of price pressure.

be underestimated as the analysis assumes that the intermediary cannot hedge any of the nonsystematic risk on her position. In practice, hedging could be possible, for example, by taking an offsetting position in stocks in the same industry.

3.4. Model-implied reservation spread

Section 3.2 shows that the intermediary's optimal control policy consists of two orthogonal decisions: one with respect to the size of the bid–ask spread and the other with respect to the amount of price pressure to apply. This allows for straightforward calculation of a reservation spread, which is the narrowest spread satisfying the intermediary's participation constraint, i.e., the spread that satisfies an intermediary's zero-profit condition.[9]

The reservation spread can be calculated based on the primitive parameters identified in Section 3.3 as the annuity value of the cost of inventory risk identified as $P\sigma_\varepsilon^2$ in Eq. (20). The model-implied marginal price pressure and inventory mean-reversion as described by Eqs. (21) and (24) identify the factor P. The remaining factor σ_ε^2 follows directly from inventory variance and mean-reversion, i.e., $\sigma_\varepsilon^2 = (1-\rho_I^2)\sigma^2(I)$, where ρ_I is the first-order autocorrelation of inventory. This daily value divided by the intermediated daily volume naturally defines the net spread. The implied reservation spread is calculated by adding back the annuity value of the price pressure subsidy that the intermediary pays to mean-revert inventory. Its value is calculated in Appendix C. Panel C of Table 4 reveals that the median model-implied reservation spread is 6.97 basis points for the largest-cap stocks and monotonically increases to 72.21 basis points for the smallest-cap stocks. The panel refers to it as net spread as it is the average net spread paid by liquidity demanders.

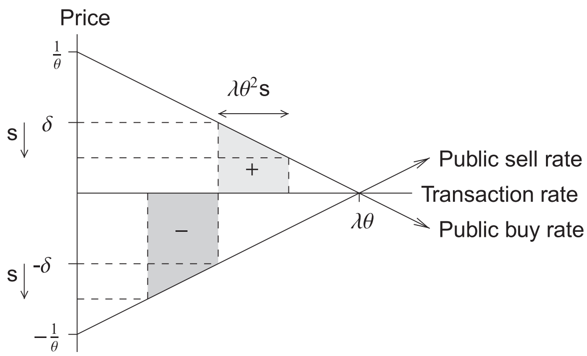

Fig. 4. Social cost of price pressure.

3.5. Deadweight loss due to Pareto inefficiency

The intermediary trades off the expected cost of price pressure against the benefit of inventory reduction. Panel A of Fig. 3 depicts the expected cost of a negative price pressure s associated with a positive inventory. The ask price is lowered and the light gray area above the horizontal axis indicates the additional expected revenue due to more public buying. The lowered price also reduces the margin on all buys, which is indicated by an expected loss equal to the dark gray area above the axis. Similar areas are drawn for the lowered bid price. Overall, the differential between the light and dark gray areas is the expected revenue decline due to price pressure. The size of this area equals the $2\lambda\theta^2 s^2$ cost in the intermediary's Bellman equation [see Eq. (18)].

Panel B shows that liquidity demanders benefit because the discount they enjoy when buying exceeds the lower price they receive when selling. The difference equals half the intermediary's cost of price pressure: $\lambda\theta^2 s^2$. The daily annuity value of this subsidy is obtained recursively (see Appendix C) and, when divided by daily intermediated volume, equals 0.56 basis points for the largest-cap stocks (7% of their reservation spread) and 15.13 basis points (17% of their reservation spread) for the smallest-cap stocks in Panel C of Table 4.

The model-implied competitive spread is the sum of the subsidy paid out by the intermediary and the reservation spread. Panel C of Table 4 finds that it is 7.63 basis points for

[9] The reservation spread could be competitively implemented by assigning the right to be the intermediary based on a second-price auction among candidate intermediaries who bid based on a bid–ask spread commitment. In such an auction, all candidate intermediaries bid their reservation value, which equals the spread that exactly compensates them for the utility cost of the stochastic dividend stream, i.e., the second term in Eq. (20).

the largest-cap stocks and it increases monotonically to 87.41 basis points for the smallest-cap stocks. It is reassuring that the model-implied spread is of the same magnitude as the observed spread. It is slightly below for the largest-cap stocks (7.63 versus 8.41 basis points) and increases to roughly 50% higher for the smallest-cap stocks (88.79 versus 46.12 basis points). The correspondence is remarkable given that the model-implied spread does not use the spread observed in the data but is identified only from the time series properties of price pressure. In addition, the cross-sectional evidence of an observed spread that is too high for large stocks and too low for small stocks is consistent with the specialist cross-subsidization practice in effect at the time (see Cao, Choe, and Hatheway, 1997).

The intermediary only experiences the cost of lost revenue due to price pressure and does not internalize the benefit it creates for liquidity demanders, which, most likely, makes the policy constrained Pareto inefficient. The adjective "constrained" is used here to emphasize that the first-best of synchronous arrivals eliminating the need for intermediation is not attained. A social planner might Pareto improve by making the intermediary suffer the net surplus destroyed by price pressure instead of her private revenue loss (i.e., the sum of Panel A and B, not Panel A). This true social cost is $\lambda\theta^2 s^2$ as depicted in Fig. 4. The social planner's dynamic program therefore equals Eqs. (16) and (17), where, in Eq. (16), the intermediary's cost of price pressure, $2\lambda\theta^2 s^2$, is replaced by its social cost equivalent: $\lambda\theta^2 s^2$. The solution equals the intermediary's solution except for replacing P in Eq. (23) by $\tilde{P}$:

$$\tilde{P} = \frac{-(1-\beta)+2\beta RQ+\sqrt{(1-\beta)^2+4\beta(1+\beta)QR+4\beta^2 Q^2 R^2}}{4Q}. \tag{28}$$

Panel C of Table 4 shows that the social planner's implied net spread is 6.20 basis points for the largest-cap stocks and increases to 54.48 basis points for the smallest-cap stocks. The gain relative to the net spread paid by the liquidity demander in the intermediary's solution (i.e., net of the subsidy) is 0.61 basis points for the largest-cap stocks and 17.66 basis points for the smallest-cap stocks. Overall, the differential is 3.16 basis points relative a net spread of 22.07 basis points, implying a social loss of approximately 10%.

4. Discussion of NYSE market structure

When interpreting our results it is worth discussing the institutional structure of the specialist intermediary at the NYSE. Historically, the NYSE granted the specialists an advantageous central position in the trading process and imposed obligations upon them. This special role could affect the interpretation of inventory's role in the price process. We first discuss the informational advantages and then other facets of the NYSE trading structure. Finally, we consider how changes in the NYSE market structure during our sample period relate to the results.

4.1. NYSE informational structure

An important privilege the specialist had in our sample period is a last-mover advantage, i.e., the specialist could decide on any incoming order to either trade against it herself by offering price improvement or to route it to the limit order book. In making this decision the specialist could condition on information about the identity of the order submitter, e.g., the brokerage firm handling the order. Her rationally exploiting this privilege by trading against uninformed orders and sending the informed orders to the book is often referred to as cream skimming (see, e.g., Benveniste, Marcus, and Wilhelm, 1992; Rock, 1990; Seppi, 1997; Parlour and Seppi, 2003). Cream skimming should reduce the adverse selection faced by the specialist, which is an important reason that this paper focuses on inventory control and price effects.

In the presence of cream skimming, if market participants can observe who is trading with whom, trades that involve the specialist should have smaller immediate price impacts, reflecting their lesser information content. A specialist would take on an inventory position and there would then be an immediate price reversal. But if other market participants cannot observe when the specialist trades, all trades would have similar immediate price impacts. The price reversals would not be immediate and would occur only as the underlying information (or lack thereof) is revealed.

While inventory control and cream skimming can both predict that inventories negatively correlate with the transitory component of prices, they make other differing predictions. First, unlike inventory control, cream skimming does not predict that prices have an unconditional transitory component. The pricing errors observed in the data through the negative autocorrelations in daily mid-quote returns (see Table 2) is predicted by inventory control and not by cream skimming. Second, the empirical evidence shows that the specialist inventories are negatively correlated with the informational (random-walk) component of price changes for all quintiles and significantly so except for the smallest-cap quintile (see Panels F and G in Table 2). In the state space model, this effect shows up through the negative coefficient on surprise inventory in the efficient price equation (see Panel C in Table 3). The adverse selection of the specialist suggests that cream skimming at the daily horizon is imperfect at best. intraday frequencies based on short-lived flow-based information, but does not operate at daily frequency studied in this sample. In fact, the specialist suffers adverse selection at lower frequencies consistent with classic market-making models where the market maker is uninformed (see, e.g., Glosten and Milgrom, 1985).

Finally, inventory control predicts that intermediaries pay a subsidy to reduce inventory risk. Hence, the information and risk management explanations make opposite predictions for the profitability of low-frequency (daily) position changes: Cream skimming predicts positive profits on daily position changes, while inventory control predicts negative profits. In a period before our sample, Hasbrouck and Sofianos (1993) find that NYSE-specialist gross profits are positive on short- and medium-term

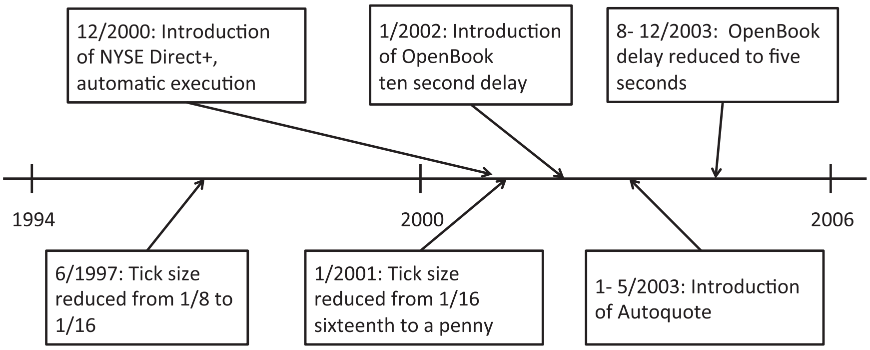

Fig. 5. NYSE market structure changes.

positions and negative on long-term positions. Studying how the 2001 change to decimalization affected specialist gross profits, Coughenour and Harris (2004) find negative specialist gross profits for medium and large stocks at horizons greater than half a day. Hence, the analysis of specialist gross profits is consistent with inventory control and not cream skimming.

4.2. Noninformational NYSE market structure

In exchange for the informational advantages discussed in Section 4.1, the NYSE imposed obligations upon specialists. Panayides (2007) shows that the most significant obligation, the Price Continuity Rule, which "requires the specialist to smooth transaction prices by providing extra liquidity as necessary to keep transaction price changes small," is important at the transaction-level horizon. Panayides finds that the rule causes specialists to accommodate investor liquidity demands via taking inventory positions to prevent "transaction prices from overshooting beyond their equilibrium levels." If this prevention of overshooting leads to prices adjusting to their efficient prices more slowly, then it causes inventories to be positively associated with pricing errors, which is the opposite of our relation between intermediary inventory and price pressure. Therefore, if the Price Continuity Rule manifests itself at a daily frequency, it causes underestimation of price pressures associated with inventory.

While the NYSE designates a single intermediary for each stock, it is likely that other investors compete with the specialist by placing limit orders to supply liquidity. This is especially important in the NYSE's 2007 market structure changes (after our sample period), which resulted in a reduced role for the specialist (Hendershott and Moulton, 2011). This change highlights an important potential weakness of our data, but also some strengths. On the positive side, the NYSE specialist system we study is the market structure for much of the data used in asset pricing. Asparouhova, Bessembinder, and Kalcheva (2010, 2013) stress the importance of noise and price pressures in asset pricing tests. Comprehensive data on the trading and the positions of other liquidity suppliers competing with (or replacing) the specialist are not available, and it is unclear when or if such data could become available.

Additional liquidity suppliers should reduce the net spread and price pressure by increasing the risk-bearing capacity in the intermediation sector. The most efficient manner to share risk among intermediaries is for the inventory to be immediately and equally shared across all liquidity suppliers, leading to perfectly correlated positions. How would this affect our estimates of price pressure? First, the marginal price pressure per unit of inventory (α) should be adjusted by the specialist's fraction of inventory, e.g., if the specialist carries one half of the total inventory, then the marginal price pressure should be multiplied by one-half. The average price pressure $[\alpha\sigma(I)]$ is unaffected by additional liquidity suppliers because the standard deviation of inventory is adjusted by the reciprocal of the adjustment to the marginal price pressure. The model-implied primitive parameters and the analysis of the spread (as summarized in Panel B and C of Table 4) are robust to this bias except for the private value rate λ, which would be adjusted by the reciprocal of the adjustment to the marginal price pressure.

4.3. NYSE market structure changes

Fig. 5 depicts a timeline of the main institutional changes at the NYSE during our sample period that have been studied in the literature. The automatic execution in Direct+ was not widely used until after our sample period (see Hendershott and Moulton, 2011). The tick size reductions in June 1997 (1/8 to 1/16) and January 2001 (1/16 to a penny) benefited the specialist. The change effectively reduced the public order precedence rule and gave the specialists more price points within the spread to quote aggressively. Coughenour and Harris (2004) study the 2001 tick change and find that specialist participation increases (participation is defined at the specialist buy volume plus her sell volume divided by twice total trading volume). The specialist participation trend in our sample confirms this result. Participation is fairly constant from 1994 through 1996. Then, surrounding the first tick change, it increases from 14.80% in 1996 to 17.75% in 1998. It stayed at about this level from 1998 through 2001. Around the second tick change, participation further increases from 17.69% in 2000 to 19.03% in 2001.

An important benefit of the ability to price aggressively in smaller increments is that the specialist could more easily mean-revert her inventory, which, in turn, could explain the increased participation rate and volatility is specialist inventories. The evidence supports this conjecture, i.e., the first-order autocorrelation decreased from 0.56 in 1996 to 0.50 in 1998 and from 0.48 in 2000 to 0.47 in 2001 (Panel D in Table 2).

The introduction of NYSE OpenBook in January 2002 and NYSE Autoquote in the first half of 2003 reduces the privileges of the specialist. OpenBook provides off-exchange traders access to the limit order book every ten seconds (see Boehmer, Saar, and Yi, 2005). Autoquote disseminates a new quote automatically whenever there was a change to the limit order book, instead of waiting for the clerk at the specialist post to type it in (see Hendershott, Jones, and Menkveld, 2011). Both changes in effect reduce the specialist's informational privilege from continuous access to the limit order book. This is consistent with our finding that the specialist experiences an increased adverse price movement when trading. The magnitude increases from 0.35 basis points per $1,000 in 2001 to 0.44 basis point in 2002 and 0.47 basis points in 2003. These changes enable limit order strategies to better compete with the specialist, and participation rates declined from 19.03% in 2001 to 14.11% in 2003.

5. Conclusion

We use empirically 12 years of NYSE intermediary data to estimate time series price dispersion due to intermediaries' inventories in financial markets. These price pressures are deviations of prices from fundamental values due to the inventory risks born by intermediaries providing liquidity to facilitate continuous marketing clearing for asynchronously arriving investors. We develop an empirical model to separate the impact of intermediaries' inventory from price effects due to information or from temporary shocks to supply and demand. We construct a theoretical model to understand and characterize the effects of price pressure. The structure of the model allows for estimation of the intermediary's risk aversion and the distribution of investors' private values. We stipulate three findings.

1. A $100,000 inventory shock causes price pressure of 0.98% for the small-capitalization stocks and 0.02% for the large-cap stocks.
2. The daily transitory volatility in stock returns due to price pressure is large: 1.18% and 0.17% for small and large stocks, respectively. For small stocks, the ratio of transitory volatility to the permanent volatility is greater than one, implying that daily price changes are driven more by noise from price pressures than from information about fundamentals.
3. The model together with the time series properties of price pressure identifies a somewhat low risk aversion for the intermediaries, a 3.96 coefficient of relative risk aversion, and deviations from constrained Pareto efficiency of 3.23 basis points of the value traded, which is approximately 10% of the cost of immediacy.

The significant size of price pressure suggests that a goal of financial market regulation should be to mitigate price pressure (see also Securities and Exchange Commission, 2010). One way to do this is by increasing capital for intermediation as the greater the risk-bearing capacity of the intermediaries the smaller the price pressure. Another approach would be to lower costs for investors to monitor the market. This would lead to investor trading being more responsive to price pressures, reducing the duration of price pressure by allowing intermediaries to mean-revert their inventories more quickly.

Appendix A. Details on the likelihood optimization

The likelihood of the state space model described by Eqs. (4), (5) and (3) is optimized in essentially three steps so as to minimize the probability of finding a local maximum. The optimization is implemented in Ox using standard optimization routines. It uses the Kalman filter and smoother from ssfpack, which is an add-on package in Ox (see Koopman, Shephard, and Doornik, 1999).

1. An ordinary least squares regression of log price difference on contemporaneous and lagged f_t yields starting values for β_i and $\beta_i^0, \ldots, \beta_i^k$ [see Eqs. (4) and (5)]. These β estimates are fixed at these values until the final step.
2. The likelihood is calculated using the Kalman filter (see Durbin and Koopman, 2012) and optimized numerically using the quasi-Newton method developed by Broyden, Fletcher, Goldfarb, and Shanno. In the optimization, all parameters are free except for the β's and $[\sigma(\varepsilon), \varphi]$. The latter runs over a nine by nine grid in which φ ranges from 0 to 0.8 and $\sigma(\varepsilon)$ ranges from 0 to a stock-specific upper bound that is calculated assuming that 80% of a stock's unconditional variance is price pressure. The likelihoods are compared across all 81 optimizations, and the $[\sigma(\varepsilon), \varphi]$ value that yields the highest likelihood is kept as the starting value for the final optimization. The rationale for this step is to prevent numerical instability of the quasi-Newton optimization. That is, if all parameters are free on arbitrary starting values, the optimization routine often runs off to a persistence parameter φ that approaches its upper bound and price pressure variance approaches the stock's unconditional variance. The optimizer starts to load the observed price series on two nonstationary series, i.e., the efficient price and the price pressure, and becomes unstable.
   The Kalman filter is initialized with a diffuse distribution for the unobserved efficient price m_0 and the unconditional price pressure distribution for s_0, i.e., $s_0 \sim N(0, \sigma^2(\varepsilon)/(1-\varphi^2))$.
3. The likelihood is optimized when all parameters are free and starting values for $[\beta_i, \beta_i^0, \ldots, \beta_i^k, \sigma(\varepsilon), \varphi]$ are equal to those found in Steps 1 and 2.

This procedure proves numerically stable as we have strong convergence in the likelihood optimization for all of our stock-year samples, i.e., convergence both in the likelihood elasticity with respect to its parameters and the one-step change in parameter values. They both become arbitrarily small.

Appendix B. Expressions for primitive parameters and spread measures

This Appendix lists expressions for the unobserved primitive parameters in terms of the empirical results. The approach for identification is matching the model-implied time series properties of price pressure with those implied by the state space model estimates. The identification strategy is described in Section 3.3. The appendix further provides a fuller set of expressions for the various spread measures developed in Section 3.4.

The expressions for the primitive parameters are

$$\lambda = \frac{(-\alpha * \nu + \delta^e * (1-\rho_I))^2}{2*(-\alpha)*(1-\rho_I)}, \tag{29}$$

$$\frac{1}{\theta} = \frac{-\alpha * \nu + \delta^e * (1-\rho_I)}{1-\rho_I}, \tag{30}$$

and

$$\gamma = \frac{2*\left(\frac{1}{\beta} - \rho_I\right) * \alpha^2 * (1-\rho_I^2) * \sigma^2(I)}{\rho_I^2 * \sigma^2(w)}, \tag{31}$$

where the γ is the coefficient of relative risk aversion, ρ_I denotes the first-order autoregressive coefficient for inventory, ν denotes intermediated volume ($specialist\_particip \times dollar\_volume$), and δ^e denotes the effective half spread. γ can be derived by multiplying the model's coefficient of absolute risk aversion ($\tilde{\gamma}$) by the intermediary's average wealth. A proxy for wealth consistent with the model is calculated by multiplying the average consumption proxy ($\overline{c}$) in Eq. (11) by the average number of stocks that the specialist firm is market maker for, i.e., (697/24), as there are 697 stocks and an average of 24 specialist firms. This number is a lower bound, as the specialist firm generates additional revenue based on stocks not in our sample. This additional revenue, however, is likely to be small as our sample contains the largest stocks that generate the highest specialist trading revenue (Sofianos, 1995).

The expressions used in the spread analysis are

$$net\_spread = \frac{-\alpha * (1-\rho_I^2) * \sigma^2(I)}{\rho_I * \nu}, \tag{32}$$

$$subsidy = \frac{-\alpha * \beta * (1-\rho_I) * (1-\rho_I^2) * \sigma^2(I)}{2*(1-\beta*\rho_I^2)*\nu}, \tag{33}$$

$$pareto\_eff\_spread = \frac{-\alpha * (-(1-\beta) + 2\beta RQ + \sqrt{(1-\beta)^2 + 4*\beta*(1+\beta)*RQ + 4*\beta^2*(RQ)^2}) * (1-\rho_I^2) * \sigma^2(I)}{4*(1-\rho_I)*\nu}, \tag{34}$$

and

$$RQ = \left(\frac{1}{\beta * \rho_I} - 1\right) * (1-\rho_I). \tag{35}$$

Appendix C. Value of price pressure discount enjoyed by liquidity demander

Let w_i be the expected value of the price pressure discount enjoyed by the liquidity demander when the intermediary starts off on an inventory of i. The Markovian law of motion for the intermediary's inventory position allows for Bellmanizing this value as

$$w_i = \mathrm{E}_i\left[\frac{1}{2} Q s^2 + \beta w_{i'}\right]. \tag{36}$$

Assume $w_i = A + Bi^2$ and calculate w_i from Eq. (36):

$$A + Bi^2 = w_i = \frac{1}{2} Q\alpha^2 i^2 + \beta \mathrm{E}_i[A + B(i + Q\alpha i + \varepsilon)^2]$$
$$= (\beta A + \beta B \sigma_\varepsilon^2) + \left(\frac{1}{2} Q\alpha^2 + \beta B(1+Q\alpha)^2\right) i^2. \tag{37}$$

A and B are solved by matching the constant and the coefficient of i^2 on both sides of the equation.

Appendix D. Supplementary material

Supplementary data associated with this paper can be found in the online version at http://dx.doi.org/10.1016/j.jfineco.2014.08.001.

References

Anand, A., Tanggaard, C., Weaver, D., 2009. Paying for market quality. Journal of Financial and Quantitative Analysis 44, 1427–1457.

Asparouhova, E., Bessembinder, H., Kalcheva, I., 2010. Liquidity biases in asset pricing tests. Journal of Financial Economics 96, 215–237.

Asparouhova, E., Bessembinder, H., Kalcheva, I., 2013. Noisy prices and inference regarding returns. Journal of Finance 68, 665–714.

Barsky, R., Juster, T., Kimball, M., Shapiro, M., 1997. Preference parameters and behavioral heterogeneity: an experimental approach in the health and retirement survey. Quarterly Journal of Economics 112, 537–579.

Benveniste, L., Marcus, M., Wilhelm, W., 1992. What's special about the specialist? Journal of Financial Economics 32, 61–86.

Boehmer, E., Saar, G., Yu, L., 2005. Lifting the veil: an analysis of pre-trade transparency at the NYSE. Journal of Finance 60, 783–815.

Born, B., Brennan, J., Engle, R., Ketchum, R., O'Hara, M., Philips, S., Ruder, D., Stiglitz, J., 2011. Recommendations regarding regulatory responses to the market events of May 6, 2010. Manuscript. Commodity Futures Trading Commission, Washington, DC.

Brogaard, J., Hendershott, T., Riordan, R., 2014. High frequency trading and price discovery. Review of Financial Studies, 27 (8), 2267–2306. http://dx.doi.org/10.1093/rfs/hhu032.

Brunnermeier, M., Pedersen, L., 2009. Market liquidity and funding liquidity. Review of Financial Studies 22, 2201–2238.

Campbell, J., Grossman, S., Wang, J., 1993. Trading volume and serial correlation in stocks returns. Quarterly Journal of Economics 108, 905–939.

Campbell, J., Lo, A.W., MacKinlay, A., 1997. The Econometrics of Financial Markets. Princeton University Press, Princeton, NJ.
Cao, C., Choe, H., Hatheway, F., 1997. Does the specialist matter? Differential trading costs and inter-security subsidization on the NYSE. Journal of Finance 52, 1615–1640.
Chetty, R., 2006. A new method of estimating risk aversion. American Economic Review 96, 1821–1834.
Cochrane, J., 2001. Asset Pricing. Princeton University Press, Princeton, NJ.
Cohen, A., Einav, L., 2007. Estimating risk preferences from deductible choice. American Economic Review 97, 745–788.
Comerton-Forde, C., Hendershott, T., Jones, C., Seasholes, M., Moulton, P., 2010. Time variation in liquidity: the role of market maker inventories and revenues. Journal of Finance 65, 295–331.
Coughenour, J., Harris, L., 2004. Specialist profits and the minimum price increment. Manuscript. University of Southern California, Los Angelas, CA.
Coval, J., Stafford, E., 2007. Asset fire sales (and purchases) in equity markets. Journal of Financial Economics 86, 479–512.
Degryse, H., de Jong, F., van Ravenswaaij, M., Wuyts, G., 2005. Aggressive orders and the resiliency of a limit order market. Review of Finance 9, 201–242.
Duffie, D., 2010. Presidential address: asset price dynamics with slow-moving capital. Journal of Finance 65, 1237–1267.
Durbin, J., Koopman, S., 2012. Time Series Analysis by State Space Models. Oxford University Press, Oxford, UK.
Gabaix, X., Gopikrishnan, P., Plerou, V., Stanley, H., 2006. Institutional investors and stock market volatility. Quarterly Journal of Economics 121, 461–504.
George, T., Hwang, C., 2001. Information flow and pricing errors: a unified approach to estimation and testing. Review of Financial Studies 14, 979–1020.
Glosten, L., Milgrom, P., 1985. Bid, ask, and transaction prices in a specialist market with heterogeneously informed agents. Journal of Financial Economics 14, 71–100.
Greenwood, R., 2005. Short- and long-term demand curves for stocks: theory and evidence on the dynamics of arbitrage. Journal of Financial Economics 75, 607–649.
Grossman, S., Miller, M., 1988. Liquidity and market structure. Journal of Finance 43, 617–633.
Harris, L., Gurel, E., 1986. Price and volume effects associated with changes in the S& P 500 list: new evidence for the existence of price pressures. Journal of Finance 41, 815–829.
Hasbrouck, J., 1991a. Measuring the information content of stock trades. Journal of Finance 46, 179–207.
Hasbrouck, J., 1991b. The summary informativeness of stock trades: an econometric analysis. Review of Financial Studies 4, 571–595.
Hasbrouck, J., 1999. The dynamics of discrete bid and ask quotes. Journal of Finance 54, 2109–2142.
Hasbrouck, J., Sofianos, G., 1993. The trades of market makers: an empirical analysis of NYSE specialists. Journal of Finance 48, 1565–1593.
Hendershott, T., Jones, C., Menkveld, A.J., 2011. Does algorithmic trading improve liquidity? Journal of Finance 66, 1–33.
Hendershott, T., Jones, C., Menkveld, A.J., 2013. Implementation shortfall and transitory price effects. In: Easley, D., de Prado, M., O'Hara, M. (Eds.), High-Frequency Trading: New Realities for Traders, Markets and Regulators, Risk Books, London, UK.
Hendershott, T., Moulton, P., 2011. Automation, speed, and stock market quality: the NYSE's hybrid. Journal of Financial Markets 14, 568–604.
Hendershott, T., Seasholes, M., 2007. Market maker inventories and stock prices. American Economic Review 97, 210–214.
Ho, T., Stoll, H., 1981. Optimal dealer pricing under transaction cost and return uncertainty. Journal of Financial Economics 9, 47–73.
Hollifield, B., Miller, R., Sandås, P., Slive, J., 2006. Estimating the gains from trade in limit order markets. Journal of Finance 61, 2753–2804.
Kirilenko, A., Kyle, A., Samadi, M., Tuzun, T., 2011. The flash crash: the impact of high-frequency trading on an electronic market. Manuscript. University of Maryland, College Park, MD.
Koopman, S., Shephard, N., Doornik, J., 1999. Statistical algorithms for models in state space using ssfpack 2.2. Econometrics Journal 2, 113–166.
Kraus, A., Stoll, H., 1972. Price impacts of block trading on the New York Stock Exchange. Journal of Finance 27, 569–588.
Lagos, R., Rocheteau, G., Weill, P., 2011. Crises and liquidity in OTC markets. Journal of Economic Theory 146, 2169–2205.
Large, J., 2007. Measuring the resiliency of an electronic limit order book. Journal of Financial Markets 10, 1–25.
Ljungqvist, L., Sargent, T., 2004. Recursive Macroeconomic Theory. MIT Press, Cambridge, MA.
Madhavan, A., Smidt, S., 1991. A Bayesian model of intraday specialist pricing. Journal of Financial Economics 30, 99–134.
Madhavan, A., Smidt, S., 1993. An analysis of changes in specialist inventories and quotations. Journal of Finance 48, 1595–1628.
Mehra, R., Prescott, E., 1985. The equity premium: a puzzle. Journal of Monetary Economics 15, 145–161.
Menkveld, A.J., 2013. High-frequency trading and the new market makers. Journal of Financial Markets 16, 712–740.
Menkveld, A.J., Koopman, S.J., Lucas, A., 2007. Modelling round-the-clock price discovery for cross-listed stocks using state space methods. Journal of Business and Economic Statistics 25, 213–225.
Menkveld, A.J., Wang, T., 2013. How do designated market makers create value for small-caps? Journal of Financial Markets 16, 571–603.
Mildenstein, E., Schleef, H., 1983. The optimal pricing policy of a monopolistic market maker in the equity market. Journal of Finance 38, 218–231.
Nagel, S., 2012. Evaporating liquidity. Review of Financial Studies 25, 2005–2039.
Nimalendran, M., Petrella, G., 2003. Do thinly traded stocks benefit from specialist intervention? Journal of Banking and Finance 27, 1823–1854.
Panayides, M., 2007. Affirmative obligations and market making with inventory. Journal of Financial Economics 86, 513–542.
Parlour, C., Seppi, D., 2003. Liquidity-based competition for order flow. Review of Financial Studies 16, 301–343.
Pastor, L., Stambaugh, R., 2003. Liquidity risk and expected returns. Journal of Political Economy 111, 642–685.
Petersen, M., 2009. Estimating standard errors in finance panel data sets: comparing approaches. Review of Financial Studies 22, 435–480.
Rock, K., 1990. The specialist's order book and price anomalies. Manuscript. Harvard University, Cambridge, MA.
Rust, J., Hall, J., 2003. Middlemen versus market makers: a theory of competitive exchange. Journal of Political Economy 111, 353–403.
Securities and Exchange Commission, 2010. Concept release on equity market structure. Release no. 34-61358. File no. S7-02-10. US Government Printing Office, Washington, DC.
Seppi, D., 1997. Liquidity provision with limit orders and a strategic specialist. Review of Financial Studies 10, 103–150.
Sofianos, G., 1995. Specialist gross trading revenues at the New York Stock Exchange. Manuscript 95-01. New York Stock Exchange, New York, NY.
Stoll, H., 1978. The supply of dealer services in securities markets. Journal of Finance 33, 1133–1151.
Townsend, R., 1978. Intermediation with costly bilateral exchange. Review of Economic Studies 45, 417–425.
Venkataraman, K., Waisburd, A., 2007. The value of the designated market maker. Journal of Financial and Quantitative Analysis 42, 735–758.
Zabel, E., 1981. Competitive price adjustment without market clearing. Econometrica 49, 1201–1221.