Pricing Variance in a Model with Fire Sales*

Albert J. Menkveld

March 25, 2025

Abstract

A Grossman-Stiglitz type model is proposed to solve two variance premium/VIX puzzles: Why do investors *pay* to hold realized variance risk, instead of earn a premium? And, why do they pay more in post-crisis months? The model consists of three periods. Agents are identical initially, then become heterogenous due to nontraded risk shocks and trade to hedge, and finally consume payoffs. The equilibrium provides a novel perspective on the time series of VIX, S&P500 returns, and SPY trading. The long time series, 1993 through 2024, includes several crises. These crises share a common pattern, which the model can generate endogenously.

*Albert J. Menkveld, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, Netherlands, albertjmenkveld@gmail.com. I thank Dion Bongaerts, Bernard Dumas, Sergei Glebkin, and Jun Wu, and (seminar) participants at AFML 2024 at Corvinus U Budapest, CBCMFM 2024 at the Banco de Mexico, EDHEC, Helsinki GSB, and at Luiss Rome. I am for financial support from the Dutch Research Council (NWO) for grant number 016.Vici.185.068.

Pricing Variance in a Model with Fire Sales

Abstract

A Grossman-Stiglitz type model is proposed to solve two variance premium/VIX puzzles: Why do investors *pay* to hold realized variance risk, instead of earn a premium? And, why do they pay more in post-crisis months? The model consists of three periods. Agents are identical initially, then become heterogenous due to nontraded risk shocks and trade to hedge, and finally consume payoffs. The equilibrium provides a novel perspective on the time series of VIX, S&P500 returns, and SPY trading. The long time series, 1993 through 2024, includes several crises. These crises share a common pattern, which the model can generate endogenously.

1 Introduction

Empirically, research on the pricing of realized variance has yielded two salient stylized facts. Unconditionally, it has been shown that investors pay to receive future realized variance, instead of being paid to hold this risk. The size of this premium is referred to as the variance risk premium (VRP) (Carr and Wu, 2009). It is commonly measured at a monthly frequency. More precisely, the VRP is defined as the *certain* amount investors commit to pay at the beginning of the month to receive the *uncertain* realized variance of daily returns at the end of the month. The former is referred to as implied variance (IV) and the latter as realized variance (RV). Implied variance has been shown to be approximated by VIX squared (Carr and Wu, 2006).

Conditionally, the VRP becomes elevated in the months after market crises (Bekaert and Hoerova, 2014; Choi, Mueller and Vedolin, 2017). The reason is that although both IV and RV jump during as a result of the crisis, IV decays more slowly than RV. Studies that have documented this pattern focus on understanding the dynamics of VIX.[1]

Theoretically, the pricing of variance risk is nontrivial. Consider a simple one-period model where a stock pays a stochastic dividend. Expected utility makes the investor value the stock less than the expected dividend, thus yielding a positive equity premium. This is an immediate result of Jensen's inequality. Similar reasoning does not hold for *squared* dividend. Jensen's inequality can no longer be applied, because the function of dividend is now a composition of a concave and a convex function (i.e., the utility function and squaring, respectively). Hence, the pricing of realized variance is a nontrivial manner.[2]

The extant theoretical literature has proposed a variety of models to generate a positive,

[1]Recent literature surveys on the variance risk premium and the real economy include Zhou (2018) and Dew-Becker and Giglio (2024).

[2]Appendix A illustrates this argument in a CARA-normal setting. In this setting, the risk premium for uncertain dividend is the well-known expression: half times the coefficient of risk aversion, times the total position, times the variance of dividend. The risk premium for squared dividend in this special case is zero (i.e., concavity and convexity offset each other perfectly in this case).

time-varying variance risk premium. All these models are representative agent models. Rational expectations models that can generate a positive VRP involve, for example, an agent who faces time-varying economic uncertainty (Bollerslev, Tauchen and Zhou, 2009) or jump risk (Todorov, 2010). An example of a model that can explain the elevated VRP post-crisis is a behavioral model that involves an agent who updates too slowly to changing volatility (Lochstoer and Muir, 2022).

I propose a rational expectations model that can generate the documented stylized facts, and two additional ones. These additional facts are the puzzling pattern of *higher* volume on *lower* liquidity post-crisis. These two additional facts I document based a intraday trading data for SPY, an actively traded exchange traded fund (ETF) that tracks the S&P500.[3] The sample runs from 1993, the year that VIX was launched and SPY started trading, until 2024. The extensive sample, encompassing ten crises, allows for reliable identification of crisis patterns.

The novel economic channel that is explored in the model is whether variance swaps are held to hedge against nontraded-risk shocks. This type of shocks were introduced by Lo, Mamaysky and Wang (2004) to generate trade in the risky asset. After several standard asset pricing tests, Carr and Wu (2009) conclude that (p. 1338) "variance risk is generated by an independent risk factor that the market prices heavily." In a sense, this paper studies if nontraded risk might be that risk factor. Receiving realized variance should be particularly attractive in states of the world where one is forced to sell in declining and illiquid markets, i.e., crisis periods. To assess whether this liquidity demand could explain a positive variance risk premium requires a model that endogenizes demand and supply of liquidity.

The structure of the model is as follows. A variance swap is added to the baseline Vayanos and Wang (2012) model, which is a Grossman-Stiglitz type model. Information is symmetric

[3]Konstantinidi and Skiadopoulos (2016) find that among four candidate models, the one with trading activity best predicts the VRP out of sample. This suggests that there is an important role for trading in a model that prices variance.

across agents in this baseline model. A fraction of agents experience the nontraded risk shock, which generates demand for the risky asset, but also an appetite for holding the variance swap. More specifically, the model consists of three stages. In the first stage, all agents are homogeneous and can buy or sell variance swaps. In the second stage, a fraction of agents suffers a nontraded risk ("endowment") shock, and can trade the risky asset to hedge.[4] The remaining agents endogenously become liquidity suppliers. In the final period, payoffs realize and agents consume. The key result that the model delivers is that variance swaps serve to mitigate the utility cost due to nontraded risk shocks. In equilibrium, implied variance exceeds expected realized variance, because the benefit to those who suffer the shock exceeds the cost to those who do not. It therefore provides the following insights on the following four main questions:

1. Does nontraded risk generate a positive VRP?

2. If so, what drives the size of the premium?

3. Can nontraded risk explain an elevated VRP post-crisis?

4. Can it explain more volume on less liquidity post-crisis?

The last question sets the nontraded risk model apart from existing models that feature a representative agent. These models cannot talk to trading patterns for the simple reason that there is no trading in such models. The model equilibrium yields relatively simple closed-form expressions for asset returns, market liquidity, and the variance swap. It provides the following insights on the four questions.

First, the model implies that the VRP is nonnegative. The reason is that variance swaps mitigate the utility cost of nontraded risk shocks.

[4]The market is incomplete in the sense that the state that determines who experiences the shock cannot be traded.

Second, the model yields a relatively simple analytical expression for the VRP, which allows for further analysis on what drives it. It turns out that the VRP increases in risk aversion and in the variance of the dividend. These results are intuitive. What is less intuitive is that the VRP also increases in the fraction of agents who experience the nontraded risk shock. The latter allows for explaining time variation in the VRP beyond what is driven by time-varying volatility (i.e., beyond "GARCH effects"). This is needed to explain an elevated VRP post-crisis, which is where I turn next.

Third, the model can explain that the VRP is slow to decay in the months after a crisis, in particular when judged against the relatively fast decay in variance. The model explains it through the level of nontraded risk. Either a higher fraction of agents experience nontraded risk shocks, or agents experience shocks that are larger in size. That is, elevated nontraded risk can be generated on the extensive or on the intensive margin, respectively.

Fourth, it turns out that the volume pattern can tell the extensive and intensive margin effects apart. If more volume is accompanied by less liquidity supply, then this is evidence of an intensive margin effect. If more agents experience the shock and demand liquidity, then fewer agents are left to supply liquidity and supply becomes more expensive. If, on the other hand, more volume is accompanied by *more* liquidity supply, then this is evidence of an intensive margin effect. One can generate this pattern by fewer agents experiencing larger shocks, because, in this case, there are more agents left to supply liquidity.

Calibration. To illustrate the model, I calibrate it to match the empirical pattern around crises. A sample with daily data on VIX and the S&P500 index (SPX) is complemented with intraday trade data on SPY. Trading due to nontraded risk shocks is identified by the correlation of daily SPY net volume and SPX return. Net volume is defined the sum of signed trade sizes, where the sign is determined by who initiates the trade. It is positive for buyer-initiated trade and negative for seller-initiated trades. The empirical analysis makes

the paper self-contained. Well documented patterns are reproduced and, as discussed, new trading patterns are added.

The calibration results show that the high post-crisis VRP is driven by fewer agents experiencing larger shocks. These agents value the variance swaps so much that the VRP is 238% higher post-crisis. SPY volume is 86% higher and the bid-ask spread is 18% higher. Liquidity supply, therefore, is lower post-crisis. All else equal, the supply should have been 165% lower, because post-crisis variance is 165% higher. Therefore, to observe a drop in supply of only 18% must mean that more agents become supplier, and therefore fewer are demander. To then match an increase in volume needs larger shocks per demander. This is why the elevated VRP is driven by an intensive margin effect.

Related literature. The nontraded risk shocks in the model endogenously generate positive covariation between net volume and permanent (log) price changes. This covariation exists in the sample between daily SPY net volume and permanent SPX changes. It is sizeable, because the flow-correlated component makes up about a third of the variance of permanent SPX changes, both before and after a crisis.

The observation that SPY flow correlates with permanent SPX returns relates to a literature on imperfectly elastic markets. Evans and Lyons (2002) document a similarly strong correlation for foreign exchange trading. They interpret the finding as due to "portfolio shifts" by some investors that need to be absorbed by other risk-averse investors. Recently, Gabaix and Koijen (2024) propose the inelastic market hypothesis that relates inelasticity to limited flexibility of institutions when responding to asset price fluctuations. Risk aversion in the Grossman-Stiglitz model could be reinterpreted as such limited flexibility.

The way nontraded risk shocks in the model drive liquidity demand further relates to a literature on "leverage-induced flow." Leverage can force investors to sell in bear markets, but it also allows them to buy in bull markets. In this spirit, Vayanos (2004) models investors as

fund managers, subject to withdrawals when performance falls below a threshold. Empirically, Jiang (2024) provides some direct evidence by showing that stocks held by highly leveraged hedge funds exhibit stronger negative skewness.

Finally, note that my finding of more agents experiencing nontraded risk shocks, i.e., more leverage-induced trading, increases the VRP sheds new light on one of the main findings in Bekaert and Hoerova (2013). They document that the VRP declines in response to lax monetary policy. Such policy might lead investors to lever up, thus leading to more agents experiencing nontraded risk shocks and, therefore, a higher VRP.

The rest of the paper is organized as follows. Section 2 presents the sample and establishes the previously mentioned empirical patterns. Section 3 presents the model and derives various theoretical results. Section 4 calibrates the model to match the VRP and trading patterns around crises. Section 5 concludes.

2 Empirical patterns

This section presents descriptive empirical results. It confirms the variance patterns that have been documented in the literature, both realized and implied variance. It does so for a longer sample that includes ten crisis periods. It complements these pricing patterns with trading patterns based on intraday trade data.

2.1 Sample

The sample is picked to illustrate how the model can produce the empirical crisis patterns. It requires data on a stock index, trading in this index, and the price of a variance swap based on it. A triplet that meets this requirement is the S&P500 index (SPX), trade data for SPY which is an exchange traded fund (ETF) that tracks the S&P500, and VIX. SPY is one of the most actively traded ETFs and can therefore be expected to be representative of index

trading. VIX squared is known to be a reasonable approximation of the price of a variance swap (Carr and Wu, 2006).

The sample starts on February 4, 1993, and ends on September 13, 2024. The start of the sample coincides with the start of SPY. 1993 was also the year when VIX was officially introduced at the Chicago Board Options Exchange (CBOE). The sample can, therefore, not be extended further back in time.

The sample is sourced from Wharton Research Data Services (WRDS) TAQ, CBOE, and Nasdaq. The following time series are constructed based on these data sources:

• *SPX*	End-of-month (EOM) value of the S&P 500 index.
• *Realized variance (RV)*	EOM sum of squared daily SPX returns.
• *Implied variance (IV)*	Beginning-of-month (BOM) squared VIX.
• *Realized VRP*	BOM implied variance minus EOM realized variance.
• *Volume*	EOM SPY volume in shares.
• *Quoted bid-ask spread*	EOM time-weighted quoted spread based on official national best bid-offer (NBBO).
• *Quoted depth*	EOM average of time-weighted quoted depth best bid and ask based on NBBO.

Details on the construction of all variables used in the analysis are in Appendix B.

[Table 1 about here.]

2.2 Analysis

This section presents various empirical results. First, it analyzes the IV, the RV, and the VRP time series to reproduce stylized facts. Second, it produces additional results based on trade data for SPY. This evidence is novel. Third, it zooms into crisis periods to characterize a pre- to post-crisis pattern in IV, RV, VRP, and trading in SPY.

2.2.1 Time series for IV, RV, and VRP

Table 1 presents summary statistics for the three key variance-based variables: implied variance, realized variance, and the variance risk premium. These statistics serve as a "reality check," because they can be compared to the ones reported in Carr and Wu (2009, Table 2). My sample extends theirs in the sense that it covers 32 years (1993-2024) instead of the eight years they analyze (1996-2003). The summary statistics lead to the following observations. First, the average realized VRP is positive. That is, the average end-of-month realized variance exceeds the beginning-of-month implied variance (i.e., squared VIX). Therefore, those who shorted the variance swap at the beginning of the month earn a premium. This premium is sizable, because it amounts to $104/342 = 30\%$ annually.

[Figure 1 about here.]

Second, both IV and RV exhibit excess kurtosis and positive skewness. These statistics are driven by occasional crisis periods, when both types of variance jump to extreme levels. The realized VRP also exhibits fat tails, but its skew is negative instead of positive. This is driven by the crisis months in which realized variance jumps unexpectedly, far exceeding beginning-of-month implied variance. As a result, the realized VRP in those months is extremely negative. Figure 1 plots all these time series and thus illustrates the excess kurtosis and skewness. The vertical lines in the graph identify ten crisis periods in the sample period. The bottom graph plots the realized VRP and thus illustrates the negative jumps in crisis months.

Third, all three series are positively autocorrelated. The autocorrelation is particularly strong for implied variance: 0.76. The autocorrelation for realized variance is 0.49. The wedge between these two can be explained by the nature of the spikes in RV and IV. Figure 1 shows that the spikes are higher for RV, and that they drop more quickly. Because short and extreme bursts push the autocorrelation of a series to zero, this force is stronger for RV than

for IV. The spikes decaying more quickly for RV, by the way, is first evidence of the pattern documented in the literature that the VRP remains elevated in post-crisis periods. I will revisit this issue when zooming in on the crisis periods.

All these observations are very similar to those documented by Carr and Wu (2009). The premium for shorting variance is higher in their sample, but skewness, kurtosis, and autocorrelation are somewhat higher in my sample. Importantly, the time series features are the same in both samples.

[Figure 2 about here.]

2.2.2 Time series SPY trading

Figure 2 depicts how SPY trading evolved throughout the sample. It plots volume along with two standard liquidity supply measures: the bid-ask spread and depth at the best quote. Volume has grown steadily since the inception of SPY and peaks at around 2015. It seems to stabilize at approximately one billion shares per month since then. In contrast, liquidity supply is monotonic throughout the sample. Both spread and depth decline steadily.

The figure further suggests that there is a particular pattern for volume and liquidity around events. Volume seems to jump in crisis months to remain elevated in the months after. This volume increase is accompanied by worse liquidity in the sense of a higher spread and lower depth. Whether these crisis patterns are real or merely noise, is analyzed next.

[Figure 3 about here.]

2.2.3 Crisis patterns

Figure 3 plots the various variance and trade quantities around the ten identified crises. This done by bucketing the quantities by month relative to the crisis, and then computing the average for each bucket. The top three plots show how IV, RV, and VRP develop in crisis

periods. RV jumps sharply in the crisis month, but declines quickly to the pre-crisis level in the three months after. IV, on the other hand, jumps in the month after the crisis, and decays relatively slowly to the pre-crisis level. Note that the one-month delayed jump is an artefact of IV being a *beginning-of-month* variable, as opposed to RV, which is an end-of-month variable.[5] The different levels of decay explain why the VRP jumps in the month after the crisis, and remains elevated in subsequent months. These patterns are statistically significant as judged by the confidence intervals.

The bottom three graphs in the figure illustrate the trading patterns. Volume jumps in the crisis month and remains elevated afterwards. The effect is not only statistically significant, it is also substantial in economic terms. Volume more than doubles in the crisis month and remains about one hundred percent higher subsequent months. The liquidity measures also show significant patterns, both in statistical and in economic terms. The bid-ask spread jumps by about 20% and remains more than 10% higher in post-crisis months. Depth at the best quotes drops by almost 50%, and remains depressed by more than 25% in the months after.

3 Model

The variance swap is priced in an equilibrium model with endowment shocks. It is a Grossman-Stiglitz type model. More specifically, the proposed model is the symmetric information variant that is used in Vayanos and Wang (2012) (VW12). The notation of VW12 is used for convenience. In a nutshell, the proposed model extends VW12 in two ways. First, all agents receive a dividend signal in the intermediate round, which creates a need to unwind or enlarge positions (due to, for example, leverage). Second, the financial market is extended with a derivative, a variance swap, which enables me to study equilibrium pricing of this

[5]The rationale for this timing throughout the paper is that one-month variance swaps are entered into at the start of each month, and paid out by the end of it.

important derivative in a relatively standard setting.

3.1 Model primitives

[Figure 4 about here.]

Figure 4 summarizes the model. It features three periods, indexed by $t \in \{1, 2, 3\}$. The financial market consists of a risky asset and a variance swap. The risky asset is in supply of $\bar{\theta}$ shares that pay v units of a consumption good in Period 3, where v is Gaussian with mean μ_v and variance σ_v^2. The price of this asset at time t is p_t.[6]

The variance swap is in zero net supply. It trades in Period 1 and pays off the realized variance of the risky asset in the final period:

$$x_1 = E\left(V\right), \tag{1}$$

where[7]

$$V = \sum_{\tau=2}^{3} \left(p_\tau - p_{\tau-1}\right)^2. \tag{2}$$

The price of this swap, y_1, is such that the Period 1 investor is indifferent between receiving a (deterministic) payment y_1, or receiving the (stochastic) realized variance at Period 3:

$$y_1 = E\left(V \times m_3\right), \tag{3}$$

where m_3 is the (state-dependent) marginal utility of an agent in the final period.[8] This equality that pins down the price of swap is a direct result of the derivative being in zero net

[6] All prices are expressed in units of Period 3 consumption good. In a more extensive model, this is the outcome of introducing a riskless asset that is used as numéraire for all asset prices (e.g., VW12). We prefer to keep the presentation brief.

[7] Price differentials in this model are henceforth referred to as returns. All empirical analysis in this manuscript uses log prices, so that differentials can indeed be interpreted as price returns.

[8] Note that the expressions for x_1 and y_1 do not feature $E_{t=1}(m_3)$ in the denominator. The reason is that prices are expressed in terms of the Period 3 consumption good as discussed in footnote 6.

supply and agents being homogeneous *ex-ante*. An equilibrium expression for y_1 is obtained when solving the model. Finally, note that the swap does not trade in Period 2. This is in line with the asset pricing literature that is focused on the payoff to variance swaps entered into at the beginning of the month, and paid out at the end of it (e.g., Carr and Wu, 2009).

Agents. The model features a measure one of agents. Agents are risk-averse and derive utility from consumption in Period 3. Their utility over the consumption good is exponential:

$$U(c_3) = -\exp\left(-\gamma c_3\right), \tag{4}$$

where $\gamma > 0$ is the coefficient of absolute risk aversion. Agents are identical in Period 1 and are, therefore, endowed with the per-capita supply of the asset and the derivative.

Trade is generated in Period 2 as follows. In this period, agents observe part of the dividend. They will learn the remainder in Period 3. Total dividend can therefore be written as:[9]

$$v = v_2 + v_3, \tag{5}$$

where

$$v_2 \coloneqq E_{t=2}\left(v\right), \quad v_2 \sim N\left(\mu_v, \sigma_{v2}^2\right), \tag{6}$$

and

$$v_3 \sim N\left(0, \sigma_{v3}^2\right), \quad \sigma_{v3}^2 = \sigma^2 - \sigma_{v2}^2, \tag{7}$$

where

$$\sigma_v^2 = \sigma_{v2}^2 + \sigma_{v3}^2. \tag{8}$$

After they receive this signal, a fraction κ of the agents learns that they will receive an

[9] One way to generate this structure is to let agents receive a noisy signal on the dividend in Period 2 as in Appendix C.

endowment shock ($z \times v_3$) of the consumption good in Period 3 with:

$$z \sim N(0, \sigma_z^2), \quad \text{Corr}\,(v_2, z) = -1, \quad \text{Corr}\,(z, v_3) = 0. \tag{9}$$

All agents observe the realization of v_2, and therefore z, in this intermediate period, but v_3 remains unknown to them until the final period. The agents who will experience the endowment shock ($z \times v_3$) can hedge it by trading the risky asset. For $z > 0$, for example, these agents can hedge this risk by selling the risky asset. These agents, therefore, initiate trades in Period 3 and become liquidity demanders. The remaining agents accommodate these trades. They become liquidity suppliers.

The perfectly negative correlation between v_2 and z could be interpreted as leverage-induced trading. Bad news on the dividend in the intermediate period coincides with positive z, which makes the shocked agents sell the risky asset to hedge the increased exposure. One might say these investors participate in fire sales as they effectively become forced to sell in a bearish market. Likewise, good news allows these investors to lever up and buy the risky asset.

CARA-normal models have become workhorse models in the literature, because they often yields tractable results. The same is true for the proposed model. With normality, however, the endowment shock ($z \times v_3$) can take extremely large negative values, which can make expected utility infinitely negative. To avoid this corner, VW12 impose the following condition on the variance of v_3 and z (c.f., VW12, Eqn. (1.2)):

$$\gamma^2 \sigma_v^2 \sigma_z^2 < 1. \tag{10}$$

3.2 Equilibrium quantities

This section presents equilibrium expressions for the quantities in the flow chart at the bottom of Figure 4. They are derived working backwards from the final period of the model. All proofs are in Appendix D.

Assumption 1 ***(Zero intercept.)*** *To focus the analysis on risk transfer only and thus avoid nonzero intercept terms in returns, in the remainder it is assumed that* $\bar{\theta} = \mu_v = 0$.

Assumption 1 allows us to ignore intercept terms in the pricing of variance (as will be discussed below, following Lemma 2). This keeps expressions clean and, thereby, maximizes economic insight. Note further that Assumption 1 is a relatively harmless one for daily returns as the squared *expected* return is an order of magnitude smaller than the *expected* squared return.[10] With Assumption 1, the security becomes a vehicle for investors to transfer *only risk*. For completeness, the proofs in Appendix D feature closed-form expressions for the general case of nonzero $\bar{\theta}$ and μ_v.

3.2.1 Period 3

The final period quantities are trivial. The price of the risky asset equals the dividend realization: $p_3 = v_2 + v_3$. The payoff to being long the variance swap equals the squared realized price differentials as per (2).

3.2.2 Period 2

The most important feature of Period 2 is trading in the risky asset. This trading yields expressions for the following three quantities: A market-clearing price p_2, net volume q_2 which is equal to $\Delta\theta_2^d$, and price impact λ_2. Since liquidity demanders initiate trades, their

[10] Remember that realized variance in the empirical analysis is computed as the sum of squared daily returns.

demand informs the sign of net volume: negative if they are selling and positive if they are buying. A standard illiquidity measure is the "Kyle λ." It is the regression coefficient that results from regressing price change on net volume (i.e., $(p_2 - p_1)$ on q_2).

Lemma 1 ***(Period 2 equilibrium quantities.)*** *This lemma presents equilibrium quantities for trading in the risky asset in Period 2:*

$$p_2 = v_2 - \gamma \sigma_{v3}^2 \kappa z, \qquad \textit{(Price)} \tag{11}$$

$$q_2 = \Delta\theta_2^d = -\kappa\left(1-\kappa\right) z, \qquad \textit{(Net volume)} \tag{12}$$

$$\lambda = \frac{\gamma \sigma_{v3}^2}{1-\kappa}. \qquad \textit{(Illiquidity)} \tag{13}$$

The results in Lemma 1 are relatively straightforward. The price equation shows that p_2 is equal to the expected dividend (v_2 is observed before trading) minus a discount to compensate investors for the remaining dividend risk due to v_3. The net volume equation shows that it scales negatively with z, because it is hedging demand. The size of the scaling factor reaches its maximum at $\kappa = 1/2$. The intuition is that values other than a half creates an imbalance in the presence of demanders and suppliers. The further κ is below a half, the fewer demanders there are relative to suppliers, which constrains trade opportunities. The same holds for values above a half, but now driven fewer suppliers relative to demanders. This all leads to the interesting result that an increase in the fraction of agents who seek liquidity (κ) leads to more volume, but only up to a point, after which volume declines.

3.2.3 Period 1

Before being able to price the variance swap, we need the Period 1 price of the risky asset.

Lemma 2 ***(Period 1 price risky asset.)*** *The price of the risky asset in Period 1 is:*

$$p_1 = 0. \tag{14}$$

Lemma 2 finds that the price of the risky asset is zero in Period 1. Since in Period 1 the expected price in later periods is also zero, this result backs up the earlier claim that intercept terms can be ignored when pricing variance (see discussion following Assumption 1). This result critically depends on the simplifying assumption that $\bar{\theta} = \mu_v = 0$. The expression for the general case along with a short discussion of the various terms is in Appendix D.

With these results in place, it is time to state the main theoretical results. They center on the price of the variance swap at time zero. Recall that the expression for realized variance is:

$$V = (p_2 - p_1)^2 + (p_3 - p_2)^2 . \tag{15}$$

The long side of the swap gets paid this variance in Period 3. In return, this side needs to pay the short side of the swap a fixed amount at maturity, which is referred to as the price of the swap: y_1. Since the swap is in zero net supply, its price in Period 1 must be equal to the reservation value of the agent. This value is well defined, because all agents are identical in Period 1.

Proposition 1 ***(Period 1 price variance swap.)*** *The Period 1 price of the variance swap is:*

$$\begin{aligned} y_1 = \kappa y_1^d + (1-\kappa)\, y_1^s = \\ = \underbrace{\left(\kappa A^2 + (1-\kappa)\, B^2\right)}_{\substack{\textit{Inflator flow-correlated} \\ \textit{return variance}}} \times \underbrace{\left(\sigma_{v2} + \gamma \sigma_{v3}^2 \sigma_z \kappa\right)^2}_{\substack{\textit{Flow-correlated} \\ \textit{return variance}}} + \\ + \underbrace{\left(\kappa A + (1-\kappa)\, B\right)}_{\substack{\textit{Inflator flow-uncorrelated} \\ \textit{return variance}}} \times \underbrace{\sigma_{v3}^2}_{\substack{\textit{Flow-uncorrelated} \\ \textit{return variance}}} , \end{aligned} \tag{16}$$

with

$$A = \frac{1}{1 - \gamma^2 \sigma_{v3}^2 \sigma_z^2 \,(2-\kappa)\,\kappa} \geq 1 \tag{17}$$

and

$$B = \frac{1}{1 + \gamma^2 \sigma_{v3}^2 \sigma_z^2 \kappa^2} \leq 1. \tag{18}$$

The following inequality characterizes the two inflators:

$$\kappa A^2 + (1 - \kappa) B^2 \geq \kappa A + (1 - \kappa) B \geq 1. \tag{19}$$

Proposition 1 presents the closed-form expression for the price of the variance swap in Period 1. The first line of (16) merely states that, for an agent in Period 1, it is equal to his expected utility, whereby with probability κ he will receive an endowment shock and his expected utility of being long the swap is y_1^s, and with probability $(1 - \kappa)$ he does not receive such shock and endogenously becomes liquidity supplier whose expected utility of being long the swap is y_1^d.

[Figure 5 about here.]

The next two lines in (16) express the value of the swap in terms of the model parameters. The expression is intuitive in the sense that implied variance is equal to expected realized variance, whereby the two components of the latter are inflated by well-defined factors. The two components are one that is correlated with net volume, i.e., with z, and one that is orthogonal to it. The inflators for these components look similar. The only difference is that the inflator for the flow-correlated term involves squared factors. The proposition states that both inflators are weakly larger than one, and therefore true *inflators*. It further states that the flow-correlated inflator weakly dominates the flow-uncorrelated one. Figure 5 illustrates this finding.

The intuition for the flow-correlated inflator being the strongest one is that the positive correlation between return[11] and flow amplifies the utility cost of being shocked. These

[11] More precisely, the positive correlation between v_2 and net volume, which is the immediate result of the perfectly negative correlation between v_2 and z (see (9) and the discussion following it).

shocked agents not only have to stomach the additional risk associated with the endowment shock, they also receive the wrong shock at the wrong time in the following sense. They need to sell in a bear market and buy in a bull market. That is, hedging the endowment shock makes them sell after a negative v_2 realization, and makes them buy after a positive one. Selling after the negative shock is particularly costly to them. This is why the flow-correlated inflator is stronger than the noncorrelated one.

The results in Proposition 1 imply that the variance risk premium cannot be negative. The model therefore delivers nonnegativity of the VRP, which is a nontrivial result as it does not follow from first principles. As argued in the introduction, Jensen's inequality cannot be invoked to sign the premium that risk-averse investors require for holding *squared* dividend. The nonnegativity result that immediately follows from Proposition 1 is therefore an important result in explaining why empirical studies consistently find a positive variance risk premium. To emphasize its importance, the result is stated in the following corollary:

Corollary 1 ***(Variance risk premium nonnegative.)*** *The variance risk premium is nonnegative:*

$$y_1 - x_1 \geq 0. \tag{20}$$

To develop more insight into what drives the price of the variance swap, it is worth exploring whether partial derivatives can be signed. The next proposition develops three useful results in this respect.

Proposition 2 ***(Variance risk premium monotonicity.)*** *The variance risk premium, $y_1 - x_1$, increases monotonically in*

- *the fraction of agents who receive the endowment shock (κ),*
- *the level of risk aversion (γ), and*
- *the size of the endowment shock (σ_z^2).*

Proposition 2 finds that the variance risk premium increases monotonically in κ, in γ, and in σ_z^2. The latter two are intuitive in the sense that more risk to be hedged, or a higher risk aversion, make the variance swap more valuable in terms of hedging against endowment shock uncertainty, and the VRP therefore rises.

[Figure 6 about here.]

The result that the VRP increases in κ is a nontrivial one. Figure 6 illustrates why this is the case. It plots the value of the variance swap as a function of κ. It plots this value unconditionally, and conditional on agent type, i.e., liquidity demander (shocked agent) or liquidity supplier. The unconditional value is a weighted average of both conditional values (see (16)). The graph further plots expected realized variance, which is equal to:

$$x_1 = \underbrace{\left(\sigma_{v2} + \gamma\sigma_{v3}^2\sigma_z\kappa\right)^2}_{\substack{\text{Flow-correlated} \\ \text{return variance}}} + \underbrace{\sigma_{v3}^2}_{\substack{\text{Flow-uncorrelated} \\ \text{return variance}}} . \tag{21}$$

Note that x_1 increases in κ due to an increase in the (transitory) price pressure which pays liquidity suppliers for their supply.

The graph shows that liquidity demanders value the swap more than expected realized variance: $y_1^d \geq x_1$. Liquidity suppliers value it less: $y_1^s \leq x_1$. The reason is that the payoff of the swap correlates positively with the cost liquidity demanders experience due to the nontraded risk shock. The states of the world in which realized variance is high Period 3 are the states in which liquidity demanders pay a lot to hedge their nontraded risk. At the same time, it is in these states where liquidity *suppliers* earn most. The payoff of the swap therefore serves as a hedge for demanders, but as an amplifier for suppliers. This explains why $y_1^d \geq x_1 \geq y_1^s$. The net effect of these two forces on y_1 is nontrivial. To prove that the net effect is positive is, therefore, a substantial result.

4 Calibration

To illustrate how the model offers an explanation for otherwise puzzling patterns, I apply it to pre- and post-crisis trading. Doing so allows me to address the two remaining overarching questions. Can the model rationalize a jump in the VRP post-crisis? And, can it rationalize *more* trading on *less* liquidity post-crisis? And, more importantly, what are the economic channels that cause these patterns?

To calibrate the model to the observed patterns, the parameters are picked to match the following four key moments in the data:

1. Pre-crisis VRP,
2. Post-crisis VRP,
3. Relative change in illiquidity from pre- to post-crisis,
4. Relative change in volume from pre- to post-crisis.

These four moments are used to infer the nature of nontraded risk pre- and post-crisis. More specifically, they are used to calibrate the pre- and post-crisis fraction of agents who experience the shock (κ) and the pre- and post-crisis size of the shock (σ_z). In sum, these four parameters are calibrated to match the four moments. Note that a perfect match is not guaranteed, because the four empirical moments might lie outside of the space generated by the model-implied moments.

The calibration further needs to the following information: the pre- and post-crisis realized variance, decomposed in a flow-correlated and a flow-uncorrelated part, and the level of risk aversion. The latter is often assumed to be in the range from one to five, so I pick $\gamma = 3$. This level implies a reasonable risk premium for holding the market portfolio.[12]

[12] A CARA investor requires a risk premium of $0.5 \times 3 \times 0.20^2 = 6\%$ for holding a market portfolio with a volatility of 20%. These values are in line with the risk and return on the US equity portfolio in recent decades.

The calibration relies critically on the monotonicity results of Proposition 2. Monotonicity ensures that, if there is a match, then there is a unique set of parameters that delivers this match. The details of the calibration procedure are in Appendix E.

[Table 2 about here.]

Table 2 presents the results of the calibration. Panel (a) quantifies the pattern to be matched, as depicted in Figure 3. The VRP jumps by a factor of 16 from the pre- to the post-crisis period, from 13.9 to 314.5 percentage squared. The NBBO spread increases by 18% and volume increases by 86%.

The calibration further takes as input the level of the two components of realized variance: the flow-correlated and the flow-uncorrelated component. This additional information is in panel (c). Realized variance increases from 364 to 964 percentage squared, an increase of 165%. The VRP increases, because implied variance increases by a larger factor: $1278/378-1 = 238\%$. The flow-correlated component accounts for 33% of the total variance pre-crisis. It drops slightly to 30% post-crisis. Finally, panel (a) shows that the model is indeed able to match the empirical patterns perfectly.

Panel (b) presents the parameter values that create the perfect match. The fraction of agents who experience a nontraded risk shock drops from 78% pre-crisis to 48% post-crisis, a drop of 38%. Therefore, there are fewer agents who experience a shock, but the shock they experience increases in size. This is evident from the standard deviation of z, which increases from 1.55 to 2.01, an increase of 29%.

The calibration yields the following insights. First, all else equal, an observed (realized) variance increase of 177% implies an illiquidity increase of 177% as per (13).[13] Against this benchmark, an *actual* increase of only 18% in the bid-ask spread suggests that other factors must have caused the market to become more liquid. The model achieves this by reducing κ

[13]More specifically, the equation shows that illiquidity increases proportionately in the flow-uncorrelated return variance, which increases in the data by $(0.70 \times 964)/(0.67 \times 364) - 1 = 177\%$.

(see, again, (13)). This means that the mass of liquidity demanders drops, but it also means that the mass of liquidity suppliers increases (as they sum to one). Less demand and more supply result in the observed relative decline in illiquidity.

Second, fewer demanders would imply a volume decrease, but by increasing the shock size the model is able to match the observed volume increase. Since volume is proportional to the standard deviation of net demand, one might expect *less* volume post-crisis for the following reason. All else equal, 38% fewer agents (κ) each experiencing a 29% larger shock (σ_z) would imply a volume change of $(1-0.38)\times(1+0.29)-1=-20\%$. The all-else-equal clause, however, is violated when trading is endogenized, as it is in the model. The increase in liquidity suppliers make the demanders demand more. On the extensive margin, fewer agents are shocked, but on the intensive margin, their shock is larger *and* they demand more liquidity because it is cheaper. These two effects result in an endogenous volume increase that matches the observed increase of 86%.[14]

Third, the (transitory) price pressure that pays liquidity suppliers for their supply, increases pre- to post-crisis. It increases both in an absolute and in a relative sense. Price pressure is meaningfully defined as:

$$\textit{PricePressure} = x_1 - \left(\sigma_{v2}^2+\sigma_{v3}^2\right) = \left(\sigma_{v2}+\gamma\sigma_{v3}^2\sigma_z\kappa\right)^2-\sigma_{v2}^2, \tag{22}$$

where the expression for x_1 is in (21). Pre-crisis, the price pressure component of monthly realized variance is 4.1 percentage squared. This increases to 13.3 post-crisis, which is an increase of 324%. This component accounts for 14% of total realized variance pre-crisis, which rises to 17% post-crisis.[15] Liquidity therefore is a nonnegligible part of total variance and

[14]Note that the mass of liquidity suppliers, i.e., $(1-\kappa)$ increases by a factor of $(1-0.48)/(1-0.78)-1=136\%$. Adding this to the earlier calculation yields $((1-0.20)\times(1+1.36)-1=+89\%$, which is closer to the observed value of +86%. Note that this is a back-of-the-envelope calculation and, therefore, does not perfectly match the endogenous volume increase of +86%.

[15]The calculations that led to these results are in Appendix F.

deserves to be recognized when pricing variance. This is one of the ways in which my model innovates over alternative models. These other models simply ignore this illiquidity-induced component of realized variance.

5 Conclusion

Previous studies document that the variance risk premium is positive in the data, it jumps after a crisis, and remains elevated for subsequent months. I replicate these findings based on a relatively long time series: 1993-2024. I add to these findings by documenting that trading the index by means of an ETF increases in the post-crisis months. This elevated volume experiences higher costs as the market is less liquid in the months after the crisis.

These empirical findings are explained by a model where agents experience fire-sale risk. The model has some attractive features. First, all results are relatively straightforward closed-form expressions. Second, the model implies that the variance risk premium is nonnegative, consistent with the data. The reason is that agents use it to hedge against fire sales. The agents who experience an endowment shock and are forced to fire sell, these agents particularly like to be paid the realized variance. Not only does it compensate for their elevated risk due to the endowment shock, it also compensates for the higher liquidity premium that needs to be paid to hedge part of it. Both are endogenous in the model. Third, the model can explain increased volume on poorer liquidity.

These findings provide a novel perspective on VIX dynamics. The model prices variance swaps and, therefore, it prices squared VIX. Regulators should become more vigilant at times when VIX is high relative to real-world volatility. The model tells them that fire-sale risk is high in these periods. This is either due to *more* agents experiencing such risk, or *fewer* agents experiencing stronger fire sell demand. Trade data lets them differentiate the two. The ten crisis experienced in 1993-2024 suggest that it is the latter. Future research should

identify policies that could mitigate these post-crisis patterns. If policies make those few agents experience less fire-sale demand, then markets might recover more quickly after a crisis. The extent to which such policies exist, and whether they are desirable from a welfare perspective is left for future research.

Appendix

A Risk premia in CARA-normal case

Consider a CARA investor with wealth W in cash. His risk aversion coefficient is γ. To find the risk premium this investor requires for holding θ units of risky assets, one needs to compute the certainty equivalent (CE). It is defined as the amount of cash he needs to be offered to become indifferent between holding these assets and receiving CE. Let dividend be stochastic: $v \sim N\left(0, \sigma_v^2\right)$.

Asset with linear payoff. The expected utility of letting the investor hold θ units of an asset that pays off v is:

$$
\begin{gathered}
E\left(-\exp\left(-\gamma\left(W+\theta v\right)\right)\right)= \\
\int_v -\frac{1}{\sqrt{2\pi}\sigma_v}\exp\left(-\gamma\left(W+\theta v\right)\right)\exp\left(-\frac{1}{2}\frac{v^2}{\sigma_v^2}\right)\partial v= \\
-\exp\left(-\gamma W+\frac{1}{2}\gamma^2\theta^2\sigma_v^2\right)\underbrace{\int_v \frac{1}{\sqrt{2\pi}\sigma_v}\exp\left(-\frac{1}{2}\frac{\left(v+\gamma\theta\sigma_v^2\right)^2}{\sigma_v^2}\right)\partial v}_{=1 \text{ (integral over PDF)}}= \\
-\exp\left(-\gamma\left(W-\frac{1}{2}\gamma\theta^2\sigma_v^2\right)\right). \qquad (23)
\end{gathered}
$$

The result in (23) shows that the CE for θ units of the risky asset is $\frac{1}{2}\gamma\theta^2\sigma_v^2$. The risk premium per unit of the asset, therefore, is:

$$\boxed{\text{Risk premium per unit of linear payoff} = \frac{1}{2}\gamma\theta\sigma_v^2.} \tag{24}$$

Asset with squared payoff. The expected utility of letting the investor hold θ units of an asset that pays off v^2 is:

$$\begin{gathered}
E\left(-\exp\left(-\gamma\left(W+\theta v^2\right)\right)\right) = \\
\int_v -\frac{1}{\sqrt{2\pi}\sigma_v}\exp\left(-\gamma\left(W+\theta v^2\right)\right)\exp\left(-\frac{1}{2}\frac{v^2}{\sigma_v^2}\right)\partial v = \\
-\exp\left(-\gamma W\right)\underbrace{\int_v \frac{1}{\sqrt{2\pi}\sigma_v}\exp\left(-\frac{1}{2}\frac{v^2\left(1+2\gamma\theta\sigma_v^2\right)}{\sigma_v^2}\right)\partial v}_{=1 \text{ (integral over PDF)}} = \\
-\exp\left(-\gamma W\right).
\end{gathered} \tag{25}$$

The result in (25) shows that the CE for θ units of such asset is zero. The risk premium per unit of the asset, therefore, is:

$$\boxed{\text{Risk premium per unit of squared payoff} = 0.} \tag{26}$$

B Sample construction

This appendix presents details on how the sample is constructed from the raw data sources. The names of these variables in the raw daily files are in typewriter font. The following variables are taken from the "Intraday Indicators" database, which is available in WRDS TAQ:

- Daily net volume in SPY, which defined as the sum of Nasdaq net volume plus

NYSE-Arca net volume. Nasdaq and NYSE-Arca run limit order markets and trades can therefore be signed reliably. The variables used to construct net volume are: `BuyVol_LR_Arca`, `SellVol_LR_Arca`, `BuyVol_LR_Nasd`, and `SellVol_LR_Nasd`.

- The bid-ask spread and depth at the best quotes for SPY are based on `QuotedSpread_Percent_tw`, `BestBidDepth_Share_tw`, and `BestOfrDepth_Share_tw`.
- Daily volume in SPY, which includes all trade volume during market hours `total_vol_m` (i.e., more than Nasdaq and NYSE-Arca).

The S&P index level, SPX, is taken from daily WRDS CRSP `spindx`. VIX was obtained from the historical daily VIX file available at the CBOE website (`Close`).

Proxy construction. For each month in the sample, proxies are constructed for the variance of v_2 and v_3. To remove any transitory effects, these proxies are constructed as follows. The proxy for the annualized variance of v_2 is:

$$\hat{\sigma}^2_{v2} = 250 \times \text{cov}(\mathit{DailyNetVolume}_t, \Delta \log \mathit{SPX}_t + \Delta \log \mathit{SPX}_{t+1}), \tag{27}$$

where tomorrow's return is added to today's return to remove the transitory impact of net volume. Following the same logic, the proxy for the annualized variance of v is:

$$\hat{\sigma}^2_{v} = 250 \times \text{cov}(\Delta \log \mathit{SPX}_t, \Delta \log \mathit{SPX}_t + \Delta \log \mathit{SPX}_{t+1}). \tag{28}$$

Therefore, the proxy for annualized variance of v_3 is:

$$\hat{\sigma}^2_{v3} = \hat{\sigma}^2_{v} - \hat{\sigma}^2_{v2}. \tag{29}$$

C Noisy signal extension

Let agents receive a noisy signal on the dividend in Period 1:

$$v + \eta, \quad \eta \sim N\left(0, \sigma_\eta^2\right), \quad \eta \perp v. \tag{30}$$

Total dividend then becomes the sum of two orthogonal Gaussian components:

$$v = v_2 + v_3, \tag{31}$$

where

$$v_2 \coloneqq E_{t=2}\left(v\right), \quad v_2 \sim N\left(\mu_v, \sigma_{v2}^2\right), \quad \sigma_{v2}^2 = \frac{\sigma_\eta^2}{\sigma_\eta^2 + \sigma^2}\sigma_v^2 \tag{32}$$

and

$$v_3 \sim N\left(0, \sigma_{v3}^2\right), \quad \sigma_{v3}^2 = \sigma_v^2 - \sigma_{v2}^2. \tag{33}$$

The deep parameters σ_v^2 and σ_η^2 could then be backed out by (32) and (33).

D Proofs

This appendix provides the proofs for all lemmas and propositions. Important variables or expressions are emphasized by placing them inside a text box.

Proof of Lemma 1

The equilibrium quantities are taken from Section 2 of VW12. In particular, the demand function for the risky asset of a liquidity demander is:

$$\theta_2^d = \frac{v_2 - p_2}{\gamma\sigma_{v3}^2} - z. \tag{34}$$

For the liquidity supplier it is:

$$\theta_2^s = \frac{v_2 - p_2}{\gamma \sigma_{v3}^2}. \tag{35}$$

Market clearing implies that the Period 2 equilibrium price is:

$$\boxed{p_2} = v_2 - \gamma \sigma_{v3}^2 \left(\bar{\theta} + \kappa z \right). \tag{36}$$

Therefore:

$$\boxed{\theta_2^d} = \bar{\theta} - (1 - \kappa) z \tag{37}$$

and

$$\boxed{\theta_2^s} = \bar{\theta} + \kappa z. \tag{38}$$

This implies that (initiator) net volume is (see also VW12, Equation (2.15)):

$$\boxed{q} = \Delta \theta_2^d = \kappa \left(\theta_2^d - \bar{\theta} \right) = -\kappa (1 - \kappa) z. \tag{39}$$

Note that this net volume is the exact opposite of net volume of liquidity suppliers:

$$(1 - \kappa) \left(\theta_2^s - \bar{\theta} \right) = (1 - \kappa) \kappa z. \tag{40}$$

Liquidity is defined as the price impact of net volume and, therefore, is equal to:

$$\begin{aligned} \boxed{\lambda_2} &= \frac{\mathrm{Cov}\left(p_2 - (p_1 + v_2), q \right)}{\mathrm{Var}\left(q \right)} = \\ &= \frac{\mathrm{Cov}\left(-\gamma \sigma_{v3}^2 \kappa z, -\kappa (1 - \kappa) z \right)}{\sigma_z^2 \kappa^2 (1 - \kappa)^2} = \\ &= \frac{\gamma \sigma_{v3}^2}{1 - \kappa}. \end{aligned} \tag{41}$$

Note that the price change in (41) is $(p_2 - (p_1 + v_2))$, because v_2 is learned before trading and this is the reason why the pre-trade price is: $(p_1 + v_2)$.

Proof of Lemma 2

This lemma presents the equilibrium price of the risky asset in Period 1, which, in the general case, is:

$$p_1 = \mu_v - \gamma\sigma_{v2}^2\bar{\theta} - \gamma\sigma_{v3}^2\bar{\theta} - \frac{\kappa M}{1-\kappa+\kappa M}\Delta_2\bar{\theta}, \tag{42}$$

where

$$\begin{aligned}
M &= \exp\left(\frac{1}{2}\gamma\Delta_3\bar{\theta}^2\right)\sqrt{\frac{1+\Delta_1\kappa^2}{1+\Delta_1(1-\kappa)^2-\gamma^2\sigma_{v3}^2\sigma_z^2}},\\
\Delta_1 &= \gamma^2\sigma_{v3}^2\sigma_z^2,\\
\Delta_2 &= \frac{\gamma\sigma_{v3}^2\Delta_1\kappa}{1+\Delta_1(1-\kappa)^2-\gamma^2\sigma_{v3}^2\sigma_z^2},\\
\Delta_3 &= \frac{\gamma\sigma_{v3}^2\Delta_1}{1+\Delta_1(1-\kappa)^2-\gamma^2\sigma_{v3}^2\sigma_z^2}.
\end{aligned}$$

The Period 1 price of the risky asset is equal to the expected dividend μ_v, minus three terms. The first two terms are the familiar CARA-normal terms (i.e., $\gamma\sigma_{v2}^2\bar{\theta}+\gamma\sigma_{v3}^2\bar{\theta}$). The final term is nonstandard. It captures an additional discount due to excess price volatility caused by Period 2 transitory price pressures, which are needed to clear the market. They compensate liquidity suppliers for holding additional risk. Note that this term disappears when there is no nontraded risk and, therefore, no trading, i.e., if $\sigma_z^2 = 0$, then $\Delta_2 = 0$. In a sense, it represents a liquidity *risk* premium. This additional variance is endogenous to the model and is, therefore, also accounted for when pricing the variance swap.

Proof of Proposition 1

The marginal utility of the swap is:

$$\frac{\partial}{\partial \theta_x}\Big(\kappa E_{t=1}\left[-\exp\left(-\gamma\left(w_d+\theta_x V\right)\right)|d\right]+ \\ +\left(1-\kappa\right)E_{t=1}\left[-\exp\left(-\gamma\left(w_s+\theta_x V\right)\right)|s\right]\Big)\Big|_{\theta_x=0}, \tag{43}$$

where the subscript d refers to the liquidity-demander type, the subscript s to the liquidity-supplier type, and final-period wealth for type i is w_i. Therefore, the price of the variance swap, expressed in units of the Period 3 consumption good, is:

$$\boxed{y_1} = \kappa\gamma E\left[V\exp\left(-\gamma w_d\right)|d\right]+\left(1-\kappa\right)\gamma E\left[V\exp\left(-\gamma w_s\right)|s\right]. \tag{44}$$

The price of the variance swap given in (44) contains expressions in the numerator and in the denominator of the form:

$$E_a\left[\left(a'B_2a+b_3'a+b_3\right)\exp\left(-\gamma\left(a'B_4a+b_5'a+b_6\right)\right)\right], \tag{45}$$

where

$$a=\begin{pmatrix} v_2 \\ z \\ v_3 \end{pmatrix}. \tag{46}$$

The reason for the shape of (45) is:

1. The quadratic expression $\left(a'B_2a+b_3'a+b_3\right)$ is the result of realized variance in (15), which is quadratic in price changes. These price changes are, in turn, affine in a.

2. The expression $\left(a'B_4a+b_5'a+b_6\right)$ corresponds to the wealth of the agent. It consists of a risk-free position, long or short, due to payment in the Period 2 trading round. It

further consists of an exposure to v_3 caused by the post-trade position in the risky asset. This latter position is affine in a, which is increased by z for the liquidity demander due to his endowment shock.

The distribution of the random variable a is:

$$a \sim N\left(\mu, \Sigma\right), \tag{47}$$

where

$$\mu = \begin{pmatrix} \mu_v & 0 & 0 \end{pmatrix}', \tag{48}$$

$$\Sigma = \begin{pmatrix} \sigma_{v2}^2 & -\sigma_{v2}\sigma_z & 0 \\ -\sigma_{v2}\sigma_z & \sigma_z^2 & 0 \\ 0 & 0 & \sigma_{v3}^2 \end{pmatrix}. \tag{49}$$

We compute (45) as follows:

$$\begin{aligned} &\int_{\mathbb{R}^3} \left(a' B_2 a + b_3' a + b_3\right) \exp\left(-\gamma\left(a' B_4 a + b_5' a + b_6\right)\right) \times \\ &\quad \times \frac{1}{\sqrt{2\kappa|\Sigma|}} \exp\left(-\frac{1}{2}\left(a-\mu\right)' \Sigma^{-1}\left(a-\mu\right)\right) da = \\ &= \sqrt{\frac{|\Sigma^*|}{|\Sigma|}} \exp\left(-\gamma b_6 + \frac{1}{2}\mu^{*\prime}\Sigma^{*-1}\mu^* - \frac{1}{2}\mu'\Sigma^{-1}\mu\right) \times \\ &\qquad \times \int_{\mathbb{R}^3} \left(a' B_2 a + b_3' a + b_3\right) \frac{1}{\sqrt{2\kappa|\Sigma^*|}} \times \\ &\qquad \times \exp\left(-\frac{1}{2}\left(a-\mu^*\right)' \Sigma^{*-1}\left(a-\mu^*\right)\right) da \\ &= \sqrt{\frac{|\Sigma^*|}{|\Sigma|}} \exp\left(-\gamma b_6 + \frac{1}{2}\mu^{*\prime}\Sigma^{*-1}\mu^* - \frac{1}{2}\mu'\Sigma^{-1}\mu\right) \times \\ &\qquad \times E\left(a^{*\prime} B_2 a^* + b_3' a^* + b_3\right), \end{aligned} \tag{50}$$

where

$$a^* \sim N\left(\mu^*, \Sigma^*\right), \tag{51}$$

and

$$\mu^* = \Sigma^* \left(\Sigma^{-1}\mu - \gamma b_5\right), \tag{52}$$

$$\Sigma^* = \left(\Sigma^{-1} + 2\gamma B_4\right)^{-1}. \tag{53}$$

What remains is to compute the expectation in (50):

$$\begin{aligned} E\left(a^{*\prime}B_2a^* + b_3'a^* + b_3\right) &= E\left(\sum_i \sum_j \left(B_2\right)_{ij} a_i^* a_j^*\right) + b_3'\mu^* + b_3 = \\ &= \sum_i \sum_j \left(B_2\right)_{ij} \left(\sigma_{ij}^* + \mu_i\mu_j\right) + b_3'\mu^* + b_3 = \\ &= \sum_i \sum_j \left(B_2\right)_{ij} \sigma_{ji}^* + \mu^{*\prime}B_2\mu^* + b_3'\mu^* + b_3 = \\ &= \sum_i \left(B_2\Sigma^*\right)_{i,i} + \mu^{*\prime}B_2\mu^* + b_3'\mu^* + b_3 = \\ &= tr\left(B_2\Sigma^*\right) + \mu^{*\prime}B_2\mu^* + b_3'\mu^* + b_3. \end{aligned} \tag{54}$$

Putting it all together yields the following expressions for the numerator and the denominator of (44):

$$\begin{aligned} \boxed{y_1} &= E\Big[\left(a'B_2a + b_3'a + b_3\right) \times \\ &\times \exp\left(-\gamma\left(a'B_4a + b_5'a + b_6\right)\right)\Big] = \\ &= \sqrt{\frac{|\Sigma^*|}{|\Sigma|}} \exp\left(-\gamma b_6 + \frac{1}{2}\mu^{*\prime}\Sigma^{*-1}\mu^* - \frac{1}{2}\mu'\Sigma^{-1}\mu\right) \times \\ &\times \left(tr\left(B_2\Sigma^*\right) + \mu^{*\prime}B_2\mu^* + b_3'\mu^* + b_3\right), \end{aligned} \tag{55}$$

with

$$\mu^* = \Sigma^* \left(\Sigma^{-1} \mu - \gamma b_5 \right) \quad \text{and} \quad \Sigma^* = \left(\Sigma^{-1} + 2\gamma B_4 \right)^{-1}. \tag{56}$$

Expressions for B_2, b_3, b_3, B_4, b_5 and b_6 in (45). Now that the equilibrium price of the variance swap has been derived analytically, what remains is some administrative work. The expressions for B_2, b_3, b_3, B_4, b_5, and b_6 in (55) need to be written down explicitly. Since all these quantities are affine in the state variables, we use subscripts to distinguish between the coefficients of the state variable a and the constant. For example, the position in the risky asset of the liquidity demander after trading in Period 1 is:

$$\theta_2^d = \bar{\theta} - (1 - \kappa) z = \theta_{1,a}^d a + \theta_{1,c}^d, \tag{57}$$

where

$$\theta_{1,a}^d = \begin{pmatrix} 0 & -(1-\kappa) & 0 \end{pmatrix}', \tag{58}$$

$$\theta_{1,c}^d = \bar{\theta}. \tag{59}$$

With this notation, we can express the constants for the liquidity demander in terms of equilibrium quantities as follows:

$$B_2 = (p_{1,a} - p_{0,a}) (p_{1,a} - p_{0,a})' + (p_{2,a} - p_{1,a}) (p_{2,a} - p_{1,a})', \tag{60}$$

$$b_3 = 2 (p_{1,c} - p_{0,c}) (p_{1,a} - p_{0,a}) + 2 (p_{2,c} - p_{1,c}) (p_{2,a} - p_{1,a}), \tag{61}$$

$$b_3 = (p_{1,c} - p_{0,c})^2 + (p_{2,c} - p_{1,c})^2, \tag{62}$$

$$B_4 = \underbrace{-\frac{1}{2} \left(\Delta\theta_{1,a}^d p_{1,a}' + p_{1,a} \Delta\theta_{1,a}^{d\prime} \right)}_{\text{Payment made}} +$$

$$+\underbrace{\frac{1}{2}\left(\left(\left(\theta_{1,a}^{d}+\begin{pmatrix}0\\1\\0\end{pmatrix}\right)\begin{pmatrix}0&0&1\end{pmatrix}+\begin{pmatrix}0\\0\\1\end{pmatrix}\left(\theta_{1,a}^{d\prime}+\begin{pmatrix}0&1&0\end{pmatrix}\right)\right)\right)}_{\text{Exposure}}, \tag{63}$$

$$b_5=-\Delta\theta_{1,c}^{d}p_{1,a}-p_{1,c}\Delta\theta_{1,a}^{d}+\theta_{1,c}^{d}\begin{pmatrix}0\\0\\1\end{pmatrix}, \tag{64}$$

$$b_6=\theta_{1,c}^{d}p_{1,c}. \tag{65}$$

For a liquidity supplier, $B1$, $b2$, and b_3 are the same, but the remaining ones need to be changed to:

$$B_4=\underbrace{\frac{1}{2}\left(\Delta\theta_{1,a}^{s}p_{1,a}^{\prime}+p_{1,a}\Delta\theta_{1,a}^{s\prime}\right)}_{\text{Payment received}}+\underbrace{\frac{1}{2}\left(\theta_{1,a}^{s}\begin{pmatrix}0&0&1\end{pmatrix}+\begin{pmatrix}0\\0\\1\end{pmatrix}\theta_{1,a}^{s\prime}\right)}_{\text{Exposure}}, \tag{66}$$

$$b_5=\Delta\theta_{1,c}^{s}p_{1,a}+p_{1,c}\Delta\theta_{1,a}^{s}+\theta_{1,c}^{s}\begin{pmatrix}0\\0\\1\end{pmatrix}, \tag{67}$$

$$b_6=\theta_{1,c}^{s}p_{1,c}. \tag{68}$$

Special case: $\bar{\theta}=\mu_v=0$.

To develop further insight, we assume $\bar{\theta}=\mu_v=0$ as per Assumption 1. That is, both the supply of the risky asset and the mean dividend are set to zero, respectively. The expressions in this subsection have been double-checked by means of a Mathematica notebook that is available at `https://bit.ly/3CM8lDZ`.

Liquidity demander. For the liquidity demander, marginal utility for the component of realized variance associated with the price change from Period 1 to 2 is equal to (55) with:

$$B_2 = \begin{pmatrix} -\left(\frac{\sigma_{v2}}{\sigma_z} + \gamma\sigma_{v3}^2\kappa\right) \\ 0 \end{pmatrix} \begin{pmatrix} -\left(\frac{\sigma_{v2}}{\sigma_z} + \gamma\sigma_{v3}^2\kappa\right) & 0 \end{pmatrix} +$$
$$+ \begin{pmatrix} \gamma\sigma_{v3}^2\kappa \\ 1 \end{pmatrix} \begin{pmatrix} \gamma\sigma_{v3}^2\kappa & 1 \end{pmatrix} = \begin{pmatrix} \frac{\sigma_{v2}^2}{\sigma_z^2} + 2\gamma\sigma_{v3}^2\kappa\frac{\sigma_{v2}}{\sigma_z} + 2\gamma^2\sigma_{v3}^4\kappa^2 & \gamma\sigma_{v3}^2\kappa \\ \gamma\sigma_{v3}^2\kappa & 1 \end{pmatrix} =$$
$$= \begin{pmatrix} \left(\frac{\sigma_{v2}}{\sigma_z} + \gamma\sigma_{v3}^2\kappa\right)^2 + \left(\gamma\sigma_{v3}^2\kappa\right)^2 & \gamma\sigma_{v3}^2\kappa \\ \gamma\sigma_{v3}^2\kappa & 1 \end{pmatrix}, \tag{69}$$

$$b_3 = \begin{pmatrix} 0 & 0 \end{pmatrix}', \tag{70}$$

$$b_3 = 0, \tag{71}$$

$$B_4 = \frac{1}{2}\begin{pmatrix} 1-\kappa \\ 0 \end{pmatrix} \begin{pmatrix} -\gamma\sigma_{v3}^2\kappa & 0 \end{pmatrix} + \frac{1}{2}\left(\begin{pmatrix} 1-\kappa \\ 0 \end{pmatrix} \begin{pmatrix} -\gamma\sigma_{v3}^2\kappa & 0 \end{pmatrix}\right)' +$$
$$+ \frac{1}{2}\begin{pmatrix} -(1-\kappa)+1 \\ 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \end{pmatrix} + \frac{1}{2}\left(\begin{pmatrix} -(1-\kappa)+1 \\ 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \end{pmatrix}\right)' =$$
$$= \begin{pmatrix} -\gamma\sigma_{v3}^2\kappa\,(1-\kappa) & \frac{\kappa}{2} \\ \frac{\kappa}{2} & 0 \end{pmatrix}, \tag{72}$$

$$b_5 = \begin{pmatrix} 0 & 0 \end{pmatrix}', \tag{73}$$

$$b_6 = 0. \tag{74}$$

Note that B_2 in (69) features *positive* off-diagonal elements. The reason is that, although price pressures mean-revert and thus do not create a wealth loss to the aggregate investor, they *do* add to realized variance. As the payoff to the variance swap is this realized variance, it adds to total variance and, therefore, the off-diagonals are positive.

The values in (69) through (74) for this special case environment imply (as per (56)):

$$\Sigma = \begin{pmatrix} \sigma_z^2 & 0 \\ 0 & \sigma_{v3}^2 \end{pmatrix}, \tag{75}$$

$$\mu = \begin{pmatrix} 0 & 0 \end{pmatrix}', \tag{76}$$

$$\begin{aligned} \Sigma^* &= \left(\Sigma^{-1} + 2\gamma B_4\right)^{-1} \\ &= \begin{pmatrix} \frac{1}{\sigma_z^2} - 2\gamma^2\sigma_{v3}^2\kappa\left(1-\kappa\right) & \gamma\kappa \\ \gamma\kappa & \frac{1}{\sigma_{v3}^2} \end{pmatrix}^{-1}, \\ &= \frac{1}{1-\gamma^2\sigma_{v3}^2\sigma_z^2\kappa\left(2-\kappa\right)} \begin{pmatrix} \sigma_z^2 & -\gamma\sigma_{v3}^2\sigma_z^2\kappa \\ -\gamma\sigma_{v3}^2\sigma_z^2\kappa & \sigma_{v3}^2\left(1-2\gamma^2\sigma_{v3}^2\sigma_z^2\kappa\left(1-\kappa\right)\right) \end{pmatrix}, \end{aligned} \tag{77}$$

$$\begin{aligned} \mu^* &= \Sigma^*\left(\Sigma^{-1}\mu - \gamma b_5\right) \\ &= \begin{pmatrix} 0 & 0 \end{pmatrix}'. \end{aligned} \tag{78}$$

Note that to value the variance swap, the real-world probability measure, often referred to as the P-measure, is replaced by a "risk-neutral measure," often known as the Q-measure. The latter features a variance of Σ^* and comparing its expression in (77), to the expression of Σ in (75), shows that the Q-measure increases variance for both z and v_3. The off-diagonal is negative, because price pressures are transitory. These price pressures, therefore, add to realized variance (as per our discussion about B_2 above), but they do not affect wealth in Period 3 for the aggregate investor.

Therefore, the marginal utility of the risky asset for d is:

$$\begin{aligned} \boxed{y_1^d} &= E\left[\left(a'B_2a + b_3'a + b_3\right)\exp\left(-\gamma\left(a'B_4a + b_5'a + b_6\right)\right)\right] = \\ &= \sqrt{\frac{|\Sigma^*|}{|\Sigma|}}tr\left(B_2\Sigma^*\right) = \end{aligned}$$

$$\boxed{= \frac{1}{1-\gamma^2\sigma_{v3}^2\sigma_z^2\kappa\left(2-\kappa\right)}\left(\frac{\left(\sigma_{v2}+\gamma\sigma_{v3}^2\sigma_z\kappa\right)^2}{1-\gamma^2\sigma_{v3}^2\sigma_z^2\kappa\left(2-\kappa\right)}+\sigma_{v3}^2\right).} \tag{79}$$

because[16]

$$\begin{aligned} |\Sigma^*| &= \frac{\sigma_{v3}^2\sigma_z^2}{1-\gamma^2\sigma_{v3}^2\sigma_z^2\kappa\left(2-\kappa\right)}, \quad |\Sigma| = \sigma_{v3}^2\sigma_z^2, \\ tr\left(B_2\Sigma^*\right) &= \frac{\left(\sigma_{v2}+\gamma\sigma_{v3}^2\sigma_z\kappa\right)^2}{1-\gamma^2\sigma_{v3}^2\sigma_z^2\kappa\left(2-\kappa\right)}+\sigma_{v3}^2. \end{aligned} \tag{80}$$

Liquidity supplier. For the liquidity supplier, all coefficients are the same except for:

$$\begin{aligned} B_4 &= \frac{1}{2}\begin{pmatrix}-\kappa\\0\end{pmatrix}\begin{pmatrix}-\gamma\sigma_{v3}^2\kappa & 0\end{pmatrix}+\frac{1}{2}\left(\begin{pmatrix}-\kappa\\0\end{pmatrix}\begin{pmatrix}-\gamma\sigma_{v3}^2\kappa & 0\end{pmatrix}\right)' \\ &+\frac{1}{2}\begin{pmatrix}\kappa\\0\end{pmatrix}\begin{pmatrix}0 & 1\end{pmatrix}+\frac{1}{2}\left(\begin{pmatrix}\kappa\\0\end{pmatrix}\begin{pmatrix}0 & 1\end{pmatrix}\right)' = \\ &= \begin{pmatrix}\gamma\kappa^2\sigma_{v3}^2 & \frac{\kappa}{2}\\ \frac{\kappa}{2} & 0\end{pmatrix}. \end{aligned} \tag{81}$$

These values imply that the following objects change relative to the liquidity demander case:

$$\begin{aligned} \Sigma^* &= \begin{pmatrix}\frac{1}{\sigma_z^2}+2\kappa^2\gamma^2\sigma_{v3}^2 & \gamma\kappa\\ \gamma\kappa & \frac{1}{\sigma_{v3}^2}\end{pmatrix}^{-1} \\ &= \frac{1}{1+\gamma^2\sigma_{v3}^2\sigma_z^2\kappa^2}\begin{pmatrix}\sigma_z^2 & -\gamma\sigma_{v3}^2\sigma_z^2\kappa\\ -\gamma\sigma_{v3}^2\sigma_z^2\kappa & \sigma_{v3}^2\left(1+2\gamma^2\sigma_{v3}^2\sigma_z^2\kappa^2\right)\end{pmatrix}, \end{aligned} \tag{82}$$

[16]Note that the outcome is well defined given the condition in (10).

therefore,

$$|\Sigma^*| = \frac{\sigma_{v3}^2 \sigma_z^2}{1+\gamma^2 \sigma_{v3}^2 \sigma_z^2 \kappa^2}, \quad tr\left(B_2 \Sigma^*\right) = \frac{\left(\sigma_{v2}+\gamma \sigma_{v3}^2 \sigma_z^2 \kappa\right)^2}{1+\gamma^2 \sigma_{v3}^2 \sigma_z^2 \kappa^2} + \sigma_{v3}^2. \tag{83}$$

The marginal utility of a variance swap to the liquidity supplier, therefore, is:

$$\boxed{y_1^s = \sqrt{\frac{|\Sigma^*|}{|\Sigma|}} tr\left(B_2 \Sigma^*\right) = \frac{1}{1+\gamma^2 \sigma_{v3}^2 \sigma_z^2 \kappa^2}\left(\frac{\left(\sigma_{v2}+\gamma \sigma_{v3}^2 \sigma_z \kappa\right)^2}{1+\gamma^2 \sigma_{v3}^2 \sigma_z^2 \kappa^2} + \sigma_{v3}^2\right).} \tag{84}$$

So, the marginal utility of the variance swap in Period 1 is obtained by taking the expected marginal utility, i.e., multiplying (79) and (84) by κ and $(1-\kappa)$, respectively:

$$\begin{aligned}
\boxed{y_1} &= \kappa y_1^d + (1-\kappa) y_1^s = \\
&= \kappa \underbrace{\left(A^2\left(\sigma_{v2}+\gamma \sigma_{v3}^2 \sigma_z \kappa\right)^2 + A \sigma_{v3}^2\right)}_{\text{Valuation liquidity demander}} + \\
&+ (1-\kappa) \underbrace{\left(B^2\left(\sigma_{v2}+\gamma \sigma_{v3}^2 \sigma_z \kappa\right)^2 + B \sigma_{v3}^2\right)}_{\text{Valuation liquidity supplier}} = \\
&= \underbrace{\left(\kappa A^2 + (1-\kappa) B^2\right)}_{\text{Inflator corr. return}} \times \underbrace{\left(\sigma_{v2}+\gamma \sigma_{v3}^2 \sigma_z \kappa\right)^2}_{\text{Variance corr. return}} + \\
&+ \underbrace{(\kappa A + (1-\kappa) B)}_{\text{Inflator noncorr. return}} \times \underbrace{\sigma_{v3}^2}_{\text{Variance noncorr. return}},
\end{aligned} \tag{85}$$

with

$$A = \frac{1}{1-\gamma^2 \sigma_{v3}^2 \sigma_z^2 (2-\kappa) \kappa} \tag{86}$$

and

$$B = \frac{1}{1+\gamma^2 \sigma_{v3}^2 \sigma_z^2 \kappa^2}. \tag{87}$$

The proof for the final statement that both inflators are larger than one follows from a result in the proof of Proposition 2. When proving that y_1 increases in κ, it is shown that

both inflators increase in κ. This result, along with the observation that both inflators are one for $\kappa = 0$, shows that both inflators are at least one. What remains to show is:

$$\kappa A^2 + (1-\kappa) B^2 \geq \kappa A + (1-\kappa) B. \tag{88}$$

Applying Jensen's inequality to the convex function $f(x) = x^2$ with the stochastic variable:

$$X = \begin{cases} A \text{ with probability } \kappa, \\ B \text{ with probability } 1-\kappa, \end{cases} \tag{89}$$

yields

$$\kappa A^2 + (1-\kappa) B^2 \geq \kappa A + (1-\kappa) B. \tag{90}$$

This completes the proof.

Proof of Proposition 2

The algebra in this subsection has been double-checked by means of a Mathematica notebook that is available at `https://bit.ly/3CM8lDZ`. The engine of the monotonicity proof is the following lemma.

Lemma 3 *Let*

$$f_{c,n}(\kappa) = \frac{1}{(1-(2-\kappa)\kappa c)^n} - \frac{1}{(1+c\kappa^2)^n}, \tag{91}$$

where

$$c \in [0,1), \quad n \in \{1,2,3,\ldots\}. \tag{92}$$

Then $f_{c,n}(0) = 1$ *and* $f_{c,n}(\kappa)$ *increases monotonically in the interval* $\kappa \in [0,1]$. *Therefore,* $f_{c,n}(\kappa)$ *is positive in this interval.*

To prove Lemma 3, first define

$$g_c(\kappa) = \frac{1}{1-(2-\kappa)\kappa c} - \frac{1}{1+c\kappa^2}, \tag{93}$$

then

$$g_c'(\kappa) = \frac{2c\left(c^2(4-3\kappa)\kappa^3 - 2c\kappa^2 + 1\right)}{(1-c(2-\kappa)\kappa)^2\left(c\kappa^2+1\right)^2}, \tag{94}$$

and the sign of this derivative therefore depends on the sign of

$$h_c(\kappa) = c^2(4-3\kappa)\kappa^3 - 2c\kappa^2 + 1. \tag{95}$$

Now, from

$$h_c'(\kappa) = -4c\kappa\left(3c(\kappa-1)\kappa + 1\right) < 0, \tag{96}$$

$c \in [0,1)$, $\kappa(1-\kappa) \leq 1/4$ for $\kappa \in [0,1]$, $h_c(0) = 1$, $h_c(1) = 0$, and $h_c(\kappa)$ being a continuous differentiable function, it follows that $h_c(\kappa)$ is indeed strictly positive on the domain $\kappa \in [0,1)$, based on the Intermediate Value Theorem applied to $\kappa \in [0, 1+\varepsilon]$ with $\varepsilon > 0$. This implies that $g_c'(\kappa) > 0$ and $g_c(\kappa)$ therefore increases monotonically in κ.

With the proof that $g_c(\kappa)$ increases monotonically in κ, what remains to be proven is that this implies that $f_{c,n}(\kappa)$ increases monotonically in κ. This follows from:

$$\begin{aligned} f_{c,n}'(\kappa) &= n\left(\frac{1}{(1-(2-\kappa)\kappa c)}\right)^{n-1}\frac{\partial}{\partial\kappa}\left(\frac{1}{(1-(2-\kappa)\kappa c)}\right) + \\ &\quad - n\left(\frac{1}{1+c\kappa^2}\right)^{n-1}\frac{\partial}{\partial\kappa}\left(\frac{1}{1+c\kappa^2}\right) && (97) \\ &\geq n\left(\frac{1}{(1-(2-\kappa)\kappa c)}\right)^{n-1}\frac{\partial}{\partial\kappa}\left(\frac{1}{(1-(2-\kappa)\kappa c)}\right) + \\ &\quad - n\left(\frac{1}{(1-(2-\kappa)\kappa c)}\right)^{n-1}\frac{\partial}{\partial\kappa}\left(\frac{1}{1+c\kappa^2}\right) && (98) \\ &= n\left(\frac{1}{(1-(2-\kappa)\kappa c)}\right)^{n-1} g_c'(\kappa). && (99) \end{aligned}$$

As $g_c'(\kappa) > 0$, it follows that $f_c'(\kappa) > 0$, which proves Lemma 3.

To prove a monotonic relationship for the VRP, i.e., $y_1 - x_1$, requires proving it for the two inflators in (16). This is what is done in the remainder of the proof. I further define:

$$c \coloneqq \gamma^2 \sigma_{v3}^2 \sigma_z^2, \tag{100}$$

which, by condition (10), is in $[0, 1)$, because $\sigma_{v3}^2 \leq \sigma_v^2$.

Proof that $y_1 - x_1$ increases monotonically in κ. Let us first the derivative of the flow-correlated inflator in (16):

$$\frac{\partial}{\partial \kappa}\left(\kappa A^2 + (1-\kappa) B^2\right) = f_{c,1}(\kappa) + 4(1-\kappa)\kappa c \times f_{c,3}(\kappa) > 0. \tag{101}$$

The derivative for the flow-uncorrelated inflator in (16) is:

$$\frac{\partial}{\partial \kappa}\left(\kappa A + (1-\kappa) B\right) = f_{c,2}(\kappa) + 2(1-\kappa)\kappa c \times f_{c,2}(\kappa) > 0. \tag{102}$$

Since both variance inflators increase monotonically in κ, y_1 increases monotonically in κ.

Proof that $y_1 - x_1$ increases monotonically in γ. This proof follows the same steps as the previous one:

$$\begin{aligned}\frac{\partial}{\partial \gamma}\left(\kappa A^2 (1-\kappa) B^2\right) &= \frac{4\kappa^2 c}{\gamma}\left(\frac{2-\kappa}{\left(1-(2-\kappa)\kappa c\right)^3} - \frac{1-\kappa}{\left(1+\kappa^2 c\right)^3}\right) \\ &\geq \left(\frac{4\kappa^2 (1-\kappa) c}{\gamma}\right) f_{c,3}(\kappa) > 0\end{aligned} \tag{103}$$

and

$$\frac{\partial}{\partial\gamma}\left(\kappa A\left(1-\kappa\right)B\right)=\frac{2\kappa^{2}c}{\gamma}\left(\frac{2-\kappa}{\left(1-\left(2-\kappa\right)\kappa c\right)^{2}}-\frac{1-\kappa}{\left(1+\kappa^{2}c\right)^{2}}\right)$$
$$\geq\left(\frac{2\kappa^{2}\left(1-\kappa\right)c}{\gamma}\right)f_{c,2}\left(\kappa\right)>0. \tag{104}$$

Therefore both variance inflators increase monotonically in γ and thus y_1 increases monotonically in γ.

Proof that $y_1 - x_1$ increases monotonically in σ_z. This proof follows the same steps as the previous one:

$$\frac{\partial}{\partial\sigma_{z}}\left(\kappa A^{2}\left(1-\kappa\right)B^{2}\right)=\frac{4\kappa^{2}c}{\sigma_{z}}\left(\frac{2-\kappa}{\left(1-\left(2-\kappa\right)\kappa c\right)^{3}}-\frac{1-\kappa}{\left(1+\kappa^{2}c\right)^{3}}\right)$$
$$\geq\left(\frac{4\kappa^{2}\left(1-\kappa\right)c}{\sigma_{z}}\right)f_{c,3}\left(\kappa\right)>0 \tag{105}$$

and

$$\frac{\partial}{\partial\sigma_{z}}\left(\kappa A\left(1-\kappa\right)B\right)=\frac{2\kappa^{2}c}{\sigma_{z}}\left(\frac{2-\kappa}{\left(1-\left(2-\kappa\right)\kappa c\right)^{2}}-\frac{1-\kappa}{\left(1+\kappa^{2}c\right)^{2}}\right)$$
$$\geq\left(\frac{2\kappa^{2}\left(1-\kappa\right)c}{\sigma_{z}}\right)f_{c,2}\left(\kappa\right)>0. \tag{106}$$

Therefore both variance inflators increase monotonically in σ_z and thus y_1 increases monotonically in σ_z.

E Details calibration procedure

The calibration involves the following steps. Fix an equidistant grid for the pre-crisis value of κ. Iterate over this grid. For each value of κ, do the following:

1. Compute the nontraded shock size σ_z^2 that delivers the observed pre-crisis VRP using (16).

2. Use the expression for illiquidity in (13), the pre-crisis κ, the observed pre- and post flow-uncorrelated price variance (σ_{v3}^2), and the observed relative change in illiquidity to compute the implied post-crisis κ.

3. Use the expression for net volume in (12) to compute expected volume, which is the expected absolute value of net volume. Expected volume, therefore, equals $\kappa(1-\kappa)\sigma_z\sqrt{\frac{2}{\kappa}}$, because z is Gaussian. The relative change in volume is used to compute post-crisis σ_z^2.

4. Use these post-crisis values for κ and σ_z^2 and the observed post-crisis flow-correlated and flow-uncorrelated price variance to compute the implied post-crisis VRP.

After all these computations are done, pick the pre-crisis value of κ that delivers a post-crisis VRP that is closest to the observed post-crisis VRP.

F Price price calculations

The pre-crisis calculation is as follows. The monthly flow-uncorrelated realized variance is $\sigma_{v3}^2 = 0.67 \times 364/12 = 20.3$ percentage squared. Therefore, $\gamma\sigma_{v3}^2\sigma_z\kappa = 3 \times (0.0244/12) \times \sqrt{2.41} \times 0.78 = 74$ basis points (see (21)). This implies $\sigma_{v2} = \sqrt{0.33 \times (0.0364/12)} - 0.0074 = 242$ basis points. Price pressure as defined in (22), therefore, is: $364/12 - (2.42^2 + (244/12)) = 4.1$, which is 14% of total realized variance. Following the same steps for the post-crisis period yields the following numbers. $\sigma_{v3}^2 = 0.70 \times 964/12 = 56.2$ percentage squared. Therefore, $\gamma\sigma_{v3}^2\sigma_z\kappa = 3 \times (0.0675/12) \times \sqrt{4.03} \times 0.48 = 163$ basis points. This implies $\sigma_{v2} = \sqrt{0.30 \times (0.0964/12)} - 0.0163 = 328$ basis points. Price pressure is: $964/12 - (3.28^2 + (675/12)) = 13.3$, which is 17% of total realized variance.

References

Bekaert, Geert and Marie Hoerova (2013). *The VIX, the Variance Premium and Stock Market Volatility*. Manuscript. NBER Working Paper 18995.

— (2014). The VIX, the Variance Premium and Stock Market Volatility. *Journal of Econometrics* 183, pp. 181–192.

Bollerslev, Tim, George Tauchen and Hao Zhou (2009). Expected Stock Returns and Variance Risk Premia. *Review of Financial Studies* 22, pp. 4464–4465.

Carr, Peter and Liuren Wu (2006). A Tale of Two Indices. *Journal of Derivatives* 13, pp. 13–29.

— (2009). Variance Risk Premiums. *Review of Financial Studies* 22, pp. 1311–1341.

Choi, Hoyong, Philippe Mueller and Andrea Vedolin (2017). Bond Variance Risk Premiums. *Review of Finance* 21, pp. 987–1022.

Dew-Becker, I. and Stefano Giglio (2024). Recent Developements in Financial Risk and the Real Economy. *Annual Reviews of Financial Economics* 16, pp. 4.1–4.22.

Evans, Martin D.D. and Richard K. Lyons (2002). Order Flow and Exchange Rate Dynamics. *Journal of Political Economy* 110, pp. 170–180.

Gabaix, Xavier and Ralph S.J. Koijen (2024). *In Search of the Origins of Financial Fluctuations: The Inelastic Markets Hypothesis*. Manuscript. University of Chicago.

Jiang, Wenxi (2024). Leveraged Speculators and Asset Prices. *Review of Finance* 28, pp. 769–804.

Konstantinidi, Eirini and George Skiadopoulos (2016). How Does the Market Variance Risk Premium Vary Over Time? Evidence From S&P 500 Variance Swap Investment Returns. *Journal of Banking and Finance* 62, pp. 62–75.

Lo, Andrew W., Harry Mamaysky and Jiang Wang (2004). Asset Prices and Trading Volume under Fixed Transaction Costs. *Journal of Political Economy* 112, pp. 1054–1090.

Lochstoer, Lars A. and Tyler Muir (2022). Volatility Expectations. *Journal of Finance* 77, pp. 1055–1096.

Todorov, Viktor (2010). Variance Risk-Premium Dynamics: The Role of Jumps. *Review of Financial Studies* 23, pp. 345–383.

Vayanos, Dimitri (2004). *Flight to Quality, Flight to Liquidity, and the Pricing of Risk.* NBER Working Paper Series #10327. London School of Economics.

Vayanos, Dimitri and Jiang Wang (2012). Liquidity and Asset Returns Under Asymmetric Information and Imperfect Competition. *Review of Financial Studies* 25, pp. 1339–1365.

Zhou, Hao (2018). Variance Risk Premia, Asset Predictability Puzzles, and Macroeconomic Uncertainty. *Annual Review of Financial Economics* 10, pp. 481–497.

Table 1: Summary statistics. This table presents summary statistics for the monthly data sample that runs from February 1993 through August 2024.

	Mean	SD	Auto	Skew	Kurt	N
Realized var SPX EOM ($\%^2$, ann.)	342	695	0.49	8.24	86.04	379
Implied var SPX BOM ($\%^2$, ann.)	447	419	0.76	3.40	16.73	379
Realized var risk premium EOM ($\%^2$, ann.)	104	560	0.15	-9.01	107.40	379

Table 2: Model calibration to match crisis patterns. This table calibrates the model to pre- and post-crisis trading. The patterns are based on ten crisis periods from 1993 through 2024. The pre-crisis period is the three-month period leading up to the crisis, and the post-crisis period is the three-month period following a crisis (see Figure 3).

	Data	Model
Panel (a): Matched moments		
Pre-crisis variance risk premium (%^{2}, ann.)	14	14
Post-crisis variance risk premium (%^{2}, ann.)	315	315
Change in NBBO quoted spread pre- to post-crisis	+18%	+18%
Change in share volume pre- to post-crisis	+86%	+86%
Panel (b): Calibrated parameters		
Pre-crisis κ (fraction of agents experiencing shock)		0.78
Post-crisis κ		0.48
Change in κ pre- to post-crisis		-38%
Pre-crisis σ_z (standard deviation nontraded risk shock)		1.55
Post-crisis σ_z		2.01
Change in σ_z pre- to post-crisis		+29%
Panel (c): Further input calibration		
Pre-crisis realized var SPX EOM (%^{2}, ann.)		364
Pre-crisis implied var SPX BOM (%^{2}, ann.)		378
Pre-crisis IV to RV ratio		1.04
Pre-crisis fraction of RV that is flow-correlated		0.33
Post-crisis realized var SPX EOM (%^{2}, ann.)		964
Post-crisis implied var SPX BOM (%^{2}, ann.)		1278
Post-crisis IV to RV ratio		1.33
Post-crisis fraction of RV that is flow-correlated		0.30
Risk-aversion parameter γ		3

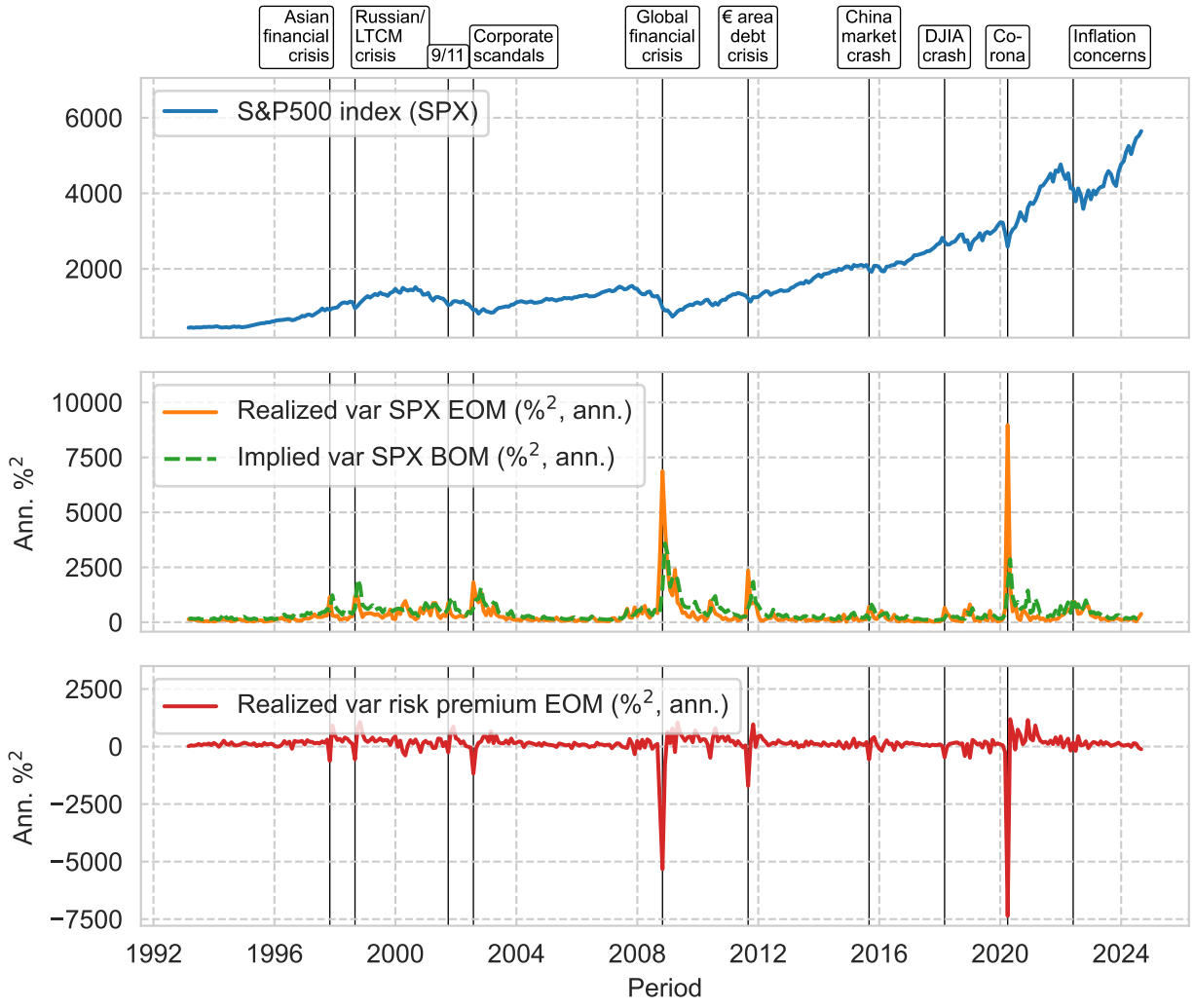

Figure 1: Time series S&P500 index returns and variances. This figure plots the S&P 500 index and several series based on its return variance.

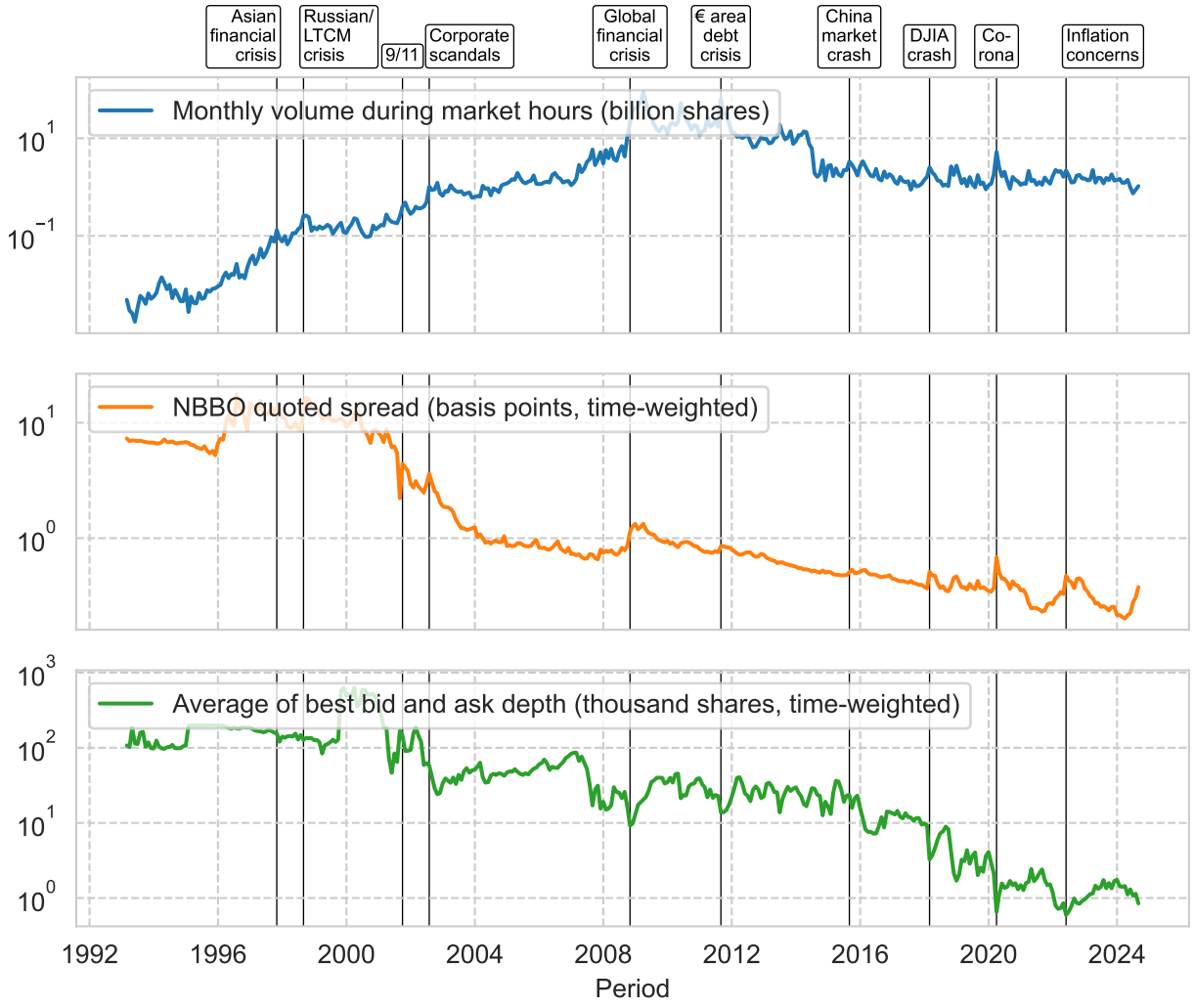

Figure 2: Time series SPY trading. This figure plots various trade statistics for SPY based on WRDS TAQ data: Volume, NBBO quoted spread, and NBBO depth.

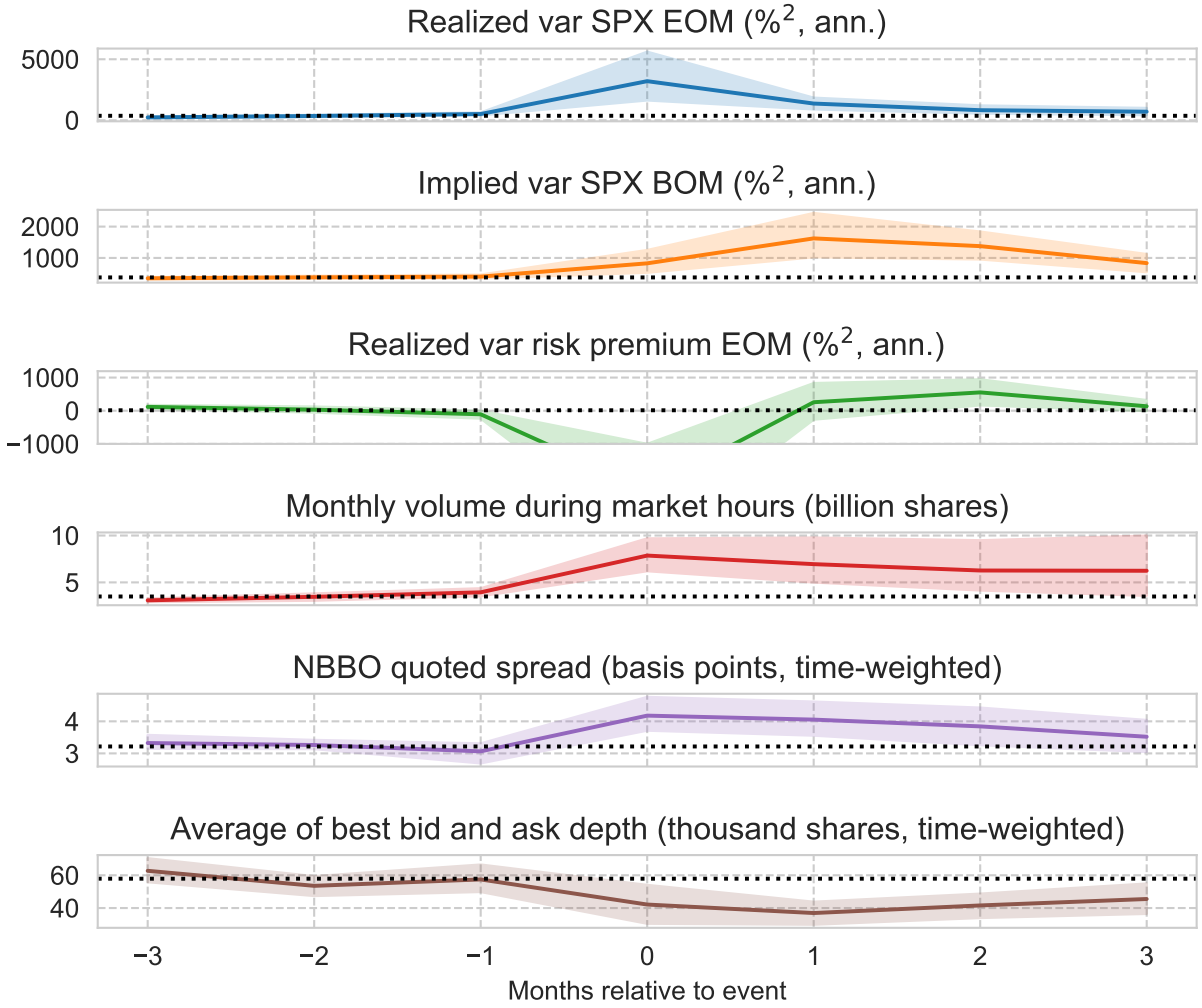

Figure 3: Crisis patterns. This figure plots various trade series evolve in crisis periods. These periods correspond with realized variance spikes (see vertical lines in Figure 1). All series are scaled by the pre-event level, which is set to 100. All plots show a 95% confidence interval.

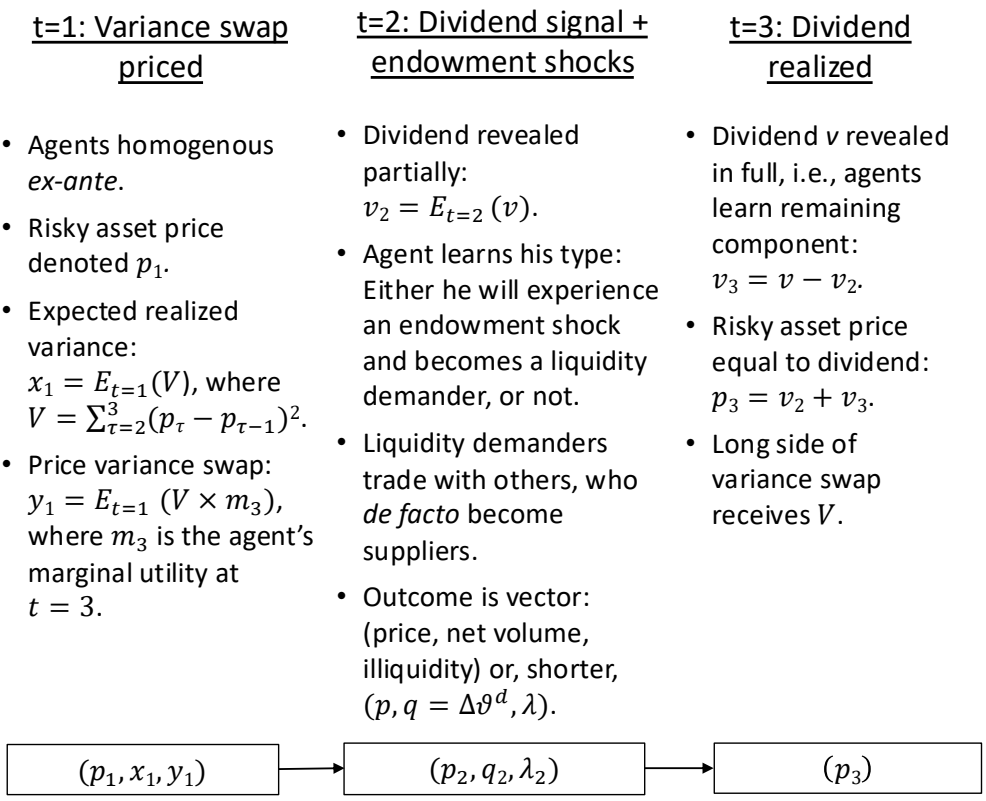

Figure 4: Model summary. This schematic summarizes the model by describing the events in the three periods. The flow diagram at the bottom shows the key variables determined in these periods.

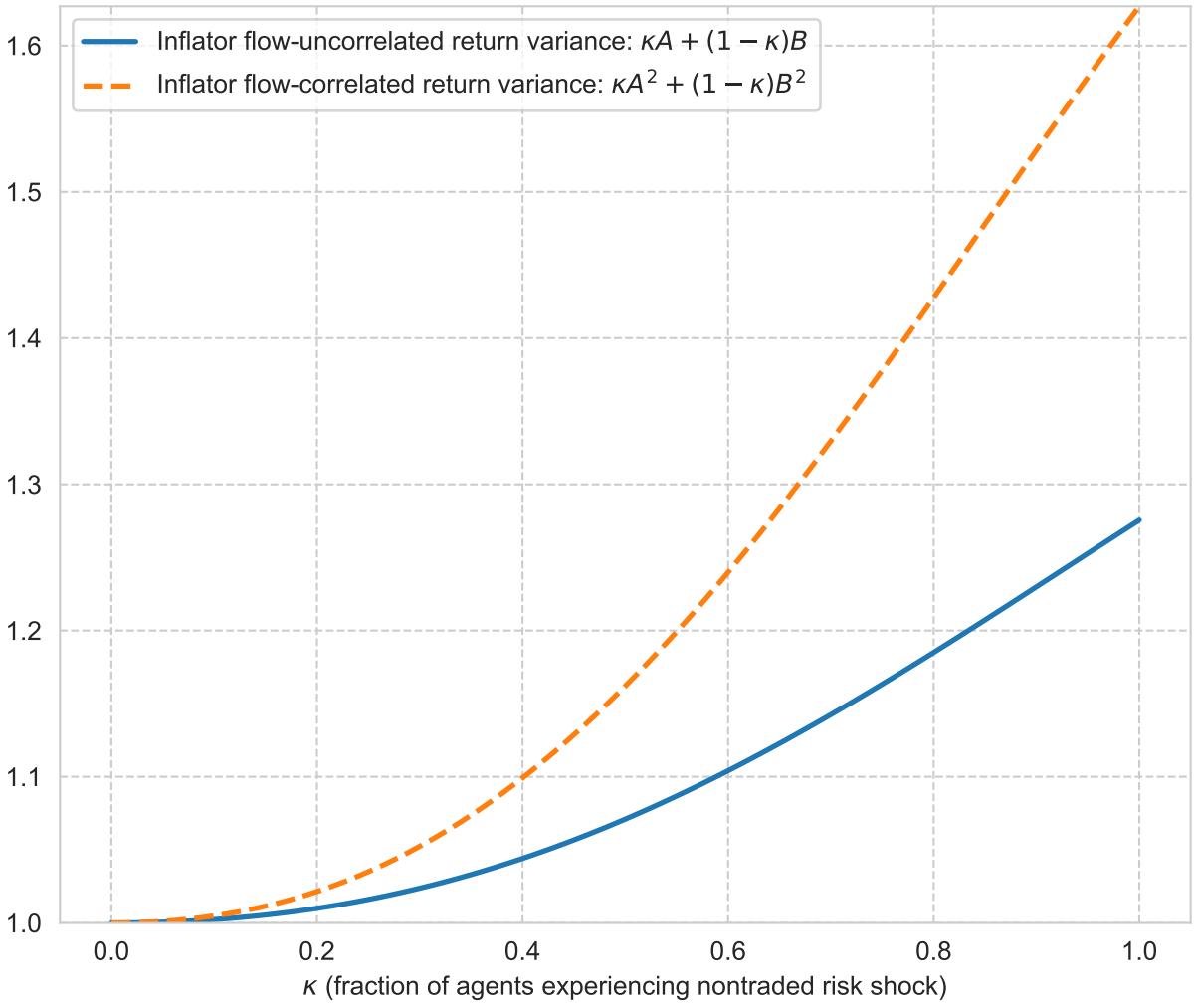

Figure 5: Variance inflators. This figure plots the variance inflators for the flow-correlated and flow-uncorrelated returns. These inflators are used in the pricing of the variance swaps for $\gamma = 3$ and $\sigma^2 = 400$ percentage squared annually. The flow-correlated part is assumed to be one fifth of the total variance, and the flow-uncorrelated part four fifths of it (i.e., $\sigma_{v2}^2 = 0.2\sigma^2$ and $\sigma_{v3}^2 = 0.8\sigma^2$). The variance of the nontraded risk shock is assumed to be nine: $\sigma_z^2 = 9$.

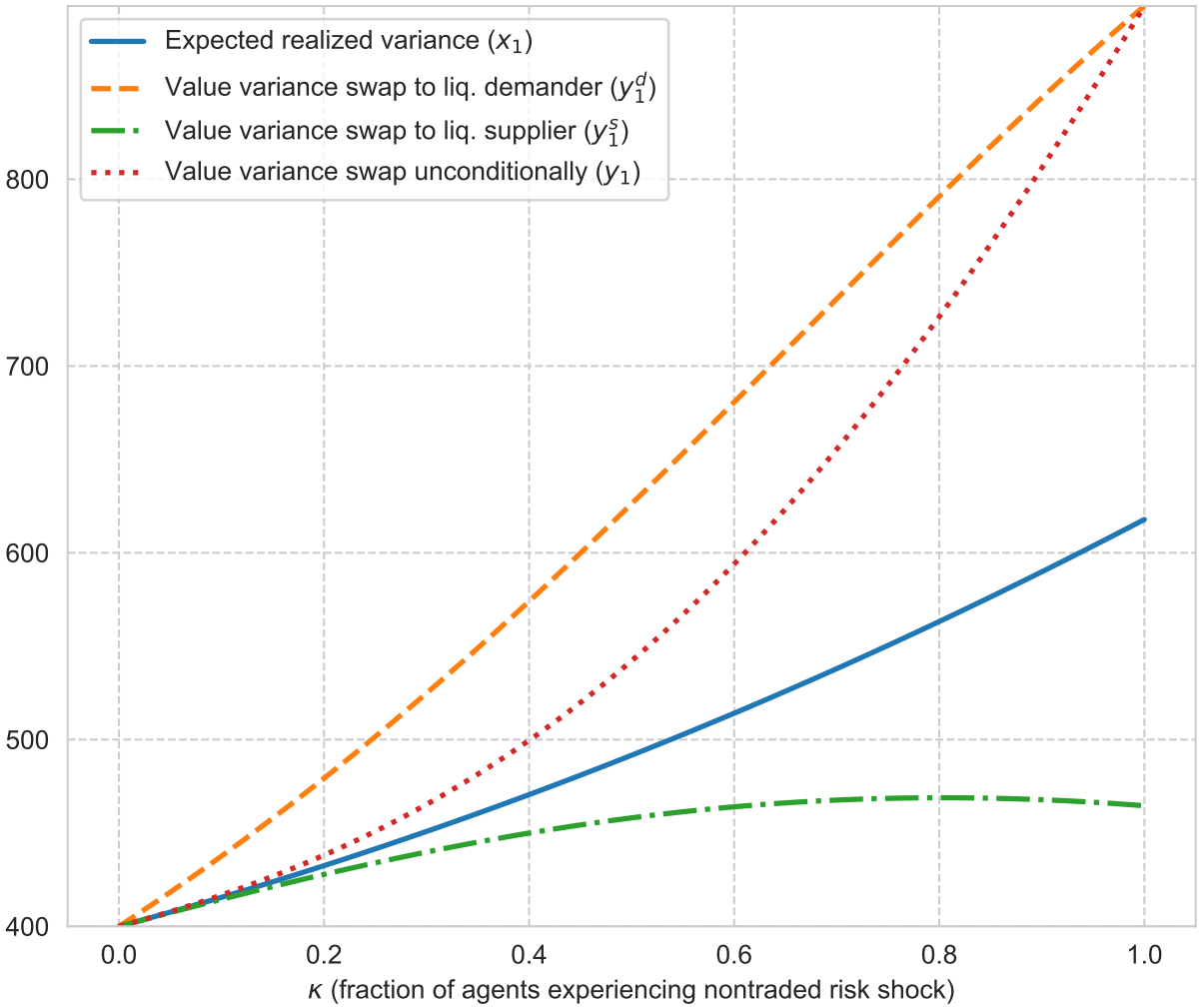

Figure 6: Pricing of the variance swap. This figure illustrates the pricing of the variance swaps for $\gamma = 3$ and $\sigma^2 = 400$ percentage squared annually. The flow-correlated part is assumed to be one fifth of the total variance, and the flow-uncorrelated part four fifths of it (i.e., $\sigma_{v2}^2 = 0.2\sigma^2$ and $\sigma_{v3}^2 = 0.8\sigma^2$). The variance of the nontraded risk shock is assumed to be nine: $\sigma_z^2 = 9$. The graph plots the valuation of a variance swap by the two types. It also plots the pricing of the variance swap as the weighted sum of the type-specific valuations. The graph further plots expected realized variance.