Internet Appendix
for
Identifying Demand and Supply in Index Option Markets

Contents

List of Figures

List of Tables

1 Identifying Demand and Supply Equations Using Sign-Restricted VARs

This section discusses the framework we use in our empirical application to identify demand and supply equations in the S&P 500 index option market. This approach is known as the pure-sign restriction vector autoregression (VAR). It was initially proposed by Uhlig (2005). Denote the vector of prices and quantities by $y_t = (p_t \quad q_t)'$. The reduced-form VAR with l lags, VAR(ℓ), is given by:

$$y_t = \alpha + B_1 y_{t-1} + B_2 y_{t-2} + ... + B_\ell y_{t-\ell} + u_t \tag{A.1}$$

where u_t is the reduced-form residual, a (2×1) vector with covariance matrix $E[u_t u_t'] = \Sigma_u$. B_i is a (2×2) matrix of coefficients, with $i \in (1, ..., \ell)$, and α is a (2×1) vector of constants. We thus have a system of two equations with price and quantity as the endogenous variables. The demand and supply equations must be identified by decomposing the residuals u_t into orthonormal structural or fundamental shocks. Specifically, the mapping from the orthonormal fundamental shocks ε_t, with $E[\varepsilon_t \varepsilon_t'] = I$, to the residual u_t is through the matrix $\mathbf{A}$, as follows:

$$u_t = \mathbf{A}\varepsilon_t \Rightarrow \Sigma_u = E[u_t u_t'] = \mathbf{A}E[\varepsilon_t \varepsilon_t']\mathbf{A'} = \mathbf{AA'}. \tag{A.2}$$

Identification of structural shocks is thus equivalent to identification of this impact matrix $\mathbf{A}$. In order to identify the first equation as a supply equation and the second equation as a demand equation, the pure-sign restriction approach imposes the following sign restrictions on the elements of the matrix $\mathbf{A}$:

$$\begin{bmatrix} u_t^P \\ u_t^Q \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} \varepsilon_t^S \\ \varepsilon_t^D \end{bmatrix} = \begin{bmatrix} - & + \\ + & + \end{bmatrix} \begin{bmatrix} \varepsilon_t^S \\ \varepsilon_t^D \end{bmatrix} \tag{A.3}$$

The first column of $\mathbf{A}$ assumes that a positive supply shock (a right shift of the supply curve) results in a lower price p_t and a higher quantity q_t. The sign restrictions on the second column of $\mathbf{A}$ assume that positive demand shock (a right shift of the demand curve) results in an increase in both price and quantity (Uhlig, 2017). These assumptions impose only the most basic economic intuition. Note also that we only impose sign restrictions on the contemporaneous responses of price and quantity to demand and supply, which is a very conservative assumption. Other applications are more restrictive. For instance, Uhlig (2005) discusses how to impose restrictions on the responses over longer horizons by checking the signs of the relevant impulse response functions.

This sign-restriction identification approach differs from the more traditional VAR identification scheme known as recursive identification, where one imposes a Cholesky decomposition on the covariance matrix Σ_u such that $chol(\Sigma_u) = \mathbf{L}$, i.e. $\Sigma_u = \mathbf{LL'}$, and $u_t = \mathbf{L}\varepsilon_t$.[1] Although this decomposition is unique, it implicitly imposes a relative exogeneity ranking on the endogenous variables in the VAR. Specifically, because the impact matrix $\mathbf{L}$ is a lower triangular matrix, if we order price before quantity in the vector Y_t, recursive identification imposes price as an exogenous variable in the system, whereas quantity is endogenous, and vice versa. When there are more than two variables in the VAR, one needs to take a stance on which variables are more or less exogenous; hence this is referred to as a relative exogeneity ranking.

In our empirical application, we do not use a relative exogeneity ranking between prices and quantities and instead, use the pure-sign restriction identification approach on the matrix $\mathbf{A}$. Every matrix $\mathbf{A}$ such that $\Sigma_u = \mathbf{AA'}$ can be expressed as a permutation of the Cholesky matrix through an orthonormal matrix $\mathbf{Q}$: $\mathbf{A} = \mathbf{LQ}$. Identifying the matrix $\mathbf{A}$ is thus equivalent to identifying the matrix $\mathbf{Q}$. After identifying the impact matrix $\mathbf{A}$, we can recover the structural shocks $\varepsilon_t = \mathbf{A}^{-1}u_t$ and the slopes of the demand and supply curves, which henceforth we refer to as β^S and β^D. However, the admissible matrix $\mathbf{A}$ is not unique; the estimation of the shocks and the coefficients of the system is based on the average of the estimates obtained for all admissible matrices $\mathbf{A}$.

We implement the resulting estimation of the pure-sign restricted VAR using a Bayesian approach, following Uhlig (2005). This approach can be summarized in six steps. First, we estimate the reduced-form VAR(ℓ) using ordinary least squares (OLS), where the optimal number of lags ℓ is chosen based on Bayesian Information Criterion (BIC):

$$\begin{bmatrix} p_t \\ q_t \end{bmatrix} = \begin{bmatrix} \alpha^S \\ \alpha^D \end{bmatrix} + \begin{bmatrix} b^P_{S1} & b^Q_{S1} \\ b^P_{D1} & b^Q_{D1} \end{bmatrix} \begin{bmatrix} p_{t-1} \\ q_{t-1} \end{bmatrix} + ... + \begin{bmatrix} b^P_{S\ell} & b^Q_{S\ell} \\ b^P_{D\ell} & b^Q_{D\ell} \end{bmatrix} \begin{bmatrix} p_{t-\ell} \\ q_{t-\ell} \end{bmatrix} + \begin{bmatrix} u^S_t \\ u^D_t \end{bmatrix} \tag{A.4}$$

Second, we assume a diffusive conjugate Normal-Wishart prior for $(\mathbf{B},\mathbf{\Sigma_u})$, where $\mathbf{B} = [\alpha, B_1, ..., B_l]$, and take 100 draws from the posterior over $\mathbf{B}$ and $\mathbf{\Sigma_u}$. Third, for each draw of $\mathbf{\Sigma_u}$, we take 1,000 draws of an orthonormal matrix $\mathbf{Q}$ that is drawn uniformly from the unit circle, using QR factorization. Fourth, we compute the candidate impact matrix $\mathbf{A_m} = \mathbf{LQ}$, where $\mathbf{L}$ is the lower triangular Cholesky decomposition for each draw of $\mathbf{\Sigma_u}$. We then verify if $\mathbf{A_m}$ satisfies the sign restrictions of Equation A.3 and discard the draws that violate the sign restriction. Fifth, for each retained $\mathbf{A_m}$, we recover the structural shocks $\varepsilon_t = \mathbf{A_m}^{-1}u_t$, and calculate the demand and supply slopes. The relation between the impact matrix $\mathbf{A}$ and the demand and supply slopes (β^S and β^D) is derived as follows. Pre-multiplying both sides of Equations (A.4) by the matrix $\mathbf{F} = \mathbf{A}^{-1}$ to arrive at the

[1] On the recursive VAR approach, see for instance Sims (1980, 1986); Bernanke (1986); Blanchard and Watson (1986). The microstructure literature employs this approach when estimating vector autoregressions to investigate the impact of order flow on prices (Hasbrouck, 1991, 1993).

structural VAR, we get:[2]

$$\begin{bmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{bmatrix} \begin{bmatrix} p_t \\ q_t \end{bmatrix} = ... + \begin{bmatrix} \varepsilon_t^S \\ \varepsilon_t^D \end{bmatrix} \tag{A.5}$$

This matrix expression corresponds to the following system of equations:

$$\begin{cases} f_{11}p_t + f_{12}q_t = ... + \varepsilon_t^S \text{ (Supply)} \\ f_{21}p_t + f_{22}q_t = ... + \varepsilon_t^D \text{ (Demand)} \end{cases} \iff \begin{cases} p_t = -\frac{f_{12}}{f_{11}}q_t + ... \text{ (Supply)} \\ p_t = -\frac{f_{22}}{f_{21}}q_t + ... \text{ (Demand)} \end{cases} \tag{A.6}$$

Once we identify the impact matrix $\mathbf{A}$, we calculate the slopes of the demand and supply curves:

- The supply curve P(Q) has the slope $\beta^S = -\frac{f_{12}}{f_{11}} = -\frac{-a_{12}}{a_{22}} = \frac{a_{12}}{a_{22}}(>= 0)$
- The demand curve P(Q) has the slope $\beta^D = -\frac{f_{22}}{f_{21}} = -\frac{a_{11}}{-a_{21}} = \frac{a_{11}}{a_{21}}(<= 0)$

Because of the imposed sign restrictions, we obtain a positive supply slope and a negative demand slope. In the final (sixth) step of the estimation procedure, we calculate the mean for all variables of interest from the retained draws.

The extended VAR which includes additional risk factors imposes the following restrictions on the impact matrix $\mathbf{A}$:

$$\begin{bmatrix} u_t^P \\ u_t^Q \\ u_t^{Return} \\ u_t^{\Delta VIX} \\ u_t^{\Delta Skew} \end{bmatrix} = \mathbf{A} \begin{bmatrix} \varepsilon_t^S \\ \varepsilon_t^D \\ \varepsilon_t^{Return} \\ \varepsilon_t^{\Delta VIX} \\ \varepsilon_t^{\Delta Skew} \end{bmatrix} = \begin{bmatrix} - & + & * & * & * \\ + & + & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \end{bmatrix} \begin{bmatrix} \varepsilon_t^S \\ \varepsilon_t^D \\ \varepsilon_t^{Return} \\ \varepsilon_t^{\Delta VIX} \\ \varepsilon_t^{\Delta Skew} \end{bmatrix} \tag{A.7}$$

We do not constrain the contemporaneous effect of the risk measures on prices and quantities because we want to infer it from the data. We also do not impose restrictions on the contemporaneous relation between the risk measures. The optimal number of lags is again determined using the BIC.

To estimate this extended VAR system, we implement the modified procedure in Arias, Rubio-Ramírez, and Waggoner (2013) and Kilian and Lütkepohl (2017) which is designed to estimate VARs identified by a mix of sign and exclusion restrictions. Compared to the baseline estimation, the only additional step is to draw 1,000 orthogonal matrices $\mathbf{Q}$ such that $\mathbf{A} = \mathbf{LQ}$ satisfies the exclusion restrictions, prior to verifying the sign restrictions on candidate matrices $\mathbf{A}$. The supply and demand slopes can again be computed using elements of the (inverse of) $\mathbf{A}$, similar to the ratios specified in Equation (A.6).

[2]Note that matrix $\mathbf{F} = \begin{bmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{bmatrix} = \mathbf{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$ and $\varepsilon_t = \mathbf{A}^{-1}u_t = \mathbf{F}u_t$.

2 The Sign-Restricted VAR Approach: Discussion

In this section, we first discuss some important issues regarding the interpretation of the results from the sign-restricted VAR approach. Then we explore the full distribution of the slopes and elasticities across the retained draws and the convergence of the estimation approach. See Kilian and Lütkepohl (2017) and Fry and Pagan (2011) for a comprehensive review of the sign-restricted VAR methodology.

2.1 Interpreting the Results from Sign-Restricted VARs

Our estimation procedure relies on rejecting the generated contemporaneous impulse responses if they fail specific sign restrictions. The question therefore arises how to interpret the rejection rate. Sign-identified VAR models are set-identified, rather than point-identified. This implies that multiple admissible structural models can generate the data. The discarding procedure sign-restricts the impulse responses and eliminates draws based on economic intuition in search of an economically plausible model. The rejection rate is therefore not a measure of the quality of the identification (Kilian and Lütkepohl, 2017). A high rejection rate may signal that alternative models may provide a better fit, but it can also indicate that the true model indeed satisfies the sign restrictions; a high rejection rate can thus be indicative of sharp identification when additional information is introduced between the models in the admissible set (Uhlig, 2017). Our retaining rate of about 45% is actually rather high compared to many other applications; for instance Kilian and Murphy (2012) report a retaining rate of approximately 2.1%.

The literature has proposed alternative refinements to reduce the set of admissible models: Informative priors on the matrix **A** (Baumeister and Hamilton, 2015, 2019); Use the model which is closest to the median (Fry and Pagan, 2011); Use the mode of the distribution of admissible models (Inoue and Kilian, 2013). However, these approaches require the researcher to have a strong prior on what the best model is and what extreme cases to be excluded. We therefore proceed by retaining all admissible models in our empirical analysis and report the means for the slope estimates, structural shocks, IRFs, and the forecast error variance decomposition. We next provide additional insight into model uncertainty by discussing the full distribution of the retained slopes and elasticities.

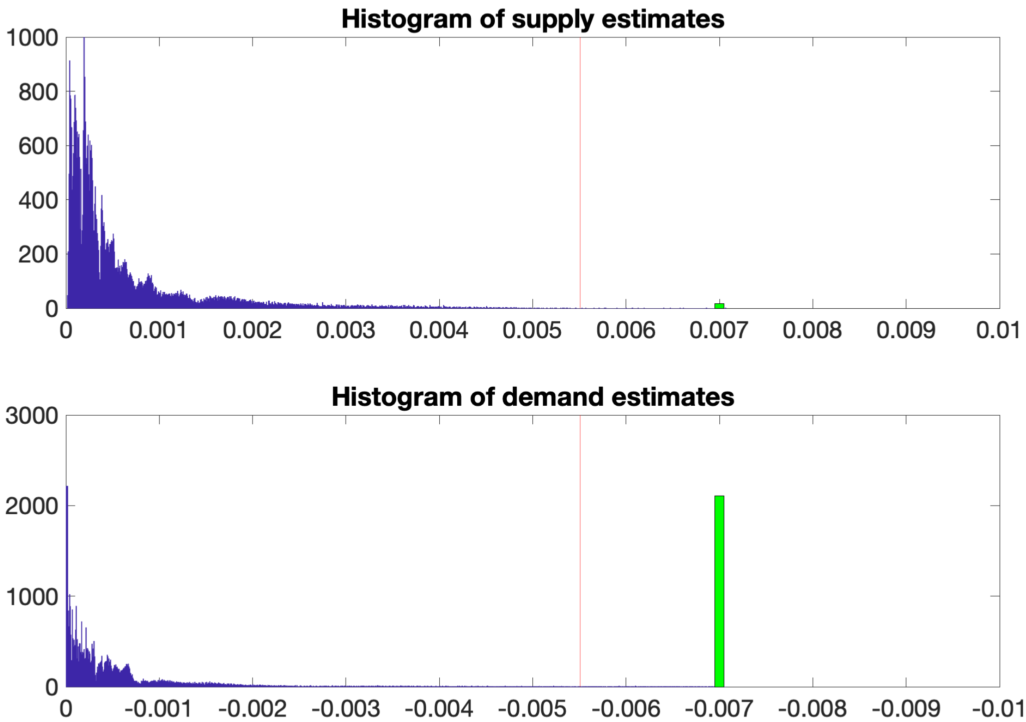
2.2 The Distribution of the Admissible Set

Figure A.1 plots the histograms of the distribution of demand and supply slopes in the ATM SPX put market from our base case analysis.[3] Recall that a slope is obtained for each draw of the matrix **A**. Note that the distribution has a very long tail, and that the

[3]We present results for put options; the findings for call options are similar.

distribution mass concentrates around positive (negative) but small values for the supply (demand) slope; both distributions appear to be highly skewed.

Figure A.1: Histograms of Supply and Demand Slopes Based on Retained Draws



We plot the histograms of the distribution of demand and supply slope estimates based on 100,000 draws for ATM put options. The red vertical line indicates the value of the slope which corresponds to unit elasticity. The demand slope distribution has a very long tail; we therefore collapse all values with elasticity smaller than one in the green bar on the left of the vertical line, to make the figure more readable.

The average demand slope estimate is almost four times bigger in magnitude than the average supply slope estimates: 22×10^{-4} versus 6×10^{-4}. The supply slope estimates that correspond to elasticities smaller than one take up only a very small portion (18 observations out of 47,627 retained draws).[4] This means that there are very few admissible models which imply an inelastic supply curve, and we conclude that the supply of market-makers is elastic. In contrast, for the demand slope we have many extreme values corresponding to elasticities smaller than one.

Figure A.2 provides more detail on the distribution of the demand and supply elastici-

[4]The red vertical line indicates the value of the slope for which we have unitary elasticity. We collapse all the values corresponding to elasticities of less than one in the green bar on the right of the vertical line to make the figure easier to read.

ties for the elastic and inelastic cases separately.[5] Figure A.2 shows that the supply curves are almost always highly elastic. The elastic demand case, which represents 95.58% of the distribution, is representative of the full retained draws sample. For inelastic demand draws, the supply curve is much more elastic (average elasticity of 225) than in the case of elastic demand draws. The case of inelastic demand and very elastic supply (the top-right plot in Figure A.2) corresponds to the null hypothesis of a flat supply curve and a vertical demand curve. Although these cases are economically meaningful, they occur infrequently (4.42% of cases). Indeed, the median estimates of the slopes are very similar across sub-samples.[6] This motivates us to focus on the mean, rather than the median, of the slope estimates. While the clusters in the tail of the slope estimates do not occur too frequently, they remain economically meaningful.

2.3 Convergence

This section investigates the convergence of our estimation strategy by focusing on estimation results for different numbers of draws. We estimate the VAR with the number of draws for the orthonormal matrix $\mathbf{Q}$ between 10 and 1,000. Because we use 100 draws on the posterior over $\mathbf{B}$ and $\mathbf{\Sigma_u}$, this results in between 1,000 and 100,000 parameter draws.[7] For a given number of parameter draws N, we repeat the estimation procedure 100 times using different seed generators and compute the standard deviation of the 100 mean estimates.

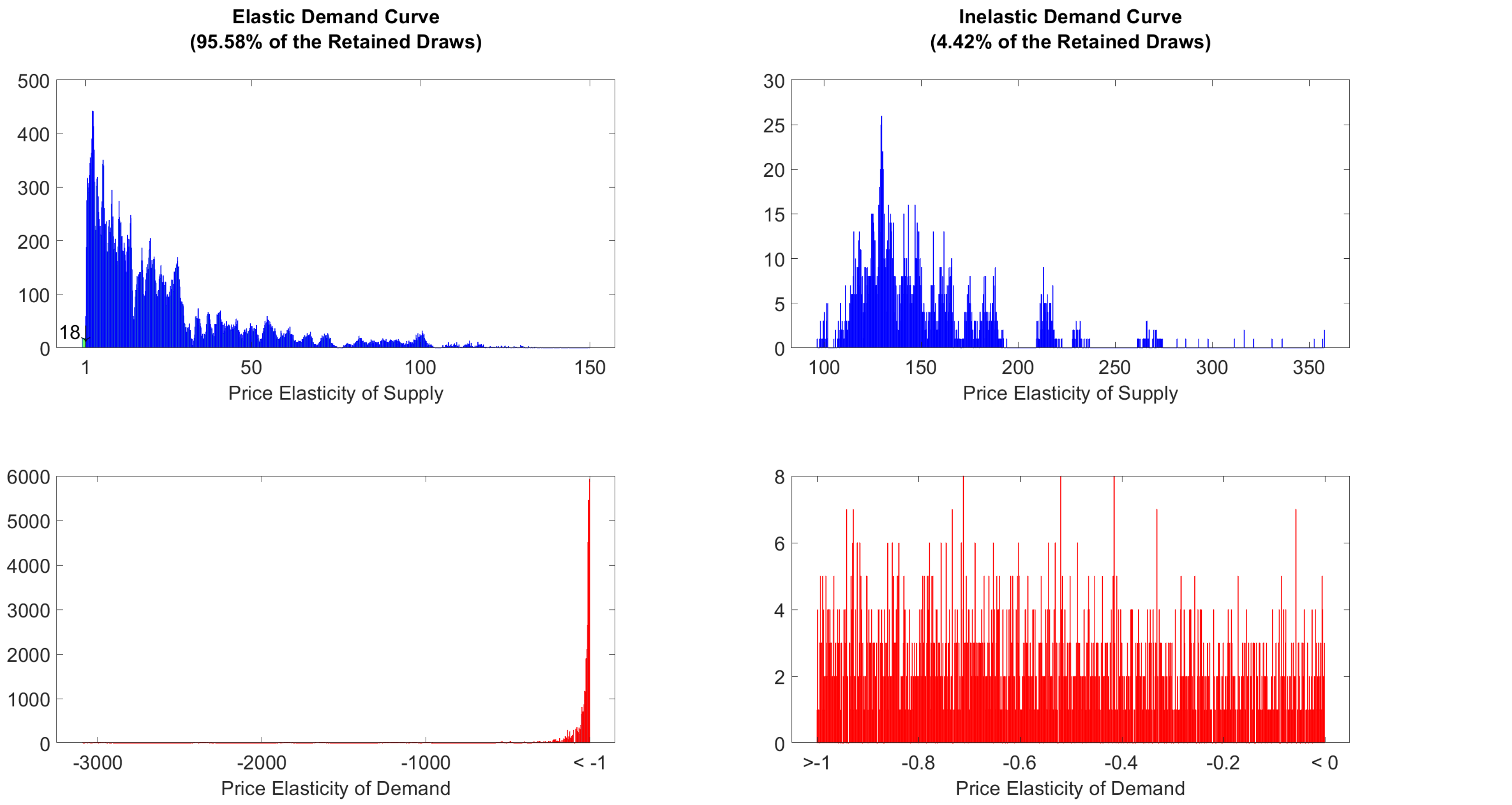
Figure A.3 plots the standard deviations of the demand and supply slope estimates across different seeds as a function of the parameter draws N, based on the ATM put option sample. The results show that our implementation converges, since the standard deviation of the estimates from different seeds decreases as the number of draws increases, both for the supply and demand slopes. The standard deviation of the demand slope estimates consistently exceeds that of the supply slope estimates; this suggests that for our empirical application and this sample, the estimates of the slope of the supply curve are more accurate than the estimates of the demand curve. We obtain similar findings for the ATM call sample. For the OTM call and put option samples, we find that the demand curves can be estimated more precisely than the supply curves. These results are presented in Figure A.9.

[5] We first compute the price elasticities of demand and supply using all retained draws. Then we split the retained sample into two groups: elastic demand (demand elasticity < -1) and inelastic demand (demand elasticity > -1). Finally, we plot the histograms of price elasticities for these two groups.

[6] Table A.2 reports on the exercise in Table 3 in the paper based on the the median estimates.

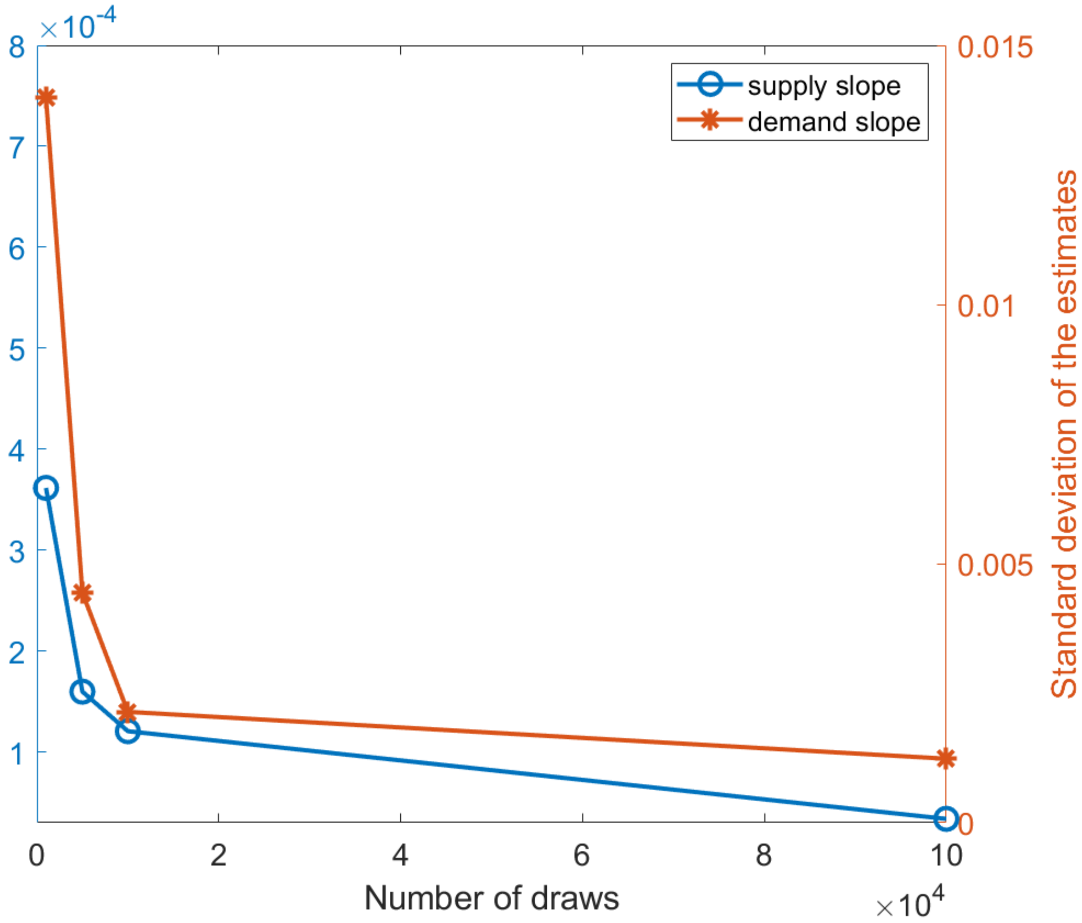
[7] As outlined in Section 1 of this Online Appendix, we take 100 draws from the posterior over $\mathbf{B}$ and $\mathbf{\Sigma_u}$; then, for each draw of $\mathbf{\Sigma_u}$, we take n draws of an orthonormal matrix $\mathbf{Q}$. The variation in our estimates largely results from the number of draws taken on the rotation matrix $\mathbf{Q}$.

Figure A.2: Histograms of Supply and Demand Elasticities in Subsamples: Demand Elasticity < -1 and > -1



We plot the histograms of the distribution of demand and supply price elasticities based on 100,000 draws for ATM put options for two subsamples. We first compute the price elasticities of demand and supply using all retained draws. Then we split the retained draws into two groups: the elastic demand group (demand elasticity < -1) and the inelastic demand group (demand elasticity > -1). We plot the histograms of the price elasticities for these two groups.

Figure A.3: Convergence of the Slope Estimates



We plot the standard deviations of the demand and supply slope estimates for ATM put options as a function of the number of parameter draws. For a given number of draws N (x-axis), we repeat the estimation procedure 100 times under different seed generators. We use the mean estimate of the slopes across the retained draws and compute the standard deviation of these 100 mean estimates. The supply slope (blue line with circle markers) is indicated on the left y-axis, while the demand slope (red line with star markers) is indicated on the right y-axis.

2.4 The Choice of the Prior for the Impact Matrix

In the VAR estimations outlined in the main paper, we employed the uninformative prior for the impact matrix $\mathbf{A}$ by assuming the standard Haar prior on the rotation matrix $\mathbf{Q}$. Nevertheless, using an uninformative prior for a parameter may result in an informative prior for a non-linearly transformed quantity of that parameter, such as the elasticity of the curves. As emphasized by Baumeister and Hamilton (2015), this situation can introduce potential bias in set-identified models like ours, where the prior does not asymptotically vanish.

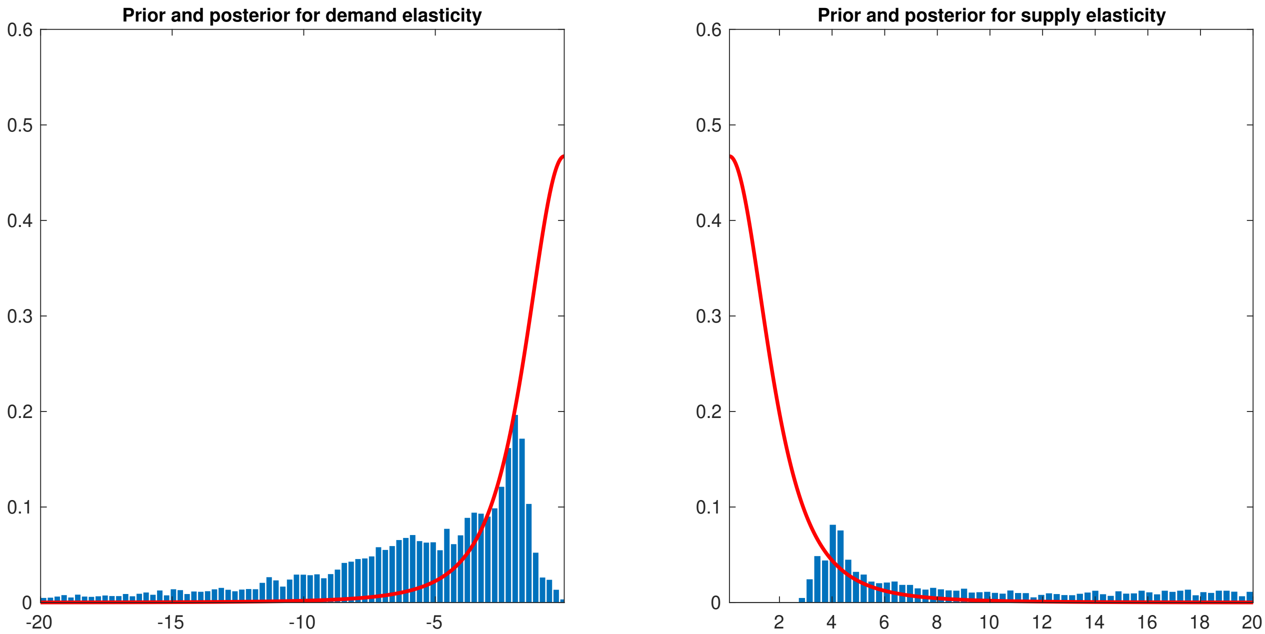
To address this concern, we adopt the robustness analysis proposed by Baumeister and Hamilton (2015). This method recommends placing the prior directly on the elasticities instead of the impact matrix and examining the behavior of the posterior distributions. Consequently, we re-estimate the baseline VARs for at-the-money (ATM) and out-of-the-money (OTM) calls and puts by assuming the truncated t-student prior proposed by Baumeister and Hamilton (2015) with a mean of 0.1, signifying an inelastic prior.

Figures A.4 and A.5 illustrate the prior and posterior distributions of elasticities for the four option samples. The data strongly rejects the inelastic prior, and the posteriors visibly deviate from zero, particularly for the supply elasticity of ATM options. Table A.1 presents the mean and mode of the posterior distributions and documents that they are all well above one.

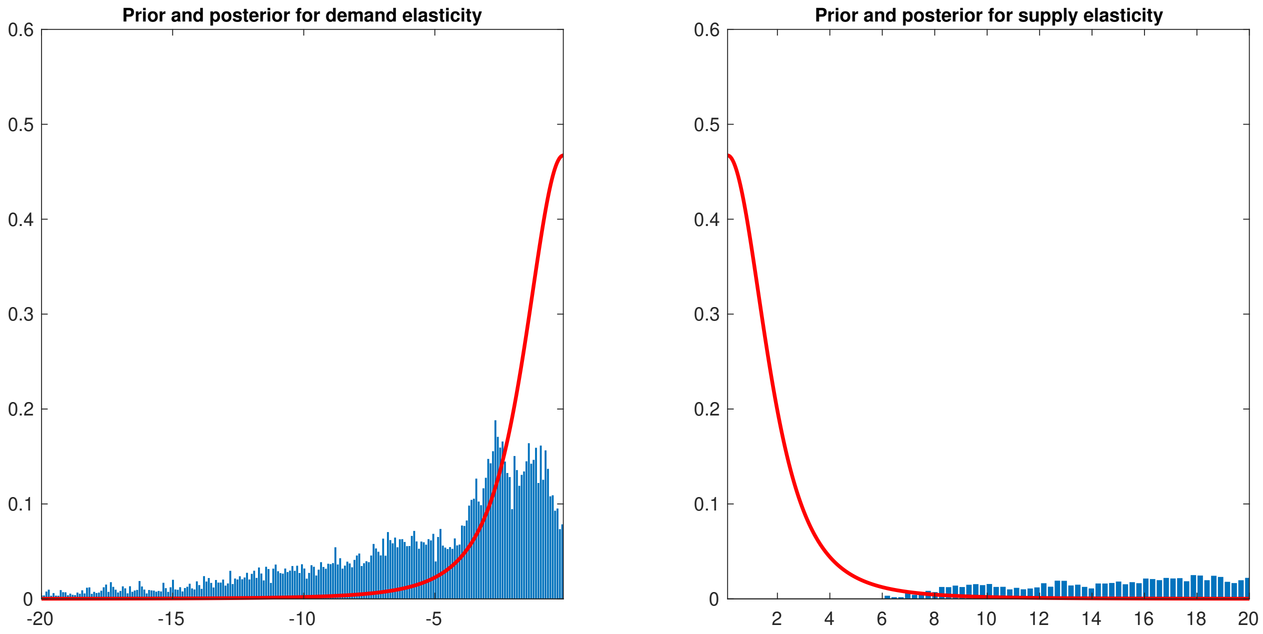
In conclusion, this analysis documents that the result regarding the elasticity of option markets remains robust with respect to the choice of prior for the impact matrix.

Figure A.4: Prior and Posterior for ATM Options

Panel B: ATM Puts



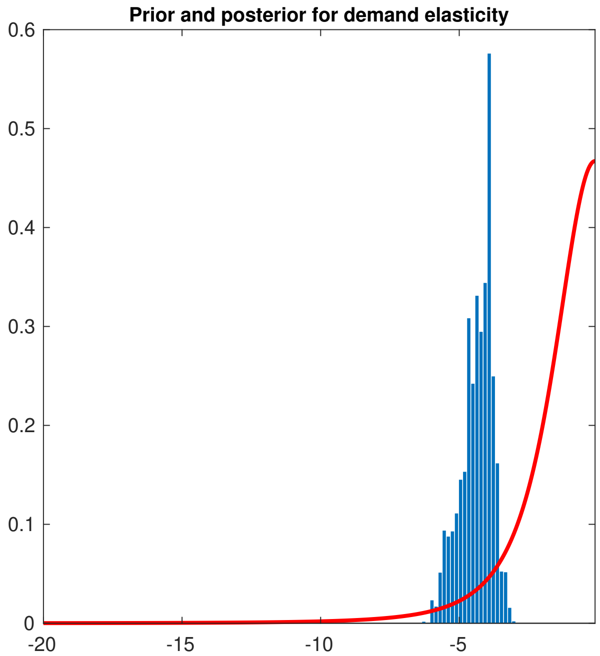
Panel A: ATM Calls

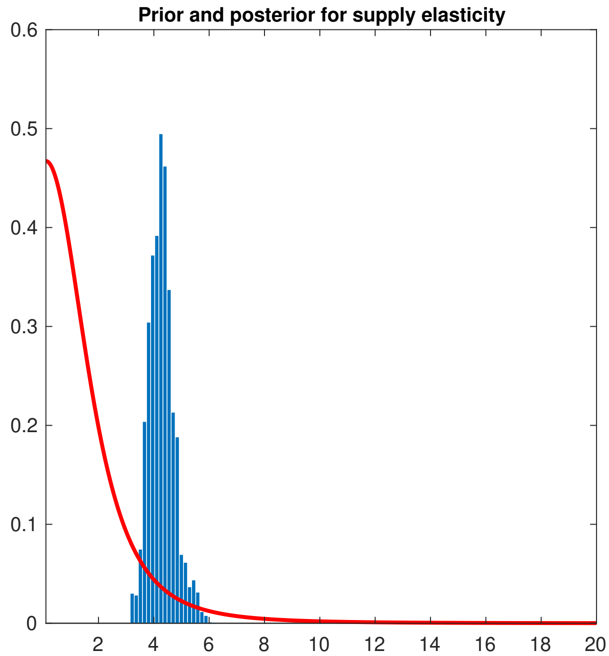


The figure reports the prior (continuous red line) and posterior distribution (histogram) of the elasticity of demand (left graphs) and the elasticity of supply (right graphs). The sample considered are at-the-money put (Panel A) and call options (Panel B).

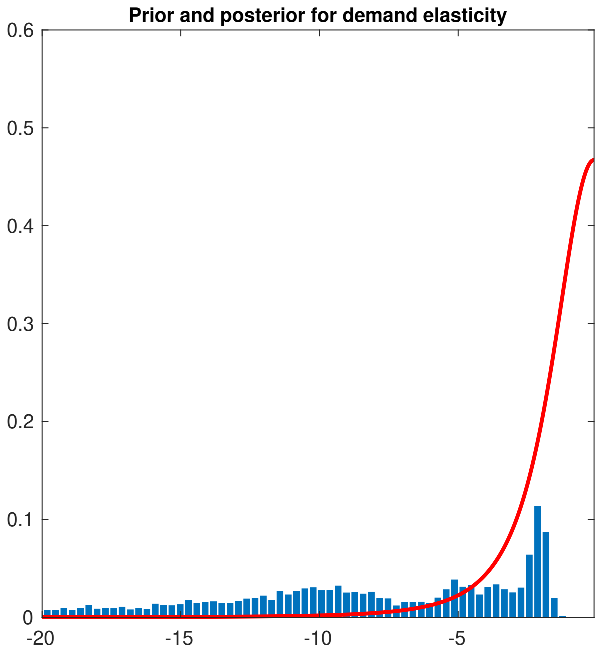
Figure A.5: Prior and Posterior for OTM Options

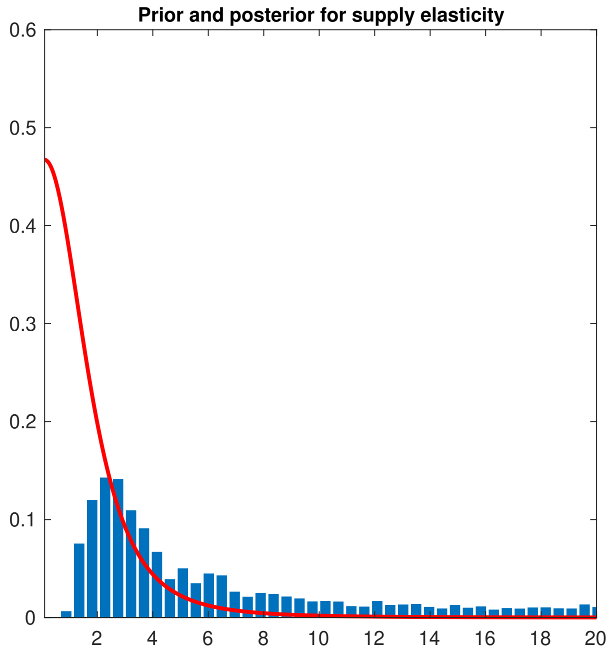
Panel B: OTM Puts





Panel A: OTM Calls





The figure reports the prior (continuous red line) and posterior distribution (histogram) of the elasticity of demand (left graphs) and the elasticity of supply (right graphs). The sample considered are out-of-the-money put (Panel A) and call options (Panel B).

Table A.1: Mean and Mode of the Posterior Distributions

	Demand Elasticity		Supply Elasticity	
	Mean	**Mode**	**Mean**	**Mode**
ATM Puts	-18.45	-1.96	41.46	4.03
ATM Calls	-6.45	-2.69	44.38	26.69
OTM Puts	-33.83	-3.92	33.43	4.25
OTM Calls	-43.61	-2.13	33.40	2.28

This table shows the mean and mode of the posterior distributions of demand and supply elasticities of Figures A.4 and A.5.

3 The CBOE Open-Close Data

The database used to construct the quantity measure is the CBOE Open-Close database, which reports daily end-user SPX order flow (buy versus sell initiated trades) for each option series between January 1, 1996 and December 30, 2020. There is a structural change in the CBOE Open-Close database in 2011 due to migration to the BATS platform and technology. The data before 2011 differentiates exclusively between trades originating from customers and firms. The sum of the trades for these two investor types constitutes end-user net demand; market maker net demand is equal to minus end-user net demand.

Post-migration, the data provides more granularity on the origin of trades: it differentiates between customers, firms, broker-dealers, professional customers, and market-makers. We merge the two databases as follows. First, we verify that the classification of firms and customers is the same across the two databases.[8] We then observe that the new dataset reports the trades of (almost) all participants in the option market. The aggregated net demands from the new traders identified in the post-migration database, i.e. broker-dealers, professional customers, and market makers, offset those from firms and customers.[9] Finally, we document that firms, customers, and market makers, collectively account for 97% of the trades, leaving broker-dealers and professional customers combined with a small part of the market.

In our analysis, we want to identify the net demand of non-market-makers. According to the CBOE classification of traders, broker-dealers and market makers are both categorized as liquidity providers. Thus, in the post-migration database, end-user net demand can be identified as the negative of the combined net demand from market makers and broker-dealers. Given the low market participation of professional customers, the negative of the combined net demand from market makers and broker-dealers is almost

[8]We analyze eight overlapping years of the new and old data (from January 2011 to March 2019, containing 3,741,172 observations); all moneyness and maturities are included. We find some discrepancies between the new and old data; however, these mismatches seem negligible as the net demand of firms and customers coincides 95% of the time between the old and new datasets. We therefore conclude that the classification of firms and customers is the same across the two databases.

[9]Net demands from all trader groups sum to zero for about 98% of the records in the sample.

identical to the sum of the net demand of firms and customers. Therefore, in order to be consistent throughout the sample period, we compute end-user net demand ND_t, both in the pre- and post-2011 data. We confirm the validity of this approximation with a robustness check by taking the negative of the combined net demand from market makers and broker-dealers as a quantity measure, which we label as ND_t^*.

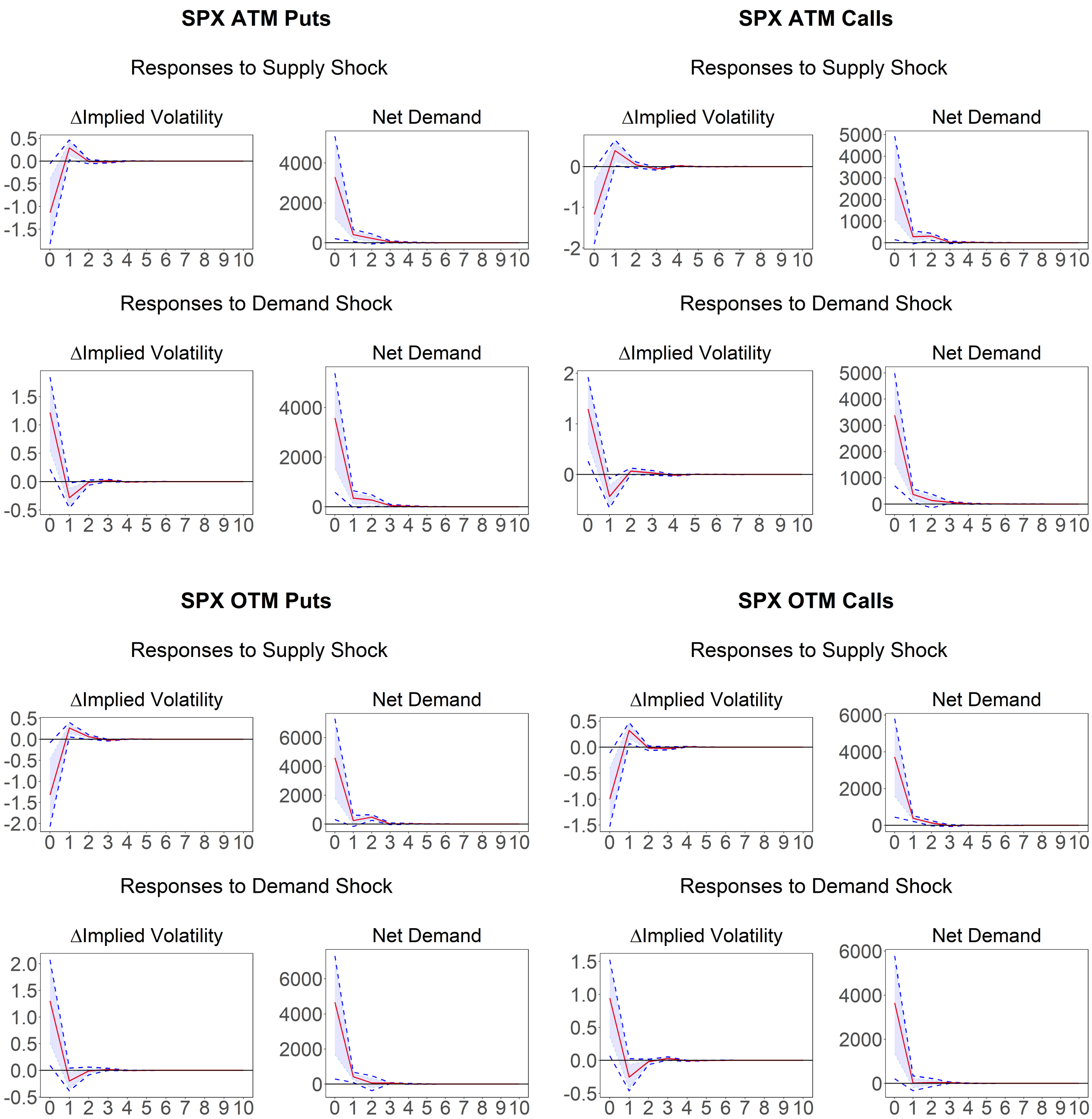
4 The Dynamic Impact of Demand and Supply

The VAR framework is also well suited for investigating the impact of demand and supply (shocks) on net option demand and changes in implied volatility over time, using the impulse response functions (IRFs) in Figure A.6. The red line in each plot reports the average response of either price or quantity to supply or demand.
The average values at lag zero in each plot in Figure A.6 correspond to the entries in the **A** matrix (averaged across retained draws): a_{11} and a_{12} are the contemporaneous responses of price to demand and supply, whereas a_{21} and a_{22} are the contemporaneous responses of quantity to changes in demand and supply.[10] Under the imposed restrictions, both types of shocks increase net demand on impact, yet the contemporaneous price impact is negative for supply and positive for demand.

The impulse response functions are very similar for the four samples, i.e. ATM puts, ATM calls, OTM puts, and OTM calls. The impacts of the shocks on net option demand and changes in implied volatility are short-lived. Net demand responses die out within a day or two. The responses of changes in option-implied volatility align with the imposed sign restrictions on impact, but these responses reverse in sign the next day and then die out quickly thereafter. This reversion in the sign of daily changes in implied volatility is consistent with the mean-reverting characteristic of volatility and with the limits-to-arbitrage hypothesis discussed in Bollen and Whaley (2004), which suggests that option price changes are due to market makers diverting from their optimal inventory levels while providing liquidity in option markets. As market makers correct their inventory the next day, prices should revert, at least partially. Our findings on price reversal following a demand shock thus support this limits-to-arbitrage hypothesis.

[10]Recall from the identification section in the paper that the slope estimates of the demand and supply curves are a_{12}/a_{22} and a_{11}/a_{21} respectively. These ratios are calculated for each retained draw of **A**. Note that the ratios of the contemporaneous IRFs in Figure A.6 do not perfectly match the slope estimates for the demand and supply curves reported in Table 3 in the paper, because the average of ratios does not equal the ratio of the averages.

Figure A.6: Impulse Responses to Demand and Supply Shocks

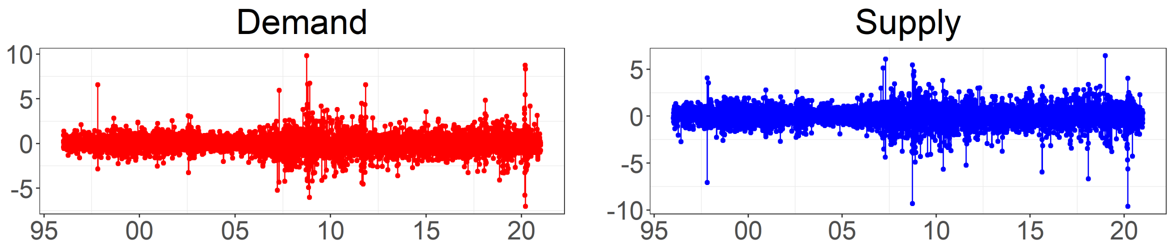


We plot the changes in implied volatility and net demand for a one-standard deviation increase in supply or demand. The mean impulse response is shown in red. The shaded area marks a pointwise 68% credible interval around the median. The dashed lines mark a pointwise 95% credible interval around the median.

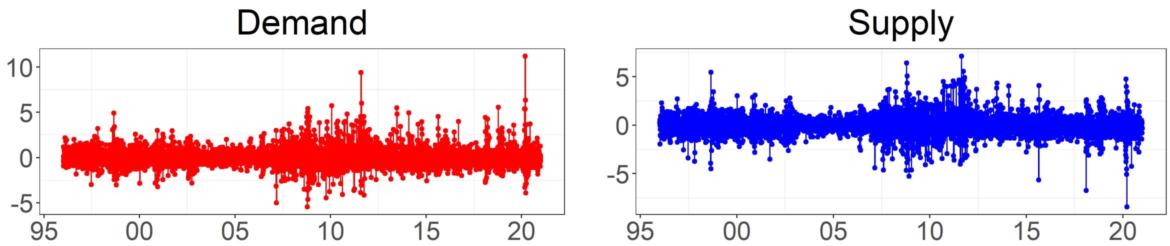
5 Other Supplementary Figures and Tables

Figure A.7: Time-Series of Daily Supply and Demand

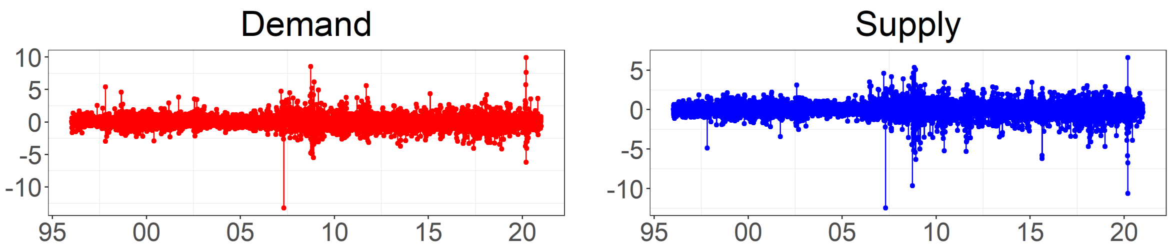
Panel A. ATM SPX Calls



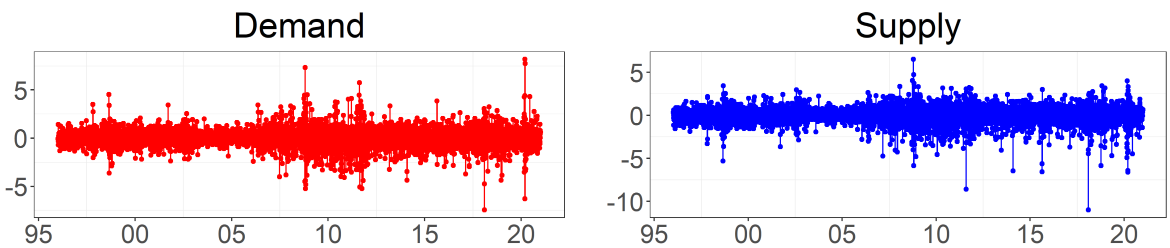
Panel B. ATM SPX Puts



Panel C. OTM SPX Calls

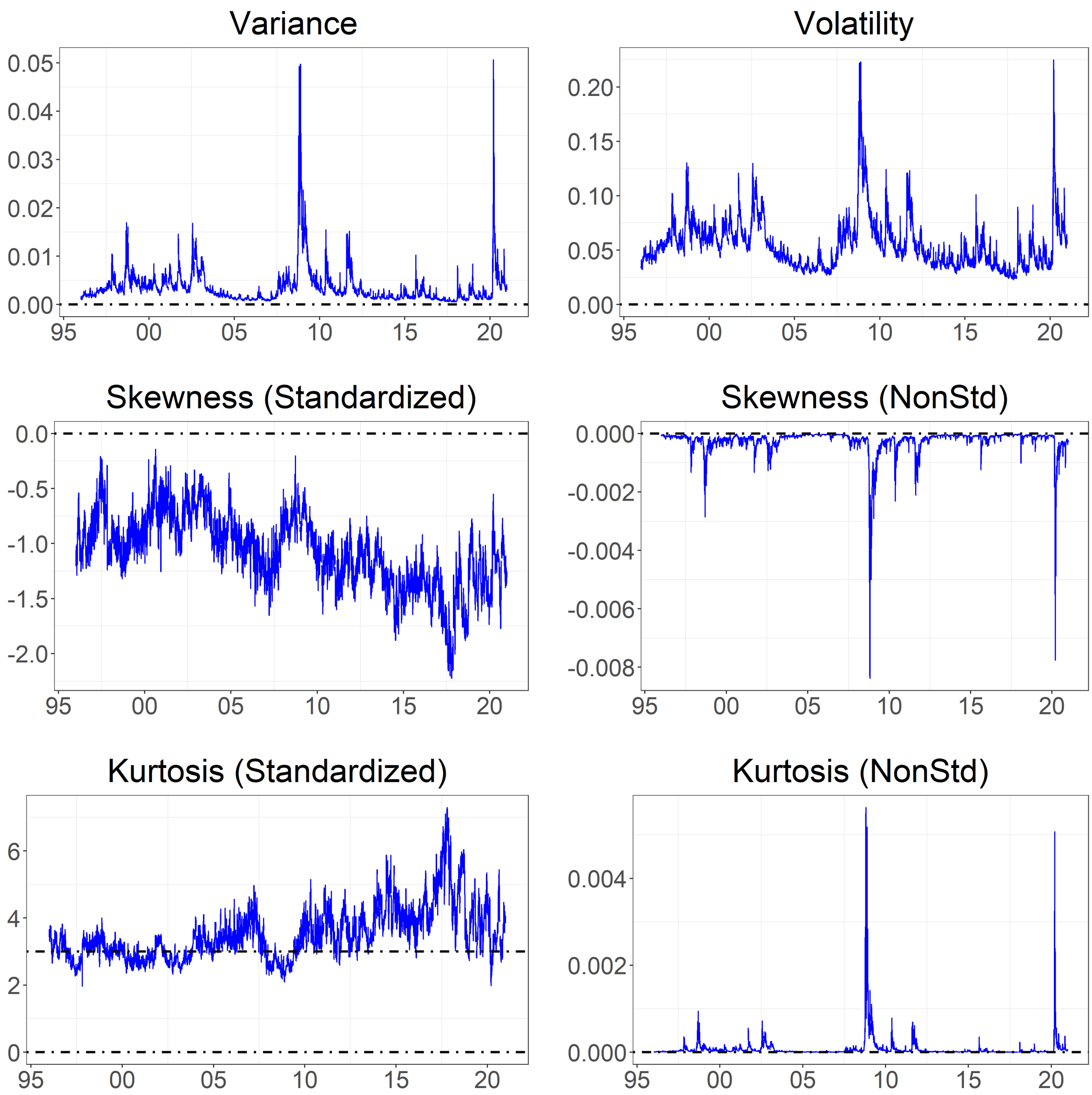


Panel D. OTM SPX Puts



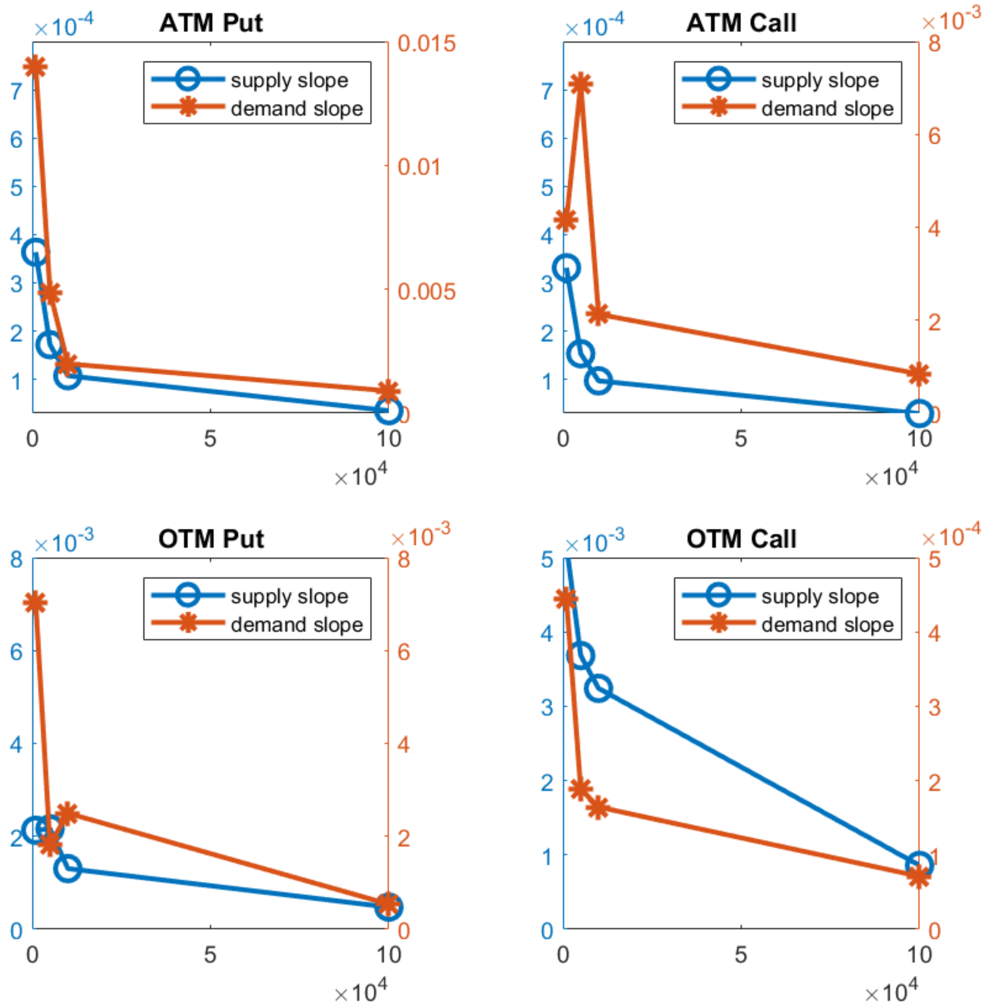
We plot the time-series of daily (latent) supply and demand in the ATM and OTM SPX call and put option markets, estimated from the signed-restricted VAR on the whole sample period (1996-2020).

Figure A.8: Daily Option Implied Moments of S&P 500 Index Returns



We plot the time-series of 30-day risk-neutral variance, volatility, skewness, and kurtosis of the S&P 500 index return from 1996 through 2020. We use the model-free methodology in Bakshi, Kapadia, and Madan (2003) to extract the 30-day option implied moments using option data available from OptionMetrics.

Figure A.9: Convergence of the Slope Estimates - Different Samples



We plot the standard deviations of the demand and supply slope estimates for different sample of options as a function of the number of parameter draws. For a given number of draws N (x-axis), we repeat the estimation procedure 100 times under different seed generators. We use the mean estimate of the slopes across the retained draws and compute the standard deviation of these 100 mean estimates. The supply slope (blue line with circle markers) is measured on the left y-axis, while the demand slope (red line with star markers) is measured on the right y-axis.

Table A.2: (Median) Demand and Supply Slopes and Price Elasticities

Panel A. Slope Estimates Based on the Sign-Restricted VAR

	Demand Slope (1)	Supply Slope (2)
ATM Puts	-37.21	31.94
	[-868.46 ; -1.05]	[4.06 ; 312.37]
ATM Calls	-42.10	36.07
	[-1311.12 ; -1.17]	[5.28 ; 275.05]
OTM Puts	-31.16	25.91
	[-656.71 ; -1.12]	[1.3 ; 705.44]
OTM Calls	-28.78	24.00
	[-339.75 ; -1.92]	[1.19 ; 728.38]

Panel B. Standardized Slopes and Price Elasticities

		1996-2020 (1)	2011-2020 (2) Using ND (a)	2011-2020 (2) Using ND* (b)	2007-2009 (3) Great Recession	2020 (4) COVID-19
			Standardized Slope Estimates			
Demand	**ATM Puts**	-1.06	-1.07	-1.07	-1.08	-0.98
	ATM Calls	-1.04	-1.02	-1.01	-1.00	-0.87
	OTM Puts	-1.08	-1.08	-1.08	-1.09	-1.07
	OTM Calls	-1.05	-1.02	-1.02	-1.05	-0.86
Supply	**ATM Puts**	0.91	0.92	0.92	0.92	0.93
	ATM Calls	0.89	0.90	0.91	0.90	0.84
	OTM Puts	0.90	0.90	0.90	0.92	0.93
	OTM Calls	0.87	0.86	0.86	0.89	0.73
Demand	**ATM Puts**	-14.81	-12.64	-12.69	-12.20	-11.70
	ATM Calls	-13.47	-10.43	-10.38	-10.42	-9.82
	OTM Puts	-15.03	-11.72	-11.73	-13.44	-11.51
	OTM Calls	-15.35	-10.78	-10.72	-11.45	-9.28
Supply	**ATM Puts**	17.24	14.69	14.73	14.35	12.30
	ATM Calls	15.73	11.81	11.50	11.63	10.23
	OTM Puts	18.08	14.05	14.06	15.94	13.26
	OTM Calls	18.41	12.84	12.72	13.51	10.95

Panel A reports the median and 95% credible interval of the supply and demand slopes for the SPX ATM and OTM call and put markets; the estimates are multiplied by 10^5. Panel B reports the standardized slope estimates and price elasticities of supply and demand for different sample periods calculated using the median.

Table A.3: Determinants of Demand and Supply – VAR with Alternative Price Proxies and VAR with Risk Factors

		Panel A Variance Risk Premium		Panel B Excess Volatility		Panel C Implied Volatility		Panel D VAR w/ Risk Factors	
Variables		Demand ε_t^D	Supply ε_t^S	Demand ε_t^D	Supply ε_t^S	Demand ε_t^D	Supply ε_t^S	Demand ε_t^D	Supply ε_t^S
$\mathbf{R}_t^M$	**ATM Puts**	-0.33***	0.27***	-0.32***	0.26***	-0.39***	0.35***	0.00031	-0.00065
	ATM Calls	-0.38***	0.3***	-0.34***	0.27***	-0.41***	0.36***	0.00066	-0.00085
	OTM Puts					-0.37***	0.41***	0.00029	-0.00079
	OTM Calls					-0.37***	0.43***	0.00047	-0.00096
$\mathbf{\Delta RV}_t$	**ATM Puts**	0.00	0.01	0.03***	-0.03***	0.04***	-0.03***	0.00002	-0.00026
	ATM Calls	0.01	0.00	0.04***	-0.03***	0.04***	-0.03***	0.00006	-0.00029
	OTM Puts					0.04***	-0.04***	0.00002	-0.00025
	OTM Calls					0.04***	-0.04***	0.00002	-0.00027
$\mathbf{\Delta Skew}_t$	**ATM Puts**	-0.21***	0.2***	-0.29***	0.27***	-0.29***	0.28***	0.00017	-0.0002
	ATM Calls	-0.16***	0.11***	-0.2***	0.16***	-0.21***	0.17***	0.00033	-0.00026
	OTM Puts					-0.3***	0.27***	0.00019	-0.00024
	OTM Calls					-0.21***	0.22***	0.00021	-0.00031
$\mathbf{\Delta Kurt}_t$	**ATM Puts**	0.23***	-0.2***	0.32***	-0.28***	0.32***	-0.29***	-0.05887	0.07029*
	ATM Calls	0.22***	-0.18***	0.29***	-0.25***	0.3***	-0.26***	0.02587	-0.04378
	OTM Puts					0.32***	-0.29***	-0.07784·	0.07977*
	OTM Calls					0.31***	-0.31***	0.04972	-0.03958
$\mathbf{\Delta TED}_t$	**ATM Puts**	0.08	0.1	1.00*	-0.92*	1.21***	-1.14**	0.00669	0.04848
	ATM Calls	-0.1	-0.38	0.72·	-1.36***	0.89*	-1.61***	-0.41462	-0.41436
	OTM Puts					1.15***	-1.12**	-0.36615	0.31261
	OTM Calls					1.39***	-1.17***	0.25207	0.23139
$\mathbf{\Delta CDS}_t$	**ATM Puts**	0.01	-0.03·	0.01	-0.03·	0.02·	-0.03*	-0.00866	-0.00702
	ATM Calls	0.04***	0.00	0.04***	0.00	0.05***	-0.01	0.01836·	0.01383
	OTM Puts					0.03*	-0.02**	0.00466	0.00138
	OTM Calls					0.03*	-0.03**	0.00314	-0.00314
$\mathbf{\Delta Wealth}_{t-1}$	**ATM Puts**	-12.58	131.20***	6.64	108.71***	20.06	94.90**	57.28*	47.06
	ATM Calls	-24.94	-98.75**	-2.91	-133.18***	-1.73	-135.37***	-11.62	-124.64***
	OTM Puts					84.91*	48.78	30.97	60.94**
	OTM Calls					-14.12	-23.96	3.02	-16.55
$\mathbf{\Delta InvRisk}_{t-1}$	**ATM Puts**	0.04	1.11.	0.11	1.21.	0.1	1.20.	0.76	0.23
	ATM Calls	-0.31	0.38	-0.26	0.41	-0.31	0.49	-0.24	0.39
	OTM Puts					1.64**	2.00**	0.69.	2.17***
	OTM Calls					0.91.	0.03	0.18	0.56
$\mathbf{\Delta BidAsk}_t$	**ATM Puts**	0.17***	-0.10**	0.21***	-0.13***	0.25***	-0.18***	0.17***	-0.11**
	ATM Calls	0.18***	-0.12**	0.20***	-0.18***	0.24***	-0.22***	0.09*	-0.07.
	OTM Puts					0.46***	-0.48***	0.32***	-0.34***
	OTM Calls					0.51***	-0.42***	0.23**	-0.12
$\mathbf{IF}_t$	**ATM Puts**	-14.40***	13.87***	-14.26***	13.43***	-18.15***	18.18***	2.09	0.16
	ATM Calls	-18.31***	14.42***	-18.50***	14.49***	-21.99***	18.76***	1.81	-1.05
	OTM Puts					-19.46***	19.75***	-0.66	-0.69
	OTM Calls					-18.31***	23.91***	1.67	0.08

We present results on univariate regressions of supply or demand in the ATM and OTM SPX call and put option markets on the realized return on the S&P 500 R_t^M, and daily changes in the following variables: realized market volatility RV_t, non-standardized risk-neutral skewness ($Skew_t$) and kurtosis ($Kurt_t$) with 30 days to maturity as in Bakshi, Kapadia, and Madan (2003), the TED spread, the CDS spread, market-maker wealth and inventory risk calculated as in Fournier and Jacobs (2020), the option bid-ask spread, and the intermediary constraint measure IF_t of He, Kelly, and Manela (2017). Demand and supply are estimated with the sign-restricted VAR using the following price variables: i) the variance risk premium (Panel A), computed as in Bekaert and Hoerova (2014) with the difference between the option implied volatility and the conditional forward looking realized volatility over the next month, ii) the excess volatility used by Gârleanu, Pedersen, and Poteshman (2009) (Panel B), computed as the difference between the option implied volatility and the risk neutral volatility estimated following the model of Bates (2006), and iii) the risk neutral implied volatility (Panel C). Panel D reports results based on the extended VAR, which includes some of the risk factors. The quantity variable is the net demand ND_t used in the baseline specification. T-statistics are computed based on Newey-West standard errors with twenty lags and *, **, and *** indicate significance at the 5%, 1%, and 0.1% level respectively. $Skew_t$, $Kurt_t$, and CDS spread are multiplied by 10^4.

Table A.4: Forecasting with Demand and Supply

Dependent Variable: R_{t+1}^{SP500}	Univariate				Demand		Supply	
Panel A. OTM Put Options								
ε_t^D	5.46				-3.43*	-3.94*		
	[1.40]				[-2.18]	[-2.28]		
ε_t^S		-7.91**					-4.30*	-2.50
		[-3.14]					[-2.37]	[-1.14]
ND_t			-0.86					
			[-0.28]					
$\Delta\sigma_t$				5.53*		0.61		7.32
				[2.26]		[0.21]		[1.54]
$\varepsilon_t^D \times D_t$					30.10**	29.76*		
					[3.08]	[2.50]		
$\varepsilon_t^S \times D_t$							-10.65.	5.72
							[-1.80]	[0.70]
D_t					4.04	3.97	4.19	4.03
					[1.12]	[1.08]	[1.16]	[1.10]
R^2 (%)	0.24	0.50	-0.01	0.87	1.89	1.99	0.70	1.25
Panel B. ATM Put Options								
ε_t^D	8.33*				3.58	0.89		
	[2.11]				[1.41]	[0.31]		
ε_t^S		-7.02.					-0.14	1.04
		[-1.71]					[-0.04]	[0.31]
ND_t			3.77					
			[0.94]					
$\Delta\sigma_t$				5.87.		7.61*		6.08
				[1.96]		[2.51]		[1.44]
$\varepsilon_t^D \times D_t$					14.5100	-1.55		
					[1.27]	[-0.16]		
$\varepsilon_t^S \times D_t$							-20.44*	-6.71
							[-2.09]	[-0.85]
D_t					6.49.	5.95	5.94	5.84
					[1.72]	[1.57]	[1.60]	[1.55]
R^2 (%)	0.49	0.32	0.01	0.80	0.85	1.22	0.98	1.25

We report results for predictive regressions using the next-day S&P500 return. The explanatory variables are: demand and supply (ε_t^D and ε_t^S) obtained from the sign-restricted VAR applied to the sample of OTM and ATM put options, the quantity variable, i.e. net demand ND_t (scaled by 10^4), and the price variable, i.e. changes in implied volatility $\Delta\sigma_t$. D_t is a dummy variable, with $D_t = 1$ when $\varepsilon_t^S \times \varepsilon_t^D < 0$ and 0 otherwise. The dependent variable is multiplied by 10^4. T-statistics computed based on Newey-West standard errors with twenty lags are in square brackets. '.', *, **, and *** indicate significance at the 10%, 5%, 1%, and 0.1% level respectively.

Table A.5: Predictability in the Chen, Joslin, and Ni (2019) Sample

	Dependent Variable: Next-Month S&P 500 Return						
	Panel A. Full Sample (1996 - 2020)						
ε_t^D	-0.06			-0.14			
	[-0.77]			[-1.32]			
ε_t^S		-0.19*			0.01		
		[-1.97]			[0.13]		
$ND_{CJN,t}$			-0.06.			-0.05	-0.05
			[-1.67]			[-1.41]	[-1.49]
$\Delta\sigma_t$							-13.10*
							[-2.46]
$\varepsilon_t^D \times D_t$				0.36			
				[1.21]			
$\varepsilon_t^S \times D_t$					-0.57***	-0.54***	-0.61***
					[-4.84]	[-5.47]	[-4.73]
D_t				-0.04	0.05	0.20	0.15
				[-0.07]	[0.10]	[0.36]	[0.30]
R^2 (%)	-0.09	2.38	1.87	0.74	7.13	8.66	10.03
	Panel B. CJN Sample (1996 - 2012)						
ε_t^D	-0.14			-0.21*			
	[-1.39]			[-2.09]			
ε_t^S		-0.26***			-0.05		
		[-3.30]			[-0.53]		
$ND_{CJN,t}$			-0.23**			-0.21**	-0.20**
			[-3.19]			[-3.17]	[-3.13]
$\Delta\sigma_t$							-6.81.
							[-1.69]
$\varepsilon_t^D \times D_t$				0.23			
				[1.19]			
$\varepsilon_t^S \times D_t$					-0.46*	-0.42**	-0.44**
					[-2.03]	[-2.61]	[-2.71]
D_t				0.39	0.53	0.78	0.72
				[0.51]	[0.68]	[1.21]	[1.16]
R^2 (%)	0.46	2.28	7.44	0.24	3.71	10.3	10.27

We present the results from regressing next month's S&P500 return on the net demand quantity variable proposed by Chen, Joslin, and Ni (2019), $ND_{CJN,t}$. We also present predictive regressions for the supply and demand from the sign-restricted VAR estimated using $ND_{CJN,t}$ as the quantity variable and $\Delta\sigma_t$ as the price variable. The sign-restricted VAR is estimated using daily data and the monthly demand and supply ε_t^D and ε_t^S are obtained by summing the estimated daily quantities over each month. D_t is a dummy that equals one in the months when demand and supply diverge on most days of the month. The market return variable is multiplied by 10^2, while the net demand variable is scaled by 10^4.

Table A.6: A VAR with Risk Variables and Robustness Tests

Panel A. Estimation of Demand and Supply Using a VAR with Risk Variables

	Demand Curve		Supply Curve			1-Step Ahead
	Slope (1)	Elasticity (2)	Slope (3)	Elasticity (4)	OLS (5)	FEVD (6)
ATM Puts	-127.98	-3.94	32.26	15.61	0.96***	5.65***
	[-436.35 ; -0.62]		[1.65 ; 170.49]		[3.85]	[36.29]
ATM Calls	-147.14	-3.40	46.93	10.66	0.73*	3.23***
	[-547.94 ; -0.83]		[1.51 ; 279.93]		[2.38]	[23.26]
OTM Puts	-54.03	-7.63	79.80	5.17	0.03	-1.02***
	[-268.71 ; -0.63]		[0.56 ; 312.16]		[0.12]	[-6.18]
OTM Calls	-74.32	-5.19	50.43	7.64	0.20	0.56***
	[-345.77 ; -0.55]		[0.72 ; 241.98]		[1.11]	[3.74]

Panel B. Robustness Tests on Various Option Maturities

	Elasticity						1-Step Ahead FEVD		
	Demand Curve			Supply Curve					
Maturity	T_1	T_2	T_3	T_1	T_2	T_3	T_1	T_2	T_3
ATM Puts	-2.44	-4.14	-10.73	8.65	10.32	32.84	9.62***	3.11***	3.99***
ATM Calls	-1.44	-4.98	-25.24	8.82	13.87	13.60	14.7***	10.22***	-0.96***
OTM Puts	-3.40	-13.10	-16.72	1.06	2.95	9.79	-0.23***	-5.94***	-1.39***
OTM Calls	-7.45	-13.05	-24.39	2.54	1.86	9.86	-6.02***	-5.04***	-0.54***

	Demand Curve		Supply Curve			1-Step Ahead
	Slope (1)	Elasticity (2)	Slope (3)	Elasticity (4)	OLS (5)	FEVD (6)
Panel C. Robustness Tests with Variance Risk Premium as Price Variable						
ATM Puts	-195.27	-0.36	50.43	1.4	2.04***	8.73***
	[-665.08 ; -0.82]		[3.08 ; 250.78]		[3.82]	[612.91]
ATM Calls	-189.98	-0.36	56.85	1.22	2.48***	9.11***
	[-800.45 ; -0.95]		[3.72 ; 276.16]		[4.48]	[629.15]
Panel D. Robustness Tests with Excesss Volatility as Price Variable						
ATM Puts	-447.43	-0.22	63.00	1.55	1.45*	7.38***
	[-895.23 ; -0.95]		[2.26 ; 360.16]		[2.38]	[590.26]
ATM Calls	-352.52	-0.28	60.32	1.66	2.89**	10.62***
	[-899.75 ; -1.02]		[4.1 ; 286.05]		[2.86]	[655.2]
Panel E. Robustness. VAR Consistent with Put-Call Parity						
ATM	-214.37	-1.99	38.43	11.09	4.18***	22.33***
	[-755.25 ; -0.79]		[5.01 ; 136.28]		[5.34]	[778.35]
All Moneyness	-131.97	-2.31	46.95	6.50	0.7*	2.53***
	[-692.11 ; -0.64]		[1.18 ; 291.68]		[2.12]	[346.77]

Panel A reports on the estimation of a VAR with risk variables. We report the mean and 95% credible interval of the supply and demand slopes for the SPX ATM and OTM call and put markets (columns (1) and (3)), the price elasticities (columns (2) and (4)), and the difference between the means of the one-step ahead forecast error variance decompositions (FEVD) of price and quantity attributable to demand versus that attributable to supply (column (6)). For comparison, column (5) presents the OLS estimates. Panel B presents the elasticities of supply and demand and the one-step ahead FEVD for samples of options with different maturities (15-45, 46-195, and 196-365 days). Panels C and D report robustness tests using different price variables. Panel E reports on the estimation of a VAR consistent with the put-call parity relation for ATM options, and on the estimation of a VAR where prices and quantities are aggregated across all options.

References

Arias, Jonas E, Juan F Rubio-Ramírez, and Daniel F Waggoner, 2013, Algorithm for inference with sign and zero restrictions, *Working Paper, Duke University.*

Bakshi, Gurdip, Nikunj Kapadia, and Dilip Madan, 2003, Stock return charachteristics, skew laws, and the differential pricing of individual equity options, *Review of Financial Studies* 16, 101–143.

Bates, David S, 2006, Maximum likelihood estimation of latent affine processes, *Review of Financial Studies* 19, 909–965.

Baumeister, Christiane, and James D Hamilton, 2015, Sign restrictions, structural vector autoregressions, and useful prior information, *Econometrica* 83, 1963–1999.

Baumeister, Christiane, and James D Hamilton, 2019, Structural interpretation of vector autoregressions with incomplete identification: Revisiting the role of oil supply and demand shocks, *American Economic Review* 109, 1873–1910.

Bekaert, Geert, and Marie Hoerova, 2014, The VIX, the variance premium and stock market volatility, *Journal of Econometrics* 183, 181–192.

Bernanke, Ben S, 1986, *Alternative explanations of the money-income correlation* vol. 25. (National Bureau of Economic Research Cambridge).

Blanchard, Olivier J, and Mark W Watson, 1986, *Are business cycles all alike?* (University of Chicago Press).

Bollen, Nicolas PB, and Robert E Whaley, 2004, Does net buying pressure affect the shape of implied volatility functions?, *Journal of Finance* 59, 711–753.

Chen, Hui, Scott Joslin, and Sophie Xiaoyan Ni, 2019, Demand for crash insurance, intermediary constraints, and risk premia in financial markets, *Review of Financial Studies* 32, 228–265.

Fournier, Mathieu, and Kris Jacobs, 2020, A tractable framework for option pricing with dynamic market maker inventory and wealth, *Journal of Financial and Quantitative Analysis* 55, 1117–1162.

Fry, Renee, and Adrian Pagan, 2011, Sign restrictions in structural vector autoregressions: A critical review, *Journal of Economic Literature* 49, 938–60.

Gârleanu, Nicolae, Lasse Heje Pedersen, and Allen M Poteshman, 2009, Demand-based option pricing, *Review of Financial Studies* 22, 4259–4299.

Hasbrouck, Joel, 1991, Measuring the information content of stock trades, *Journal of Finance* 46, 179–207.

Hasbrouck, Joel, 1993, Assessing the quality of a security market: A new approach to transaction-cost measurement, *Review of Financial Studies* 6, 191–212.

He, Zhiguo, Bryan Kelly, and Asaf Manela, 2017, Intermediary asset pricing: New evidence from many asset classes, *Journal of Financial Economics* 126, 1–35.

Inoue, Atsushi, and Lutz Kilian, 2013, Inference on impulse response functions in structural VAR models, *Journal of Econometrics* 177, 1–13.

Kilian, Lutz, and Helmut Lütkepohl, 2017, *Structural vector autoregressive analysis.* (Cambridge University Press).

Kilian, Lutz, and Daniel P Murphy, 2012, Why agnostic sign restrictions are not enough: understanding the dynamics of oil market VAR models, *Journal of the European Economic Association* 10, 1166–1188.

Sims, Christopher A, 1980, Macroeconomics and reality, *Econometrica: Journal of the Econometric Society* 48, 1–48.

Sims, Christopher A, 1986, Are forecasting models usable for policy analysis?, *Quarterly Review* 10, 2–16.

Uhlig, Harald, 2005, What are the effects of monetary policy on output? Results from an agnostic identification procedure, *Journal of Monetary Economics* 52, 381–419.

Uhlig, Harald, 2017, Shocks, sign restrictions, and identification, *Advances in Economics and Econometrics* 2, 95.