ISSN 0956-8549-750

What is the expected Return on the Market?

By
Ian Martin

DISCUSSION PAPER NO 750

DISCUSSION PAPER SERIES

March 2016

What is the Expected Return on the Market?

Ian Martin*

March, 2016

Abstract

This paper presents a new lower bound on the equity premium in terms of a volatility index, SVIX, that can be calculated from index option prices. This bound, which relies only on very weak assumptions, implies that the equity premium is extremely volatile, and that it rose above 20% at the height of the crisis in 2008. More aggressively, I argue that the lower bound—whose time-series average is about 5%—is approximately tight and that the high equity premia available at times of stress largely reflect high expected returns over the very short run.

*London School of Economics; http://personal.lse.ac.uk/martiniw. I thank John Campbell, John Cochrane, George Constantinides, Darrell Duffie, Bernard Dumas, Lars Hansen, Bryan Kelly, Gordon Liao, Stefan Nagel, Lubos Pastor, Christopher Polk, José Scheinkman, Dimitri Vayanos, and to seminar participants at Stanford University, Northwestern University, the NBER Summer Institute, INSEAD, Swiss Finance Institute, MIT Sloan, Harvard University, Morgan Stanley, Princeton University, London School of Economics, the Federal Reserve Bank of Atlanta, Toulouse School of Economics, FIME, BI Business School, Copenhagen Business School, Washington University in St Louis, Stockholm School of Economics, the Brazilian Finance Society, Tuck School of Business, the University of Chicago, Warwick Business School, the Bank of England, Cambridge University, and the European Central Bank for their comments. I am very grateful to John Campbell and Robin Greenwood for sharing data, and to Brandon Han for excellent research assistance. I also thank the ERC for their support under Starting Grant 639744.

The expected excess return on the market, or equity premium, is one of the central quantities of finance. Aside from its obvious intrinsic interest, the equity premium is a key determinant of the risk premium required for arbitrary assets in the CAPM and its descendants; and time-variation in the equity premium lies at the heart of the literature on excess volatility.

The starting point of this paper is an identity that relates the market's expected return to its risk-neutral variance. Under the weak assumption of no-arbitrage, the latter can be measured unambiguously from index option prices. I call the associated volatility index SVIX and use the identity (coupled with a minimal assumption, the *negative correlation condition*, introduced in Section 1) to derive a lower bound on the equity premium in terms of the SVIX index. The bound implies that the equity premium is extremely volatile, and that it rose above 21% at the height of the crisis in 2008. At horizons of less than a year, the equity premium fluctuates even more wildly: the lower bound on the *monthly* equity premium exceeded 4.5% (unannualized) in November 2008.

I go on to argue, more aggressively, that the lower bound appears empirically to be approximately tight, so that the SVIX index provides a direct measure of the equity premium. While it is now well understood that the equity premium is time-varying, this paper deviates from the literature in its basic aim, which is to use theory to motivate a signal of whether expected returns are high or low at a given point in time that is based *directly on asset prices*. The distinctive features of my approach, relative to the literature, are that (i) the predictor variable, $SVIX^2$, is motivated by asset pricing theory; (ii) no parameter estimation is required, so concerns over in-sample/out-of-sample fit do not arise; and (iii) since the $SVIX^2$ index is an asset price, I avoid the need to use infrequently-updated accounting data. My approach therefore allows the equity premium to be measured *in real time*.

The $SVIX^2$ index can be interpreted as the equity premium perceived by an unconstrained rational investor with log utility who is fully invested in the market. This is a sensible benchmark even if there are many investors who are constrained and many investors who are irrational, and it makes for a natural comparison with survey evidence on investor expectations, as studied by Shiller (1987) and Ben-David, Graham and Harvey (2013), among others. In particular, Greenwood and Shleifer (2014) emphasize the unsettling fact that the 'expectations of returns' extracted from surveys are

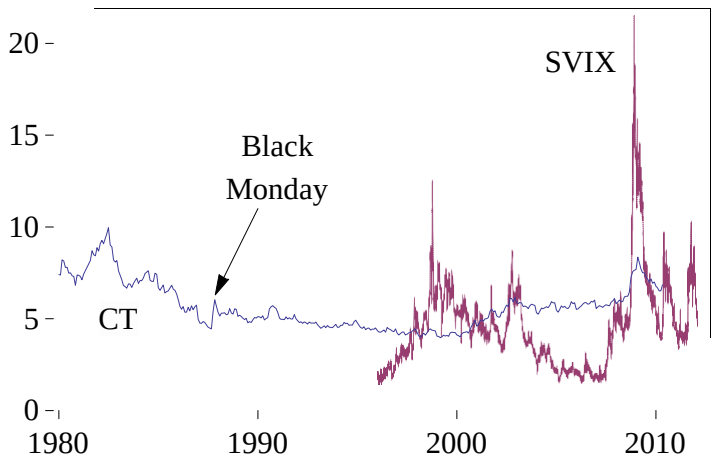


Figure 1: Equity premium forecasts based on Campbell–Thompson (CT, 2008) and on SVIX. Annual horizon.

negatively correlated with subsequent realized returns. Greenwood and Shleifer also document the closely related fact that a range of survey measures of return expectations are negatively correlated with the leading predictor variables used in the literature to forecast expected returns. I show that the SVIX-based equity premium forecast is also negatively correlated with the survey measures of return expectations. But the SVIX forecast is positively correlated with subsequent returns—a minimal requirement for a measure of rationally expected returns.

The view of the equity premium that emerges from the SVIX measure deviates in several interesting ways from the conventional view based on valuation-ratio-based measures. Figure 1 plots the SVIX equity premium measure on the same axes as the smoothed earnings yield predictor of Campbell and Thompson (2008), whose work I take as representative of the vast predictability literature because their approach, like mine, avoids the in-sample/out-of-sample critique of Goyal and Welch (2008).[1] The figure illustrates the results of the paper: I argue that the equity premium is more volatile, more right-skewed, and that it fluctuates at a higher frequency than the literature has acknowledged.

I sharpen the distinction between the SVIX and valuation-ratio views of the world

[1]Early papers in this literature include Keim and Stambaugh (1986), Campbell and Shiller (1988), and Fama and French (1988). A more recent paper that also argues for volatile discount rates is Kelly and Pruitt (2013). I thank John Campbell for sharing an updated version of the dataset used in Campbell and Thompson (2008).

by focussing on two periods in which their predictions diverge. Valuation-ratio-based measures of the equity premium were famously bearish throughout the late 1990s (and as noted by Ang and Bekaert (2007) and Goyal and Welch (2008), that prediction is partially responsible for the poor performance of valuation-ratio predictors in recent years); in contrast, the SVIX index suggests that, at horizons up to one year, expected returns were high in the late 1990s. I suggest that this distinction reflects the fact that valuation ratios should be thought of as predictors of very long run returns, whereas the SVIX index aims to measure short-run expected returns. The most striking divergence in predictions, however, occurs on one of the most dramatic days in stock market history, the great crash of October 1987, when option prices soared as the market collapsed.[2] On the valuation-ratio view of the world, the equity premium barely changed on Black Monday; on the SVIX view, it exploded.

1 Expected returns and risk-neutral variance

If we use asterisks to denote quantities calculated with risk-neutral probabilities, and M_T to denote the stochastic discount factor (SDF) that prices time-T payoffs from the perspective of time t, then we can price any time-T payoff X_T either via the SDF or by computing expectations with risk-neutral probabilities and discounting at the (gross) riskless rate, $R_{f,t}$, which is known at time t. The SDF notation,

$$\text{time-}t \text{ price of a claim to } X_T \text{ at time } T = \mathrm{E}_t(M_T X_T), \tag{1}$$

is commonly used in equilibrium models or, more generally, whenever there is an emphasis on the real-world distribution (whether from the subjective perspective of an agent within a model, or from the 'objective' perspective of the econometrician).

The risk-neutral notation,

$$\text{time-}t \text{ price of a claim to } X_T \text{ at time } T = \frac{1}{R_{f,t}} \mathrm{E}_t^* X_T, \tag{2}$$

is commonly used in derivative pricing, or more generally whenever the underlying logic is that of no-arbitrage. The choice of whether to use SDF or risk-neutral notation

[2] Figure 15, in the appendix, shows that the VXO index—that is, 1-month at-the-money implied volatility on the S&P 100—rose *extremely* sharply on October 19, 1987. (The VIX index itself did not exist at that time.) As it turned out, the annualized return on the S&P 500 index was 81.2% over the month, and 23.2% over the year, following Black Monday.

is largely a matter of taste; I will tend to follow convention by using the risk-neutral notation when no-arbitrage logic is emphasized.

Equations (1) and (2) can be used to translate between the two notations; thus, for example, the conditional risk-neutral variance of a gross return R_T is

$$\operatorname{var}_t^* R_T = \mathrm{E}_t^* R_T^2 - (\mathrm{E}_t^* R_T)^2 = R_{f,t}\,\mathrm{E}_t\left(M_T R_T^2\right) - R_{f,t}^2. \tag{3}$$

Expected returns and risk-neutral variance are linked by the following identity:

$$\begin{aligned} \mathrm{E}_t R_T - R_{f,t} &= \left[\mathrm{E}_t(M_T R_T^2) - R_{f,t}\right] - \left[\mathrm{E}_t(M_T R_T^2) - \mathrm{E}_t R_T\right] \\ &= \frac{1}{R_{f,t}} \operatorname{var}_t^* R_T - \operatorname{cov}_t(M_T R_T, R_T). \end{aligned} \tag{4}$$

The first equality adds and subtracts $\mathrm{E}_t(M_T R_T^2)$; the second exploits (3) and the fact that $\mathrm{E}_t M_T R_T = 1$.

The identity (4) decomposes the asset's risk premium into two components. It applies to any asset return R_T, but in this paper I will focus on the case in which R_T is the return on the S&P 500 index. In this case the first component, risk-neutral variance, can be computed directly given time-t prices of S&P 500 index options, as will be shown in Section 3. The second component is a covariance term that can be controlled: under a weak condition (discussed in detail in Section 2), it is negative.

Definition 1. *Given a gross return R_T and stochastic discount factor M_T, the* negative correlation condition (NCC) *holds if* $\operatorname{cov}_t(M_T R_T, R_T) \le 0$.

Together, the identity (4) and the NCC imply the following inequality, from which the results of the paper flow:

$$\mathrm{E}_t R_T - R_{f,t} \ge \frac{1}{R_{f,t}} \operatorname{var}_t^* R_T. \tag{5}$$

This inequality can be compared to the Hansen–Jagannathan (1991) bound. The two inequalities place opposing bounds on the equity premium:

$$\frac{1}{R_{f,t}} \operatorname{var}_t^* R_T \le \mathrm{E}_t R_T - R_{f,t} \le R_{f,t} \cdot \sigma_t(M_T) \cdot \sigma_t(R_T),$$

where $\sigma_t(\cdot)$ denotes conditional (real-world) standard deviation. The left-hand inequality is (5). It has the advantage that it relates the unobservable equity premium to a *directly observable* quantity, risk-neutral variance; but the disadvantage that it requires

the NCC to hold. In contrast, the right-hand inequality, the Hansen–Jagannathan bound, has the advantage of holding completely generally; but the disadvantage (noted by Hansen and Jagannathan) that it relates two quantities neither of which can be directly observed. Time-series averages must therefore be used as proxies for the true quantities of interest, forward-looking means and variances. This procedure requires assumptions about the stationarity and ergodicity of returns over appropriate sample periods and at the appropriate frequency. Such assumptions are not completely uncontroversial—see, for example, Malmendier and Nagel (2011).

The inequality (5) is reminiscent of the approach of Merton (1980), based on the equation

$$\text{instantaneous risk premium} = \gamma\sigma^2\,, \tag{6}$$

where γ is a measure of aggregate risk aversion, and σ^2 is the instantaneous variance of the market return, and of a closely related calculation carried out by Cochrane (2011, p. 1082).

There are some important differences between the two approaches, however. The first is that Merton assumes that the level of the stock index follows a geometric Brownian motion, thereby ruling out the effects of skewness and of higher cumulants by construction.[3] In contrast, we need no such assumption. Related to this, there is no distinction between risk-neutral and real-world (instantaneous) variance in a diffusion-based model: the two are identical, by Girsanov's theorem. Once we move beyond geometric Brownian motion, however, the appropriate generalization relates the risk premium to *risk-neutral* variance. As a bonus, this will have the considerable benefit that—unlike forward-looking real-world variance—forward-looking risk-neutral variance at time t can be directly and unambiguously computed from asset prices at time t, as I show in Section 3.

A second difference is that (6) requires that there is a representative agent with constant relative risk aversion γ. The NCC holds under considerably more general circumstances, as shown in Section 2.

Third, Merton implements (6) using realized historical volatility rather than by exploiting option price data, though he notes that volatility measures can be calculated "by 'inverting' the Black–Scholes option pricing formula." However, Black–Scholes

[3]Cochrane's calculation also implicitly makes this assumption; I will argue in Section 6.1 that it is inconsistent with the data.

implied volatility would only provide the correct measure of σ if we really lived in a Black–Scholes (1973) world in which prices followed geometric Brownian motions. The results of this paper show how to compute the right measure of variance in a more general environment.

2 The negative correlation condition

This section examines the NCC more closely in the case in which R_T is the return on the market; it is independent of the rest of the paper. I start by laying out various sufficient conditions for the NCC to hold. It is worth emphasizing that these conditions are not *necessary*: the NCC may hold even if none of the conditions below applies. The sufficient conditions cover many of the leading macro-finance models, including Campbell and Cochrane (1999), Bansal and Yaron (2004), Bansal, Kiku, Shaliastovich and Yaron (2012), Campbell, Giglio, Polk and Turley (2012), Barro (2006), and Wachter (2013).[4]

The NCC is a convenient and flexible way to restrict the set of stochastic discount factors under consideration. It may be helpful to note that the NCC would fail badly in a risk-neutral economy—that is, if M_T were deterministic. We will need the SDF to be volatile, as is the case empirically (Hansen and Jagannathan (1991)). We will also need the SDF to be negatively correlated with the return R_T; this will be the case for any asset that even roughly approximates the idealized notion of 'the market' in economic models.[5]

The first example of this section indicates, in a conditionally lognormal setting, why the NCC is likely to hold in practice. It shows, in particular, that the NCC holds in several leading macro-finance models. (All proofs for this section are in the appendix.)

Example 1. Suppose that the SDF M_T and return R_T are conditionally lognormal and write $r_{f,t} = \log R_{f,t}$, $\mu_{R,t} = \log \mathsf{E}_t R_T$, and $\sigma^2_{R,t} = \operatorname{var}_t \log R_T$. Then the NCC

[4] In fact, I am not aware of any model that attempts to match the data quantitatively in which the NCC does not hold.

[5] The NCC would fail for hedge assets (such as gold or, in recent years, US Treasury bonds) whose returns tend to be high at times when the marginal value of wealth is high—that is, for assets whose returns are positively correlated with the SDF. Indeed, it may be possible to exploit this fact to derive *upper* bounds on the returns on such assets.

is equivalent to the assumption that the conditional Sharpe ratio of the asset, $\lambda_t \equiv (\mu_{R,t} - r_{f,t})/\sigma_{R,t}$, exceeds its conditional volatility, $\sigma_{R,t}$.

The NCC therefore holds in any conditionally lognormal model in which the market's conditional Sharpe ratio is higher than its conditional volatility. Empirically, the Sharpe ratio of the market is on the order of 50% while its volatility is on the order of 16%, so it is unsurprising that this property holds in the calibrated models of Campbell and Cochrane (1999), Bansal and Yaron (2004), Bansal, Kiku, Shaliastovich and Yaron (2012) and Campbell, Giglio, Polk and Turley (2012), among many others.

The special feature of the lognormal setting is that real-world volatility and risk-neutral volatility are one and the same thing.[6] So if an asset's Sharpe ratio is larger than its (real-world or risk-neutral) volatility, then its expected excess return is larger than its (real-world or risk-neutral) variance. That is, by (4), the NCC holds.

Unfortunately, the lognormality assumption is inconsistent with well-known properties of index option prices. The most direct way to see this is to note that equity index options exhibit a volatility smile: Black–Scholes implied volatility varies across strikes, holding option maturity constant. (See also Result 4 below.) This concern motivates the next example, which provides an interpretation of the NCC that is not dependent on a lognormality assumption.

Example 2. Suppose that there is an unconstrained investor who maximizes expected utility over next-period wealth, whose wealth is fully invested in the market, and whose relative risk aversion (which need not be constant) is at least one at all levels of wealth. Then the NCC holds for the market return. Moreover, if (but *not* only if) the investor has log utility, the covariance term in (4) is identically zero; then, the inequality (5) holds with *equality*, and $\mathrm{E}_t R_T - R_{f,t} = \frac{1}{R_{f,t}} \mathrm{var}^*_t R_T$.

Example 2 does not require that the identity of the investor whose wealth is fully invested in the market should be fixed over time; thus it allows for the possibility that the portfolio holdings and beliefs of (and constraints on) different investors are highly heterogeneous over time. Moreover, it does not require that all investors are fully invested in the market, that all investors are unconstrained, or that all investors are rational. In view of the evidence presented by Greenwood and Shleifer (2014), this is

[6] More precisely, $\mathrm{var}_t \log R_T = \mathrm{var}^*_t \log R_T$ if M_T and R_T are conditionally jointly lognormal under the real-world measure.

an attractive feature. Under the interpretation of Example 2, the question answered by this paper is this: What expected return must be perceived by an unconstrained investor with log utility who chooses to hold the market? This is a natural benchmark: there are many ways to be constrained, but only one way to be unconstrained. For reasons that will become clear in Sections 4.2 and 6.1, I prefer to interpret the data from the perspective of a log investor who holds the market, rather than the familiar representative investor who consumes aggregate consumption. Thus my approach has nothing to say about—in particular, it does not resolve—the equity premium puzzle. In fact, on the contrary, the paper documents yet another dimension on which existing equilibrium models fail to fit the data; see Section 6.1.

By focussing on a one-period investor, Example 2 abstracts from intertemporal issues, and therefore from the presence of state variables that affect the value function. To the extent that we are interested in the behavior of long-lived utility-maximizing investors, we may want to allow for the fact that investment opportunities vary over time, as in the framework of Merton (1973). When will the NCC hold in (a discrete-time analog of) Merton's framework? Example 1 provided one answer to this question, but we can also frame sufficient conditions directly in terms of the properties of preferences and state variables, as in the next example (in which the driving random variables are Normal, as in Example 1; this assumption will shortly be relaxed).

Example 3a. Suppose, in the notation of Cochrane (2005, pp. 166–7), that the SDF takes the form

$$M_T = \beta \frac{V_W(W_T, z_{1,T}, \ldots, z_{N,T})}{V_W(W_t, z_{1,t}, \ldots, z_{N,t})},$$

where W_T is the time-T wealth of a risk-averse investor whose wealth is fully invested in the market, so that $W_T = (W_t - C_t)R_T$ (where C_t denotes the investor's time-t consumption and R_T the return on the market); V_W is the investor's marginal value of wealth; and $z_{1,T}, \ldots, z_{N,T}$ are state variables, with signs chosen so that V_W is weakly decreasing in each (so a high value of $z_{1,T}$ is good news, just as a high value of W_T is good news). Suppose also that

(i) Risk aversion is sufficiently high: $-WV_{WW}/V_W \geq 1$ at all levels of wealth W and all values of the state variables.

(ii) The market return, R_T, and state variables, $z_{1,T}, \ldots, z_{N,T}$, are increasing functions of conditionally Normal random variables with (weakly) positive pairwise

correlations.

Then the NCC holds for the market return.

Condition (i) imposes an assumption that risk aversion is at least one, as in Example 2; again, risk aversion may be wealth- and state-dependent. Condition (ii) ensures that the movements of state variables do not undo the logic of Example 1. To get a feel for it, consider a model with a single state variable, the price-dividend ratio of the market (perhaps as a proxy for the equity premium, as in Campbell and Viceira (1999)).[7] For consistency with the sign convention on the state variables, we need the marginal value of wealth to be weakly decreasing in the price-dividend ratio. It is intuitively plausible that the marginal value of wealth should indeed be high in times when valuation ratios are low; and this holds in Campbell and Viceira's setting, in the power utility case, if risk aversion is at least one.[8] Then condition (iii) amounts to the (empirically extremely plausible) requirement that the correlation between the wealth of the representative investor and the market price-dividend ratio is positive. Equivalently, we need the return on the market and the market price-dividend ratio to be positively correlated. Again, this holds in Campbell and Viceira's calibration.

Example 3a assumes that the investor is fully invested in the market. Roll (1977) famously criticized empirical tests of the CAPM by pointing out that stock market indices are imperfect proxies for the idealized notion of 'the market' that may not fully capture risks associated with labor or other sources of income. Without denying the force of this observation, the implicit position taken is that although the S&P 500 index is not the sum total of all wealth, it *is* reasonable to ask, as a benchmark, what equity premium would be perceived by someone fully invested in the S&P 500. (In contrast, it would be much less reasonable to assume that some investor holds all of his wealth in gold in order to estimate the expected return on gold.)

Nonetheless, one may want to allow part of the investor's wealth to be held in assets other than the equity index. The next example generalizes Example 3a to do so. It also generalizes in another direction, by allowing the driving random variables to be

[7]The price-dividend ratio is positive, so evidently cannot be Normally distributed; this is why condition (ii) allows the state variables to be arbitrary increasing functions of Normal random variables. For instance, we may want to assume that the log price-dividend ratio is conditionally Normal, as Campbell and Viceira do.

[8]Campbell and Viceira also allow for Epstein–Zin preferences, which I handle separately below.

non-Normal.

Example 3b. Modify Example 3a by assuming that only a fraction α_t of wealth net of consumption is invested in 'the market' (that is, in the equity index that is the focus of this paper), with the remainder invested in some other asset or portfolio of assets that earns the gross return $R_T^{(i)}$:

$$W_T = \underbrace{\alpha_t(W_t - C_t)R_T}_{\text{market wealth, } W_{M,T}} + \underbrace{(1-\alpha_t)(W_t - C_t)R_T^{(i)}}_{\text{non-market wealth}}.$$

If the signs of state variables are chosen as in Example 3a, and if

(i) Risk aversion is sufficiently high: $-WV_{WW}/V_W \geq W_T/W_{M,T}$.

(ii) $R_T, R_T^{(i)}, z_{1,T}, \ldots, z_{N,T}$ are *associated* random variables.[9]

then the NCC holds for the market return.

Condition (i) shows that we can allow the investor's wealth to be less than fully invested in the market (for example, in bonds, housing, and human capital), so long as he cares more about the position he does have—that is, has higher risk aversion. If, say, at least a third of the investor's time-T wealth is invested in the market, then the NCC holds so long as risk aversion is at least three.

The next example handles models, such as Wachter (2013), that are neither conditionally lognormal nor feature investors with time-separable utility.

Example 4a. Suppose that there is a representative agent with Epstein–Zin (1989) preferences. If (i) risk aversion $\gamma \geq 1$ and elasticity of intertemporal substitution $\psi \geq 1$, and (ii) the market return R_T and wealth-consumption ratio W_T/C_T are associated, then the NCC holds for the market return.

As special cases, condition (ii) would hold if, say, the log return $\log R_T$ and log wealth-consumption ratio $\log W_T/C_T$ are both Normal and nonnegatively correlated; or if the elasticity of intertemporal substitution $\psi = 1$, since then the wealth-consumption

[9]The concept of *associated* random variables (Esary, Proschan and Walkup (1967)) extends the concept of nonnegative correlation in a manner that can be extended to the multivariate setting. In particular, jointly Normal random variables are associated if and only if they are nonnegatively correlated (Pitt (1982)), and increasing functions of associated random variables are associated; thus Example 3a is a special case of Example 3b.

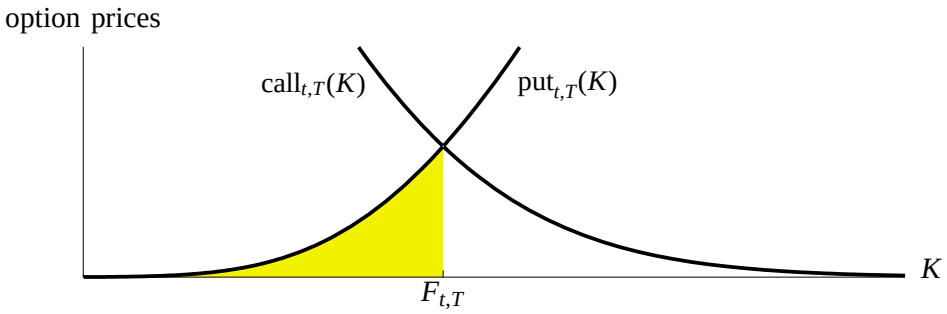


Figure 2: The prices, at time t, of call and put options expiring at time T.

ratio is constant (and hence, trivially, associated with the market return). This second case covers Wachter's (2013) model with time-varying disaster risk.

Example 4b. If there is a representative investor with Epstein–Zin (1989) preferences, with risk aversion $\gamma = 1$ and arbitrary elasticity of intertemporal substitution then the NCC holds *with equality* for the market return. This case was considered (and not rejected) by Epstein and Zin (1991) and Hansen and Jagannathan (1991).

3 Risk-neutral variance and the SVIX index

We now turn to the question of measuring the risk-neutral variance that appears on the right-hand side of (5). The punchline will be that risk-neutral variance is uniquely pinned down by European option prices, by a static no-arbitrage argument. To streamline the exposition, I will temporarily assume that the prices of European call and put options expiring at time T on the asset with return R_T are perfectly observable at all strikes K; this unrealistic assumption will be relaxed below.

Figure 2 plots a generic collection of time-t prices of calls expiring at time T with strike K (written call$_{t,T}(K)$) and of puts expiring at time T with strike K (written put$_{t,T}(K)$). The figure illustrates two well-known facts that will be useful. First, call and put prices are convex functions of strike. (Any non-convexity would provide a static arbitrage opportunity.) This property will allow us, below, to deal with the issue that option prices are only observable at a limited set of strikes. Second, the forward price of the underlying asset, $F_{t,T}$, which satisfies

$$F_{t,T} = \mathsf{E}^*_t S_T, \tag{7}$$

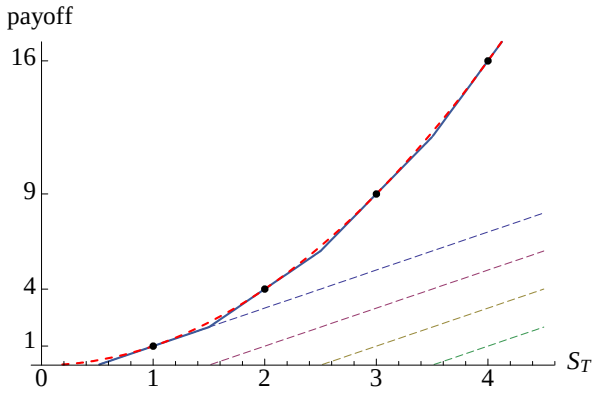


Figure 3: The payoff S_T^2 (dotted line); and the payoff on a portfolio of options (solid line), consisting of two calls with strike $K = 0.5$, two calls with $K = 1.5$, two calls with $K = 2.5$, two calls with $K = 3.5$, and so on. Individual option payoffs are indicated by dashed lines.

can be determined by observing the strike at which call and put prices are equal, i.e., $F_{t,T}$ is the unique solution x of the equation $\text{call}_{t,T}(x) = \text{put}_{t,T}(x)$. This fact follows from put-call parity; it means that the forward price can be backed out from time-t option prices.

We want to measure $\frac{1}{R_{f,t}} \text{var}_t^* R_T$. I assume that the dividends earned between times t and T are known at time t and paid at time T,[10] so that

$$\frac{1}{R_{f,t}} \text{var}_t^* R_T = \frac{1}{S_t^2} \left[\frac{1}{R_{f,t}} \text{E}_t^* S_T^2 - \frac{1}{R_{f,t}} (\text{E}_t^* S_T)^2 \right]. \tag{8}$$

We can deal with the second term inside the square brackets using equation (7), so the challenge is to calculate $\frac{1}{R_{f,t}} \text{E}_t^* S_T^2$. This is the price of the 'squared contract'—that is, the price of a claim to S_T^2 paid at time T.

How can we price this contract, given put and call prices as illustrated in Figure 2? Suppose we buy two call options with a strike of $K = 0.5$; two calls with a strike of $K = 1.5$; two calls with a strike of $K = 2.5$; two calls with a strike of $K = 3.5$; and so on, up to arbitrarily high strikes. The payoffs on the individual options are shown as dashed lines in Figure 3, and the payoff on the portfolio of options is shown as a solid line. The idealized payoff S_T^2 is shown as a dotted line. The solid and dotted lines

[10] If dividends are not known ahead of time, it is enough to assume that prices and dividends are (weakly) positively correlated, since then $\text{var}_t^* R_T \geq \text{var}_t^*(S_T/S_t)$, so that using $\frac{1}{R_{f,t}} \text{var}_t^*(S_T/S_t)$ instead of the ideal lower bound, $\frac{1}{R_{f,t}} \text{var}_t^* R_T$, is conservative.

almost perfectly overlap, illustrating that the payoff on the portfolio is almost exactly S_T^2 (and it is *exactly* S_T^2 at integer values of S_T). Therefore, the price of the squared contract is approximately the price of the portfolio of options:

$$\frac{1}{R_{f,t}} \mathrm{E}_t^* S_T^2 \approx 2 \sum_{K=0.5,1.5,\ldots} \mathrm{call}_{t,T}(K). \tag{9}$$

I show in the appendix that the squared contract can be priced exactly by replacing the sum with an integral:

$$\frac{1}{R_{f,t}} \mathrm{E}_t^* S_T^2 = 2 \int_{K=0}^{\infty} \mathrm{call}_{t,T}(K)\, dK. \tag{10}$$

This is an application of the classic result of Breeden and Litzenberger (1978).

In practice, however, option prices are not observable at *all* strikes K, so we will need to approximate the idealized integral (10) by a sum along the lines of (9). To see how this will affect the results, notice that Figure 3 also demonstrates a subtler point: the option portfolio payoff is not just equal to the 'squared payoff' at integers, it is *tangent* to it, so that the payoff on the portfolio of options very closely approximates *and is always less than or equal to* the ideal squared payoff. As a result, the sum over call prices in (9) will be slightly *less* than the integral over call prices in (10). This implies that the bounds presented are robust to the fact that option prices are not observable at all strikes: they would be *even higher* if all strikes were observable. Section 3.1 expands on this point.

Finally, since deep-in-the-money call options are neither liquid in practice nor intuitive to think about, it is convenient to split the range of integration into two and use put-call parity to replace in-the-money call prices with out-of-the-money put prices. Doing so, and substituting the result back into (8), we find that

$$\frac{1}{R_{f,t}} \mathrm{var}_t^* R_T = \frac{2}{S_t^2} \left[\int_0^{F_{t,T}} \mathrm{put}_{t,T}(K)\, dK + \int_{F_{t,T}}^{\infty} \mathrm{call}_{t,T}(K)\, dK \right]. \tag{11}$$

The expression in the square brackets is the shaded area shown in Figure 2.

The right-hand side of (11) is strongly reminiscent of the definition of the VIX index, and indeed there are links that will be explored in Section 6. To bring out the connection it will be helpful to define an index, SVIX_t, via the formula

$$\mathrm{SVIX}_t^2 = \frac{2R_{f,t}}{(T-t)F_{t,T}^2} \left[\int_0^{F_{t,T}} \mathrm{put}_{t,T}(K)\, dK + \int_{F_{t,T}}^{\infty} \mathrm{call}_{t,T}(K)\, dK \right]. \tag{12}$$

horizon	mean	s.d.	skew	kurt	min	1%	10%	25%	50%	75%	90%	99%	max
1 mo	5.00	4.60	4.03	24.6	0.83	1.03	1.54	2.44	3.91	5.74	8.98	25.7	55.0
2 mo	5.00	3.99	3.37	17.5	1.01	1.20	1.65	2.61	4.11	5.91	8.54	23.5	46.1
3 mo	4.96	3.60	3.01	14.0	1.07	1.29	1.75	2.69	4.24	5.95	8.17	21.4	39.1
6 mo	4.89	2.97	2.37	9.13	1.30	1.53	1.95	2.88	4.39	6.00	7.69	16.9	29.0
1 yr	4.64	2.43	1.87	5.99	1.47	1.64	2.07	2.81	4.35	5.72	7.19	13.9	21.5

Table 1: Mean, standard deviation, skewness, excess kurtosis, and quantiles of the lower bound on the equity premium, $R_{f,t} \cdot \text{SVIX}^2_t$ at various horizons (annualized and measured in %).

The SVIX index measures the annualized risk-neutral variance of the realized excess return: comparing equations (11) and (12), we see that

$$\text{SVIX}^2_t = \frac{1}{T-t} \text{var}^*_t(R_T/R_{f,t}). \tag{13}$$

Inserting (11) into inequality (5), we have a lower bound on the expected excess return of any asset that obeys the NCC:

$$\text{E}_t R_T - R_{f,t} \geq \frac{2}{S_t^2} \left(\int_0^{F_{t,T}} \text{put}_{t,T}(K)\, dK + \int_{F_{t,T}}^{\infty} \text{call}_{t,T}(K)\, dK \right) \tag{14}$$

or, in terms of the SVIX index,

$$\frac{1}{T-t}(\text{E}_t R_T - R_{f,t}) \geq R_{f,t} \cdot \text{SVIX}^2_t. \tag{15}$$

The bound will be applied in the case of the S&P 500; from now on, R_T always refers to the gross return on the S&P 500 index. I construct a time series of the lower bound from January 4, 1996 to January 31, 2012 using option price data from *OptionMetrics*; Appendix B.1 contains full details of the procedure. I compute the bound for time horizons $T-t = 1$, 2, 3, 6, and 12 months. I report results in annualized terms; that is, both sides of the above inequality are multiplied by $\frac{1}{T-t}$ with t and T measured in years (so, for example, monthly expected returns are multiplied by 12 to convert them into annualized terms).

Figure 4a plots the lower bound, annualized and in percentage points, at the 1-month horizon. Figures 4b and 4c repeat the exercise at 3-month and 1-year horizons. Table 1 reports the mean, standard deviation, and various quantiles of the distribution of the lower bound in the daily data for horizons between 1 month and 1 year.

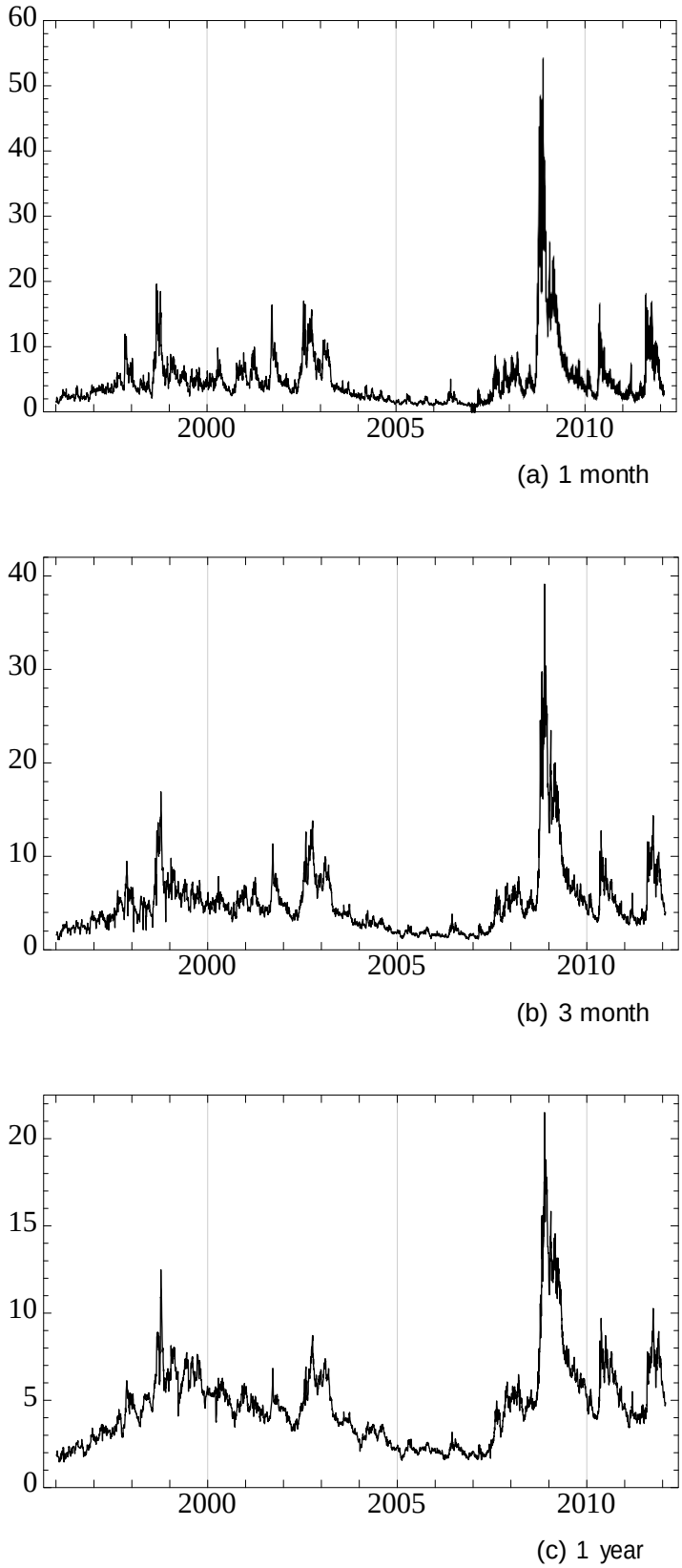


Figure 4: The lower bound on the annualized equity premium at different horizons.

The mean of the lower bound over the whole sample is 5.00% at the monthly horizon. This number is strikingly close to typical estimates of the unconditional equity premium, which suggests that the bound may be fairly tight: that is, it seems that the inequality (14) may approximately hold with equality. Below, I provide further tests of this possibility and develop some of its implications.

The time-series average of the lower bound is lower at the annual horizon than it is at the monthly horizon where the data quality is best (perhaps because of the existence of trades related to VIX, which is itself a monthly index). It is likely that this reflects a less liquid market in 1-year options, with a smaller range of strikes traded, rather than an interesting economic phenomenon. I discuss this further in Section 3.1 below.

The lower bound is volatile, right-skewed, and fat-tailed. At the annual horizon the equity premium varies from a minimum of 1.22% to a maximum of 21.5% over my sample period. But variation at the one-year horizon masks even more dramatic variation over shorter horizons. The monthly lower bound averaged only 1.86% (annualized) during the "Great Moderation" years 2004–2006, but peaked at 55.0%—more than 10 standard deviations above the mean—in November 2008, at the height of the subprime crisis. Indeed, the lower bound hit peaks at all horizons during the recent crisis, notably from late 2008 to early 2009 as the credit crisis gathered steam and the stock market fell, but also around May 2010, coinciding with the beginning of the European sovereign debt crisis. Other peaks occur during the LTCM crisis in late 1998; during the days following September 11, 2001; and during a period in late 2002 when the stock market was hitting new lows following the end of the dotcom boom.

Figure 13, in the appendix, shows that there was an increase in daily volume and open interest in S&P 500 index options over my sample period. The peaks in SVIX in 2008, 2010, and 2011 are associated with spikes in volume.

Consider, finally, a thought experiment. Suppose you find the lower bound on the equity premium in November 2008 implausibly high. What *trade* should you have done to implement this view? You should have sold a portfolio of options, namely an at-the-money-forward straddle and (equally weighted) out-of-the-money calls and puts. Such a position means that you end up short the market if the market rallies and long the market if the market sells off: essentially, you are taking a contrarian position, providing liquidity to the market. At the height of the credit crisis, extraordinarily high risk premia were available for investors who were able and prepared to take on

this position.

3.1 Robustness of the lower bound

Were option markets illiquid during the subprime crisis? One potential concern is that option markets may have been illiquid during periods of extreme stress. If so, one would expect to see a significant disparity between bounds based on mid-market option prices, such as those shown in Figure 4, and bounds based on bid or offer prices, particularly in periods such as November 2008. Thus it is possible in principle that the lower bounds would decrease significantly if bid prices were used. Figure 14, in the appendix, plots bounds calculated from bid prices. Reassuringly, the results are very similar: the lower bound is high at all horizons whether mid or bid prices are used.

Option prices are only observable at a discrete range of strikes. Two issues arise when implementing the lower bound. Fortunately, both issues mean that the numbers presented in this paper are conservative: with ideal data, the lower bound would be even higher.

First, we do not observe option prices at all strikes K between 0 and ∞. This means that the range of integration in the integral we would ideally like to compute—the shaded area in Figure 2—is truncated. Obviously, this will cause us to underestimate the integral in practice. This effect is likely to be strongest at the 1-year horizon, because (in my dataset) 1-year options are less liquid than shorter-dated options.

Second, even within the range of observable strikes, prices are only available at a discrete set of strikes. Thus the idealized lower bound that emerges from the theory in the form of an integral (over option prices at all strikes) must be approximated by a sum (over option prices at observable strikes). What effect will this have? In the discussion of Figure 2, I provided an example in which the price of a particular portfolio of calls with a discrete set of strikes would very slightly underestimate the idealized measure, and hence be conservative. The general case, using out-of-the-money puts and calls, is handled in Appendix B.2. The conclusion is the same: discretization leads to underestimates of risk-neutral variance, and hence to a conservative bound.

horizon	α	s.e.	β	s.e.	R^2	R^2_{OS}
1 mo	0.012	[0.064]	0.779	[1.386]	0.34%	0.42%
2 mo	−0.002	[0.068]	0.993	[1.458]	0.86%	1.11%
3 mo	−0.003	[0.075]	1.013	[1.631]	1.10%	1.49%
6 mo	−0.056	[0.058]	2.104	[0.855]	5.72%	4.86%
1 yr	−0.029	[0.093]	1.665	[1.263]	4.20%	4.73%

Table 2: Coefficient estimates for the regression (16).

4 SVIX as predictor variable

The time-series average of the lower bound in recent data is approximately 5% in annualized terms, a number close to conventional estimates of the equity premium. Over the period 1951–2000, Fama and French (2002) estimate the unconditional average equity premium to be 3.83% or 4.78%, based on dividend and earnings growth respectively.[11] It is therefore natural to wonder whether the lower bound might in fact be tight. We want to test the hypothesis that $\frac{1}{T-t}(\mathrm{E}_t R_T - R_{f,t}) = R_{f,t} \cdot \mathrm{SVIX}^2_{.t}$ Table 2 shows the results of regressions

$$\frac{1}{T-t}(R_T - R_{f,t}) = \alpha + \beta \times R_{f,t} \cdot \mathrm{SVIX}^2_t + \varepsilon_T\,, \tag{16}$$

together with robust Hansen–Hodrick standard errors that account for heteroskedasticity and overlapping observations. The null hypothesis that $\alpha = 0$ and $\beta = 1$ is not rejected at any horizon. The point estimates on β are close to 1 at all horizons, lending further support to the possibility that the lower bound is tight. This is encouraging because, as Goyal and Welch (2008) emphasize, this period is one in which conventional predictive regressions fare poorly.

One might worry that these results are entirely driven by the period in 2008 and 2009 in which volatility spiked and the stock market crashed before recovering strongly. To address this concern, Table 5, in the appendix, shows the result of deleting all observations that overlap with the period August 1, 2008–July 31, 2009. Over horizons

[11] These are the 'bias-adjusted' figures presented in their Table IV. In an interview with Richard Roll available on the AFA website at http://www.afajof.org/details/video/2870921/Eugene-Fama-Interview.html, Fama says, "I always think of the number, the equity premium, as five per cent."

of 1, 2, and 3 months, deleting this period in fact *increases* the forecastability of returns by SVIX, reflecting the fact that the market continued to drop for a time after volatility spiked up in November 2008. On the other hand, the subsequent strong recovery of the market means that this was a period in which 1-year options successfully predicted 1-year returns, so by removing the crash from the sample, the forecasting power deteriorates at the 1-year horizon.

We now have seen from two different angles that the lower bound (14) appears to be approximately tight: (i) as shown in Table 1 and Figure 4, the average level of the lower bound over my sample is close to conventional estimates of the average equity premium; and (ii) Table 2 shows that the null hypothesis that $\alpha = 0$ and $\beta = 1$ in the forecasting regression (16) is not rejected at any horizon. These observations suggest that SVIX can be used as a measure of the equity premium without estimating any parameters—that is, imposing $\alpha = 0$, $\beta = 1$ in (16), so that

$$\frac{1}{T-t}\left(\mathrm{E}_t R_T - R_{f,t}\right) = R_{f,t} \cdot \mathrm{SVIX}^2_t. \tag{17}$$

To assess the performance of the forecast (17), I follow Goyal and Welch (2008) in computing an out-of-sample R-squared measure

$$R^2_{OS} = 1 - \frac{\sum \varepsilon_t^2}{\sum v_t^2}, \tag{18}$$

where ε_t is the error when SVIX (more precisely, $R_{f,t} \cdot \mathrm{SVIX}^2_t$) is used to forecast the equity premium and v_t is the error when the historical mean equity premium (computed on a rolling basis) is used to forecast the equity premium.[12]

The rightmost column of Table 2 reports the values of R^2_{OS} at each horizon. These out-of-sample R^2_{OS} values can be compared with corresponding numbers for forecasts based on valuation ratios, which are the subject of a vast literature.[13] Goyal and Welch (2008) consider return predictions in the form

$$\text{equity premium}_t = a_1 + a_2 \times \text{predictor variable}_t, \tag{19}$$

where a_1 and a_2 are constants estimated from the data, and argue that while conventional predictor variables perform reasonably well in-sample, they perform worse

[12] More detail on the construction of the rolling mean is provided in the appendix.

[13] Among many others, Campbell and Shiller (1988), Fama and French (1988), Lettau and Ludvigson (2001), and Cochrane (2008) make the case for predictability. Other authors, including Ang and Bekaert (2007), make the case against.

out-of-sample than the rolling mean. Over their full sample (which runs from 1871 to 2005, with the first 20 years used to initialize estimates of a_1 and a_2, so that predictions start in 1891), the dividend-price ratio, dividend yield, earnings-price ratio, and book-to-market ratio have negative out-of-sample R^2s of −2.06%, −1.93%, −1.78% and −1.72%, respectively. The performance of these predictors is particularly poor over Goyal and Welch's 'recent sample' (1976 to 2005), with R^2s of −15.14%, −20.79%, −5.98% and −29.31%, respectively.[14]

Campbell and Thompson (2008) confirm Goyal and Welch's finding, and respond by suggesting that the coefficients a_1 and a_2 be fixed based on *a priori* considerations. Motivated by the Gordon growth model $D/P = R - G$ (where D/P is the dividend-price ratio, R the expected return, and G expected dividend growth), Campbell and Thompson suggest making forecasts of the form

$$\text{equity premium}_t = \text{dividend-price ratio}_t + \text{dividend growth}_t - \text{real interest rate}_t$$

or, more generally,

$$\text{equity premium}_t = \text{valuation ratio}_t + \text{dividend growth}_t - \text{real interest rate}_t, \quad (20)$$

where in addition to the dividend-price ratio, Campbell and Thompson also consider earnings yields, smoothed earnings yields, and book-to-market as valuation ratios. Since these forecasts are drawn directly from the data without requiring estimation of coefficients, they are a natural point of comparison for the forecast (17) suggested in this paper.

Over the full sample, the out-of-sample R^2s corresponding to the forecasts (20) range from 0.24% (using book-to-market as the valuation ratio) to 0.52% (using smoothed earnings yield) in monthly data; and from 1.85% (earnings yield) to 3.22% (smoothed earnings yield) in annual data.[15] The results are worse over Campbell and Thompson's most recent subsample, from 1980–2005: in monthly data, R^2 ranges from −0.27% (book-to-market) to 0.03% (earnings yield). In annual data, the forecasts do even more poorly, each underperforming the historical mean, with R^2s ranging from −6.20% (book-to-market) to −0.47% (smoothed earnings yield).

[14]Goyal and Welch show that the performance of an out-of-sample version of Lettau and Ludvigson's (2001) *cay* variable is similarly poor, with R^2 of −4.33% over the full sample and −12.39% over the recent sample.

[15]Out-of-sample forecasts are from 1927 to 2005, or 1956 to 2005 when book-to-market is used.

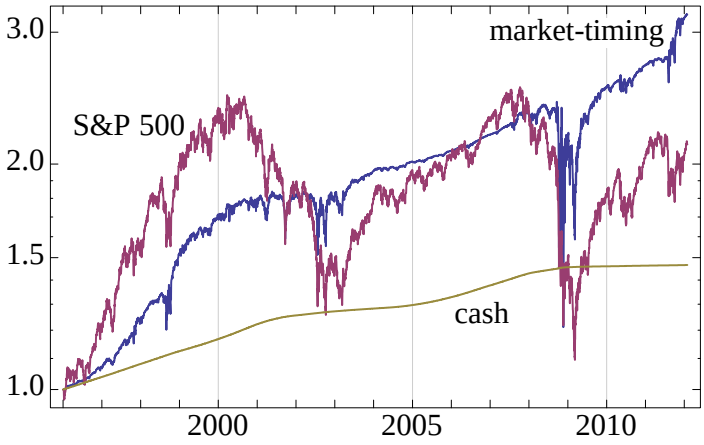


Figure 5: Cumulative returns on $1 invested in cash, in the S&P 500 index, or in a market-timing strategy whose allocation to the market is proportional to $R_{f,t} \cdot \text{SVIX}^2_t$. Log scale.

In relative terms, therefore, the out-of-sample *R*-squareds shown in Table 2 compare very favorably with the corresponding *R*-squareds for predictions based on valuation ratios. But are they too small to be interesting in absolute terms? No. Ross (2005, pp. 54–57) and Campbell and Thompson (2008) point out that high R^2 statistics in predictive regressions translate into high attainable Sharpe ratios, for the simple reason that the predictions can be used to formulate a market-timing trading strategy; and if the predictions are very good, the strategy will perform extremely well. If Sharpe ratios above some level are 'too good to be true,' then one should not expect to see R^2s from predictive regressions above some upper limit.

With this thought in mind, consider using risk-neutral variance in a contrarian market-timing strategy: invest, each day, a fraction α_t in the S&P 500 index and the remaining fraction $1-\alpha_t$ at the riskless rate, where α_t is chosen proportional to 1-month SVIX^2 (scaled by the riskless rate, as on the right-hand side of (15)). The constant of proportionality has no effect on the strategy's Sharpe ratio, so I choose it such that the market-timing strategy's mean portfolio weight in the S&P 500 is 35%, with the remaining 65% in cash; the resulting median portfolio weight is 27% in the S&P 500, with 73% in cash. Figure 5 plots the cumulative return on an initial investment of $1 in this market-timing strategy and, for comparison, on strategies that invest in the short-term interest rate or in the S&P 500 index. In my sample period, the daily

Sharpe ratio of the market is 1.35%, while the daily Sharpe ratio of the market-timing strategy is 1.97%; in other words, the out-of-sample R^2 of 0.42% reported in Table 2 is enough to deliver a 45% increase in Sharpe ratio for the market-timing strategy relative to the market itself. This exercise also illustrates the attractive feature that since risk-neutral variance is an asset price, it can be computed in daily data, or at even higher frequency, and so permits high-frequency market-timing strategies to be considered.

As illustrated in Figure 1, valuation ratios and SVIX tell qualitatively very different stories about the equity premium. First, option prices point toward a far more volatile equity premium than do valuation ratios. Second, SVIX is much less persistent than are valuation ratios, and so the SVIX predictor variable is less subject to Stambaugh (1999) bias. It is also noteworthy that SVIX forecasts a relatively high equity premium in the late 1990s. In this respect it diverges sharply from valuation-ratio-based forecasts, which predicted a low or even negative 1-year equity premium at the time.

But perhaps the most striking aspect of Figure 1 is the behavior of the Campbell–Thompson predictor variable on Black Monday, October 19, 1987. This was by far the worst day in stock market history. The S&P 500 index dropped by over 20%—more than twice as far as on the second-worst day in history—and yet the valuation-ratio approach suggests that the equity premium barely responded. In sharp contrast, option prices exploded on Black Monday, implying that the equity premium was even higher than the peaks attained in November 2008.

4.1 The term structure of equity premia

Campbell and Shiller (1988) showed that any dividend-paying asset satisfies the approximate identity

$$d_t - p_t = \text{constant} + \mathrm{E}_t \sum_{j=0}^{\infty} \rho^j \left(r_{t+1+j} - \Delta d_{t+1+j}\right),$$

which relates its log dividend yield $d_t - p_t$ to expectations of future log returns r_{t+1+j} and future log dividend growth Δd_{t+1+j}. Empirically, dividend growth is approximately unforecastable; to the extent that this is the case, we can absorb the terms $\mathrm{E}_t \Delta d_{t+1+j}$ into the constant, giving

$$d_t - p_t = \text{constant} + \mathrm{E}_t \sum_{j=0}^{\infty} \rho^j r_{t+1+j}. \tag{21}$$

This points a path toward reconciling the differing predictions of SVIX and valuation ratios. We can think of dividend yield as providing a measure of expected returns over the very long run. In contrast, the SVIX index measures expected returns over the short run.[16] The gap between the two is therefore informative about the gap between long-run and short-run expected returns. In the late 1990s, for example, $d_t - p_t$ was extremely low, indicating low expected long-run returns (Shiller (2000));[17] but Figures 4a–4c show that SVIX, and hence expected *short-run* returns, were relatively *high* at that time.

We can also compare expected returns across shorter horizons. For example, Figures 4a–4c suggest that an unusually large fraction of the elevated 1-year equity premium available in late 2008 was expected to materialize over the first few months of the 12-month period. To analyze this more formally, define the annualized *forward equity premium from T_1 to T_2* (calculated from the perspective of time t) by the formula

$$\mathrm{EP}_{T_1 \to T_2} \equiv \frac{1}{T_2 - T_1} \left(\log \frac{\mathrm{E}_t R_{t \to T_2}}{R_{f,t \to T_2}} - \log \frac{\mathrm{E}_t R_{t \to T_1}}{R_{f,t \to T_1}} \right), \tag{22}$$

and the corresponding 'spot' equity premium from time t to time T by

$$\mathrm{EP}_{t \to T} \equiv \frac{1}{T - t} \log \frac{\mathrm{E}_t R_{t \to T}}{R_{f,t \to T}}.$$

Using (17) to substitute out for $\mathrm{E}_t R_{t \to T_1}$ and $\mathrm{E}_t R_{t \to T_2}$ in (22), we can write

$$\mathrm{EP}_{T_1 \to T_2} = \frac{1}{T_2 - T_1} \log \frac{1 + \mathrm{SVIX}^2_{t \to T_2}(T_2 - t)}{1 + \mathrm{SVIX}^2_{t \to T_1}(T_1 - t)} \quad \text{and} \quad \mathrm{EP}_{t \to T} = \frac{1}{T - t} \log \left(1 + \mathrm{SVIX}_{t \to T}(T - t) \right).$$

(I have modified previous notation to accommodate the extra time dimension: for example, $R_{t \to T_2}$ is the simple return on the market from time t to time T_2, $R_{f,t \to T_1}$ is

---

[16] It would be interesting to narrow the gap between 'long' and 'short' run by exploring, in future research, expected returns over the intermediate horizons that should be most relevant for macroeconomic aggregates such as investment. How do risk premia at, say, five- or ten-year horizons behave? Data availability is a major challenge here: long-dated options are relatively illiquid.

[17] There is an important caveat. The discussion surrounding equation (21) follows much of the literature in blurring the distinction between expected arithmetic returns and the expected *log* returns that appear in the Campbell–Shiller loglinearization. Since $\mathrm{E}_t r_{t+1+j} = \log \mathrm{E}_t R_{t+1+j} - \frac{1}{2} \mathrm{var}_t r_{t+1+j} - \sum_{n=3}^{\infty} \frac{\kappa_t^{(n)}(r_{t+1+j})}{n!}$, where $\kappa_t^{(n)}(r_{t+1+j})$ is the nth conditional cumulant of r_{t+1+j}, the gap between the two depends on the cumulants of log returns. So a low dividend yield may be associated with *high* expected arithmetic returns at times when log returns are highly volatile, right-skewed, or fat-tailed.

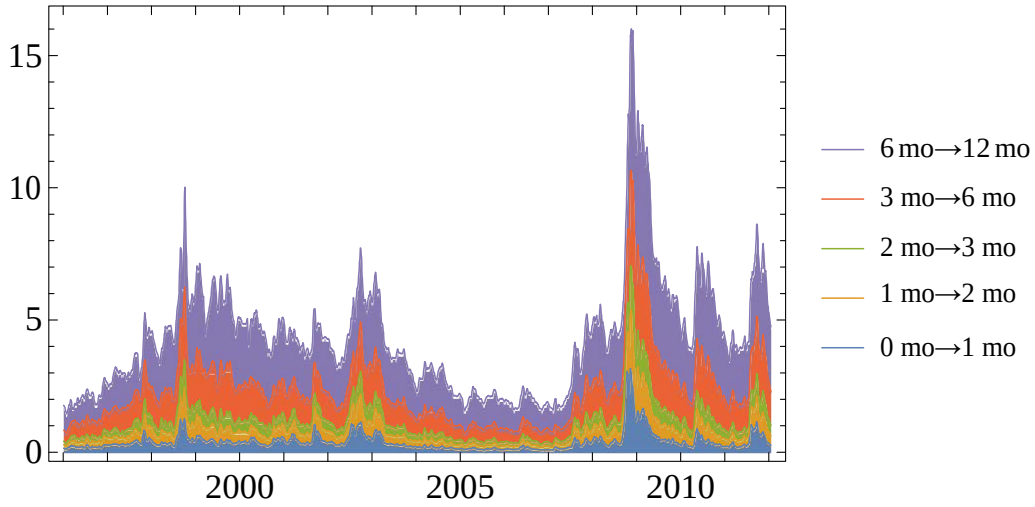


Figure 6: The term structure of equity premia. 10-day moving average.

the riskless return from time t to time T_1, and $\text{SVIX}^2_{t\to T_2}$ is the time-t level of the SVIX index calculated using options expiring at T_2.)

The definition (22) is chosen so that, for arbitrary $T_1, \ldots, T_N$, we have the decomposition

$$\text{EP}_{t\to T_N} = \frac{T_1 - t}{T_N - t}\,\text{EP}_{t\to T_1} + \frac{T_2 - T_1}{T_N - t}\,\text{EP}_{T_1\to T_2} + \cdots + \frac{T_N - T_{N-1}}{T_N - t}\,\text{EP}_{T_{N-1}\to T_N}, \qquad (23)$$

which expresses the long-horizon equity premium $\text{EP}_{t\to T_N}$ as a weighted average of forward equity premia, exactly analogous to the relationship between spot and forward bond yields.

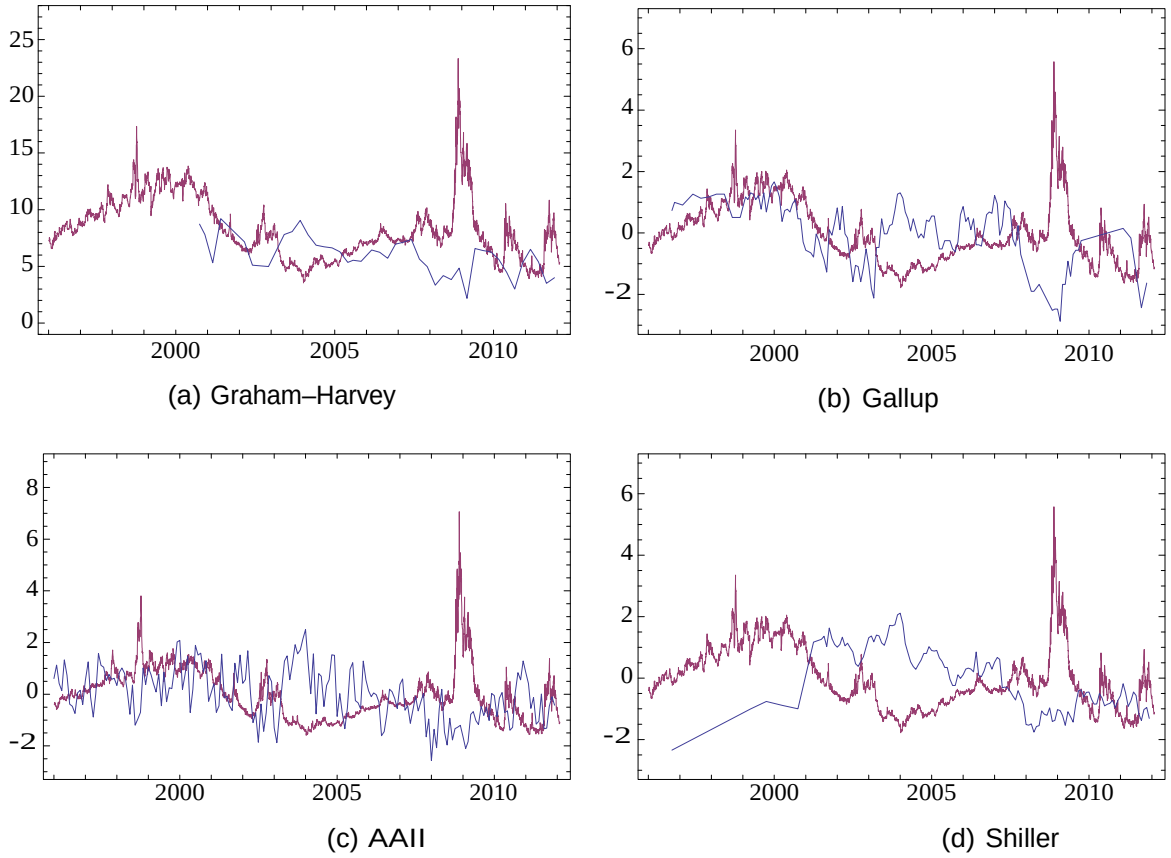
Figure 6 shows how the annual equity premium previously plotted in Figure 4c decomposes into a one-month spot premium plus forward premia from one to two, two to three, three to six, and six to twelve months. The figure stacks the *un*annualized forward premia—terms of the form $(T_n - T_{n-1})/(T_N - t)\,\text{EP}_{T_{n-1}\to T_n}$—which add up to the annual equity premium, as shown in (23). For example, on any given date t, the gap between the top two lines represents the contribution of the unannualized 6-month-6-month-forward equity premium, $\frac{1}{2}\,\text{EP}_{t+6\text{mo}\to t+12\text{mo}}$, to the annual equity premium, $\text{EP}_{t\to t+12\text{mo}}$.

In 'normal' times, the 6-month-6-month-forward equity premium contributes about half of the annual equity premium, as might have been expected. More interestingly, the figure shows that at times of stress, much of the annual equity premium is compressed into the first few months. For example, about a third of the equity premium over the

year from November 2008 to November 2009 can be attributed to the (unannualized) equity premium over the two months from November 2008 to January 2009.

4.2 Expectations of returns and expected returns

The view of the equity premium proposed above can usefully be compared with the expectations reported in surveys of market participants, as studied by Shiller (1987), Ben-David, Graham and Harvey (2013), and others. In particular, Greenwood and Shleifer (2014) emphasize that survey-based return expectations are negatively correlated with expected return forecasts based on conventional predictor variables. We will now see that this is also true when SVIX is used as a predictive variable.



(a) Graham–Harvey (b) Gallup

(c) AAII (d) Shiller

Figure 7: Expectations of returns (blue) and SVIX-implied expected returns (red). The units in panel (a) are percentage points. The time series in panels (b), (c), and (d) are normalized to have zero mean and unit variance. The forecasting horizon is one year for panels (a), (b), and (d), and six months for panel (c).

Figure 7 shows four of the survey measures considered by Greenwood and Shleifer: the Graham–Harvey Chief Financial Officer surveys, the Gallup investor survey measure, the American Association of Individual Investors (AAII) survey, and Robert Shiller's investor survey. The Graham–Harvey survey is based on the expectations of market returns reported by the chief financial officers of major US corporations; this survey can be compared directly with the expected return implied by SVIX. The other three measures are not in the same units, so panels (b), (c), and (d) of Figure 7 show time series standardized to have zero mean and unit variance. The Gallup survey measure is the percentage of investors who are "optimistic" or "very optimistic" about stock returns over the next year, minus the percentage who are "pessimistic" or "very pessimistic." The AAII survey measure is the percentage of surveyed individual investors (members of the AAII) who are "bullish" about stock returns over the next six months, minus the corresponding "bearish" percentage. The Shiller measure reports the percentage of individual investors surveyed who expected the market to go up over the following year.

Each panel also shows the time series of expected returns implied by the SVIX index (calculated by adding the riskless rate to the right-hand side of (17)). To be consistent with the phrasing of each survey, I compare the the Gallup, Graham–Harvey and Shiller surveys to the SVIX-implied equity premium (or expected return) at the 1-year horizon, and the AAII survey to the 6-month SVIX-implied equity premium (or expected return).

The most notable feature of Figure 7 is that survey expectations tend to move in the opposite direction from the rational measure of expected returns based on SVIX, as emphasized by Greenwood and Shleifer. As Table 3 reports, all four survey series are negatively correlated with the SVIX-implied equity premium.[18] This is true whether one measures correlations in levels or in differences, and whether one compares the surveys to the expected return on the market (that is, including the riskless rate, as in the series shown in Figure 7) or to the expected *excess* return on the market. There is also a contrast in that the skewness and excess kurtosis of the return expectations series are negative or close to zero, whereas they are strongly positive for SVIX, as

[18] I convert the SVIX-implied equity premium into a monthly series by averaging within months, and calculate correlations over all dates that are shared by SVIX and the appropriate survey-based measure.

	Gallup	Graham–Harvey	AAII	Shiller
skewness	−0.73	−0.10	0.04	0.06
excess kurtosis	0.04	−0.28	−0.53	−1.03
corr(survey,ER)	−0.06	−0.29	−0.20	−0.45
	[0.232]	[0.030]	[0.003]	[0.000]
corr(survey, EER)	−0.53	−0.50	−0.37	−0.46
	[0.000]	[0.000]	[0.000]	[0.000]
corr(Δsurvey, ΔER)	−0.40	−0.21	−0.29	−0.16
	[0.000]	[0.097]	[0.000]	[0.035]
corr(Δsurvey, ΔEER)	−0.44	−0.22	−0.27	−0.16
	[0.000]	[0.083]	[0.000]	[0.032]

Table 3: Skewness and excess kurtosis of return expectation measures; and correlations between return expectations and SVIX-implied expected returns (ER), and between return expectations and SVIX-implied expected *excess* returns (EER), in levels and in differences (denoted by Δ). Numbers in square brackets indicate *p*-values on the hypothesis that the correlation between the two series is zero.

shown in Table 1.[19] Moreover, the *lowest* points in the Graham–Harvey and Gallup series coincide with the *highest* point in the SVIX series.

Consistent with the thesis of Greenwood and Shleifer, it is implausible, given this evidence, that the surveyed investors have rational expectations.[20] This fact is unsettling for proponents of rational-expectations representative-agent models. (To compound

[19]The negative kurtosis of the Gallup and AAII measures may reflect the design of the surveys, each of which provides a fixed scale of possible responses.

[20]On the other hand, the findings of Shiller (1987)—reporting the results of investor surveys that were sent out in the immediate aftermath of the crash in October 1987—are potentially consistent with the thesis that subjectively expected returns may have been very high at short horizons following Black Monday. Although the survey questions Shiller asked are hard to compare directly with the results of this paper, he documents that a substantial fraction of investors expected a market rebound from the crash. Shiller also reports that some investors had more nuanced expectations of market returns: for instance, some thought that the market would perform better over shorter horizons than over long horizons, consistent with the results of Section 4.1.

the problem, I consider a range of leading representative-agent models in Section 6.1, and show that none can match the behavior of VIX and SVIX quantitatively, or even qualitatively.) Seen in a certain light, however, this cloud may have a silver lining: the fact that there is a systematic—albeit negative—relationship between (rationally) expected returns and the expectations of surveyed investors points to a pattern that may be amenable to modelling. Barberis, Greenwood, Jin and Shleifer (2015) take a first step in this direction by presenting an equilibrium model in which irrational extrapolators interact with rational investors. It is the latter class of investors whose expectations should be thought of as reflected in the SVIX index.

5 What is the probability of a crash?

The theory presented in Section 1 was based on a rather minimal assumption, the NCC. I argued in subsequent sections that the NCC may hold with equality, that is, that we may have $\text{cov}_t(M_T R_T, R_T) = 0$. I now strengthen this latter condition further by taking the perspective of an investor with log utility who chooses to invest fully in the market. The next result shows how to convert the problem of inferring the subjective expectations of such an investor into a *derivative pricing problem*.

Result 1. *Let X_T be some random variable of interest whose value becomes known at time T, and suppose that we can price a claim to $X_T R_{t\to T}$ delivered at time T. Then we can compute the expected value of X_T from the perspective of an investor with log utility whose wealth is invested in the market by pricing an asset:*

$$\mathbb{E}_t X_T = \textit{time-t price of a claim to the time-T payoff } X_T R_{t\to T}. \tag{24}$$

Proof. Such an investor must perceive the market as growth-optimal. The reciprocal of the growth-optimal return is an SDF (Long (1990)), so from the perspective of this log investor, $1/R_{t\to T}$ is an SDF. The right-hand side of (24) therefore equals $\mathbb{E}_t\left[\frac{1}{R_{t\to T}} X_T R_{t\to T}\right]$; the result follows immediately. □

If the payoff $X_T R_{t\to T}$ can be replicated, and hence priced, then we are done. The next result applies Result 1 to calculate a measure of the probability of a market crash.[21]

[21] The link between option prices and tail probabilities has been studied by several authors using

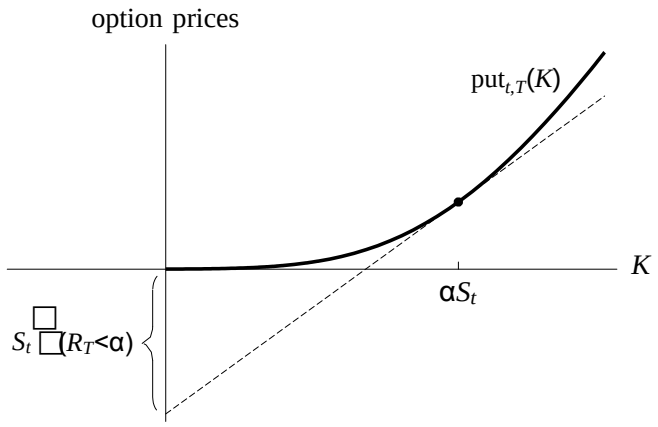


Figure 8: Calculating the probability of a crash, $\mathbb{P}(R_T < \alpha)$, from put prices.

Result 2. *For simplicity, assume there are no dividend payments between times t and T, so that $R_T = S_T/S_t$. Then the log investor's subjective probability that the return on the market is less than α is*

$$\mathbb{P}(R_T < \alpha) = \alpha \left(\text{put}'_{t,T}(\alpha S_t) - \frac{\text{put}_{t,T}(\alpha S_t)}{\alpha S_t} \right). \tag{25}$$

Proof. Since $\mathbb{P}(R_T < \alpha) = \mathbb{E}(\mathbf{1}_{\{R_T<\alpha\}})$, we must (by Result 1) price a claim to the payoff $R_T \mathbf{1}_{\{R_T<\alpha\}}$. The result follows because

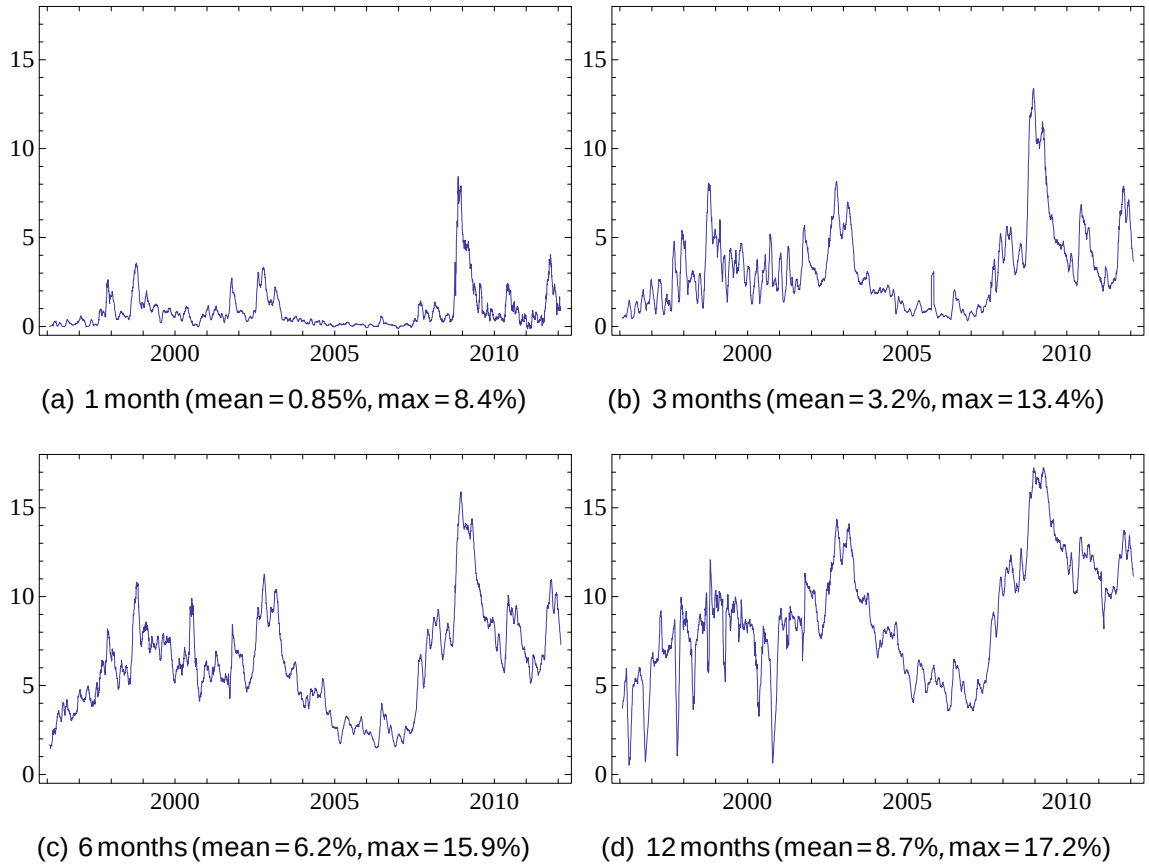
$$R_T \mathbf{1}_{\{R_T<\alpha\}} = \frac{S_T}{S_t} \mathbf{1}_{\{S_T<\alpha S_t\}} = \alpha \Big(\underbrace{\mathbf{1}_{\{S_T<\alpha S_t\}}}_{\text{digital put payoff}} - \frac{1}{\alpha S_t} \underbrace{\max\{0, \alpha S_t - S_T\}}_{\text{put payoff}} \Big),$$

since (as is well-known) the price of a *digital put* with strike αS_t—that is, the price of a claim to \$1 paid if and only if $S_T < \alpha S_t$—is $\text{put}'_{t,T}(\alpha S_t)$. □

The crash probability index (25) has a geometrical interpretation that is illustrated in Figure 8: the tangent to $\text{put}_{t,T}(K)$ at $K = \alpha S_t$ cuts the y-axis at $-S_t \mathbb{P}(R_T < \alpha)$. Thus the crash probability is high when put prices exhibit significant convexity (as a function of strike) at and below $K = \alpha S_t$.

Figure 9 shows the (log investor's subjective) probability of a 20% market crash at various horizons, smoothed by taking a 20-day moving average. Over my sample period, the probability of a crash in the next month averages 0.85% and peaks at

various different approaches; see, for example, Bates (1991), Backus, Chernov and Martin (2011), Bollerslev and Todorov (2011), and Barro and Liao (2016).



(a) 1 month (mean = 0.85%, max = 8.4%)

(b) 3 months (mean = 3.2%, max = 13.4%)

(c) 6 months (mean = 6.2%, max = 15.9%)

(d) 12 months (mean = 8.7%, max = 17.2%)

Figure 9: The probability, in percent, of a 20% market crash at various horizons (20-day moving average).

8.4%, while that of a crash in the next year averages 8.7% and peaks at 17.2%. It is interesting to note, in panel (d), that the run-up in the 1-year crash probability started in 2007, at a time when the S&P 500 index was at or near historic highs, well before the onset of the crash proper.

6 VIX, SVIX, and variance swaps

The SVIX index, defined in equation (12), can usefully be compared to the VIX index:

$$\text{VIX}_t^2 \equiv \frac{2R_{f,t}}{T-t}\left(\int_0^{F_{t,T}} \frac{1}{K^2}\,\text{put}_{t,T}(K)\,dK + \int_{F_{t,T}}^{\infty} \frac{1}{K^2}\,\text{call}_{t,T}(K)\,dK\right) . \qquad (26)$$

We saw in equation (13) that the SVIX index measures the risk-neutral volatility of the return on the market. What does VIX measure? Since option prices are equally

weighted by strike in the definition of SVIX, but weighted by $1/K^2$ in the definition of VIX, it is clear that VIX places relatively more weight on out-of-the-money puts and less weight on out-of-the-money calls; and hence places more weight on left-tail events.

Result 3 (What does VIX measure?)**.** *If the underlying asset does not pay dividends, so that $R_T = S_T/S_t$, then* VIX *measures the risk-neutral entropy of the simple return:*

$$\mathrm{VIX}_t^2 = \frac{2}{T-t} \mathsf{L}_t^*(R_T/R_{f,t}), \tag{27}$$

where entropy is defined by $\mathsf{L}_t^*(X) \equiv \log \mathsf{E}_t^* X - \mathsf{E}_t^* \log X$.

Proof. As an application of the result of Breeden and Litzenberger (1978), the price of a claim to $\log R_T$ is

$$\frac{1}{R_{f,t}} \mathsf{E}_t^* \log R_T = \frac{\log R_{f,t}}{R_{f,t}} - \int_0^{F_{t,T}} \frac{1}{K^2} \mathrm{put}_{t,T}(K)\, dK - \int_{F_{t,T}}^{\infty} \frac{1}{K^2} \mathrm{call}_{t,T}(K)\, dK.$$

The result follows by combining this with the fact that $\mathsf{E}_t^* R_T = R_{f,t}$. □

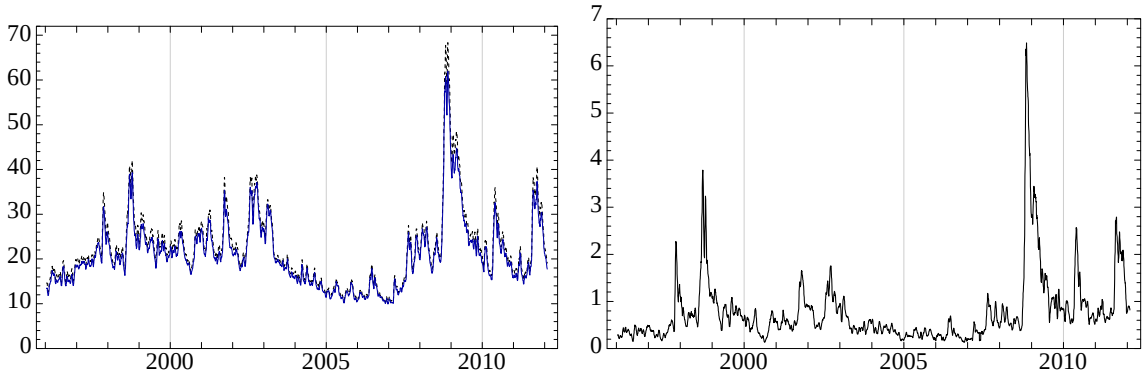
Entropy is a measure of the variability of a positive random variable.[22] Like variance it is nonnegative by Jensen's inequality, and like variance it measures variability by the extent to which a concave function of an expectation of a random variable exceeds an expectation of a concave function of a random variable.

If the VIX index measures entropy, and the SVIX index measures variance, which is a better measure of return variability? The answer is that both are of interest. Entropy is more sensitive to the left tail of the return distribution, while variance is more sensitive to the right tail, as can be seen by comparing the entropy measure (26), which loads more strongly on out-of-the-money puts, with the variance measure (12), which loads equally on options of all strikes.

The next result shows that VIX and SVIX take a particularly simple form in conditionally lognormal models.

Result 4. *If the SDF M_T and return R_T are conditionally jointly lognormal, then* $\mathrm{SVIX}_t^2 = \frac{1}{T-t}(e^{\sigma_t^2(T-t)} - 1)$ *and* $\mathrm{VIX}_t^2 = \sigma_t^2$*, where* $\sigma_t^2 = \frac{1}{T-t} \mathrm{var}_t \log R_T$*. In particular,* $\mathrm{SVIX}_t > \mathrm{VIX}_t$.

[22] Entropy makes appearances elsewhere in the finance literature: see, for example, Alvarez and Jermann (2005), Backus, Chernov and Martin (2011), and Backus, Chernov and Zin (2013). The Hansen–Jagannathan (1991) bound relates to the variance of the stochastic discount factor, while the Alvarez–Jermann (2005) bound relates to its entropy.



(a) VIX and SVIX (b) VIX minus SVIX

Figure 10: Left: Time series of closing prices of VIX (dotted line) and SVIX (solid line). Right: VIX minus SVIX. Both figures show 10-day moving averages.

Proof. The claims in the first sentence are proved in Appendix D. Since $e^x - 1 > x$ for any real number x, it follows that $\text{SVIX}_t > \text{VIX}_t$ under lognormality. □

It would also follow under lognormality that the difference between SVIX and VIX—which the above result shows would be positive—should be negligible for empirically relevant values of σ_t and $T - t$: if for example $\sigma_t = 20\%$ and $T - t = 1/12$ (i.e., at a 1-month horizon) then we would have $\text{VIX}_t = 20\%$ and $\text{SVIX}_t = 20.02\%$. Figure 10 shows (also at a 1-month horizon) that these predictions are dramatically violated in the data. The gap between VIX and SVIX is particularly large at times of market stress, but VIX is higher than SVIX on *every single day* in my sample. This is direct, model-free evidence that the market return and SDF are not conditionally lognormal at the 1-month horizon. It is not that nonlognormality only matters at times of crisis; it is a completely pervasive feature of the data. It is also worth emphasizing that this evidence is much stronger than the familiar observation that histograms of log returns are not Normal, since that leaves open the possibility that log returns are *conditionally* Normal (with, perhaps, time-varying conditional volatility). Figure 10b excludes that possibility.

6.1 VIX and SVIX as diagnostics of equilibrium models

The characterizations of VIX and SVIX in terms of risk-neutral variance and entropy can be read in reverse, as a way to calculate implied VIX and SVIX indices within equilibrium models: it is far easier to calculate risk-neutral entropy and variance than it is to compute option prices and then integrate over strikes.

Can equilibrium models account for the behavior of VIX and SVIX? It might seem that there is room for optimism, given that consumption growth spiked downwards in late 2008 as SVIX spiked upwards (see Figure 17, in the Appendix); but as we will now see, leading consumption-based models are unable to match the properties of the two series.

The top panel of Table 4 reports various summary statistics of VIX, SVIX, and VIX minus SVIX: namely, the mean, median, standard deviation, maximum, minimum, skewness, excess kurtosis, and autocorrelation of each series (computed on a monthly basis; full details are provided in Appendix D). The panels below report the corresponding quantities calculated within six leading consumption-based models: the Campbell–Cochrane (CC, 1999) habit formation model, the long-run risk model in the original stochastic volatility calibration of Bansal–Yaron (BY, 2004) and in the more recent calibration of Bansal, Kiku and Yaron (BKY, 2012), Wachter's (2013) model with a time-varying disaster arrival rate, and two models that explicitly address the properties of option prices, Bollerslev, Tauchen and Zhou (BTZ, 2009), and Drechsler and Yaron (DY, 2011). These numbers are generated by simulating 1,000,000 sample paths of VIX and SVIX within each model, and computing the average value of the mean, median, etc., across the paths. I also generate an empirical *p*-value for each statistic: this represents the proportion of the 1,000,000 paths that generate values that are as, or more, extreme as observed in the data.

The results are easily summarized. None of the models comes close to matching the properties of either VIX or SVIX. The difference between the two is particularly problematic: for all six models, the mean (and median) level of VIX minus SVIX observed in the data lies outside the support of the 1,000,000 trials. In the case of the CC, BY, BKY, and BTZ models, which are approximately conditionally lognormal, this failure is a consequence of Result 4. The DY model is not lognormal, but still does not generate a sufficiently large mean gap between VIX and SVIX. The Wachter model, with its extreme disasters, generates too *large* a mean gap. The models also fail on the

Data	mean	median	s.d.	min	max	skewness	kurtosis	AC(1)
VIX	21.73	20.36	8.54	10.08	67.60	1.91	6.25	0.802
SVIX	20.96	19.77	7.89	9.91	62.31	1.77	5.51	0.803
VIX − SVIX	0.77	0.56	0.75	0.10	6.08	3.60	18.32	0.714
CC	mean	median	s.d.	min	max	skewness	kurtosis	AC(1)
VIX	18.56	18.72	2.33***	**13.46***	22.74***	−0.24***	−0.49**	0.959***
SVIX	18.58	18.75	2.34***	**13.47***	22.79***	−0.24**	−0.49**	0.959***
VIX − SVIX	**−0.03*****	**−0.03*****	0.01***	**−0.05*****	**−0.01*****	−0.09**	−0.28**	0.957***
BY	mean	median	s.d.	min	max	skewness	kurtosis	AC(1)
VIX	17.20***	17.26*	**1.38*****	14.08*	**19.90*****	−0.15***	−0.49***	0.962***
SVIX	17.21**	17.27*	**1.38*****	14.08*	**19.92*****	−0.15***	−0.49***	0.962***
VIX − SVIX	**−0.01*****	**−0.01*****	**0.00*****	**−0.02*****	**−0.01*****	**−0.14*****	**−0.53*****	0.962***
BKY	mean	median	s.d.	min	max	skewness	kurtosis	AC(1)
VIX	16.64	16.76	**2.18*****	11.97	**20.65*****	−0.19***	−0.54***	0.967**
SVIX	16.65	16.77	**2.18*****	11.98	**20.67*****	−0.19***	−0.54***	0.967**
VIX − SVIX	**−0.01*****	**−0.01*****	**0.00*****	**−0.02*****	**−0.01*****	**−0.28*****	−0.47***	0.968***
BTZ	mean	median	s.d.	min	max	skewness	kurtosis	AC(1)
VIX	19.50	19.72	8.63	0.90*	38.02***	−0.08***	−0.50***	0.907*
SVIX	19.53	19.74	8.65	0.90*	38.15**	−0.08***	−0.50***	0.907*
VIX − SVIX	**−0.03*****	**−0.02*****	**0.03*****	**−0.13*****	**0.00*****	**−1.37*****	2.08**	0.911**
DY	mean	median	s.d.	min	max	skewness	kurtosis	AC(1)
VIX	16.71*	15.27**	4.34*	**13.06*****	37.78**	2.48	7.95	0.825
SVIX	16.62*	15.20**	4.25	**13.06*****	37.33*	2.50	8.05	0.824
VIX − SVIX	**0.09*****	**0.06*****	**0.09*****	**0.00*****	**0.45*****	1.81**	3.97**	0.838
W	mean	median	s.d.	min	max	skewness	kurtosis	AC(1)
VIX	38.86	38.79	9.42	19.85	58.25	0.02***	−0.59***	0.966***
SVIX	32.41	32.37	7.69	16.92	48.18	0.02***	−0.59***	0.966***
VIX − SVIX	**6.44*****	**6.42*****	1.73**	2.93	10.06	**0.04*****	**−0.56*****	0.965***

Table 4: One asterisk: p-value < 0.05. Two asterisks: p-value < 0.01. Three asterisks: p-value = 0.000 to three d.p. Figures in bold indicate that the value observed in the data lies outside the range generated in 1,000,000 trials of the given model (so p-value < 10^{-6}). "Kurtosis" refers to excess kurtosis, which equals zero for a Normal random variable.

other statistics of VIX minus SVIX: its volatility (the Wachter model generates too much, the others not enough), its spikiness (all the models generate too little skewness and kurtosis), and its autocorrelation (higher in the models than in the data). As for VIX and SVIX themselves, only the DY model can match their high skewness and kurtosis and relatively low autocorrelation, and it fails on the other dimensions.

6.2 Variance swaps and simple variance swaps

The equation underpinning the VIX index (26) is a definition rather than a statement about asset pricing, but the form of the definition originally emerged from the theory of variance swap pricing. This section explores this connection in further detail, and proposes a definition of a tradable contract, the *simple variance swap*, that is to SVIX as variance swaps are to VIX. As we will see, simple variance swaps are considerably more robust than conventional variance swaps. In particular, they can be hedged even if the underlying asset is subject to jumps. This is an attractive feature, because the variance swap market collapsed during the events of 2008.

A variance swap is an agreement (initiated, say, at time 0) to exchange

$$\left(\log \frac{S_\Delta}{S_0}\right)^2 + \left(\log \frac{S_{2\Delta}}{S_\Delta}\right)^2 + \cdots + \left(\log \frac{S_T}{S_{T-\Delta}}\right)^2 \tag{28}$$

for some fixed "strike" V at time T. Here Δ is some small time-increment; typically, $\Delta = 1$ day. The market convention is to set V so that no money needs to change hands at initiation of the trade:

$$V = \mathrm{E}_0^*\left[\left(\log \frac{S_\Delta}{S_0}\right)^2 + \left(\log \frac{S_{2\Delta}}{S_\Delta}\right)^2 + \cdots + \left(\log \frac{S_T}{S_{T-\Delta}}\right)^2\right]. \tag{29}$$

The following result, which is due to Carr and Madan (1998) and Demeterfi, Derman, Kamal, and Zou (1999), building on an idea of Neuberger (1994), shows how to price a variance swap—that is, how to compute the expectation on the right-hand side of (29)—under some assumptions that are standard in the variance swap literature but that were not required in preceding sections:

A1 the continuously-compounded interest rate is constant, at r;

A2 the underlying asset does not pay dividends; and

A3 the underlying asset's price follows an Itô process $dS_t = rS_t\,dt + \sigma_t S_t\,dZ_t$ under the risk-neutral measure (so that, in particular, there are no jumps).

Result 5. *Under Assumptions A1–A3, the strike on a variance swap is*

$$\mathcal{V} = 2e^{rT}\left(\int_0^{F_{0,T}} \frac{1}{K^2}\mathrm{put}_{0,T}(K)\,dK + \int_{F_{0,T}}^{\infty} \frac{1}{K^2}\mathrm{call}_{0,T}(K)\,dK\right) \tag{30}$$

in the limit as $\Delta \to 0$*; and this quantity has the interpretation*

$$\mathcal{V} = \mathrm{E}^*\left[\int_0^T \sigma_t^2\,dt\right]. \tag{31}$$

The variance swap can be hedged by holding

(i) *a static position in* $(2/K^2)\,dK$ *puts expiring at time* T *with strike* K*, for each* $K \le F_{0,T}$*,*

(ii) *a static position in* $(2/K^2)\,dK$ *calls expiring at time* T *with strike* K*, for each* $K \ge F_{0,T}$*, and*

(iii) *a dynamic position in* $2(F_{0,t}/S_t - 1)/F_{0,T}$ *units of the underlying asset at time* t*,*

financed by borrowing.

Sketch proof of (30) *and* (31)*.* In the limit as $\Delta \to 0$, the right-hand side of (29) converges to

$$\mathcal{V} = \mathrm{E}^*\left[\int_0^T (d\log S_t)^2\right].$$

(Jarrow et al. (2010) provide a rigorous analysis.) Neuberger (1994) observed that, by Itô's lemma and Assumption A5, $d\log S_t = (r - \frac{1}{2}\sigma_t^2)dt + \sigma_t\,dZ_t$ under the risk-neutral measure, so $(d\log S_t)^2 = \sigma_t^2\,dt$, and

$$\begin{aligned}\mathcal{V} &= \mathrm{E}^*\left[\int_0^T \sigma_t^2\,dt\right] \\ &= 2\,\mathrm{E}^*\left[\int_0^T \frac{1}{S_t}\,dS_t - \int_0^T d\log S_t\right] \\ &= 2rT - 2\,\mathrm{E}^*\log\frac{S_T}{S_0}.\end{aligned} \tag{32}$$

This shows that the strike on a variance swap is determined by pricing a notional contract that pays, at time T, the logarithm of the underlying asset's simple return

$R_T = S_T/S_0$. Carr and Madan (1998) and Demeterfi et al. (1999) then showed how to use the approach of Breeden and Litzenberger (1978) to find the price of this contract, $P_{\log}$, in terms of the prices of European call and put options on the underlying asset:

$$P_{\log} \equiv e^{-rT}\,\mathrm{E}^*\log R_T = rTe^{-rT} - \int_0^{F_{0,T}} \frac{1}{K^2}\,\mathrm{put}_{0,T}(K)\,dK - \int_{F_{0,T}}^{\infty} \frac{1}{K^2}\,\mathrm{call}_{0,T}(K)\,dK.$$

Substituting back into (32), we have the result. □

This result is often referred to as "model-free," since it applies if the underlying asset's price follows any sufficiently well-behaved Itô process. But this is a very strong condition. In reality, the market does not follow an Itô process, so VIX^2 does not correspond to the fair strike on a variance swap, V;[23] the replicating portfolio provided in Result 5 does not replicate the variance swap payoff; and neither V nor VIX^2 has the interpretation (31).

Since variance swaps cannot be hedged at times of jumps, market participants have had to impose caps on their payoffs. These caps—which have become, since 2008, the market convention in index variance swaps as well as single-name variance swaps—limit the maximum possible payoff on a variance swap, but further complicate the pricing and interpretation of the contract. A fundamental problem with the definition of a conventional variance swap can be seen very easily: if the underlying asset—an individual stock, say—goes bankrupt, so that S_t hits zero at some point before expiry T, then the payoff (28) is *infinite*.

Simple variance swaps do not suffer from this deficiency. A simple variance swap is an agreement to exchange

$$\left(\frac{S_\Delta - S_0}{F_{0,0}}\right)^2 + \left(\frac{S_{2\Delta} - S_\Delta}{F_{0,\Delta}}\right)^2 + \cdots + \left(\frac{S_T - S_{T-\Delta}}{F_{0,T-\Delta}}\right)^2 \tag{33}$$

for a pre-arranged strike V at time T. (Recall that $F_{0,t}$ is the forward price of the underlying asset to time t, which is known at time 0.) The choice to put forward prices in the denominators is important: below we will see that this choice leads to a huge simplification of the formula for the strike V, and of the associated hedging strategy, in the limit as the period length Δ goes to zero. In an idealized frictionless market,

[23] Aït-Sahalia, Karaman and Mancini (2012) document a large gap between index variance swap strikes and VIX-type indices (squared) at all horizons: on the order of 2% in volatility units, compared to an average volatility level around 20%.

this simplification of the hedging strategy would merely be a matter of analytical convenience; in practice, with trade costs, it acquires far more importance.

The following result shows how to price a simple variance swap (i.e. how to choose V so that no money need change hands initially) in the $\Delta \to 0$ limit. From now on, I write V for the fair strike on a simple variance swap in this limiting case, and write $V(\Delta)$ when the case of $\Delta > 0$ is considered. The result depends on weaker assumptions than were required for the conventional variance swap. Most important, there is no need to assume that the underlying asset follows a diffusion.

B1 the continuously-compounded interest rate is constant, at r; and

B2 the underlying asset pays dividends continuously at rate δS_t per unit time.

Given these assumptions, $F_{0,t} = S_0 e^{(r-\delta)t}$. Dividends should be interpreted broadly: if the underlying asset is a foreign currency then δ corresponds to the foreign interest rate. Appendix E.4 considers other ways of dealing with dividend payouts.

Result 6 (Pricing and hedging a simple variance swap in the $\Delta \to 0$ limit)**.** *Under Assumptions B1 and B2, the strike on a simple variance swap is*

$$V = \frac{2e^{rT}}{F_{0,T}^2}\left(\int_0^{F_{0,T}} \text{put}_{0,T}(K)\,dK + \int_{F_{0,T}}^{\infty} \text{call}_{0,T}(K)\,dK\right), \tag{34}$$

and the payoff on a simple variance swap can be replicated by holding

- ***(i)*** *a static position in $(2/F_{0,}^2)\,dK$ puts expiring at time T with strike K, for each $K \le F_{0,T}$,*
- ***(ii)*** *a static position in $(2/F_{0,T}^2)\,dK$ calls expiring at time T with strike K, for each $K \ge F_{0,T}$, and*
- ***(iii)*** *a dynamic position in $2e^{-\delta(T-t)}(1 - S_t/F_{0,t})/F_{0,T}$ units of the underlying asset at time t,*

financed by borrowing.

Proof. The derivation of (34) divides into two steps. *Step 1*: The absence of arbitrage implies that there are stochastic discount factors $M_\Delta, M_{2\Delta}, \ldots$ such that a payoff $X_{j\Delta}$ at time $j\Delta$ has price $\mathrm{E}_{i\Delta}\left(M_{(i+1)\Delta}M_{(i+2)\Delta}\cdots M_{j\Delta}X_{j\Delta}\right)$ at time $i\Delta$. The subscript on

the expectation operator indicates that it is conditional on time-$i\Delta$ information. I abbreviate $M_{(j\Delta)} \equiv M_\Delta M_{2\Delta} \cdots M_{j\Delta}$.

V is chosen so that the swap has zero initial value, i.e.,

$$\mathrm{E}\left[M_{(T)} \left(\left(\frac{S_\Delta - S_0}{F_{0,0}} \right)^2 + \cdots + \left(\frac{S_T - S_{T-\Delta}}{F_{0,T-\Delta}} \right)^2 - V \right) \right] = 0. \tag{35}$$

We have

$$\begin{aligned} \mathrm{E}[M_{(T)}(S_{i\Delta} - S_{(i-1)\Delta})^2] &= e^{-r(T-i\Delta)}\, \mathrm{E}[M_{(i\Delta)}(S_{i\Delta} - S_{(i-1)\Delta})^2] \\ &= e^{-r(T-i\Delta)} \left\{ \mathrm{E}[M_{(i\Delta)} S_{i\Delta}^2] - (2e^{-\delta\Delta} - e^{-r\Delta})\, \mathrm{E}[M_{((i-1)\Delta)} S_{(i-1)\Delta}^2] \right\}, \end{aligned}$$

using (i) the law of iterated expectations; (ii) the fact that the interest rate r is constant, so that $\mathrm{E}_{(i-1)\Delta} M_{i\Delta} = e^{-r\Delta}$; and (iii) the fact that if dividends are continuously reinvested in the underlying asset, then an investment of $e^{-\delta\Delta} S_{(i-1)\Delta}$ at time $(i-1)\Delta$ is worth $S_{i\Delta}$ at time $i\Delta$, which implies that $\mathrm{E}_{(i-1)\Delta} M_{i\Delta} S_{i\Delta} = e^{-\delta\Delta} S_{(i-1)\Delta}$. If we define $\Pi(i)$ to be the time-0 price of a claim to S_i^2, paid at time i, then

$$\mathrm{E}\left[M_{(T)} \left(S_{i\Delta} - S_{(i-1)\Delta} \right)^2 \right] = e^{-r(T-i\Delta)} \left[\Pi(i\Delta) - \left(2 - e^{-(r-\delta)\Delta} \right) e^{-\delta\Delta} \Pi((i-1)\Delta) \right].$$

Substituting this into (35), we find that

$$V(\Delta) = \sum_{i=1}^{T/\Delta} \frac{e^{ri\Delta}}{F_{0,(i-1)\Delta}^2} \left[\Pi(i\Delta) - (2 - e^{-(r-\delta)\Delta}) e^{-\delta\Delta} \Pi((i-1)\Delta) \right]. \tag{36}$$

As we have already seen,

$$\Pi(t) = 2 \int_0^\infty \mathrm{call}_{0,t}(K)\, dK \tag{37}$$

or, using put-call parity to express $\Pi(t)$ in terms of out-of-the-money options,

$$\Pi(t) = 2 \int_0^{F_{0,t}} \mathrm{put}_{0,t}(K)\, dK + 2 \int_{F_{0,t}}^\infty \mathrm{call}_{0,t}(K)\, dK + e^{-rt} F_{0,t}^2. \tag{38}$$

Step 2. Observe that (36) can be rewritten

$$V(\Delta) = \sum_{i=1}^{T/\Delta} \left(\frac{e^{ri\Delta}}{F_{0,(i-1)\Delta}^2} P(i\Delta) - (2 - e^{-(r-\delta)\Delta}) e^{-\delta\Delta} P((i-1)\Delta) \right) + \frac{T}{\Delta} \left(e^{(r-\delta)\Delta} - 1 \right)^2,$$

where

$$P(t) \equiv 2 \left(\int_0^{F_{0,t}} \mathrm{put}_{0,t}(K)\, dK + \int_{F_{0,t}}^\infty \mathrm{call}_{0,t}(K)\, dK \right).$$

For $0 < j < T/\Delta$, the coefficient on $P(j\Delta)$ in this equation is

$$\frac{e^{rj\Delta}}{F^2_{0,(j-1)\Delta}} - \frac{e^{r(j+1)\Delta}}{F^2_{0,j\Delta}}(2 - e^{-(r-\delta)\Delta})e^{-\delta\Delta} = \frac{e^{rj\Delta}}{F^2_{0,j\Delta}}\left(e^{(r-\delta)\Delta} - 1\right)^2.$$

(The definition (33) was originally found by viewing the normalizing constants $F_{0,j\Delta}$, for $j = 0, \ldots T/\Delta$, as arbitrary, and choosing them so that the above equation would hold.) We can therefore rewrite

$$V(\Delta) = \frac{e^{rT}}{F^2_{0,T-\Delta}}P(T) + \underbrace{\sum_{j=1}^{T/\Delta-1} \frac{e^{rj\Delta}}{F^2_{0,j\Delta}}\left(e^{(r-\delta)\Delta} - 1\right)^2 P(j\Delta)}_{O(1/\Delta) \text{ terms of size } O(\Delta^2)} + \frac{T}{\Delta}\left(e^{(r-\delta)\Delta} - 1\right)^2. \quad (39)$$

The second term on the right-hand side is a sum of $T/\Delta - 1$ terms, each of size on the order of Δ^2; all in all, the sum is $O(\Delta)$. The third term is also $O(\Delta)$, so both tend to zero as $\Delta \to 0$. The first term tends to $e^{rT}P(T)/F^2_{0,}$, as required.

The above argument implicitly supplies the dynamic trading strategy that replicates the payoff on a simple variance swap. Appendix E.1 describes the strategy in detail. □

The derivation of the pricing result (34) has two main components. The first is the exact expression (36), which applies for fixed $\Delta > 0$. It shows that the strike on a simple variance swap is dictated by the prices of options across all strikes and the whole range of expiry times Δ, 2Δ, . . . , T. But, correspondingly, the hedge portfolio requires holding portfolios of options of each of these maturities. Although this is not a serious issue if Δ is large relative to T, it raises the concern that hedging a simple variance swap may be extremely costly in practice if Δ is very small relative to T. Fortunately, the second component shows that this concern is misplaced: by choosing forward prices as the normalizing weights in the definition (33), both the pricing formula (36) and the hedging portfolio simplify nicely in the limit as $\Delta \to 0$. In principle, we could have put any other constants known at time 0 in the denominators of the fractions in (33). Had we done so, we would have to face the unappealing prospect of a hedging portfolio requiring positions in options of all maturities between 0 and T. Using forward prices lets us sidestep this problem, meaning that the hedge calls only for a single static portfolio of options expiring at time T, and equally weighted by strike.

The dynamic position in the underlying can be thought of as a delta-hedge: if, say, the underlying's price at time t happens to exceed $F_{0,t} = S_0e^{(r-\delta)t}$, then the replicating

portfolio is short the underlying in order to offset the effects of increasing delta as calls go in-the-money and puts go increasingly out-of-the-money.

Robustness. What happens if sampling and trading occurs at discrete intervals $\Delta > 0$, rather than continuously? What if deep-out-of-the-money options cannot be traded? What are the effects of different dividend payout policies? I show in Appendix E that simple variance swaps have good robustness properties in each case.

7 Conclusion

The starting point of this paper is the identity (4), which shows that the expected excess return on any asset equals the risk-neutral variance of the asset's return minus a covariance term. If options are traded on the asset, then risk-neutral variance can be unambiguously measured without requiring any assumptions other than the absence of arbitrage. I apply the identity to the return on the market. In this case, risk-neutral variance is equal to the square of a volatility index, SVIX, that is similar to the VIX index, and I argue that the covariance term is weakly negative. The square of the SVIX index is therefore a lower bound on the equity premium.

I construct the SVIX index using S&P 500 index option data from 1996 to 2012. The index is strikingly volatile; it implies that in late 2008, the equity premium rose above 21% at the 1-year horizon and above 55% (annualized) at the 1-month horizon. More aggressively, I argue that the lower bound is approximately *tight*—that is, risk-neutral variance is not merely a lower bound on the equity premium, it is approximately *equal* to the equity premium.

The implications of this fact represent a challenge to finance theory: I have shown that none of the leading equilibrium models of financial markets can generate the sudden shifts in VIX and SVIX that are observed in the data. More broadly, the results point to a novel view of the equity premium, with important implications for finance and macroeconomics.

First, they suggest that the equity premium is far more volatile than implied by the valuation-ratio predictors of Campbell and Thompson (2008). The distinction between the two views is sharpest on days such as Black Monday, in 1987, when the S&P 500 and Dow Jones indices experienced very severe declines, with daily returns roughly twice as negative as the next-worst day in history. On the Campbell–Thompson view

of the world, the equity premium rose on the order of two or three percentage points during this episode. In sharp contrast, option prices are known to have exploded on Black Monday, which I argue implies also that the equity premium exploded.

Second, this volatility often reflects movements in the equity premium at weekly, daily, or even higher frequency. The macro-finance literature, which seeks to rationalize market gyrations at the business cycle frequency, typically has not acknowledged or attempted to address such movements.

Third, the equity premium is strongly right-skewed: the median equity premium is on the order of 3 or 4%, but there are occasional opportunities for unconstrained investors to earn a much higher equity premium.

Fourth, the term structure of the equity premium reveals that during such episodes, a disproportionate fraction of the equity premium is concentrated in the form of extremely high expected returns over the very short run.

8 References

Aït-Sahalia, Y., M. Karaman, and L. Mancini (2012), "The Term Structure of Variance Swaps, Risk Premia and the Expectation Hypothesis," working paper.

Alvarez, F., and U. J. Jermann (2005), "Using Asset Prices to Measure the Persistence of the Marginal Utility of Wealth," *Econometrica*, 73:6:1977–2016.

Ang, A., and G. Bekaert (2007), "Stock Return Predictability: Is It There?" *Review of Financial Studies*, 20:3:651–707.

Backus, D. K., Chernov, M. and I. W. R. Martin (2011), "Disasters Implied by Equity Index Options," *Journal of Finance*, 66:6:1969–2012.

Bansal, R., D. Kiku, I. Shaliastovich, and A. Yaron (2012), "Volatility, the Macroeconomy, and Asset Prices," working paper.

Bansal, R., D. Kiku, and A. Yaron (2012), "An Empirical Evaluation of the Long-Run Risks Model for Asset Prices," *Critical Finance Review*, 1:183–221.

Bansal, R. and A. Yaron (2004), "Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles," *Journal of Finance*, 59:4:1481–1509.

Barberis, N., R. Greenwood, L. Jin, and A. Shleifer (2015), "X-CAPM: An Extrapolative Capital Asset Pricing Model," *Journal of Financial Economics*, 115:1–24.

Barro, R. J. (2006), "Rare Disasters and Asset Markets in the Twentieth Century," *Quarterly Journal of Economics*, 121:3:823–866.

Barro, R. J., and G. Liao (2016), "Options-Pricing Formula with Disaster Risk," NBER Working Paper 21888.

Bates, D. S. (1991), "The Crash of '87: Was It Expected? The Evidence from Options Markets," *Journal of Finance*, 46:3:1009–1044.

Ben-David, I., J. R. Graham, and C. R. Harvey (2013), "Managerial Miscalibration," *Quarterly Journal of Economics*, 128:1547–1584.

Black, F., and M. Scholes (1973), "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, 81:637–659.

Bollerslev, T., G. Tauchen, and H. Zhou (2009), "Expected Stock Returns and Variance Risk Premia," *Review of Financial Studies*, 22:11:4463–4492.

Bollerslev, T., and V. Todorov (2011), "Tails, Fears, and Risk Premia," *Journal of Finance*, 66:6:2165–2221.

Breeden, D. T., and R. H. Litzenberger (1978), "Prices of State-Contingent Claims Implicit in Option Prices," *Journal of Business*, 51:4:621–651.

Brunnermeier, M. and S. Nagel (2004), "Hedge Funds and the Technology Bubble," *Journal of Finance*, 59(5), 2013-2040.

Campbell, J. Y. and J. H. Cochrane (1999), "By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior," *Journal of Political Economy*, 107:2:205–251.

Campbell, J. Y., S. Giglio, C. Polk, and R. Turley, "An Intertemporal CAPM with Stochastic Volatility," working paper.

Campbell, J. Y., and R. J. Shiller (1988), "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors," *Review of Financial Studies*, 1:3:195–228.

Campbell, J. Y., and S. B. Thompson (2008), "Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average?" *Review of Financial Studies*, 21:4:1509–1531.

Cochrane, J. H. (2005), *Asset Pricing*, Princeton University Press, Princeton, NJ.

Cochrane, J. H. (2008), "The Dog That Did Not Bark: A Defense of Return Predictability," *Review of Financial Studies*, 21:4:1533–1575.

Cochrane, J. H. (2011), "Discount Rates," *Journal of Finance*, 66:4:1047–1108.

Drechsler, I., and A. Yaron (2011), "What's Vol Got to Do with It," *Review of Financial Studies*, 24:1:1–45.

Epstein, L., and S. Zin (1989), "Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: A Theoretical Framework," *Econometrica*, 57:937–969.

Fama, E. F., and K. R. French (1988), "Dividend Yields and Expected Stock Returns," *Journal of Financial Economics*, 22:3–25.

Fama, E. F., and K. R. French (2002), "The Equity Premium," *Journal of Finance*, 57:2:637–659.

Goyal, A., and I. Welch (2008), "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction," *Review of Financial Studies*, 21:4:1455–1508.

Greenwood, R., and A. Shleifer (2014), "Expectations of Returns and Expected Returns," *Review of Financial Studies*, 27:3:714–746.

Hansen, L. P. and R. Jagannathan (1991), "Implications of Security Market Data for Models of Dynamic Economies," *Journal of Political Economy*, 99:2:225–262.

Keim, D. B., and R. F. Stambaugh (1986), "Predicting returns in the stock and bond markets," *Journal of Financial Economics* 17, 357–390.

Kelly, B., and S. Pruitt (2013), "Market Expectations in the Cross-Section of Present Values," *Journal of Finance*, 68:5:1721–1756.

Lettau, M., and S. Ludvigson (2001), "Consumption, Aggregate Wealth, and Expected Stock Returns," *Journal of Finance*, 56:3:815–849.

Malmendier, U., and S. Nagel (2011), "Depression Babies: Do Macroeconomic Experiences Affect Risk Taking?" *Quarterly Journal of Economics*, 126:1:373–416.

Merton, R. C. (1980), "On Estimating the Expected Return on the Market," *Journal of Financial Economics*, 8:323–361.

Roll, R. (1977), "A Critique of the Asset Pricing Theory's Tests I: On Past and Potential Testability of the Theory," *Journal of Financial Economics*, 4:129–176.

Ross, S. A. (2005), *Neoclassical Finance*, Princeton University Press.

Shiller, R. J. (1987), "Investor Behavior in the October 1987 Stock Market Crash: Survey Evidence," NBER Working Paper 2446.

Shiller, R. J. (2000), *Irrational Exuberance*, Princeton University Press.

Stambaugh, R. F. (1999), "Predictive Regressions," *Journal of Financial Economics*, 54:375–421.

Wachter, J. (2013), "Can time-varying risk of rare disasters explain aggregate stock market volatility?" *Journal of Finance*, 68:987–1035

A The negative correlation condition

This section contains proofs that the examples in Section 2 satisfy the NCC.

Example 1. Write $M_T = e^{-r_{f,t}+\sigma_{M,t}Z_{M,T}-\sigma_{M,t}^2/2}$ and $R_T = e^{\mu_{R,t}+\sigma_{R,t}Z_{R,T}-\sigma_{R,t}^2/2}$, where $Z_{M,T}$ and $Z_{R,T}$ are (potentially correlated) standard Normal random variables. The requirement that $\mathrm{E}_t M_T R_T = 1$ implies that $\mu_{R,t} - r_{f,t} + \mathrm{cov}_t(\log M_T, \log R_T) = 0$. This fact, together with some straightforward algebra, implies that $\mathrm{E}_t M_T R_T^2 \le \mathrm{E}_t R_T$ if and only if $\lambda_t \ge \sigma_{R,t}$, where λ_t is the conditional Sharpe ratio $(\mu_{R,t} - r_{f,t})/\sigma_{R,t}$.

Example 2. By assumption, there is an investor with wealth W_t and utility function $u(\cdot)$ who chooses, at time t, from the available menu of assets with returns $R_T^{(i)}$, $i = 1, 2, \ldots$. In other words, he chooses portfolio weights $\{w_i\}$ to solve the problem

$$\max_{\{w_i\}} \mathrm{E}_t\, u\left(W_t \sum_i w_i R_T^{(i)}\right) \quad \text{subject to} \quad \sum_i w_i = 1. \tag{40}$$

The first-order condition for (say) w_j is that

$$\mathrm{E}_t\left[W_t u'\left(W_t \sum_i w_i R_T^{(i)}\right) R_T^{(j)}\right] = \lambda_t,$$

where $\lambda_t > 0$ is the Lagrange multiplier associated with the constraint in (40). Since the investor chooses to hold the market, we have $\sum_i w_i R_T^{(i)} = R_T$. Thus,

$$\mathrm{E}_t\left[\frac{W_t}{\lambda_t} u'(W_t R_T) R_T^{(j)}\right] = 1$$

for any return $R_T^{(j)}$. It follows that the SDF is proportional (with a constant of proportionality that is known at time t) to $u'(W_t R_T)$.

To show that the NCC holds, we must show that $\mathrm{cov}_t(u'(W_t R_T)R_T, R_T) \leq 0$. This holds because $u'(W_t R_T)R_T$ is decreasing in R_T: its derivative is $u'(W_t R_T) + W_t R_T u''(W_t R_T) = -u'(W_t R_T)[\gamma(W_t R_T) - 1]$, which is negative because risk aversion $\gamma(x) \equiv -xu''(x)/u'(x)$ is at least one.

If the investor has log utility, then $\gamma(x) \equiv 1$, so the inequality holds with equality. But it is not *necessary* for the investor to have log utility for the inequality to hold with equality: all we require is that $M_T R_T$ is uncorrelated with R_T. That is, we merely need that $M_T = I_T/R_T$ where I_T and R_T are uncorrelated (and $\mathrm{E}_t I_T = 1$ since $\mathrm{E}_t M_T R_T$ must equal one). Log utility is the special case in which $I_T \equiv 1$.

Examples 3a and 3b. For reasons given in the text, Example 3a is a special case of Example 3b, which we now prove. We must check that $\mathrm{cov}_t(M_T R_T, R_T) \leq 0$, or equivalently that

$$\mathrm{cov}_t(-R_T V_W(W_T, z_{1,T}, \ldots, z_{N,T}), R_T) \geq 0. \tag{41}$$

That is, we must prove that the covariance of two functions of $R_T, R_T^{(i)}, z_{1,T}, \ldots, z_{N,T}$ is positive. The two functions are

$$f(R_T, R_T^{(i)}, z_{1,T}, \ldots, z_{N,T}) = -R_T V_W(\alpha_t(W_t - C_t)R_T + (1-\alpha_t)(W_t - C_t)R_T^{(i)}, z_{1,T}, \ldots, z_{N,T}) \tag{42}$$

and

$$g(R_T, R_T^{(i)}, z_{1,T}, \ldots, z_{N,T}) = R_T.$$

(Since the covariance is conditional on time-t information, α_t and $(W_t - C_t)$ can be treated as known constants.) By the defining property of associated random variables,

(41) holds so long as f and g are each weakly increasing functions of their arguments. This is obviously true for g, so it only remains to check that the first derivatives of $\boldsymbol{f}$ are all nonnegative.

Differentiating (42) with respect to R_T, we need $-V_W(W_T, z_{1,T}, \ldots, z_{N,T}) - \alpha_t(W_t - C_t)R_T V_{WW}(W_T, z_{1,T}, \ldots, z_{N,T}) \geq 0$, or equivalently

$$-\frac{W_T V_{WW}(W_T, z_{1,T}, \ldots, z_{N,T})}{V_W(W_T, z_{1,T}, \ldots, z_{N,T})} \geq \frac{W_T}{W_{M,T}},$$

where W_T and $W_{M,T}$ are as given in the main text. This is the constraint on risk aversion.

Differentiating (42) with respect to $R_T^{(j)}$, we need $-R_T(1-\alpha_t)(W_t - C_t)V_{WW}(W_T, z_{1,T}, \ldots, z_{N,T}) \geq 0$, which follows because $V_{WW} < 0$.

Differentiating (42) with respect to $z_{j,T}$, we need $-R_T V_{Wj}(W_T, z_{1,T}, \ldots, z_{N,T}) \geq 0$, which follows because V_{Wj} (the cross derivative of the value function with respect to wealth and the jth state variable) is weakly negative due to the choice of sign on the state variables.

Examples 4a and 4b. With Epstein–Zin preferences, the SDF is proportional (up to quantities known at time t) to $(W_T/C_T)^{(\gamma-1)/(1-\psi)}R_T^{-\gamma}$, so the desired inequality, $\mathrm{cov}_t(M_T R_T, R_T) \leq 0$, is equivalent to

$$\mathrm{cov}_t\left[-\left(\frac{W_T}{C_T}\right)^{(\gamma-1)/(1-\psi)} R_T^{1-\gamma}, R_T\right] \geq 0.$$

If $\gamma = 1$, as in Example 4b, then this holds with equality.

If W_T/C_T and R_T are associated, as assumed in Example 4a, then we need to check that the first derivatives of $\boldsymbol{f}(x, y) = -x^{(\gamma-1)/(1-\psi)}y^{1-\gamma}$ are nonnegative. That is, we need $\gamma \geq 1$ and $\psi \geq 1$, as claimed.

B Calculating risk-neutral variance

Note that for any $x \geq 0$, we have $x^2 = 2\int_0^\infty \max\{0, x - K\}\, dK$. Setting $x = S_T$, taking risk-neutral expectations, and multiplying by $\frac{1}{R_{f,t}}$,

$$\begin{aligned}\frac{1}{R_{f,t}} \mathrm{E}_t^* S_T^2 &= 2\int_0^\infty \frac{1}{R_{f,t}} \mathrm{E}_t^* \max\{0, S_T - K\}\, dK \\ &= 2\int_0^\infty \mathrm{call}_{t,T}(K)\, dK. \end{aligned} \tag{43}$$

It follows from (7), (8), and (43) that risk-neutral variance can be calculated from option prices:

$$\frac{1}{R_{f,t}} \operatorname{var}_t^* R_T = \frac{1}{S_t^2} \left[2 \int_0^\infty \operatorname{call}_{t,T}(K)\, dK - \frac{F_{t,T}^2}{R_{f,t}} \right]. \tag{44}$$

This expression incorporates the prices of in-the-money calls, which are usually illiquid. But by put-call parity, $\operatorname{call}_{t,T}(K) = \operatorname{put}_{t,T}(K) + \frac{1}{R_{f,t}}(F_{t,T} - K)$, so

$$\begin{aligned}
\int_0^\infty \operatorname{call}_{t,T}(K)\, dK &= \int_0^{F_{t,T}} \operatorname{call}_{t,T}(K)\, dK + \int_{F_{t,T}}^\infty \operatorname{call}_{t,T}(K)\, dK \\
&= \int_0^{F_{t,T}} \operatorname{put}_{t,T}(K) + \frac{1}{R_{f,t}}(F_{t,T} - K)\, dK + \int_{F_{t,T}}^\infty \operatorname{call}_{t,T}(K)\, dK \\
&= \int_0^{F_{t,T}} \operatorname{put}_{t,T}(K)\, dK + \frac{F_{t,T}^2}{2R_{f,t}} + \int_{F_{t,T}}^\infty \operatorname{call}_{t,T}(K)\, dK.
\end{aligned}$$

Substituting this into (44), we have the formula (11) for risk-neutral variance:

$$\frac{1}{R_{f,t}} \operatorname{var}_t^* R_T = \frac{2}{S_t^2} \left[\int_0^{F_{t,T}} \operatorname{put}_{t,T}(K)\, dK + \int_{F_{t,T}}^\infty \operatorname{call}_{t,T}(K)\, dK \right].$$

B.1 Construction of the lower bound

The data are from *OptionMetrics*, runnning from January 4, 1996, to January 31, 2012; they include the closing price of the S&P 500 index, and the expiration date, strike price, highest closing bid and lowest closing ask of all call and put options with fewer than 550 days to expiry. I clean the data in several ways. First, I delete all replicated entries. Second, for each strike, I select the option—call or put—whose mid price is lower. Third, I delete all options with a highest closing bid of zero. Finally, I delete all Quarterly options, which tend to be less liquid than regular S&P 500 index options and to have a smaller range of strikes. Having done so, I am left with 1,165,585 option-day datapoints. I compute mid-market option prices by averaging the highest closing bid and lowest closing ask, and using the resulting prices to compute the lower bound by discretizing the right-hand side of inequality (14).

On any given day, I compute the lower bound at a range of time horizons depending on the particular expiration dates of options traded on that day, with the constraint that the shortest time to expiry is never allowed to be less than 7 days; this is the same procedure that the CBOE follows. I then calculate the bound for $T = 30$, 60, 90, 180, and 360 days by linear interpolation. Occasionally, extrapolation

is necessary, for example when the nearest-term option's time-to-maturity first dips below 7 days, requiring me to use the two expiry dates further out; again, this is the procedure followed by the CBOE.

B.2 The effect of discrete strikes

The integrals that appear throughout the paper are idealizations: in practice we only observe options at some finite set of strikes. Write $\Omega_{t,T}(K)$ for the price of an out-of-the-money option with strike K, that is,

$$\Omega_{t,T}(K) \equiv \begin{cases} \text{put}_{t,T}(K) & \text{if } K < F_{t,T} \\ \text{call}_{t,T}(K) & \text{if } K \geq F_{t,T} \end{cases};$$

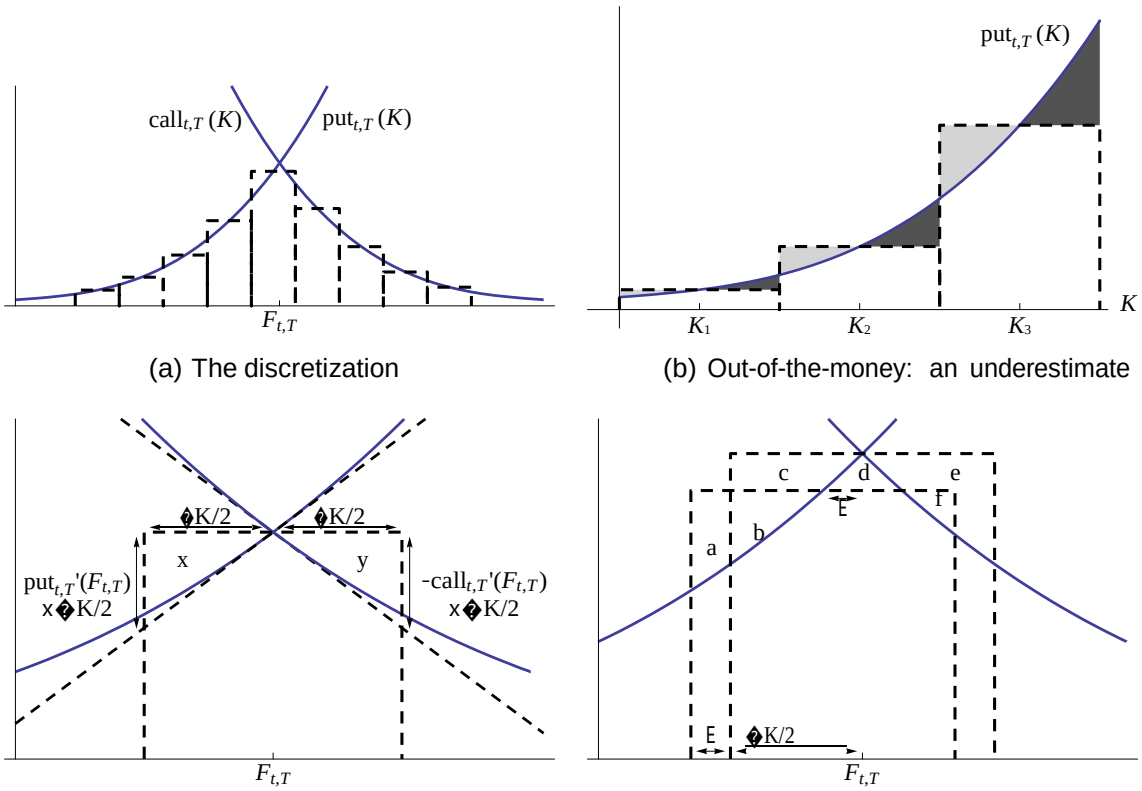
write $K_1, \ldots, K_N$ for the strikes of observable options; write K_j for the strike that is nearest to the forward price $F_{t,T}$;[24] and define $\Delta K_i \equiv (K_{i+1} - K_{i-1})/2$. Then the idealized integral $\int_0^\infty \Omega_{t,T}(K)\,dK$ is replaced, in practice, by the observable sum $\sum_{i=1}^{N} \Omega_{t,T}(K_i)\,\Delta K_i$. (This is the CBOE's procedure in calculating VIX, and I follow it in this paper.) Figure 11a illustrates.

The question is, how well does the sum approximate the integral? The next result shows that there are two forces pushing in the direction of underestimation (of the integral by the sum) and one pushing in the direction of overestimation. But the latter effect is very minor in practice, so one should think of discretization as leading to underestimation of the integral.

Result 7 (The effect of discretization by strike)**.** *Discretizing by strike will tend to lead to an underestimate of the idealized lower bound, in that*

$$\underbrace{\frac{2}{(T-t)R_{f,t}S^2}\sum_{i=1}^{N}\Omega_{t,T}(K_i)\,\Delta K_i}_{\text{discretization}} \leq \underbrace{\frac{2}{(T-t)R_{f,t}S^2}\int_0^\infty \Omega_{t,T}(K)\,dK}_{\text{idealized lower bound}} - \underbrace{\frac{(\Delta K)^2}{4(T-t)\cdot R_{f,t}^2\cdot S^2}\cdot\ldots}_{\text{very small}}$$

[24] For simplicity, I assume that strikes are evenly spaced near-the-money, $K_{j+1} - K_j = K_j - K_{j-1}$. This is not essential, but it is almost always the case in practice and lets me economize slightly on notation.



(a) The discretization

(b) Out-of-the-money: an underestimate

(c) At-the-money: an overestimate

(d) The worst-case scenario in Lemma 1

Figure 11: The effect of discretization. Different panels use different scales.

Proof. Non-observability of deep-out-of-the-money options obviously leads to an underestimate of the lower bound.

Consider, first, the out-of-the-money puts with strikes $K_1, \ldots, K_{j-1}$. The situation is illustrated in Figure 11b: by convexity of $\text{put}_{t,T}(K)$, the light grey areas that are included (when they should be excluded) are smaller than the dark grey areas that are excluded (when they should be included). The same logic applies to the out-of-the-money calls with strikes $K_{j+1}, K_{j+2}, \ldots$. Thus the observable options—excluding the nearest-the-money option—will always underestimate the part of the integral which they are intended to approximate.

It remains to consider the nearest-the-money option with strike K_j, which alone can lead to an overestimate. Lemma 1, below, shows that the worst case is if the strike of the nearest-the-money option happens to be exactly *equal* to the forward price $F_{t,T}$, as in Figure 11c. For an upper bound on the overestimate in this case we must find

an upper bound on the sum of the approximately triangular areas (x) and (y) that are shown in the figure. We can do so by replacing the curved lines in the figure by the (dashed) tangents to $\text{put}_{t,T}(K)$ and $\text{call}_{t,T}(K)$ at $K = F_{t,T}$. The areas of the resulting triangles provide the desired upper bound, by convexity of $\text{put}_{t,T}(K)$ and $\text{call}_{t,T}(K)$:

$$\text{area (x)} + \text{area (y)} \leq \frac{1}{2}\left(\frac{\Delta K}{2}\right)^2 \text{put}'_{t,T}(K) - \frac{1}{2}\left(\frac{\Delta K}{2}\right)^2 \text{call}'_{t,T}(K).$$

But, by put-call parity, $\text{put}'_{t,T}(K) - \text{call}'_{t,T}(K) = 1/R_{f,t}$. Thus, the overestimate due to the at-the-money option is at most

$$\frac{1}{2}\left(\frac{\Delta K}{2}\right)^2 \frac{1}{R_{f,t}}.$$

Since the contributions from out-of-the-money and missing options led to underestimates, the overall overestimate is at most this amount. Finally, since the definition scales the integral by $2/((T-t)R_{f,t}S_t^2)$, the result follows. □

The maximal overestimate provided by this result is *extremely small*: for the S&P 500 index, the interval between strikes near-the-money is $\Delta K_j = 5$. If, say, the forward price of the S&P 500 index is $F_{t,T} = 1000$ and we are considering a monthly horizon, $T - t = 1/12$, then the discretization leads to an overestimate of SVIX^2 that is *at most* $7.5 \times 10^{-5} < 0.0001$. By comparison, the average level of SVIX^2 is on the order of 0.05, as shown in Table 1. Since the non-observability of deep-out-of-the-money options causes underestimation, there is therefore a very strong presumption that the sum underestimates the integral.

It only remains to establish the following lemma, which is used in the proof of Result 7. The goal is to consider the largest possible overestimate that the option whose strike is nearest to the forward price, $F_{t,T}$, can contribute. Figure 11d illustrates. The dotted rectangle in the figure is the contribution if the strike happens to be *equal* to $F_{t,T}$; I will call this *Case 1*. The dashed rectangle is the contribution if the strike equals $F_{t,T} - \varepsilon$, for some $\varepsilon > 0$ (for concreteness—the case $\varepsilon < 0$ is essentially identical); I will call this *Case 2*.

Lemma 1. *The option with strike closest to the forward overestimates most in the case in which its strike is equal to the forward.*

Proof. The overestimate in Case 1 is greater than that in Case 2 if

$$\text{area (b)} + \text{area (c)} + \text{area (e)} + \text{area (f)} \geq \text{area (a)} + \text{area (b)} + \text{area (f)} - \text{area (d)}$$

in Figure 11d, or equivalently,

$$\text{area (c)} + \text{area (d)} + \text{area (e)} \geq \text{area (a)}. \tag{45}$$

But, by convexity of $\text{put}_{t,T}(K)$, area (b) + area (c) ≥ area (a) + area (b), from which (45) follows. An almost identical argument applies if $\varepsilon < 0$. □

C Supplementary tables and figures

Table 5 reproduces the results in Table 2, but excludes the period August 1, 2008–July 31, 2009.

horizon	α	s.e.	β	s.e.	R^2
1 mo	−0.095	[0.061]	3.705	[1.258]	3.36%
2 mo	−0.081	[0.062]	3.279	[1.181]	4.83%
3 mo	−0.076	[0.067]	3.147	[1.258]	5.98%
6 mo	−0.043	[0.072]	2.319	[1.276]	4.94%
1 yr	0.045	[0.088]	0.473	[1.731]	0.27%

Table 5: Coefficient estimates for the regression (16), excluding the crisis period August 1, 2008–July 31, 2009 from the sample.

horizon	α	s.e.	β_1	s.e.	β_2	s.e.	R^2
1 mo	−0.086	[0.063]	2.048	[1.273]	3.908	[1.053]	4.96%
2 mo	−0.113	[0.061]	2.634	[1.007]	3.884	[0.761]	8.54%
3 mo	−0.086	[0.071]	2.273	[1.407]	2.749	[0.346]	6.79%
6 mo	−0.051	[0.076]	1.992	[1.132]	−0.525	[1.259]	6.56%
1 yr	−0.073	[0.078]	2.278	[0.909]	−0.694	[0.680]	10.34%

Table 6: Coefficient estimates for the regression (46).

Table 6 reports results for regressions

$$R_T - R_{f,t} = \alpha + \beta_1 \times R_{f,t} \cdot \text{SVIX}^2_t + \beta_2 \times VRP_t + \varepsilon_T \tag{46}$$

horizon	α	s.e.	β_1	s.e.	β_2	s.e.	R^2
1 mo	−0.103	[0.061]	3.333	[1.292]	1.548	[1.125]	3.61%
2 mo	−0.097	[0.063]	3.137	[1.353]	1.532	[1.801]	6.04%
3 mo	−0.083	[0.068]	2.902	[1.451]	1.133	[1.855]	6.34%
6 mo	0.016	[0.071]	0.797	[1.560]	0.360	[2.095]	0.74%
1 yr	0.008	[0.061]	0.331	[2.274]	1.761	[3.760]	3.10%

Table 7: Coefficient estimates for the regression (46), excluding the crisis period August 1, 2008 to July 31, 2009.

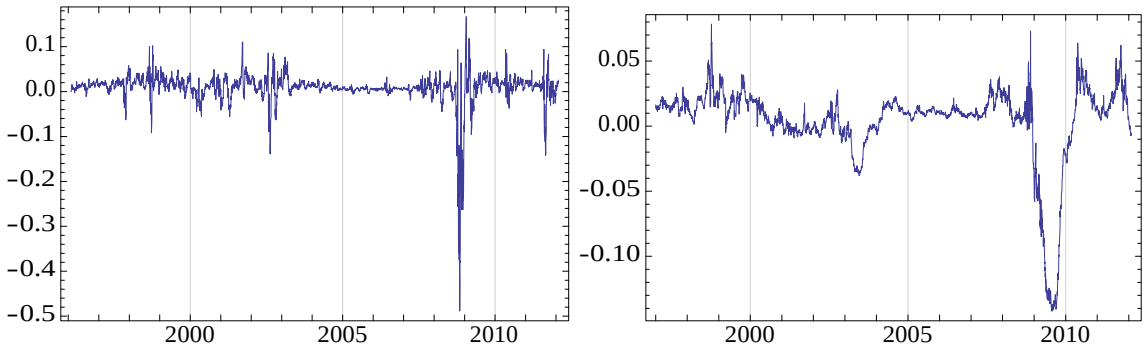
of realized returns onto risk-neutral variance and a measure of the variance risk premium, $VRP_t \equiv R_{f,t} \cdot \text{SVIX}_t^2 - \text{SVAR}_t$. Realized daily return variance, SVAR_t, is computed at time t by looking backwards over the same horizon-length, $T - t$, as the corresponding forward-looking realized return (so, for example, I use 1-month backward-looking realized variances to predict 1-month forward-looking realized returns). If realized variance is a good proxy for forward-looking real-world variance, this is a measure of the 'variance risk premium.'

Consistent with the empirical findings of Bollerslev, Tauchen and Zhou (2009) and Drechsler and Yaron (2011), the coefficient on VRP_t is positive and strongly significant at predictive horizons out to 3 months.[25] This predictive success reflects the fact that implied and realized volatility, SVIX_t and SVAR_t, rose sharply as the S&P 500 dropped in late 2008; implied volatility then fell relatively quickly, while SVAR_t declined more sluggishly. VRP_t therefore turned dramatically negative in late 2008, as shown in Figure 12 below. Since the market then continued to fall, this sluggish response of VRP_t helps fit the data. At the 6-month and 1-year horizons, however, VRP_t responds too sluggishly—it remains strongly negative even as the market starts to rally in March, 2009—so there is a sign-flip, with *negative* estimates of the coefficient on VRP_t at the 6-month and 1-year horizons. The empirical facts are therefore hard to interpret: the sign-flip raises the concern that the apparent success of VRP_t as a predictor variable may be an artefact of this particular sample period. Table 7 therefore repeats the

[25] My approach follows that of Bollerslev, Tauchen and Zhou (2009) rather than that of Drechsler and Yaron (2011), who use predictive regressions to forecast the evolution of variance itself. I follow the former approach to avoid in-sample/out-of-sample issues.

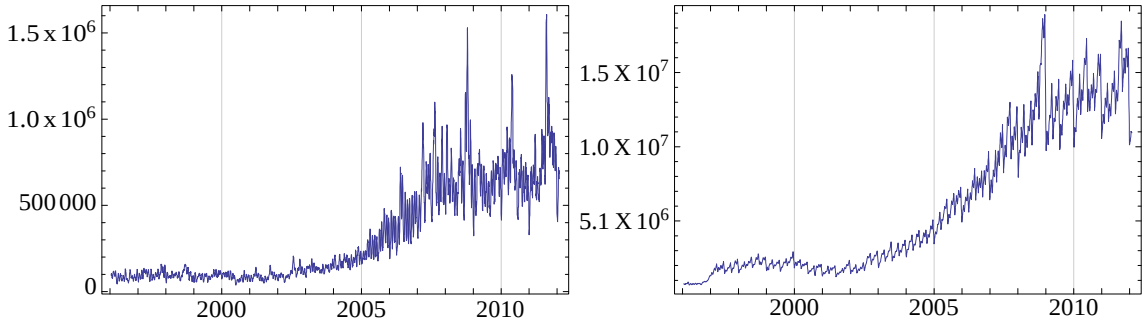
regression (46), but excludes the period from August 1, 2008 to July 31, 2009. Once this crisis period is excluded, VRP_t does not enter significantly at any horizon.

From a theoretical point of view, it is hard to rationalize a negative equity premium forecast within any equilibrium model. It is also implausible that the correctly-measured variance risk premium should ever be negative. More specifically, Bollerslev, Tauchen and Zhou (2009) show that within their own preferred equilibrium model, the variance risk premium would always be positive.



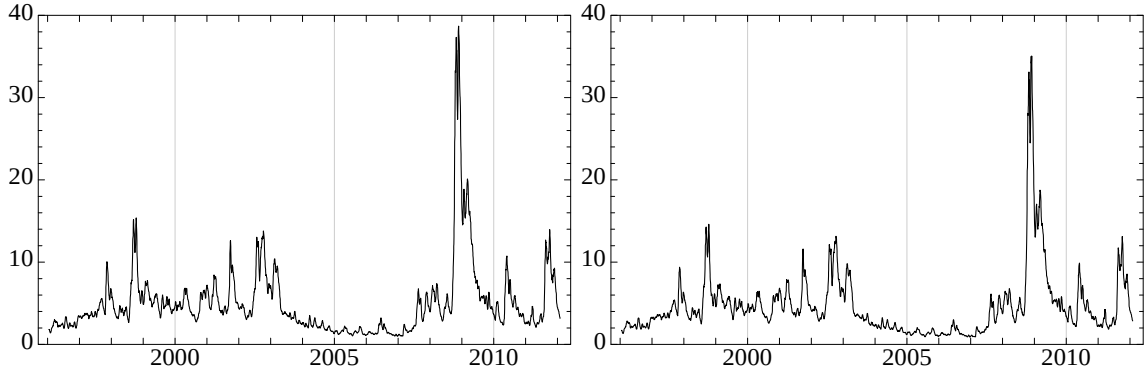
(a) 1-month (b) 1-year

Figure 12: The variance risk premium, calculated as $R_{f,t} \cdot \text{SVIX}_t^2 - \text{SVAR}_t$.



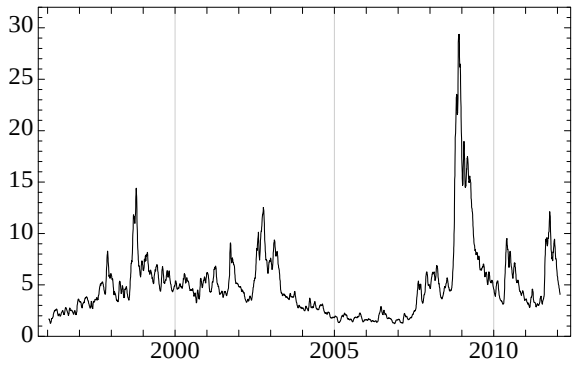
(a) Volume (b) Open interest

Figure 13: Volume and open interest in S&P 500 index options. The figures show 10-day moving averages.

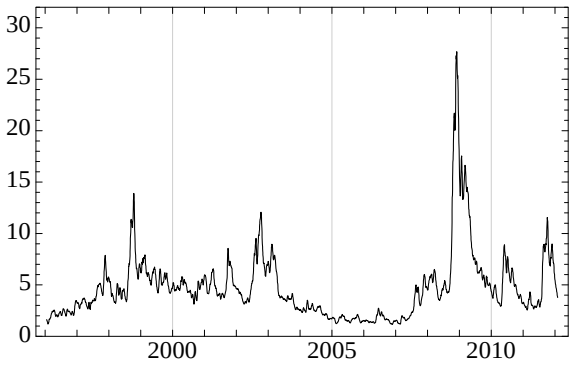


(a) 1 month, mid prices

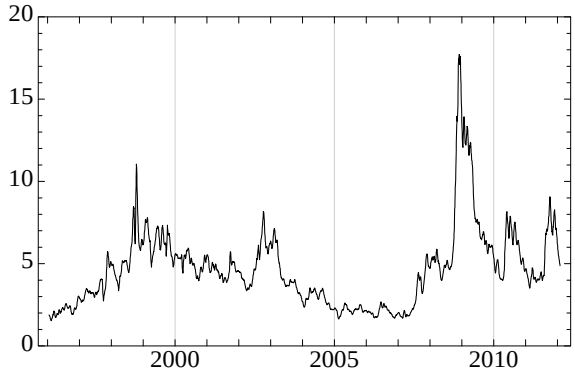
(b) 1 month, bid prices



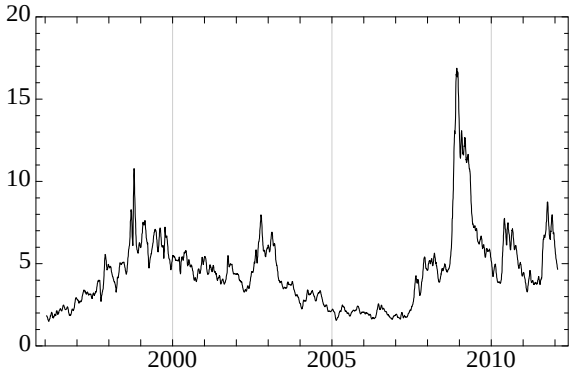
(c) 3 month, mid prices



(d) 3 month, bid prices



(e) 1 year, mid prices



(f) 1 year, bid prices

Figure 14: The lower bound on the annualized equity premium at different horizons (10-day moving averages, in %). Mid prices on left; bid prices on right.

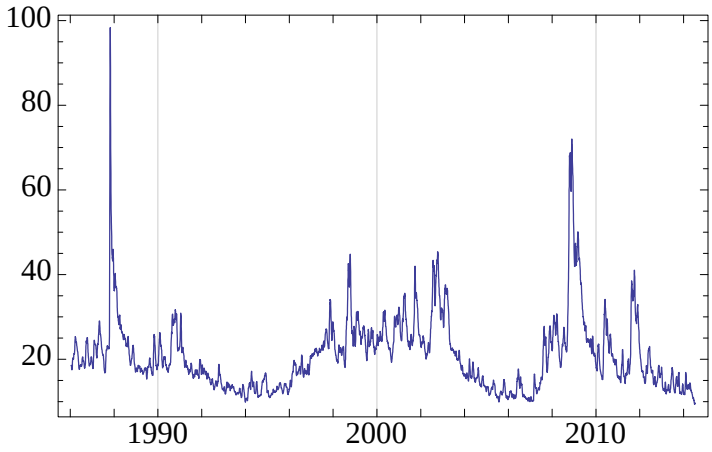


Figure 15: The VXO index, which exploded on Black Monday, October 19, 1987. 10-day moving average.

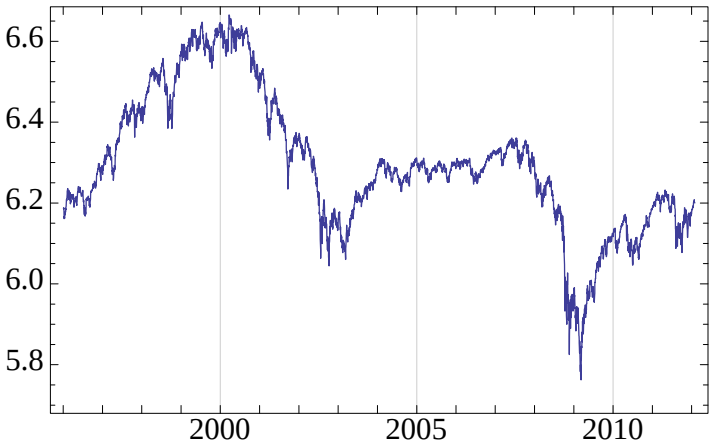


Figure 16: The calculations of R^2_{OS} in Table 2 depend on the rolling mean historical equity premium (shown here on an annualized basis). The rolling mean is computed using the data series used by Campbell and Thompson (2008), which itself is based on S&P 500 total returns from February 1871, with the data prior to January 1927 obtained from Robert Shiller's website.

D VIX, SVIX, and equilibrium models

Proof of Result 4. I write $\tau = T - t$ to make the notation easier to handle. Let $R_T = e^{\mu_{R,t}\tau + \sigma_t\sqrt{\tau}Z_R - \sigma_t^2\tau/2}$ and $M_T = e^{-r_{f,t}\tau + \sigma_{M,t}\sqrt{\tau}Z_M - \sigma_{M,t}\tau/2}$, where Z_R and Z_M are Normal random variables with mean zero, variance one, and correlation ρ_t, and $r_{f,t} = \log R_{f,t}$. Since $\mathrm{E}_t M_T R_T = 1$, we must have $\mu_{R,t} - r_{f,t} = -\rho_t\sigma_{M,t}\sigma_t$.

From (13), $\mathrm{SVIX}_t^2 = \frac{e^{-2r_{f,t}\tau}}{\tau}[\mathrm{E}_t^*(R_T^2) - (\mathrm{E}_t^* R_T)^2] = \frac{1}{\tau}(e^{-r_{f,t}\tau}\mathrm{E}_t M_T R_T^2 - 1)$. Now, using the fact that $\mu_{R,t} - r_{f,t} = -\rho_t\sigma_{M,t}\sigma_t$, we have

$$\begin{aligned}\mathrm{E}_t M_T R_T^2 &= \mathrm{E}\, e^{-r_{f,t}\tau + \sigma_{M,t}\sqrt{\tau}Z_M - \frac{1}{2}\sigma_{M,t}^2\tau + 2\mu_{R,t}\tau + 2\sigma_t\sqrt{\tau}Z_R - \sigma_t^2\tau} \\ &= e^{r_{f,t}\tau + \sigma_t^2\tau}.\end{aligned}$$

Thus, $\mathrm{SVIX}_t^2 = \frac{1}{\tau}(e^{\sigma_t^2\tau} - 1)$ as required.

The calculation for VIX is slightly more complicated. Using (27), $\mathrm{VIX}_t^2 = \frac{2}{\tau}\mathrm{L}_t^* R_T = \frac{2}{\tau}[\log \mathrm{E}_t^* R_T - \mathrm{E}_t^* \log R_T] = \frac{2}{\tau}[r_{f,t}\tau - e^{r_{f,t}\tau}\mathrm{E}_t M_T \log R_T]$. Now,

$$\begin{aligned}\mathrm{E}_t[M_T \log R_T] &= \mathrm{E}_t\left[(\mu_{R,t}\tau + \sigma_t\sqrt{\tau}Z_R - \tfrac{1}{2}\sigma_t^2\tau)e^{-r_{f,t}\tau + \sigma_{M,t}\sqrt{\tau}Z_M - \sigma_{M,t}^2\tau/2}\right] \\ &= (\mu_{R,t} - \tfrac{1}{2}\sigma_t^2)\tau \cdot e^{-r_{f,t}\tau} + \sigma_t\sqrt{\tau}e^{-r_{f,t}\tau - \sigma_{M,t}^2\tau/2}\,\mathrm{E}_t\left[Z_R e^{\sigma_{M,t}\sqrt{\tau}Z_M}\right].\end{aligned}$$

We can write $Z_R = \rho_t Z_M + \sqrt{1-\rho^2}Z$, where Z is uncorrelated with Z_M (conditional on time-t information) and hence, since they are both Normal, independent of Z_M. The expectation in the above expression then becomes

$$\begin{aligned}\mathrm{E}_t\left[Z_R e^{\sigma_{M,t}\sqrt{\tau}Z_M}\right] &= \mathrm{E}_t\left[(\rho_t Z_M + \sqrt{1-\rho_t^2}Z)e^{\sigma_{M,t}\sqrt{\tau}Z_M}\right] \\ &= \rho_t\,\mathrm{E}_t\left[Z_M e^{\sigma_{M,t}\sqrt{\tau}Z_M}\right].\end{aligned}$$

By Stein's lemma,

$$\begin{aligned}\mathrm{E}_t\left[Z_M e^{\sigma_{M,t}\sqrt{\tau}Z_M}\right] &= \mathrm{E}_t\left[\sigma_{M,t}\sqrt{\tau}e^{\sigma_{M,t}\sqrt{\tau}Z_M}\right] \\ &= \sigma_{M,t}\sqrt{\tau}e^{\sigma_{M,t}^2\tau/2}.\end{aligned}$$

These results, together with the fact that $\mu_{R,t} - r_{f,t} = -\rho_t\sigma_{M,t}\sigma_t$, imply that $\mathrm{E}_t M_T \log R_T = (\mu_{R,t} - \frac{1}{2}\sigma_t^2 + \rho_t\sigma_{M,t}\sigma_t)\tau e^{-r_{f,t}\tau} = (r_{f,t} - \frac{1}{2}\sigma_t^2)\tau e^{-r_{f,t}\tau}$. Thus $\mathrm{VIX}_t^2 = \frac{2}{\tau}r_{f,t}\tau - (r_{f,t} - \frac{1}{2}\sigma_t^2)\tau = \sigma_t^2$, as required. □

Figure 17 plots log real consumption growth and the 1-year SVIX-implied equity premium on the same axes, with each time series scaled to have zero mean and unit

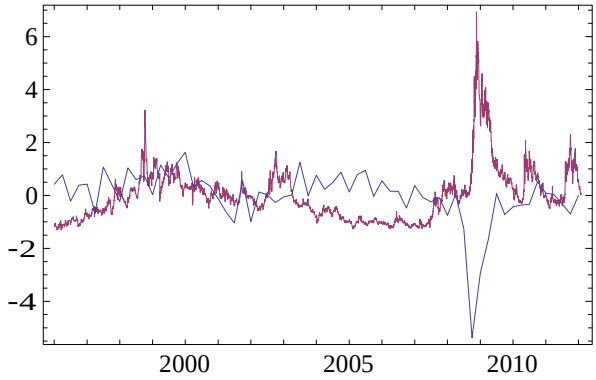


Figure 17: Log quarterly consumption growth (blue) and monthly-averaged 1-year SVIX (red). Each series is scaled to have zero mean and unit variance.

variance. The data series in the figure is quarterly, seasonally-adjusted Personal Consumption Expenditures, taken from the Bureau of Economic Analysis.

The top panel of Table 4 reports a variety of summary statistics for VIX, SVIX, and VIX minus SVIX in the data, at the 1-month horizon. (For comparison with the models, which are simulated at monthly frequency, I generate a monthly series from the daily series of VIX and SVIX by taking the 1st, 22nd, 43rd, 64th, . . . , elements and compute the mean, median, etc. Then I repeat using the 2nd, 23rd, . . . elements; the 3rd, 24th, . . . ; and so on, up to the 21st, 42nd, ... Finally, I average each statistic over the 21 choices of initial element.) The panels below report corresponding statistics computed within six equilibrium models, namely the Campbell–Cochrane (CC, 1999) habit formation model, the Bansal–Yaron (BY, 2004) long-run risk model in its original calibration with stochastic volatility, the updated calibration of the long-run risk model studied in Bansal, Kiku and Yaron (BKY, 2012), Wachter's (W, 2013) model with a time-varying disaster arrival rate, and two models that explicitly address the properties of option prices: Bollerslev, Tauchen and Zhou (BTZ, 2009), and Drechsler and Yaron (DY, 2011).

Within each model, I simulate 1,000,000 16-year-long sample paths of VIX, SVIX, and VIX minus SVIX. Each sample path is generated by initializing state variables at their long-run averages, then computing a 32-year sample realization. I discard the first 16-year "burn-in" period and use the second 16 years (for comparability with the 16 years of data). I compute VIX and SVIX at the 1-month horizon for all models apart from Wachter's (2013) continuous-time model, for which I compute an instantaneous

measure, i.e. I report the limiting case in which the time horizon $T - t$ approaches zero, rather than one month.[26] Along each sample path, I compute the mean, standard deviation, median, minimum, maximum, skewness, excess kurtosis, and monthly autocorrelation for VIX, SVIX, and VIX minus SVIX. The numbers reported in Table 4 are the averages, across 1,000,000 sample paths in each model, of each of these quantities. Asterisks in Table 4 indicate statistics for which a given model struggles to match the data. The models find it so difficult to match the data that I use fewer asterisks than is conventional to indicate significance levels: one asterisk denotes a *p*-value of 0.05 (fewer than 5% of the 1,000,000 trials gave statistics as extreme as are observed in the data), two asterisks a *p*-value of 0.01, three asterisks a *p*-value of 0.000 to three decimal places. Boldface font indicates that the observed statistic in the data lies completely outside the support of the 1,000,000 model-generated statistics (i.e., an empirical *p*-value of zero). It goes without saying that a successful model should not have any boldface statistics; unfortunately, there are multiple such examples for all six models.

[26] I do so for tractability, because this quantity can be computed in closed form. Since the Wachter model generates too little skewness and kurtosis in VIX and SVIX, and too much persistence, it is likely that using the 1-month measure would make the results even worse. It is also possible to solve for VIX and SVIX in closed form within Barro's (2006) model; in this case, the term structures of VIX and SVIX are flat, so there is no distinction between the instantaneous and 1-month measures. Following Martin (2013) by defining $\kappa(\theta) \equiv \log \mathrm{E}_t \left[(C_{t+1}/C_t)^{\theta} \right]$, it can be shown that within Barro's model—or indeed within any consumption-based model with an Epstein–Zin representative agent and i.i.d. consumption growth—we have

$$\begin{aligned} \mathrm{VIX}^2 &= 2\left[\kappa(1-\gamma) - \kappa(-\gamma) - \kappa^1(-\gamma)\right] \\ \log\left(1 + \mathrm{SVIX}^2\right) &= \kappa(2-\gamma) - 2\kappa(1-\gamma) + \kappa(-\gamma), \end{aligned}$$

where γ is relative risk aversion. Using Barro's (2006) calibrated parameters and empirical distribution of disaster sizes, one finds that VIX = 23.8% and SVIX = 18.4%, and that the difference between the two is 5.4%, well above the value observed in the data. (VIX and SVIX are constant in Barro's model.) These calculations assume that there is no default on index options, and model 'equity' as an unlevered claim to consumption. Allowing for default would move the numbers in the right direction—the gap between VIX and SVIX would decline—since VIX loads more heavily on the deep out-of-the-money puts that would be most vulnerable to default. Allowing for leverage would move the numbers in the wrong direction, expanding the gap between VIX and SVIX.

E Simple variance swaps

This section provides the details of how to hedge a simple variance swap, and collects together some robustness results regarding simple variance swaps.

E.1 Hedging a simple variance swap

The proof of Result 6 implicitly supplies the dynamic trading strategy that replicates the payoff on a simple variance swap. Tables 8 and 9 describe the strategy in detail. Each row of Table 8 indicates a sequence of dollar cashflows that is attainable by investing in the asset indicated in the leftmost column. Negative quantities indicated that money must be invested; positive quantities indicate cash inflows. Thus, for example, the first row indicates a time-0 investment of $\$e^{-rT}$ in the riskless bond maturing at time T, which generates a time-T payoff of \$1. The second and third rows indicate a short position in the underlying asset, held from 0 to Δ with continuous reinvestment of dividends, and subsequently rolled into a short bond position. The fourth row represents a position in a portfolio of call options of all strikes expiring at time Δ, as in equation (37); this portfolio has simple return $S_\Delta^2/\Pi(\Delta)$ from time 0 to time Δ. The fifth, sixth, and seventh rows indicate how the proceeds of this option portfolio are used after time Δ. One part of the proceeds is immediately invested in the bond until time T; another part is invested from Δ to 2Δ in the underlying asset, and subsequently from 2Δ to T in the bond. The replicating portfolio requires similar positions in options expiring at times $2\Delta, 3\Delta, \ldots, T-2\Delta$. These are omitted from Table 8, but the general such position is indicated in Table 9, together with the subsequent investment in bonds and underlying that each position requires.

The self-financing nature of the replicating strategy is reflected in the fact that the total of each of the intermediate columns from time Δ to time $T-\Delta$ is zero. The last column of Table 8 adds up to the desired payoff,

$$\left(\frac{S_\Delta - S_0}{F_{0,0}}\right)^2 + \left(\frac{S_{2\Delta} - S_\Delta}{F_{0,\Delta}}\right)^2 + \ldots + \left(\frac{S_T - S_{T-\Delta}}{F_{0,T-\Delta}}\right)^2 - V.$$

Therefore, the first column must add up to the cost of entering the simple variance swap. Equating this cost to zero, we find the value of V provided in equation (36).

The replicating strategy simplifies nicely in the $\Delta \to 0$ limit. The dollar investment in each of the option portfolios expiring at times $\Delta, 2\Delta, \ldots, T-\Delta$ goes to zero at

rate $O(\Delta^2)$. We must account, however, for the dynamically adjusted position in the underlying, indicated in rows beginning with a U. As shown in Table 9, this calls for a short position in the underlying asset of $2e^{-r(T-(j+1)\Delta)}S^2_{j\Delta}e^{-\delta\Delta}/F^{\,2}_{0,j\Delta}$ in dollar terms at time $j\Delta$, that is, a short position of $2e^{-r(T-(j+1)\Delta)}S_{j\Delta}e^{-\delta\Delta}/F^2_{0,j\Delta}$ units of the underlying. In the limit as $\Delta \to 0$, holding $j\Delta = t$ constant, this equates to a short position of $2e^{-r(T-t)}S_t/F^2_{0,t}$ units of the underlying asset at time t.

The static position in options expiring at time T, shown in the penultimate line of Table 8, does not disappear in the $\Delta \to 0$ limit. We can think of the option portfolio as a collection of calls of all strikes, as in (37). It is more natural, though, to use put-call parity to think of the position as a collection of calls with strikes above $F_{0,T}$ and puts with strikes below $F_{0,T}$, together with a long position in $2e^{-\delta(T-t)}/F_{0,T}$ units of the underlying asset—after continuous reinvestment of dividends—and a bond position. Combining this static long position in the underlying with the previously discussed dynamic position, the overall position at time t is long $2e^{-\delta(T-t)}/F_{0,T} - 2e^{-r(T-t)}S_t/F^2_{0,t} = 2e^{-\delta(T-t)}(1 - S_t/F_{0,t})/F_{0,T}$ units of the asset and long out-of-the-money-forward calls and puts, all financed by borrowing.

E.2 Pricing and hedging with $\Delta > 0$

The hedging strategy provided in Tables 8 and 9 perfectly replicates the desired payoff when $\Delta > 0$, but requires positions in options at all expiry dates $\Delta, \ldots, T - \Delta$. Discretizing the continuous-time strategy provided in the statement of Result 6 (which is exactly valid in the limit as $\Delta \to 0$) is equivalent to ignoring all such positions in options with intermediate expiry dates. The cashflows in these rows contribute a term of size $O(\Delta)$ at time 0, and terms of size $O(\Delta^2)$ at dates between 1 and $T - \Delta$. Thus the overall replication error is of size $O(\Delta)$, so the limiting strike is a good approximation to the truth for sampling intervals $\Delta > 0$. The next result makes this formal.

Result 8. *For $\Delta > 0$, the exact simple variance swap strike $V(\Delta)$, given by equation* (36)*, is very well approximated by V, given in equation* (34)*:*

$$|V(\Delta) - V| \le \frac{T}{\Delta}\left(e^{(r-\delta)\Delta} - 1\right)^2 (1 + V) + \left(e^{2(r-\delta)\Delta} - 1\right) V. \qquad (47)$$

If $T = 1$, $r - \delta = 0.02$, $V = 0.05$, then the right-hand side of (47) *is less than 0.00001 with daily sampling ($\Delta = 1/252$), less than 0.00005 with weekly sampling ($\Delta = 1/52$), and less than 0.0002 with monthly sampling ($\Delta = 1/12$).*

asset	0	Δ	2Δ	...	$T-\Delta$	T
B	$-e^{-rT}$			...		$\frac{S_0^2}{S^2}$
U	$2e^{-r(T-\Delta)}e^{-\delta\Delta}$	$-2e^{-r(T-\Delta)}\frac{S_\Delta}{S_0}$		...		
B		$2e^{-r(T-\Delta)}\frac{S_\Delta}{S_0}$		...		$\frac{S_0 S_\Delta}{S^2}$
Δ	$-\frac{(e^{(r-\delta)\Delta}-1)^2\Pi_\Delta}{e^{r(T-\Delta)}F^2}$	$\frac{(e^{(r-\delta)\Delta}-1)^2 S^2}{e^{r(T-\Delta)}F^2}$		...		
B		$e^{-r(T-\Delta)}\left[\frac{-S^2}{0,\Delta}\right.$		...		$\frac{S^2}{S_0}+\frac{S^2}{F}$
U		$\frac{2e^{-r(T-2\Delta)}S^2}{F^2}$	$\frac{r(T-2\Delta)}{F^2}$	...		
B			$\frac{2e^{-r(T-2\Delta)}S_\Delta S_{2\Delta}}{F^2}$	...		$\frac{-2S_\Delta S_{2\Delta}}{S^2}$
:	:			...		:
$T-\Delta$	$-\frac{(e^{(r-\delta)\Delta}-1)^2\Pi_{T-\Delta}}{e^{r\Delta}F^2}$			...	$\frac{(e^{(r-\delta)\Delta}-1)^2 S^2}{e^{r\Delta}F^2}$	
B				...	$e^{-r\Delta}\left[=S_{T-\Delta}-\frac{S_{T-\Delta}}{}\right.$	$\frac{S_{T\ \Delta}}{}+\frac{S_{T\ \Delta}}{}$
U				...	$\frac{2S^2}{F^2}\ \Delta\quad\delta\Delta$	$\frac{-2S_{T\ \Delta}S_T}{F^2}$
T	$-\frac{\Pi_T}{}$			...		$\frac{S^2}{F^2}$
B	Ve^{-rT}			...		$-V$

Table 8: Replicating the simple variance swap. In the left column, B indicates dollar positions in the bond, U indicates dollar positions in the underlying with dividends continuously reinvested, and $j\Delta$, for $j = 1, 2, \ldots, T/\Delta$, indicates a position in the portfolio of options expiring at time $j\Delta$ that replicates the payoff $S^2_{j\Delta}$, whose price at time 0 is $\Pi_{j\Delta}$.

asset	0	$j\Delta$	$(j+1)\Delta$	T
$j\Delta$	$\frac{-(e^{(r-\delta)\Delta}-1)^2\Pi_{j\Delta}}{e^{r(T-j\Delta)}F^2_{0,j\Delta}}$	$\frac{(e^{(r-\delta)\Delta}-1)^2 S^2_{j\Delta}}{e^{r(T-j\Delta)}F^2_{0,j\Delta}}$		
B		$e^{-r(T-j\Delta)}\left(\frac{-S^2_{j\Delta}}{F^2_{0,(j-1)\Delta}} - \frac{S^2}{F^2_{0,j\Delta}}\right)$		$\frac{S^2}{F^2_{0,(j-1)\Delta}} + \frac{S^2}{F^2_{0,j\Delta}}$
U		$\frac{2S^2 e^{-\delta\Delta}}{e^{r(T-(j+1)\Delta)}F^2_{0,j\Delta}}$	$\frac{-2S_{j\Delta}S}{e^{r(T-(j+1)\Delta)}F^2_{0,j\Delta}}$	
B			$\frac{2S_{j\Delta}S_{(j+1)\Delta}}{e^{r(T-(j+1)\Delta)}F^2_{0,j\Delta}}$	$\frac{-2S_{j\Delta}S_{(j+1)\Delta}}{F^2_{0,j\Delta}}$

Table 9: Replicating the simple variance swap. The generic position in options of intermediate maturity, together with the associated trades required after expiry. In the left column, B indicates a position in the bond, U indicates a position in the underlying with dividends continuously reinvested, and $j\Delta$ indicates a position in options expiring at $j\Delta$.

Proof. Result 6 implies that for $j < T/\Delta$,

$$\frac{e^{rj\Delta}P(j\Delta)}{F^2_{0,j\Delta}} = \lim_{\Delta\to 0} \mathrm{E}^* \sum_{i=1}^{j} \left(\frac{S_{i\Delta} - S_{(i-1)\Delta}}{F_{0,(i-1)\Delta}}\right)^2 \leq \lim_{\Delta\to 0} \mathrm{E}^* \sum_{i=1}^{T/\Delta} \left(\frac{S_{i\Delta} - S_{(i-1)\Delta}}{F_{0,(i-1)\Delta}}\right)^2 = \frac{e^{rT}P(T)}{F^2_{0,T}}.$$

Combining this observation with (39), we find that

$$\begin{aligned} V(\Delta) - \frac{e^{rT}P(T)}{F^2_{0,T-\Delta}} &= \sum_{j=1}^{T/\Delta-1} \left(e^{(r-\delta)\Delta} - 1\right)^2 \frac{e^{rj\Delta}P(j\Delta)}{F^2_{0,j\Delta}} + \frac{T}{\Delta}\left(e^{(r-\delta)\Delta} - 1\right)^2 \\ &\leq \frac{T}{\Delta}\left(e^{(r-\delta)\Delta} - 1\right)^2 \frac{e^{rT}P(T)}{F^2_{0,T}} + \frac{T}{\Delta}\left(e^{(r-\delta)\Delta} - 1\right)^2. \end{aligned}$$

Now, by definition of V, we have $|e^{rT}P(T)/F^2_{0,T-\Delta} - V| = e^{2(r-\delta)\Delta} - 1\ V$. Since $|V(\Delta) - V| \leq |V(\Delta) - e^{rT}P(T)/F^2_{0,T-\Delta}| + |e^{rT}P(T)/F^2_{0,T-\Delta} - V|$, by the triangle inequality, the result follows. □

E.3 Pricing and hedging when deep-out-of-the-money strikes are not tradable

Options that are sufficiently deep-out-of-the-money have prices so close to zero that they are not traded. Thus the idealized replicating portfolio, which comprises options of all strikes, is not attainable in practice. This issue affects both conventional variance swaps and simple variance swaps. Fortunately there is a practical solution to this problem. Suppose that, at time 0, options with strikes between A and B are tradable; the idealized scenario in which *all* strikes are tradable corresponds to $A = 0$, $B = \infty$. Then we can define the modified payoff

$$\left(\frac{S_\Delta - S_0}{F_{0,0}}\right)^2 + \left(\frac{S_{2\Delta} - S_\Delta}{F_{0,\Delta}}\right)^2 + \cdots + \left(\frac{S_T - S_{T-\Delta}}{F_{0,T-\Delta}}\right)^2 - \varphi(S_T), \qquad (48)$$

where the correction term $\varphi(S_T)$ is zero unless the underlying asset's price happens to end up outside the original strike range (A, B):

$$\varphi(S_T) = \begin{cases} \left(\frac{A - S_T}{F_{0,T-\Delta}}\right)^2 & \text{if } S_T < A. \\ 0 & \text{if } A \le S_T \le B. \\ \left(\frac{S_T - B}{F_{0,T-\Delta}}\right)^2 & \text{if } S_T > B. \end{cases}$$

The modified payoff (48) *can* be replicated without needing to trade options with strikes outside the range (A, B), by holding

(i) a static position in $2/F_{0,T}^2\, dK$ puts expiring at time T with strike K, for each $A < K \le F_{0,T}$,

(ii) a static position in $2/F_{0,T}^2\, dK$ calls expiring at time T with strike K, for each $F_{0,T} \le K < B$, and

(iii) a dynamic position in $2e^{-\delta(T-t)}(1 - S_t/F_{0,t})/F_{0,T}$ units of the underlying asset at time t,

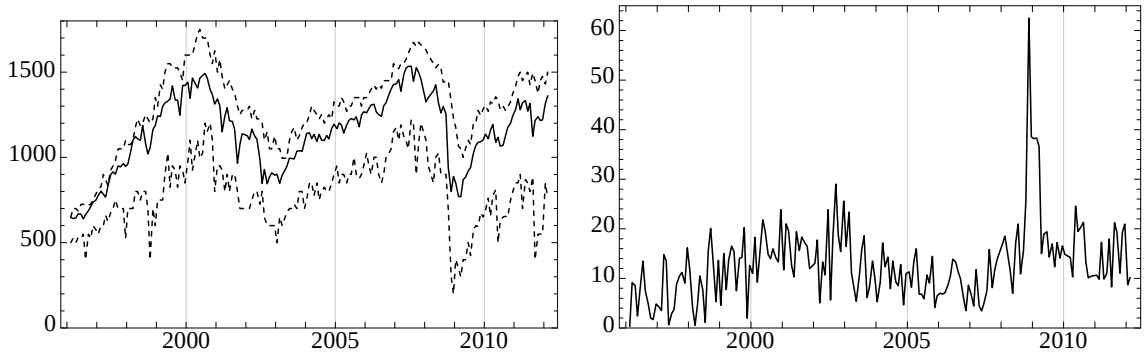
financed by borrowing. To see this, simply note that the payoff $\varphi(S_T)$ is precisely the payoff on the "missing" options with strikes less than A and greater than B that are

not included in the above position.

In the limit as $\Delta \to 0$, the fair strike that should be exchanged for the payoff (48) at time T is

$$V \equiv \frac{2e^{rT}}{F_{0,T}^2}\left(\int_A^{F_{0,T}} \text{put}_{0,T}(K)\,dK + \int_{F_{0,T}}^{B} \text{call}_{0,T}(K)\,dK\right) .$$

To explore how large the adjustment term $\varphi(S_T)$ is in practice in the case of the S&P 500 index, I looked at every day in the sample on which *OptionMetrics* had data for options expiring in 30 days. On each such day, I recorded the lowest tradable strike the strike of the most deep-out-of-the-money put option) and the highest tradable strike (i.e. the strike of the most deep-out-of-the-money call option), together with the subsequently realized level of the market at expiry time T.



(a) S_T, A, and B (b) $\frac{\min\{S_T - A, B - S_T\}}{S_T}$

Figure 18: Left: Upper and lower strike boundaries, A and B (dotted lines) and subsequent realized level of the market at expiry, S_T (solid line). Right: The distance at expiry from the edge of the strike range, expressed as a percentage of the terminal level S_T.

The results are shown in Figure 18. Over the sample period, the underlying asset's price *never* ended up outside the range of tradable strikes. In other words, the correction term $\varphi(S_T)$ was zero in every case: in Figure 18a, the value of S_T at expiry is within the range of strikes that were tradable at initiation on every day in sample. Figure 18b shows how far the underlying ended from the closer of the two boundaries, expressing the result as a percentage of S_T; the graph is always positive, reflecting the fact that the strike boundary was never crossed over the sample period. The spike in the figure occurred on 21 November, 2008, when the S&P 500 happened to close near

the middle of the strike range that had prevailed 30 days previously; moreover, this occurred at a time when implied volatilities were very high, so that an extremely wide range of strikes had been traded. The low point in the figure occurred at the very beginning of the sample, on January 18, 1996, when the S&P 500 closed at 608.24. On that day, the highest strike tradable on options expiring in 30 days—on Saturday, February 17, 1996—was 650; in the event, the S&P 500 closed just two points lower, at 647.98, on Friday, February 16.

As is apparent from Figure 18a, the width of the range of tradable strikes has tended to increase over time. The mean value of the percentage distance to the edge of the strike range, as illustrated in Figure 18b, is 12.9%; the median value is 11.9%. In other words, on the median day in sample, the S&P 500 would have had to move a *further* 11.9% in the appropriate direction in order to exit the relevant range of tradable strikes.

E.4 Pricing and hedging under different assumptions on dividends

This section shows what happens to pricing and hedging of simple variance swaps under various different assumptions about dividend payout policies.

E.4.1 The case of completely unanticipated dividend payouts

Result 6 continues to hold if the asset makes unanticipated dividend payouts. Consider an extreme case in which the simple variance swap is priced and hedged, at time zero, as though $\delta = 0$; but immediately after inception of the trade, at time $t = \Delta$, the underlying asset is suddenly liquidated via an extraordinary dividend, causing its (ex-dividend) price to equal 0 from time Δ onwards. The payout that must be made by the counterparty who is short variance is given by equation (33): in this extreme example, it will equal 1. Meanwhile, the hedge portfolio given in the above result will generate a positive payoff due to the put options going in-the-money. (The dynamic position will have zero payoff: it was neither long nor short at time 0, and subsequently the asset's price never moved from zero.) Since $S_T = 0$, the total payoff will be

$$\frac{2}{F_{0,T}^2} \int_0^{F_{0,T}} \max\{0, K - S_T\}\, dK = \frac{2}{F^{0,T}} \int_0^{F_{0,T}} K\, dK = 1.$$

In other words, the strategy perfectly replicates the desired payoff. This applies more generally: once the strike V is set and the replicating portfolio is in place, it does not matter why the price path moves around subsequently, whether due to the payment of unanticipated dividends or not.

E.4.2 The case of perfectly anticipated dividends

For simplicity, consider the case in which the asset pays a single dividend $D_{k\Delta}$ at time $k\Delta$ for some k, and no dividends at any other time up to and including the expiry date, T. The price of a portfolio whose payoff is S_i^2 at time i continues to equal $\Pi(i)$, given by equation (37).

In this section, it will be important to distinguish between $F_{0,t}$, the forward price of the dividend-paying asset to time t, and $\tilde{F}_{0,t} \equiv S_0 e^{rt}$, the appropriate normalization for the definition of a simple variance swap in this case. A standard no-arbitrage argument implies that the forward price is $F_{0,t} = S_0 e^{rt}$ if $t < k\Delta$, and $F_{0,t} = S_0 e^{rt} - D_{k\Delta} e^{r(t-k\Delta)}$ if $t \geq k\Delta$, so $F_{0,t}$ and $\tilde{F}_{0,t}$ coincide for times t before the payment of the dividend, but differ thereafter. It turns out that $\tilde{F}_{0,t}$ is the appropriate normalization so that the intermediate option positions are negligibly small, as was the case in the main text.

The definition of the payoff on the simple variance swap must be modified to allow for the presence of the dividend. At time T, the counterparties to the simple variance swap now exchange V for

$$\left(\frac{S_\Delta - S_0}{\tilde{F}_{0,0}}\right)^2 + \cdots + \left(\frac{S_{(k-1)\Delta} - S_{(k-2)\Delta}}{\tilde{F}_{0,(k-2)\Delta}}\right)^2 + \left(\frac{S_{k\Delta} + D_{k\Delta} - S_{(k-1)\Delta}}{\tilde{F}_{0,(k-1)\Delta}}\right)^2 + \left(\frac{S_{(k+1)\Delta} - S_{k\Delta}}{\tilde{F}_{0,k\Delta}}\right)^2 + \cdots + \left(\frac{S_T - S_{T-\Delta}}{\tilde{F}_{0,T-\Delta}}\right)^2 . \quad (49)$$

If the stock price happens to track the forward price at all points in time, then the payoff (49) will be zero in the $\Delta \to 0$ limit, as is the case with variance swaps and simple variance swaps in the absence of dividends.

The starting point of the replicating strategy will be to carry out precisely the trades listed in Tables 8 and 9 with δ set equal to zero (and replacing $F_{0,t}$ with $\tilde{F}_{0,t}$ wherever it occurs in the tables). This replicating strategy generates the payoff (49) minus V, plus an extra payoff of $(D_{k\Delta}/\tilde{F}_{0,(k-1)\Delta})^2 - 2D_{k\Delta}(S_{k\Delta} + D_{k\Delta})/\tilde{F}^2_{0,(k-1)\Delta}$. To offset this extra payoff, two new positions are required: (i) a short position of $e^{-rT}(D_{k\Delta}/\tilde{F}_{0,(k-1)\Delta})^2$

(measured in dollars) in bonds, and (ii) a long position of $2D_{k\Delta}e^{-r(T-k\Delta)}/F^2_{0,(k-1)\Delta}$ units of the underlying held until time $k\Delta$, then rolled into bonds.

After some algebra (and up to terms of order Δ, as usual) this implies that the simple variance swap strike is given by

$$V = \frac{2e^{rT}}{F^2_{0,T}}\left(\int_0^{F_{0,T}} \text{put}_{0,T}(K)\,dK + \int_{F_{0,T}}^{\infty} \text{call}_{0,T}(K)\,dK\right),$$

and that the replicating portfolio is equivalent to holding

(i) a static position of $2/F^2_{0,}$ puts expiring at time T with strike K, for each $K \leq F_{0,T}$,

(ii) a static position of $2/F^2_{0,}$ calls expiring at time T with strike K, for each $K \geq F_{0,T}$, and

(iii) a dynamic position of $2(F_{0,t} - S_t)/(F_{0,t}F_{0,T})$ units of the underlying asset at time t,

financed by borrowing.

E.4.3 The case of imperfectly anticipated dividends

The fully general case in which dividends are potentially anticipated but of unknown size and timing is, of course, the most challenging (and to the best of my knowledge, it is ignored in the variance swap literature). Even so, there is an elegant solution in this case too if *total return* options can be traded: these are options on a claim to the underlying asset with dividends reinvested. Such options have recently started to trade over-the-counter. I will call the underlying with dividends reinvested the *dividend-adjusted underlying*. Then we can price and hedge a simple variance swap on the dividend-adjusted underlying directly from Result 6 simply by reinterpreting the inputs. The price S_t corresponds to the price of the dividend-adjusted underlying (so S_0 is the spot price of the underlying asset); the instantaneous dividend yield $\delta = 0$; $F_{0,t}$ is the forward price of the dividend-adjusted underlying, which equals S_0e^{rt} for all t by a static no-arbitrage argument; and $\text{put}_{0,T}(K)$ and $\text{call}_{0,T}(K)$ are the prices of total return options expiring at time T.