Journal of Financial Econometrics, 2022, 1–40
https://doi.org/10.1093/jjfinec/nbac008
Article

Arbitrage Pricing Theory for Idiosyncratic Variance Factors*

Eric Renault[1,2], Thijs Van Der Heijden[3], and Bas J. M. Werker [4]

[1]Department of Economics, University of Warwick, Coventry CV4 7AL, UK, [2]Department of Econometrics, Monash University, Wellington Road, Clayton, VIC 3800, Australia, [3]Department of Finance, The University of Melbourne, 198 Berkeley Street, Carlton, VIC 3010, Australia and [4]Econometrics and Finance Group, Netspar, Tilburg University, P.O. Box 90153, 5000 LE, Tilburg, The Netherlands

Address correspondence to Bas J. M. Werker, Econometrics and Finance Group, Netspar, Tilburg University, P.O. Box 90153, 5000 LE, Tilburg, The Netherlands, or e-mail:Werker@TilburgUniversity.edu.

Received March 16, 2021; revised February 28, 2022; editorial decision March 3, 2022; accepted March 4, 2022

Abstract

We develop an arbitrage pricing theory framework extension to study the pricing of squared returns/volatilities. We analyze the interplay between factors at the return level and those in idiosyncratic variances. We confirm the presence of a common idiosyncratic variance factor, but do not find evidence that this represents a missing risk factor at the (linear) return level. Thereby, we consistently identify idiosyncratic returns. The price of the idiosyncratic variance factor identified by squared returns is small relative to the price of market variance risk. The quadratic pricing kernels induced by our model are in line with standard economic intuition.

Key words: arbitrage pricing theory, common volatility factors, nonlinear pricing kernels, option prices

JEL classification: C58, G12

Recently, several papers have documented the presence of a (priced) common factor in idiosyncratic volatilities or variances from a linear return factor model, for example, Connor,

* We thank conference participants at the Econometrics of High-Dimensional Risk Networks at the Stefanovich Center, the Sofie 2016 conference, the Vienna–Copenhagen 2017 conference, the Australian Finance and Banking Conference 2018, seminar participants at Tilburg University and the University of Lugano, and Jonathan Dark, Rob Engle, Dacheng Xiu, and Chu Zhang for helpful comments and suggestions. Two referees and the editor provided valuable comments to improve the paper as well.

Korajczyk, and Linton (2006); Duarte et al. (2014); and Herskovic et al. (2016).[1] We revisit this result within the Ross (1976) arbitrage pricing theory (APT) framework. Specifically, we extend the APT to squared returns, which provide a natural setting to study (idiosyncratic) variance factors. Using industry portfolios, individual stocks, implied volatility (IVOL) portfolios, and size-B/M portfolios as test assets, we find evidence of the idiosyncratic variance factor being priced for portfolios of squared returns. However, in contrast to the existing literature on the idiosyncratic volatility (IV) puzzle, we do not find evidence that the idiosyncratic variance factor is priced for excess returns themselves. Hence, like van der Heijden, Zeng, and Zhu (2019), we conclude that the effects of the commonality of idiosyncratic volatilities appear limited to volatility-sensitive claims.[2]

Our theoretical contribution is a reformulation and extension of the APT that not only prices excess returns, but also squares of excess returns. Our reformulation of the APT starts from a continuum of assets and the returns on portfolios of sets of assets. Therefore, our work is related to other extensions of the APT using a continuum of assets, in particular Al-Najjar (1998) and Gagliardini, Ossola, and Scaillet (2016). Using a continuum of assets has the advantage that we can replace the traditional conclusion of APT "most assets have small pricing errors" by the more testable statement that "outside a set of Lebesgue measure zero, every asset is exactly factor-priced." Our framework encompasses the concept of an approximate factor structure recently put forward by Gagliardini et al. (2016). We formulate our approach in terms of continuous stochastic processes, which may be more familiar to the dynamic asset pricing literature. Note, however, that in our static setting the continuous index does not represent "time," but rather a specific portfolio. Our new formulation of an approximate factor structure easily extends the APT for linear returns to squared returns and, thereby, to (idiosyncratic) variances.[3]

The notion of "idiosyncratic" return depends on which factors have been included in the model in the first place. Different factor choices at the linear return level lead to different "idiosyncratic" returns and, hence, different notions of "idiosyncratic" variance factors. More precisely, an omitted factor at the linear return level will generally show up as a common factor in the "idiosyncratic" variances. This distinguishes our paper from, for instance, Connor, Korajczyk, and Linton (2006); Ang et al. (2006b); Duarte et al. (2014); and Herskovic et al. (2016). In those studies, starting from a set of factors at the linear return level, the authors identify a common factor in the induced idiosyncratic variances. Duarte et al. (2014) and Herskovic et al. (2016) then add, in a second step, this idiosyncratic variance factor to the *initial* set of factors and show it to be priced. It cannot be ruled out that the common idiosyncratic variance factor simply picks up an omitted factor at the linear return level from the first step.

Our main theoretical result (Corollary 1) identifies the effect of possible omitted factors at the linear return level on the idiosyncratic variances. This allows us to disentangle the

1 See also Ang et al. (2006b) and the literature that sprung from that paper focusing on the "idiosyncratic volatility puzzle" in which stocks with a high current idiosyncratic volatility earn low average returns in the future.

2 van der Heijden, Zeng, and Zhu (2019) argue that the commonality is driven by time-varying financial leverage. In this paper, we are agnostic about the economic drivers of the commonality.

3 To make the distinction clear, we will refer explicitly to "linear" and "squared" returns in the remainder.

effect of possibly omitted factors at the linear return level from factors in idiosyncratic variances. Quadratic functions of the original risk factors show up as implied risk factors in the APT equation for the squared returns, in addition to any idiosyncratic variance factors. We show that each of those risk factors is associated with their own price of risk, separate from the prices of risk of the original factors. In our empirical analysis, we use this to directly examine the price of market variance risk.[4]

The motivation to study squared excess returns is the direct link to idiosyncratic variance factors as well as their connection to returns on variance swaps, which have gained popularity in portfolio choice problems recently.[5] The realized variance of a variance swap is usually computed from daily squared returns on the underlying asset. This setting gives rise to a pricing kernel that is quadratic in the factors of the excess return process, and linear in any idiosyncratic variance factors.

In addition to the link with variance swaps, the framework we develop to study the pricing of squared returns also has close links to the literature on skewness in asset pricing through the quadratic pricing kernel we obtain. This literature is formalized in Kraus and Litzenberger (1976) and Harvey and Siddique (2000) who write down a model in which the pricing kernel is linear in the market return and its square. Chabi-Yo, Leisen, and Renault (2014) study the aggregation of preferences in the presence of skewness risk and show how the risk premium for skewness is linked to the portfolio that optimally hedges the squared market return. However, since this hedge is not perfect, an additional factor may appear in case of heterogeneous preferences for skewness. Therefore, the results of Chabi-Yo, Leisen, and Renault (2014) provide some structural underpinnings to our working hypothesis that, in case of a linear factor model, investors' preferences may lead to not only the squared market return as a factor, but also an additional factor due to the tracking error on the squared market return. Using the market factor as the sole factor for linear returns, the pricing kernel in our model reduces to that in Harvey and Siddique (2000) albeit allowing for a non-zero price of risk on the squared market return, like in Chabi-Yo, Leisen, and Renault (2014).

More generally, our extension of the APT to squared excess returns may be seen in line with the literature on the APT with (possible) misspecification in the underlying factor models. As shown by Uppal, Zaffaroni, and Zviadadze (2018), the APT may be misspecified not only due to omission of pervasive factors, but also through non-zero asset-specific pricing errors. While they put forward an extension of the APT where at least one of the eigenvalues of the idiosyncratic variance matrix is unbounded (which allows for asset-specific pricing errors), we restore the standard APT pricing formulas by introducing a second-order approximate factor structure in the squared excess returns. However, we acknowledge that this strategy would lead to a rather *ad hoc* hierarchy of factor structures if we had to price any nonlinear payoff. An alternative is to acknowledge the impossibility of perfect hedging of any nonlinear payoff and to rather look for APT-consistent minimum

4 Technically we consider squared excess market return risk.

5 See, for example, Carr and Lee (2009) for an overview and Daigler and Rossi (2006); Dash and Moran (2005); and Signori, Briere, and Burgues (2010) for portfolio problems adding volatility exposure as a separate asset class. Trading in (derivatives on) the VIX index has grown exponentially over the past 10 years.

dispersion smart stochastic discount factors following Korsaye, Quaini, and Trojani (2020).

Compared with the existing literature which relies on multi-step estimation procedures for prices of risk, our framework allows us to adopt a single-step GMM procedure to estimate all prices of risk and factor loadings jointly. We use the Bakshi and Madan (2000) methodology, motivated by results in Breeden and Litzenberger (1978), to compute the price of squared excess returns from option prices. The requirement that option prices be available limits our sample to the set of optionable stocks over the period 1996–2013. As test assets we use both Fama–French industry portfolios and individual optionable stocks. We find evidence for a common factor in idiosyncratic volatilities when using either Fama–French factors or principal components as risk factors. Despite our relative short sample period, we find that this idiosyncratic variance factor is priced for squared returns, in addition to the market variance risk factor. Both carry a negative unconditional price of risk, but for the average portfolio in our sample, the magnitude of the price of market variance risk (45 bps/month) is an order of magnitude larger than that of the idiosyncratic variance factor (5 bps/month).

This result should not be confused with extant literature on the so-called IV puzzle. As stressed by Lehmann (1990), in any beta pricing model, "residual variances should in part reflect squared loadings on omitted risk factors, and, hence, should be associated with a significant risk premium in large samples." However, while Lehman (1990) predicts that stocks with higher residual variation should earn higher expected returns, a first puzzle has been pointed out by Ang, Chen, and Xing (2006a) that documents underperformance of stocks with high IV. Chen and Petkova (2012) attribute the low return of high IV stocks to the fact that they serve as a hedge to increases in market variance. In a thorough empirical analysis, Duarte et al. (2014) confirms the Chen and Petkova's (2012) interpretation while, in accordance with Avramov et al. (2013), they also note that it is related to default spreads and, thus, to financial distress. The new incomplete markets interpretation recently proposed by Herskovic et al. (2016), where the IV factor related to income risk faced by households is arguably similar in spirit.

We do not pretend to bring added value to the above IV puzzle, namely the occurrence of the IV factor priced in the return themselves. Confirming Lehman's (1990) observation that there is a long history of "the inability to obtain statistically reliable estimates of a linear residual risk effect," we actually do not find evidence of the IV factor being a missing priced factor in the linear return model. The risk premium for this factor, which we denote by λ_G below, is not found to be significantly different from zero for any of the sets of test assets we analyze. Again following Lehman's (1990), another possible explanation is "the information loss typically incurred by grouping securities into a smaller number of portfolios." As far as sample size is concerned, we are limited by the set of options available in the OptionMetrics Volatility Surface data which are used to compute the price of the squared excess return.

The pricing kernel induced by the model, evaluated as a function of the market factor/first principal component while fixing the other factors at their conditional means, takes a conventional convex shape when principal components are used as factors. On the other hand, when using Fama and French (1993) factors one obtains a concavely shaped pricing kernel which is inconsistent with risk aversion. We identify this pricing kernel without assumptions on the agent's preferences.

Our paper is related to the literature on the factor structure in option prices, for example, Kadan, Liu, and Tang (2017) and Christoffersen, Fournier, and Jacobs (2018). These papers study a direct factor structure in option prices/returns, while we follow a different route. We study a factor structure in *squared* excess returns and use options to price these squared returns. The advantage of this approach is that it is more strongly founded in APT, using our main result Corollary 1.

1 First and Second-Order Approximate Factor Structures

1.1 General Framework

Let $(\Omega, \mathcal{F}, P)$ be the probability space on which all random variables and processes below are defined. The financial market is modeled in terms of a collection of random variables indexed by $U = [0, 1]$. In particular, like Al-Najjar (1998) and Gagliardini, Ossola, and Scaillet (2016), we consider sequences of stochastic processes of excess returns[6] $R^{(n)} = \{R^{(n)}(u); u \in U\}$. Formally, for each $n = 1, 2, \ldots,$ $R^{(n)}$ is a stochastic process in $C_0(U)$, the space of continuous functions on U vanishing at zero, that is, $R^{(n)}(0) \equiv 0$. The continuity assumption on the paths of $R^{(n)}$ does not imply any level of dependence between individual returns. We just see $R^{(n)}(u)$ as the excess return on a *portfolio* of the first u-fraction of assets in the economy. In particular, $R^{(n)}(1)$ represents the (equally weighted) market portfolio. The simplest setting of a market with (cross-sectional) i.i.d. excess returns $R^{(n)}$ can be thought of as a linearly interpolated random walk (converging to a Brownian motion as $n \to \infty$). To formalize this, we first discuss the notion of convergence we use in more detail.

We consider a weak form of convergence for stochastic processes on $C_0(U)$. More precisely, we say that a sequence $Z^{(n)}$ of stochastic processes on $C_0(U)$ converges to zero if, for all finite-variation functions h on U, we have

$$\int_0^1 Z^{(n)}(u) \mathrm{d}h(u) \xrightarrow{p} 0. \tag{1}$$

We denote this convergence as $Z^{(n)} \xrightarrow{w^*} 0$. Note that, precisely because $Z^{(n)}$ is continuous and h of finite-variation, the above integral can be defined path-by-path as a Riemann–Stieltjes integral. By partial integration, we may also write, for any $R^{(n)} \in C_0(U)$,

$$\int_0^1 h(u) \mathrm{d}R^{(n)}(u) = h(1) R^{(n)}(1) - \int_0^1 R^{(n)}(u) \mathrm{d}h(u). \tag{2}$$

Integrals of the form $\int_0^1 h(u) \mathrm{d}R^{(n)}(u)$ will be considered below since they can be interpreted as the return on portfolio h. For example, the equally weighted market portfolio corresponds to $h \equiv 1$. Finally, observe that $R^{(n)} \xrightarrow{w^*} 0$ implies $R^{(n)}(u_0) \xrightarrow{p} 0$ for all $u_0 \in U$ (take $h(u) = 1\{u < u_0\}$). This shows that the convergence $\xrightarrow{w^*}$ we adopt is stronger than point-wise convergence in probability. At the same time, it is weaker than uniform (over U) convergence in probability.

With the above, our notion of a (first-order) approximate factor structure for $R^{(n)}$ is formalized in the following definition.

6 We use the term excess return for any asset that has zero price.

Definition 1: *The sequence of excess return processes $R^{(n)} \in C_0(U)$ satisfies a first-order approximate factor structure if there exists a K-dimensional random factor F and deterministic finite-variation function α_R and β_R such that we may write, for all $u \in U$,*

$$R^{(n)}(u) = \int_0^u \alpha_R(v)\mathrm{d}v + \int_0^u \beta_R(v)^\top F\mathrm{d}v + Z_R^{(n)}(u), \tag{3}$$

with $Z_R^{(n)} \xrightarrow{w^*} 0$ as $n \to \infty$.

Comparing our mathematical setup to Gagliardini, Ossola, and Scaillet (2016), there are three differences. Firstly, we interpret each $R^{(n)}(u)$ as the excess return on a portfolio of the first u fraction in the collection of primitive assets. Thus, when Gagliardini, Ossola, and Scaillet (2016) write a linear factor model (see their definition APR1), this should be seen as an assumption on the infinitesimal increment $\mathrm{d}R^{(n)}(u)$ in our setup. In other words, the bounds maintained on the cross-sectional dependence in an approximate factor structure (condition APR3 in Gagliardini, Ossola, and Scaillet 2016) must be defined between properly scaled increments $R^{(n)}(k\Delta_n) - R^{(n)}((k-1)\Delta_n)$, with $\Delta_n \to 0$ and for $k = 1, 2, \ldots$. We will see below why our formulation in terms of cumulative returns $R^{(n)}(u) = \int_0^u \mathrm{d}R^{(n)}(v)$ is well-suited for the definition of second-order factor structures. Also note that the measurability condition APR.2 in Gagliardini, Ossola, and Scaillet (2016) on $\mathrm{d}R^{(n)}$ implies that $R^{(n)} \in C_0(U)$ as integrals of Lebesgue measurable functions are continuous (actually, it even implies $R^{(n)} \in C_1(U)$ introduced below). Secondly, our setup does not require any moment conditions on the excess returns as our formalization merely builds on convergence in probability. As such, Equation (3) does not necessarily specify a regression equation and there is no moment-based notion of orthogonality between F and $Z_R^{(n)}$.[7] Thirdly, to be equipped with an approximate decomposition of risk in terms of a finite number of factors, we will adopt a logic germane to Al-Najjar (1995) to see all asset returns as well-approximated by a "finitely generated process," that is, a process linearly spanned by a finite number of factors; see also Section 1.2.

The classical arbitrage pricing result now follows directly.

Theorem 1: *Assume that the excess return process $R^{(n)}$ satisfies an approximate factor structure as in Definition 1. Furthermore, assume that there are no arbitrage opportunities in the sense that it is not possible to construct a portfolio h whose excess return (2) converges, in probability and as $n \to \infty$, to a non-zero constant. Then, there exists a K-dimensional vector with prices of risk π such that*

$$\alpha_R(u) = -\beta_R(u)^\top \pi, \tag{4}$$

for almost every $u \in U$.

Proof of Theorem 1: *Let h denote any portfolio without exposure to the factors, that is, h is a finite-variation function with $\int_0^1 h(u)\beta(u)\mathrm{d}u = 0$. Then, from Equation (3), the induced portfolio returns are*

$$\int_0^1 h(u)\mathrm{d}R^{(n)}(u) = \int_0^1 h(u)\alpha_R(u)\mathrm{d}u + \int_0^1 h(u)\mathrm{d}Z_R^{(n)}(u). \tag{5}$$

7 While Definition 1 is sufficient to derive the APT, an empirical analysis will rely on a regression interpretation of Equation (3) in order to identify α_R and β_R statistically; see Section 3.

In view of Equation (2), we thus find

$$\int_0^1 h(u)\mathrm{d}R^{(n)}(u) \xrightarrow{p} \int_0^1 h(u)\alpha_R(u)\mathrm{d}u. \tag{6}$$

In the absence of arbitrage, we must have $\int_0^1 h(u)\alpha_R(u)\mathrm{d}u = 0$. As this must hold for any finite-variation function h orthogonal to all components of β_R (and α_R and β_R are of finite variation themselves), we have (up to a set of Lebesgue measure zero) $\alpha_R(u) = -\beta_R(u)^\top \pi$ for some vector π. □

In order to analyze squared (excess) returns, we denote by $C_1(U)$ the linear subspace of $C_0(U)$ of those functions that are absolutely continuous, that is, for $R^{(n)} \in C_1(U)$ its derivative $\nabla R^{(n)}(u)$ exists a.e. For excess returns processes in $C_1(U)$, we define the so-called excess squared excess return[8] process $S^{(n)}$ by

$$S^{(n)}(u) = \int_0^u \left(\left[\nabla R^{(n)}(v)\right]^2 - p^{(n)}(v) \right)\mathrm{d}v, \tag{7}$$

where $p^{(n)}$ denotes the price of squared returns, that is, $S^{(n)} \in C_1(U)$ is indeed an *excess* return and has itself zero price. In our empirical analysis, we will use plain vanilla options traded on individual assets to reconstruct the no-arbitrage price $p^{(n)}$ of squared excess returns using a well-known technique going back to Breeden and Litzenberger (1978) and Bakshi and Madan (2000), see Section 3.1.1 for details.

We impose throughout

$$\int_0^u p^{(n)}(v)\mathrm{d}v \xrightarrow{w^*} \int_0^u p(v)\mathrm{d}v, \tag{8}$$

as $n \to \infty$, for some integrable function p. A sufficient condition for Equation (8) is that the $p^{(n)}$ are uniformly bounded in n together with the point-wise convergence of $p^{(n)}$ to p. Indeed, in that case, we have for any finite-variation function h (using the dominated convergence theorem)

$$\begin{aligned} \textstyle\int_0^1 h(u)\mathrm{d}\int_0^u p^{(n)}(v)\mathrm{d}v &= \textstyle\int_0^1 h(u)p^{(n)}(u)\mathrm{d}u \\ &\to \textstyle\int_0^1 h(u)p(u)\mathrm{d}u. \end{aligned} \tag{9}$$

This brings us to our definition of a second-order factor structure.

Definition 2: *The sequence of excess return processes $R^{(n)} \in C_1(U)$ satisfies a second-order approximate factor structure if, in addition to the conditions in Definition 1, we have, as $n \to \infty$,*

$$\int_0^u \left[\nabla Z_R^{(n)}(v)\right]^2 \mathrm{d}v \xrightarrow{w^*} \int_0^u \omega_S(v)\mathrm{d}v + \int_0^u \varphi_S(v)^\top G\mathrm{d}v, \tag{10}$$

for deterministic finite-variation functions ω_S and φ_S and a K_S-dimensional random factor G.

8 The precise term for $S^{(n)}$ is indeed "excess squared excess return." When there is no risk of confusion, we also simply refer to it as "squared return."

A key insight of this paper is that a second-order approximate factor structure for excess returns $R^{(n)}$ gives rise to a first-order factor structure for the excess squared excess returns $S^{(n)}$. The factors appearing in this excess squared excess return process intuitively include all the quadratic functions of the initial factors F. At the same time, they contain the new factors G induced by Equation (10). This is formalized in the following proposition. All proofs are provided in Appendix A.

Proposition 1: *If the sequence of excess return processes $R^{(n)}$ satisfies a second-order approximate factor structure, then the squared excess return process $S^{(n)}$ defined in Equation (7) satisfies the first-order approximate factor structure*

$$\begin{aligned} S^{(n)}(u) = {} & \int_0^u [\alpha_R(v) + \beta_R^\top(v)F]^2 \mathrm{d}v \\ & + \int_0^u [\omega_S(v) - p(v)]\mathrm{d}v + \int_0^u \varphi_S^\top(v) G \mathrm{d}v + Z_S^{(n)}(u), \end{aligned} \tag{11}$$

with $Z_S^{(n)} \overset{w^*}{\to} 0$ as $n \to \infty$.

The key message of Proposition 1 is that the initial second-order approximate factor structure for the excess returns $R^{(n)}$ induces a first-order approximate factor structure for squared returns. Combined with Theorem 1, we immediately find our pricing result for squared returns.

Corollary 1: *Under the conditions of Proposition 1 and the no-arbitrage assumption in Theorem 1, there exists a K-dimensional vector with prices of risk π for F, a $K \times K$-dimensional (symmetric) matrix with prices of risk δ for$FF^\top$, and a K_S-dimensional vector with prices of risk η for G such that, for almost every $u \in U$,*

$$\alpha_R(u) = -\beta_R(u)^\top \pi, \tag{12}$$

$$\omega_S(u) - p(u) = -\beta_R(u)^\top \delta \beta_R(u) - \varphi_S(u)^\top \eta. \tag{13}$$

The above result precisely identifies the consequence of the no-arbitrage condition for the prices of squared returns and, thereby, for the prices of common factors in (idiosyncratic) variances. The first term in Equation (13) gives the effect of the linear return factors F on the prices of squared returns. It's intuitively clear that this effect exists, but the present paper seems to be the first to make this precise. Alternatively stated, the first term in Equation (13) also gives the consequences for pricing "idiosyncratic" variances in case some factors have been omitted in the linear return factor model. In case of omitted linear return factors, the term "idiosyncratic" is a misnomer. This means that existing results in the literature on common volatility factors must always be discussed relative to the linear return factors they take into account (be it PCA or Fama–French type factors). Also observe that the price of risk for squared (excess) returns related to the squares of the factor loadings $\beta_R(u)$ are given by a parameter δ that is unrelated to the prices of risk at the linear return factor model π. In Section 3, we will relate δ to the price of variance risk, in particular the price of market variance risk for the first element of δ.

Assuming that no factors for the linear returns have been omitted, the second term in Equation (13), $\varphi_S(u)^\top \eta$, gives the pricing effect of common factors G in idiosyncratic variances. Quadratic returns command a risk premium from exposure to the common idiosyncratic variance factor G. This risk premium is, as in the standard APT, linear in the exposure of the individual squared return to the common idiosyncratic variance factor, that is, linear in φ_S. Note that the idiosyncratic variance factor G may be correlated with the return factors F or their squares $FF^\top$. The no-arbitrage condition does neither impose nor exclude this.[9]

The analysis so far considered a continuum of assets indexed by $u \in U$. Historically, however, the APT literature emerged from considering (large) countable sets of individual asset returns. Therefore, we discuss in Section 1.2 how countable sets of (linear) returns approximate a continuum of assets. Section 1.3 studies the same question for squared excess returns. Together with Corollary 1 and the β-pricing formulation to be discussed in Section 2.5, these results pave the way to study the price of idiosyncratic variance risk in Section 3.

1.2 Approximation by Countable Sets of (Linear) Returns

Ross (1976)'s initial formalization of the APT was based on a countable set of asset returns. Even though Al-Najjar (1998) has put forward compelling arguments to work with a continuum of assets (as we do), some justification through a characterization by "finitely generated processes" (Al-Najjar 1995) is worth considering for both a better understanding and also for the purpose of econometric inference. We describe such a characterization, first for linear portfolios based on the excess return process $R^{(n)}$ and, second, in Section 1.3, for the (excess) squared (excess) return process $S^{(n)}$. It must be kept in mind that this finite-dimensional characterization, albeit useful for interpretation and econometric inference, is not necessary for our development of the APT above.

A sequence of excess returns $R_i^{(n)}$, $i = 1, \ldots, n$, can be embedded in a piecewise linear excess return process $R^{(n)}$. The following definition introduces the notion of a *simple excess return process*.

Definition 3: *The excess return process $R^{(n)} \in C_1(U)$ is called simple if there exists intervals*

$$U_i^{(n)} = (\Delta_{i-1}^{(n)}, \Delta_i^{(n)}], \quad i = 1, \ldots, n \tag{14}$$

that form a finite partition of U, such that $R^{(n)}$ is piecewise linear on each $U_i^{(n)}$, $i = 1, \ldots, n$.

Definition 3 is our equivalent of the notion of a "simple process" as defined in Section 1.6 in Al-Najjar (1995). A simple excess return process $R^{(n)}$ has sample paths in $C_1(U)$ and $\nabla R^{(n)}$ is constant on each interval $U_i^{(n)}$. Therefore, we can embed the sequence of excess returns $R_i^{(n)}$, $i = 1, \ldots, n$, in a sequence of simple excess return processes $R^{(n)}$ via

$$\nabla R^{(n)}(u) \equiv R_i^{(n)} \ \text{ for } \ u \in U_i^{(n)}. \tag{15}$$

9 In empirical work, the factors F, $FF^\top$, and G should be linearly independent to uniquely identify all prices of risk. Moreover, identification of the factors requires the factor loadings β_R, $\beta_R\beta_R^\top$, and φ_S to be linearly independent; see Section 2.3 for details.

We introduce $\Delta_i^{(n)}$ in Equation (14) for ease of exposition. The notation suggests a prescribed ordering of assets, however, this is not required for the remainder of our results. Ordering of assets will be discussed in more detail in Section 2.2.[10]

In the spirit of Al-Najjar (1995), we see simple excess return processes as an approximation of the financial market with a continuum of assets as discussed in Section 1. Such approximation is precisely what will be employed in empirical analyses, that is, also in our Section 3. Thus, for a sequence of excess returns $R_i^{(n)}$, $i = 1, \dots, n$, and a partition $U_1^{(n)}, \dots, U_n^{(n)}$ of U as in Equation (14), we define the simple excess return process $R_\Delta^{(n)}$ by integrating Equation (15), that is,

$$R_\Delta^{(n)}(u) = \int_0^u \sum_{i=1}^n R_i^{(n)} 1_{\left\{v \in U_i^{(n)}\right\}} \mathrm{d}v. \tag{16}$$

Standard APT theory usually starts from more primitive (and testable) assumptions with respect to (first-order) factor structures imposed on the sequence of excess returns $R_i^{(n)}$, $i = 1, \dots, n$. Generally, a factor decomposition is considered, that is,

$$R_i^{(n)} = \alpha_i^{(n)} + \beta_i^{(n)\top} F + \varepsilon_i^{(n)}, \quad i = 1, \dots, n, \tag{17}$$

with $(\alpha_i^{(n)})_{1 \le i \le n}$ and $(\beta_i^{(n)})_{1 \le i \le n}$ deterministic sequences and $E\{\varepsilon_i^{(n)}\} = 0$. Note that, for instance, Gagliardini et al. (2016) assumes that Equation (17) is a regression equation, that is, $\varepsilon_i^{(n)}$ is uncorrelated with F. While this is needed for statistical identification, it is immaterial to derive the APT and we did not use this assumption in Theorem 1. This is similar to Gagliardini et al. (2016), who do not use that assumption either in proving their Proposition 1.

By plugging Equation (17) into Equation (16), we obtain

$$R_\Delta^{(n)}(u) = \sum_{i=1}^n (\alpha_i^{(n)} + \beta_i^{(n)\top} F) \int_0^u 1_{\{v \in U_i^{(n)}\}} \mathrm{d}v + \sum_{i=1}^n \varepsilon_i^{(n)} \int_0^u 1_{\{v \in U_i^{(n)}\}} \mathrm{d}v, \tag{18}$$

where $\int_0^u 1_{\{v \in U_i^{(n)}\}} \mathrm{d}v$ is linear on $U_i^{(n)}$.

The following proposition gives sufficient conditions on the regression coefficients and residuals in Equation (17) such that the simple excess return process $R_\Delta^{(n)}$ satisfies a first-order approximate factor structure in the sense of Definition 1.

Proposition 2: *The simple excess return process $R_\Delta^{(n)}$ based on Equation (17) satisfies a first-order approximate factor structure in case there exist finite-variation functions α_R and β_R such that*

$$\sum_{i=1}^n \alpha_i^{(n)} \int_0^u 1_{\{v \in U_i^{(n)}\}} \mathrm{d}v \xrightarrow{w^*} \int_0^u \alpha_R(v) \mathrm{d}v, \tag{19}$$

$$\sum_{i=1}^n \beta_i^{(n)} \int_0^u 1_{\{v \in U_i^{(n)}\}} \mathrm{d}v \xrightarrow{w^*} \int_0^u \beta_R(v) \mathrm{d}v, \tag{20}$$

10 Al-Najjar (1995) only requires simple processes to be constant on a partition of measurable sets, not necessarily intervals. We do not consider that level of generality for ease of interpretation.

and the idiosyncratic excess returns $\varepsilon_i^{(n)}$ *satisfy a functional law of large numbers such that* $Z_\Delta^{(n)}$ *defined by*

$$Z_\Delta^{(n)}(u) = \sum_{i=1}^{n} \varepsilon_i^{(n)} \int_0^u 1_{\{v \in U_i^{(n)}\}} \mathrm{d}v, \tag{21}$$

converges to zero in $\xrightarrow{w^*}$.

For the commonly analyzed case of equally weighted returns, that is, $U_i^{(n)} = ((i-1)/n, i/n]$, we find the conditions

$$\frac{1}{n}\sum_{i=1}^{\lfloor un \rfloor} \alpha_i^{(n)} \xrightarrow{w^*} \int_0^u \alpha_R(v)\mathrm{d}v, \tag{22}$$

$$\frac{1}{n}\sum_{i=1}^{\lfloor un \rfloor} \beta_i^{(n)} \xrightarrow{w^*} \int_0^u \beta_R(v)\mathrm{d}v, \tag{23}$$

as $n \to \infty$. Note that these convergences are deterministic and point-wise convergence is again sufficient provided that $\alpha_i^{(n)}$ and $\beta_i^{(n)}$ are bounded.

For the idiosyncratic errors $Z_\Delta^{(n)}$ defined in Equation (21), the equally weighted portfolio condition implies that, for any finite-variation function h, we have

$$\int_0^1 h(u)\mathrm{d}Z_\Delta^{(n)}(u) = \sum_{i=1}^{n} \varepsilon_i^{(n)} \int_{U_i^{(n)}} h(u)\mathrm{d}u. \tag{24}$$

While we use convergence in probability, seminal papers as Al-Najjar (1995), Al-Najjar (1998), and Gagliardini, Ossola, and Scaillet (2016) consider convergence in L_2, the space of square-integrable random variables.

In this setup, sufficient conditions for $Z_\Delta^{(n)}$ in Equation (21) to converge weakly to zero are that $\varepsilon_i^{(n)}$ has mean zero and finite variance, and, as $n \to \infty$,

$$\frac{\rho\left(\mathrm{Var}\left\{\varepsilon_i^{(n)}\right\}\right)}{n} \to 0, \tag{25}$$

$$n\left(\int_{U_i^{(n)}} h(u)\mathrm{d}u\right)^\top \left(\int_{U_i^{(n)}} h(u)\mathrm{d}u\right) = O(1), \tag{26}$$

where ρ denotes the spectral radius of a symmetric matrix, that is, its largest absolute eigenvalue, implying that Equation (25) is essentially Assumption APR.3 in Gagliardini, Ossola, and Scaillet (2016).

If we consider, without loss of generality, a portfolio bounded by 1, that is, $|h(u)| \leq 1$, Condition (26) can be rewritten as

$$\begin{aligned} n\left(\int_{U_i^{(n)}} h(u)\mathrm{d}u\right)^\top \left(\int_{U_i^{(n)}} h(u)\mathrm{d}u\right) &= n\sum_{i=1}^{n}\left(\int_{U_i^{(n)}} h(u)\mathrm{d}u\right)^2 \\ &\leq n\sum_{i=1}^{n}\left(\Delta_i^{(n)} - \Delta_{i-1}^{(n)}\right)^2 \\ &= O(1). \end{aligned} \tag{27}$$

Note that, in another context, Mykland and Zhang (2006) dubbed the limit of the latter sum the "Asymptotic Quadratic Variation in Time," illustrating an analogy between the mathematics of a continuum of assets and continuous time finance.

In view of the above, as far as a (linear) excess returns $R^{(n)}$ are considered, our maintained assumptions regarding the approximate factor structure, in particular the strength of correlations allowed between idiosyncratic errors $\varepsilon_i^{(n)}$, are implied by assumption APR.3 in Gagliardini, Ossola, and Scaillet (2016). Analogously to Equation (25), a sufficient condition would be that the covariance matrix of the idiosyncratic errors $\varepsilon_i^{(n)}$ has eigenvalues that are uniformly bounded (and not only $o(n)$). This assumption of uniformly bounded eigenvalues actually coincides with the definition of an approximate factor structure as maintained in Chamberlain and Rothschild (1983); see their Definition 2.

1.3 Portfolios of Squared Returns

In Section 3, we use the theory developed in this paper to calculate the market-consistent price of idiosyncratic variances based on (observed) prices of squared excess returns. To that extent, we consider in the present section the pricing of squared returns for simple excess return process.

Consider the simple excess return process $R_\Delta^{(n)}$ in Equation (16). The induced excess squared excess return process

$$\begin{aligned} S_\Delta^{(n)}(u) &= \int_0^u [\nabla R^{(n)}(v)]^2 - p^{(n)}(v)\mathrm{d}v \\ &= \int_0^u \sum_{i=1}^n [(R_i^{(n)})^2 - p_i^{(n)}] 1_{\{v \in U_i^{(n)}\}} \mathrm{d}v. \end{aligned} \tag{28}$$

is a simple excess return process as well where $p_i^{(n)}$ is the arbitrage-free price of $(R_i^{(n)})^2$. Note that we do need to assume the existence of an explicit liquid market for squared excess returns. In the empirical section, we restrict attention to optionable stocks. Then, using static portfolios of these options, we compute an arbitrage-free market price $p_i^{(n)}$ for the squared excess return $(R_i^{(n)})^2$, so we use the (excess) squared (excess) return

$$S_i^{(n)} = (R_i^{(n)})^2 - p_i^{(n)}. \tag{29}$$

Just as we motivated the first-order approximate factor structure (3) for a continuum of assets by the factor structure (17) for the sequence of returns $R_i^{(n)}$, $n = 1, 2, \ldots$, we justify the second-order factor structure in Definition 2 by a factor structure for the sequence of squared idiosyncratic returns $[\varepsilon_i^{(n)}]^2$, $n = 1, 2, \ldots$. More precisely, we define the errors $\nu_i^{(n)}$ relative to a common factor G by

$$[\varepsilon_i^{(n)}]^2 = \omega_i^{(n)} + \varphi_i^{(n)\top} G + \nu_i^{(n)}. \tag{30}$$

Similar to the discussion below Equation (17), we do not need, for pricing, the condition

$$E\{\nu_i^{(n)} | F, G\} = 0, \tag{31}$$

which, under $E\{\varepsilon_i^{(n)} | F, G\} = 0$, is equivalent to

$$\mathrm{Var}\{\varepsilon_i^{(n)} | F, G\} = \omega_i^{(n)} + \varphi_i^{(n)\top} G. \tag{32}$$

In empirical work, the regression condition (31) is usually imposed. In that case, the price of squared excess returns $[\varepsilon_i^{(n)}]^2$ can also be identified with the price of the idiosyncratic variances $\mathrm{Var}\{\varepsilon_i^{(n)} | F, G\}$, that is, the price of the variance factor G.

Now, parallel to Proposition 2, we have the following result.

Proposition 3: *In addition to the assumptions of Proposition 2, assume*

$$\sum_{i=1}^{n} \omega_i^{(n)} \int_0^u 1_{\{v \in U_i^{(n)}\}} \mathrm{d}v \xrightarrow{w^*} \int_0^u \omega_S(v)\mathrm{d}v, \tag{33}$$

$$\sum_{i=1}^{n} \varphi_i^{(n)} \int_0^u 1_{\{v \in U_i^{(n)}\}} \mathrm{d}v \xrightarrow{w^*} \int_0^u \varphi_S(v)\mathrm{d}v, \tag{34}$$

and that the errors $\nu_i^{(n)}$ satisfy the functional law of large numbers

$$\sum_{i=1}^{n} \nu_i^{(n)} \int_0^u 1_{\{v \in U_i^{(n)}\}} \mathrm{d}v \xrightarrow{w^*} 0. \tag{35}$$

Then the simple excess return process $R_\Delta^{(n)}$ based on Equation (17) satisfies a second-order approximate factor structure.

While we state our main theory in this section in terms of a continuum of assets, indexed by $u \in U$, Propositions 2 and 3 give sufficient conditions in terms of sequences of assets. Such conditions are more common in the APT literature and used in empirical applications. We will study the price of idiosyncratic variance risk in Section 3 after providing some comments on APT in general.

2 Some Comments on APT

The literature on APT is huge. In this section, we reconsider some often-discussed issues. The reader may, when desired, jump immediately to Section 3 where we study the empirical question of interest in this paper: are idiosyncratic variances priced?

2.1 Factor-Mimicking Portfolios

If the excess returns process $R^{(n)}$ satisfies a first-order approximate factor structure according to Definition 1, we can define a K-dimensional finite-variation function H on U such that

$$\int_0^1 H(u)\beta_R(u)^\top \mathrm{d}u = I_K, \tag{36}$$

the $K \times K$ identity matrix. This is possible as long as the components of β are linearly independent.[11,12] Then the K portfolios (excess returns) induced by H, that is,

$$\tilde{F} = \int_0^1 H(u)\mathrm{d}R^{(n)}(u), \tag{37}$$

can also be used as factors. To see this, note that Equation (3) implies

$$\tilde{F} = \int_0^1 H(u)\alpha_R(u)\mathrm{d}u + F + \int_0^1 H(u)\mathrm{d}Z_R^{(n)}(u).$$

Hence, we may write, again using Equation (3),

11 Formally, the K components of the function H are obtained by Gramm–Schmidt orthogonalization (and normalization) of the linearly independent components of the function β using the scalar product $\langle f, g\rangle = \int_0^1 f(u)g(u)\mathrm{d}u$.

12 Section 2.3 shows that this assumption is not restrictive in general.

$$\begin{aligned}R^{(n)}(u) &= \int_0^u \alpha_R(v)\mathrm{d}v\\&+\int_0^u \beta_R(v)^\top[\tilde{F}-\int_0^1 \alpha_R(w)H(w)\mathrm{d}w-\int_0^1 H(w)\mathrm{d}Z^{(n)}(w)]\mathrm{d}v+Z_R^{(n)}(u)\\&=\int_0^u[\alpha_R(v)-\beta_R(v)^\top\int_0^1 \alpha_R(w)H(w)\mathrm{d}w]\mathrm{d}v+\int_0^u \beta_R(v)^\top\tilde{F}\mathrm{d}v\\&+[Z^{(n)}(u)-\int_0^u \beta_R(v)^\top\mathrm{d}v\int_0^1 H(w)\mathrm{d}Z^{(n)}(w)],\end{aligned}$$

where the last term again converges to zero. Observe that switching to factor mimicking portfolios in this way does not affect the factor loadings β_R, but the intercept α_R and the notion of idiosyncratic return are affected.

2.2 Repackaging

An important point in theoretical foundations of the APT is that its assumptions should be invariant under so-called "repackaging," see, for example, Al-Najjar (1999). Loosely speaking this means that the assumptions should be invariant with respect to reordering the assets and with respect to forming portfolios. Our Definition 1 indeed obeys to this invariance which can be seen as follows.

Firstly, consider a reordering of the assets. Note that Equation (3) implies, for given $w \in U$,

$$R^{(n)}(u)-R^{(n)}(w)=\int_w^u \alpha(v)\mathrm{d}v+\int_w^u \beta^\top(v)\mathrm{d}v+Z^{(n)}(u)-Z^{(n)}(w). \tag{38}$$

Now consider $p+1$ fixed constants $0=u_0<u_1<\cdots<u_p=1$. A reordering of assets can be obtained by permuting the p intervals $[u_{j-1},u_j]$, $j=1,\ldots,p$. Reordering the assets by pasting together the increments of the excess return processes $R^{(n)}$ over each of the permuted intervals satisfies the conditions of Definition 1 in case the original excess return processes $R^{(n)}$ do. Alternatively said, as discussed below Definition 3, we have never assumed any ordering of the intervals $U_i^{(n)}$ in the first place.

Secondly, consider forming portfolios of the available assets. This is formalized by a fixed finite-variation function h^* and by considering the excess return process $\int_0^u h^*(v)\mathrm{d}R^{(n)}(v)$. Such a process again satisfies Definition 1 as soon as $R^{(n)}$ does. Indeed, we have, in view of Equation (3),

$$\int_0^u h^*(v)\mathrm{d}R^{(n)}(v)=\int_0^u h^*(v)\alpha_R(v)\mathrm{d}v+\int_0^u h^*(v)\beta_R(v)^\top\mathrm{d}v+\int_0^u h^*(v)\mathrm{d}Z^{(n)}(v), \tag{39}$$

where the last term converges to zero. Moreover, $h^*\alpha_R$ and $h^*\beta_r$ are both of finite variation (as the product of finite variation functions). Consequently, $\int_0^u h^*(v)\mathrm{d}R^{(n)}(v)$ satisfies an approximate factor structure as well.

2.3 Factor Space Dimension and Omitted Factors

An important empirical question relates to the appropriate number of factors. We address this question here from a theoretical point of view. First note that, as usual, the relevant notion in Definition 1 is the space spanned by the (random) components of F and the constant. Thus, we can always specify a vector F of factors such that no linear combination of the components of F is deterministic, that is, such that, if it exists, the variance matrix of F is non-singular. However, it may be the case that a strictly smaller factor space is also valid. Indeed, assume that one of the components of the factor loadings β_R is a linear combination of the other components; for instance suppose that, for almost every $u \in U$, we have

$$\beta_{R1}(u) = \sum_{k=2}^{K} \zeta_k \beta_{Rk}(u).$$

In that case, we can also write down a $K-1$-dimensional factor structure using the factor $\tilde{F}$ defined by

$$\tilde{F} = [F_k + \zeta_k F_1]_{k=2}^{k=K}.$$

Therefore, one usually maintains the assumption that no linear combination of the K components of F is deterministic and that no linear combination of the K components of β_R is zero. Under this maintained assumption, it is not possible to write an approximate factor structure with less than K factors as we now show.

Consider the situation of possibly omitted factors. Suppose that the excess return process satisfies an approximate factor structure with factors (F, F_o), that is,

$$R^{(n)}(u) = \int_0^u \alpha_R(v)\mathrm{d}v + \int_0^u \beta_R(v)^\top F \mathrm{d}v + \int_0^u \beta_{Ro}(v)^\top F_o \mathrm{d}v + Z_R^{(n)}(u), \tag{40}$$

where $Z_r^{(n)} \xrightarrow{w^*} 0$. Assume now that the researcher omits the factors F_o from the analysis. This researcher effectively considers the "idiosyncratic" errors $\int_0^u \beta_{Ro}(v)^\top F_o \mathrm{d}v + Z_R^{(n)}(u)$. This will only converge to zero if $\beta_{Ro} = 0$. Consequently, Definition 1 precisely identifies the correct number of factors. The identification of these factors is the content of the next section.

2.4 Factor Identification

A subtle and sometimes overlooked point refers to the regression interpretation of a factor model as assumed in Definition 1. Note that this definition does not impose orthogonality (in L_2 sense) of F and $Z^{(n)}$, but merely that the process $Z^{(n)}$ vanishes asymptotically. Nowhere in laying out the framework do we assume the existence of any moments for $R^{(n)}$, F, or $Z_R^{(n)}$. Similarly to the discussion following Proposition 2, one may formulate sufficient conditions in terms of L_2 convergence for our results to hold.

However, it is a standard practice in the empirical literature to assume the idiosyncratic errors $\varepsilon_i^{(n)}$ to be uncorrelated with the factors for each individual asset in Equation (17). We also impose this identifying condition in Section 3. This orthogonality, in itself, does not imply that errors have zero price. In general, this empirical approach has been criticized, going as far as questioning the testability of Ross (1976)'s APT. This criticism motivated the approach of characterizing the APT in an economy with a continuum of assets, as in Al-Najjar (1998) and Gagliardini, Ossola, and Scaillet (2016). Our paper addresses the issue by conveniently considering cumulative portfolios of assets.

Besides its conceptual relevance, the above remarks also concern estimation of factor models using time-series regressions. In this paper, we focus on a single-period formulation, but the APT framework developed here can readily be extended to multiple periods, assuming that each period the no-arbitrage assumption holds and that trading is possible. Once the factor loadings β_R have been identified, Theorem 1 can be applied period-by-period, where a "non-zero constant payoff" is to be understood as a deterministic function of past conditioning information. Both factor loadings β_R and prices of risk π may become stochastically, but predictably, time-varying in that case. This is in contrast with multi-period *equilibrium* models which generally feature hedge demands as well.

2.5 Beta Pricing

The result in Corollary 1 can also be written in the form of a beta-pricing relation for squared excess returns, much akin to the standard beta-pricing relation for (linear) returns.

Using Equation (17) and the pricing result (12), we find the beta pricing relation

$$R_i^{(n)} = \beta_i^{(n)\top}(F-\pi) + \varepsilon_i^{(n)}. \tag{41}$$

Combining with Equations (29), (30), and (13), the result (41) in turn implies

$$\begin{aligned}
(R_i^{(n)})^2 - p_i^{(n)} &= (\beta_i^{(n)\top}(F-\pi))^2 + (\varepsilon_i^{(n)})^2 + 2\beta_i^{(n)\top}(F-\pi)\varepsilon_i^{(n)} - p_i^{(n)} \\
&= (\beta_i^{(n)\top}(F-\pi))^2 + \omega_i^{(n)} + \varphi_i^{(n)\top}G + \nu_i^{(n)} + 2\beta_i^{(n)\top}(F-\pi)\varepsilon_i^{(n)} - p_i^{(n)} \\
&= \beta_i^{(n)\top}[(F-\pi)(F-\pi)^\top - \delta]\beta_i^{(n)} + \varphi_i^{(n)\top}(G-\eta) + \nu_i^{(n)} + 2\beta_i^{(n)\top}(F-\pi)\varepsilon_i^{(n)} \\
&= \beta_i^{(n)\top}[(F-\pi)(F-\pi)^\top - \delta]\beta_i^{(n)} + \varphi_i^{(n)\top}(G-\eta) + \zeta_i^{(n)},
\end{aligned} \tag{42}$$

where $E\{\zeta_i^{(n)}|F,G\} = 0$ has zero expectation conditionally on F and G precisely under the regression conditions $E\{\varepsilon_i^{(n)}|F,G\} = 0$ and $E\{\nu_i^{(n)}|F,G\} = 0$. Thus, while our APT does not need any moment conditions, they are inherently needed when considering beta pricing relations. Also note that the above derivation assumes that the pricing relations in Corollary 1 are valid for each asset individually.

Equation (42) is a pricing equation for squared (excess) returns that identifies both the prices of risk of the squared factors $(F-\pi)(F-\pi)^\top$ and the price of risk of the common idiosyncratic variance factors G. Observe that the prices of risk δ of the squared factors $(F-\pi)(F-\pi)^\top$ are allowed to vary freely, separate of the prices of risk π of F. In particular, a factor F that is not priced at the linear return level (i.e., its corresponding price of risk π equals zero) may very well feature a non-zero price for its square (i.e., for its corresponding components in δ). The prices of risk π, δ, and η are thus, ultimately, to be determined empirically. This will be studied in Section 3, using the GMM approach detailed in Appendix B.

Note that Equation (42) provides a pricing equation for a non-linear transformation of returns (in this case their squares). Conceptually the same machinery may be used to find the (non-linear) factor structure implied by other transformations, for example, it may be applied to option prices leading to an induced (again, non-linear) factor model for the cross section of option prices. Both Kadan, Liu, and Tang (2017) and Christoffersen, Fournier, and Jacobs (2018) analyze this factor structure of option prices using alternative frameworks. We follow a route based on our APT for squared returns and use option prices to identify empirically the price of these squared returns.

2.6 Induced Pricing Kernel

A natural question at this point is what pricing kernels are compatible with the pricing Equations (12) and (13). As the market is incomplete, there exist infinitely many pricing kernels. In view of the analysis above, we consider pricing kernels that are quadratic in the factors F, and linear in G. More precisely, consider a candidate pricing kernel

$$M = a + (F-\pi)^\top b + (F-\pi)^\top c(F-\pi) + (G-\eta)^\top d, \tag{43}$$

with $a \in \mathbb{R}$, $b \in \mathbb{R}^K$, c a symmetric matrix in $\mathbb{R}^{K\times K}$, and $d \in \mathbb{R}^{K_S}$.

We now have the following result.

Proposition 4: *Under the additional assumptions*

$$\mathrm{Cov}\{\varepsilon_i^{(n)}, (F-\pi)(F-\pi)^\top\} = 0, \tag{44}$$

$$\mathrm{Cov}\{\varepsilon_i^{(n)}, G-\eta\} = 0, \tag{45}$$

$$\mathrm{Cov}\{\zeta_i^{(n)}, F-\pi\} = 0, \tag{46}$$

$$\mathrm{Cov}\{(F-\pi)(F-\pi)^\top, G-\eta\} = 0, \tag{47}$$

a valid pricing kernel that prices both linear and squared returns is obtained in Equation (43) for b, c, and d that solve the linear equations

$$\begin{aligned} 0 = \frac{EF-\pi}{R_f} &+ \mathrm{Var}\{F-\pi\}b \\ &+\mathrm{Cov}\Big\{F-\pi, (F-\pi)^\top c(F-\pi)\Big\} + \mathrm{Cov}\Big\{F-\pi, (G-\eta)^\top\Big\}d \end{aligned} \tag{48}$$

$$0 = \frac{E\{G-\eta\}}{R_f} + \mathrm{Cov}\Big\{(G-\eta), (F-\pi)^\top\Big\}b + \mathrm{Cov}\Big\{(G-\eta), (G-\eta)^\top\Big\}d, \tag{49}$$

$$\begin{aligned} 0 = \frac{E\Big[(F-\pi)(F-\pi)^\top - \delta\Big]}{R_f} &+ \mathrm{Cov}\Big\{\Big[(F-\pi)(F-\pi)^\top - \delta\Big], (F-\pi)^\top b\Big\} \\ &+\mathrm{Cov}\Big\{\Big[(F-\pi)(F-\pi)^\top - \delta\Big], (F-\pi)^\top c(F-\pi)\Big\}. \end{aligned} \tag{50}$$

The constant a in Equation (43) is furthermore identified by the relation$EM = 1/R_f$, where R_f denotes the risk-free rate.

Assumptions (44) and (45) essentially state that no factors have been omitted at the linear return level. Assumption (46) states that the residual squared returns and linear return factors are uncorrelated. Assumption (47) states that G is orthogonal to the induced common variance factors $(F-\pi)(F-\pi)^\top$. These assumptions are also consistent with the empirical analysis below. Proposition 4 provides a linear system of equations in the unknowns b, c, and d that describe the quadratic pricing kernel (43). The number of unknowns is $K + K(K+1)/2 + K_S$ for, respectively, b, c, and d. The number of equations also equals $K + K_S + K(K+1)/2$ and, thus, the system is exactly identified.

Proposition 4 has two important implications.

Corollary 2: *Let M^* denote the pricing kernel defined in Proposition 4. Then,*

1. *the coefficients a, b, c, and d in Equation (43) are uniquely determined by correctly pricing the risk-free rate and the factors, that is, by* $EM = 1/R_f$, $EM(F-\pi) = EM^*(F-\pi)$, $EM(F-\pi)(F-\pi)^\top = EM^*(F-\pi)(F-\pi)^\top$, *and* $EM(G-\eta) = EM^*(G-\eta)$;

2. *if N denotes a valid pricing kernel that is affine in the excess returns* $R_i^{(n)}$ *and the squared excess returns* $(R_i^{(n)})^2 - p_i^{(n)}$, *then*
   2a. *N assigns zero price to the idiosyncratic excess returns* $\varepsilon_i^{(n)}$ *and* $\zeta_i^{(n)}$;
   2b. *the projection of N on (the components of)* $F - \pi$, $(F - \pi)(F - \pi)^\top$, $G - \eta$, *and a constant equals* M^*.

Proof: The first statement follows from the imposed orthogonality of idiosyncratic excess returns and the factors. For the second statement, observe that N is an affine combination of (the components of) the factors and the idiosyncratic excess returns $\varepsilon_i^{(n)}$ and $\zeta_i^{(n)}$. As N and M^* give the same price to the factors, $N - M^*$ assigns price zero to $\varepsilon_i^{(n)}$ and $\zeta_i^{(n)}$. As M^* assign price zero to $\varepsilon_i^{(n)}$ and $\zeta_i^{(n)}$, so does N. Statement 2b follows from 2a by observing that $N - M^*$ assigns price zero to all factors. □

The pricing kernel (43) is not necessarily positive. Like the pricing kernels that are affine in the excess returns in linear factor models, M can be interpreted as the projection of any affine pricing kernel on the space spanned by the constant and, for all i, R_i, and R_i^2. The APT pricing Equations (41) and (42) imply that such projection is spanned by the constant and the factors F, $FF^\top$, and G. We will see later (Figures 5 and 4) that the empirical pricing kernel (43) takes a plausible form.

We note that a quadratic pricing kernel is tightly related to pricing skewness and in particular the so-called three-moment CAPM of Kraus and Litzenberger (1976). The key insight of the three-moment CAPM is that investors' preferences are not fully characterized by a trade-off between mean and variances, but that they also care about skewness in returns. As a result, they may optimally pick a portfolio that is not on the mean–variance frontier. One then resorts to a three-fund separation theorem that, besides the risk-free investment and the market portfolio, contains what Chabi-Yo, Leisen, and Renault (2014) have dubbed the skewness portfolio. This skewness portfolio is the projection of the squared market return on the space of linear payoffs. Since the squared market return cannot be perfectly hedged by linear instruments, it forms a source of incompleteness that leads, depending on the point of view, either to APT pricing errors controlled by smart SDFs or to additional common factors for squared returns. The latter are then not only squared values of common factors for linear pricing but also possibly contain common factors for idiosyncratic risk.

3 The Pricing of Idiosyncratic Variance

In this section, we examine the idiosyncratic variance factor within the APT model developed in Section 1. We use Fama–French industry portfolios and individual S&P500 stocks as test assets. Our focus is on squared excess returns which are not directly traded, but for which we can compute their prices using traded options. This restriction limits our sample to the set of optionable stocks. We present evidence in support of a priced idiosyncratic variance factor in addition to a priced market variance factor in squared returns. These results confirm recent results about the price of (market) variance risk in, for example, Carr and Wu (2009) and Bollerslev, Tauchen, and Zhou (2009). Our results contrast with recent research on the price of idiosyncratic variance risk in Chen and Petkova (2012), Duarte et al. (2014), and Herskovic et al. (2016): while they argue that idiosyncratic variance factor is priced in linear returns, potentially reflecting a missing risk factor in the model for linear returns, we find that idiosyncratic variance risk is only priced for squared returns. In economic terms, the risk premium for idiosyncratic variance risk is low at about 5 bps/month for the average test portfolio of squared returns in our sample, compared with about 45 bps/month for market variance risk.[13]

3.1 Sample Description

Our sample spans the period from January 1996 to December 2013, for all optionable stocks (5,064 unique firms), obtained by merging CRSP with OptionMetrics.[14] From this set of stocks, we construct test assets by sorting stocks according to the Fama–French 49 industry classifications at the end of each month. Requiring that the portfolio has a complete time series of monthly returns and at least five stocks in the portfolio at any point in time, we end up with 34 industry portfolios as test assets (henceforth, referred to as FF34). Later, we will also use the 100 individual S&P500 stocks with a full return history over the sample period as test assets. On the last trading day of each month, we measure value-weighted holding period returns in excess of the risk-free rate over the next 30 calendar days. This period matches the maturity of the options in the OptionMetrics Volatility Surface data which we use to compute the price of the squared excess return. The risk-free rate is taken from the OptionMetrics Zero Coupon Yield Curve data set and linearly interpolated to match the 30-day maturity.

We extract principal components from the subset of 291 firms in the intersection of the CRSP and OptionMetrics data sets with a full return history over the 216 months between 1996 and 2013.[15] The first principal component is closely aligned with the Fama–French market factor; the correlation between the two factors equals 0.972. The proportion of the variance of the matrix of returns explained by the first four principal components equals 29.6%, 10.6%, 5.2%, and 4.2%, respectively. For the models using the Fama–French factors, we construct 30-day factor returns by aggregating the daily factor returns available on Ken French's website. Throughout this section, an excess return should be read as any payoff with zero price: the return minus the risk-free rate for linear returns or minus its option-implied price for squared excess returns.

To motivate a model with an idiosyncratic/residual variance factor, we analyze the correlation structure of the residual returns of the 291 stocks with a full return history over the sample period.[16] Figure 1(a) displays the density of pairwise correlations of residual returns. Figure 1(b) does the same for the squared residual returns. In both plots, the density is almost identical for the three different factor models. This is quite surprising given the differences in methodology and objective.[17] On average, residual returns are uncorrelated

13 The average excess squared excess return on the test asset portfolios is −34 bps/month.

14 A common objection in the literature on (the pricing of) idiosyncratic variance risk is that it is driven by small firms, see for example, Bali and Cakici (2008). Since optionable stocks tend to be larger on average, our sample biases against this criticism.

15 The principal components are extracted using a singular value decomposition directly on the panel of 30-day returns in excess of the risk free rate and each component is scaled to have unit variance. We use the uncentered principal components, which can be interpreted as (scaled) excess returns.

16 Residual returns are specific to each factor specification.

17 The PCA factors can be viewed as fixed portfolios of the set of 291 stocks used to extract them, while the Fama–French factors are annually rebalanced portfolios. The objective of PCA is to

although there is a small tail of substantial pairwise correlations. Compared with the correlations of the residuals, substantially positive pairwise correlations are considerably more prevalent for the squared residuals and motivates the construction of the residual/idiosyncratic variance factor G to examine whether these pairwise correlations have a common driver that may be priced. We will analyze the pricing of G by studying squared returns following the results in Section 1. To do so, we need to construct the panel of prices of squared returns as well.

3.1.1 The price of squared excess returns

The OptionMetrics Volatility Surface data contain smoothed implied volatilities for a standardized set of deltas ranging from −0.8 to −0.2 for puts and 0.2–0.8 for calls, as well as an implied option premium and implied strike price for each standardized option contract. We retain only those observations for which the implied option premium and the implied strike price are larger than zero, and for which the smoothed implied volatility is finite. We compute a stock's forward price using realized dividends over the life of the option from the OptionMetrics dividend file, discounted using the interpolated risk-free rates.

Call and put implied volatility smiles are not always identical and the standardized call and put deltas yield slightly different implied strike prices, so we compute an average implied volatility smile.[18,19]

Following Bakshi and Madan (2000), any twice differentiable payoff function of the stock, $H(S)$, can be spanned as a static portfolio of plain vanilla (European) put ($P(K)$) and call ($C(K)$) options with the same maturity as the payoff, a bond, and a forward contract.

$$\begin{aligned} H(S) = H(K_0) + (S - K_0)H_S(K_0) + \mathrm{e}^{r\tau} \textstyle\int_0^{K_0} H_{SS}(K)P(K)\mathrm{d}K \\ + \mathrm{e}^{r\tau} \textstyle\int_{K_0}^{\infty} H_{SS}(K)C(K)\mathrm{d}K, \end{aligned} \tag{51}$$

with K_0 a predetermined cut-off level separating the strike space into put and call options, r the continuously compounded risk-free rate, τ the relevant maturity, and H_S (H_{SS}) the (second) partial derivative of the payoff function with respect to the strike

minimize the residual variance, while the Fama–French factors are constructed to explain average returns.

18 Goncalves-Pinto et al. (2020) show how call and put Black–Scholes implied volatilities can differ if there is price pressure in the stock market.

19 We obtain one implied volatility smile per stock-date as follows. First, we interpolate the smoothed call implied volatilities at the put option implied strike prices and vice versa to obtain call and put smoothed implied volatilities for all observed implied strike prices. Then we average the put and call implied volatility for each strike price. We linearly interpolate those implied volatilities to a fine grid of moneyness levels, defined as the ratio of the strike price over the forward price. Outside the observed range of moneyness levels, we assume the implied volatility is constant at the endpoints in the observed data. This adjustment is common in the literature, see for example, Rehman and Vilkov (2012) and allows us to specify a function returning implied volatility for any strike price. As part of the numerical approximation of the integral below, we then use the Black–Scholes formula to translate the implied volatilities back to dollar option prices as inputs in calculating the price of the squared excess return using the Bakshi and Madan (2000) methodology.

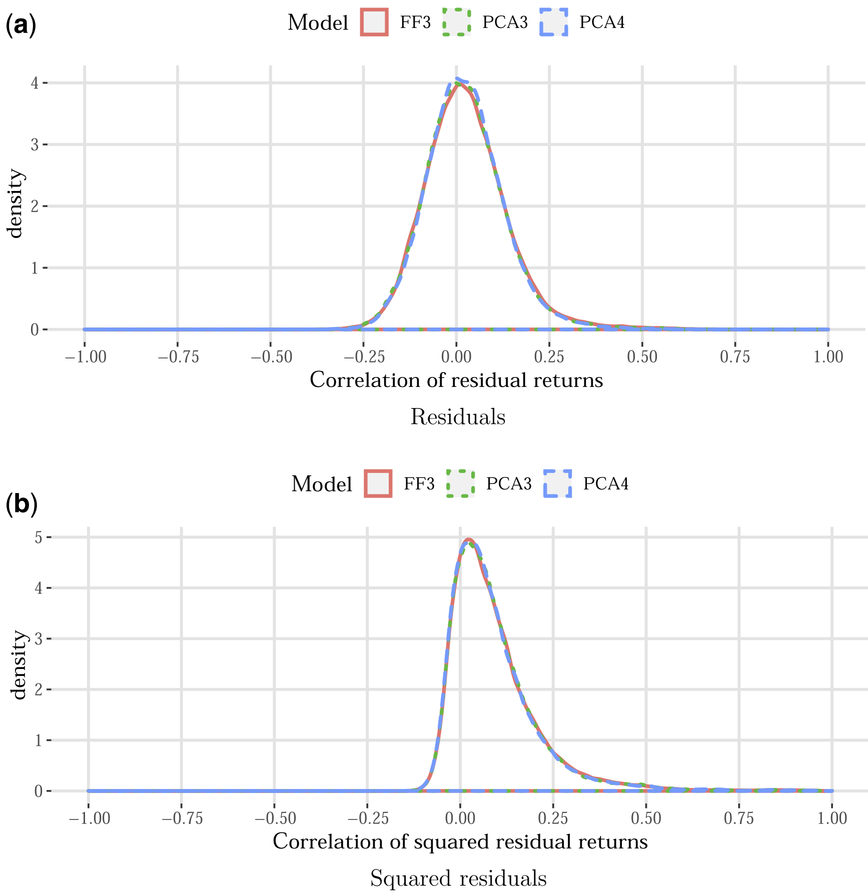

Figure 1 Density of residual return correlations. These figures plot the density of the distribution of pairwise correlations between residual returns (a) and squared residual returns (b). Returns are taken from the panel of 291 stocks with a full return history over the sample period covering January 1996 to December 2013. Returns are measured over 30 calendar days from the last trading day of each month. Residual returns are defined as the residuals from stock-level time-series regressions of 30-day returns on either the first three principal components of the full panel of 30-day returns (PCA3), the first four principal components (PCA4), or the three Fama–French factors as factors. The maximum pairwise correlation is 25.9% with an average of 0.2% and standard deviation 0.5%.

price.[20] With X_{it} the forward share price for firm i on the relevant horizon as of time t, S_{it} the spot price of firm i's shares, and $H(S_{i,t+1}) = \frac{S_{i,t+1}}{S_{it}} - 1$, we seek to compute the price of the discretely compounded squared excess return p_{it}, given as

20 Individual equity options are American rather than European, however, the early exercise premium is likely to be small. Ofek, Richardson, and Whitelaw (2004) report a median early exercise premium equal to 70 bps of the option's value for at-the-money put options. The bid–ask spread of those options is an order of magnitude larger than that.

$$p_{it} = (1 - e^{r\tau})^2 + \frac{2}{S_{it}}(X_{it} - S_{it})(1 - e^{r\tau}) + \frac{2}{S_{it}^2}\int_0^{S_{it}} P(K)\mathrm{d}K + \frac{2}{S_{it}^2}\int_{S_{it}}^{\infty} C(K)\mathrm{d}K, \tag{52}$$

which shows that the price of a squared excess return can be replicated using an equal-weighted portfolio of options with varying strike prices, a short forward position, and a position in bonds.[21] We numerically approximate the integrals in Equation (52) by a standard adaptive quadrature routine.

Figure 2 plots the time series of monthly observations of the annualized, equal-weighted cross-sectional average price of squared excess returns and the CBOE VIX index squared, which proxies for the price of market variance. That the average price of the squared return exceeds the price of the average squared return or market variance follows from Jensen's inequality and is consistent with results reported in Driessen, Maenhout, and Vilkov (2009). The time-series correlation between the two series is 0.72, confirming the link between market variance and average idiosyncratic variance reported in Herskovic et al. (2016). However, this high correlation should not be interpreted as co-movement in the cross-section of idiosyncratic variances being subsumed by the market variance, as we will show below and as also demonstrated in Chen and Petkova (2012), Duarte et al. (2014), and Herskovic et al. (2016).

3.1.2 The variance factor G

We analyze several sets of factors F for the linear returns. In addition to the first three or four principal components as described in Section 3.1, we also use the three Fama–French factors as factors. Once the factors F for the linear returns have been determined, we construct the variance factor G as follows. As before, we use the sample of 291 optionable stocks with a full return history and options traded throughout our sample period. First, we estimate constant beta loadings on each stock by regressing the time series of 30-day returns on the factor realizations F, and retain the panel of residual returns from this regression, $(\epsilon_{i,t})_{i=1,\ldots,291;t=1,\ldots,216}$.[22]

Before constructing portfolios of squared returns, we winsorize the panel of squared returns at the 99.95-th percentile to reduce the influence of outliers.[23] Then we extract the first principal component from the panel of squared residual returns using singular value decomposition, fit an AR(1) model to the principal component time series, and retain the innovations after scaling them to have unit variance. In the remainder, G refers to these scaled innovations. The magnitude of the correlation between G and the Herskovic et al. (2016) CIV factor is 0.55 in the different factor models we consider below. Figure 3 plots the time series of the G factor for the different factor models. If our model is correctly specified, the

21 The integrals in Equation (52) are defined properly. The put price is integrable as a function of the strike price over any interval of the form $[a, b)$ for $a \geq 0, b < \infty$ and in particular up to the current spot price that we use as a cut-off. The call price is integrable as a function of the strike price over any interval on the positive real axis.

22 We only perform this procedure once for each model we estimate. Formally, this should be repeated with each iteration, but empirically the GMM estimates are indistinguishable from the initial beta estimates.

23 Following the model in Section 1, we stress our analysis is on portfolios of squared returns, not squared returns of portfolios.

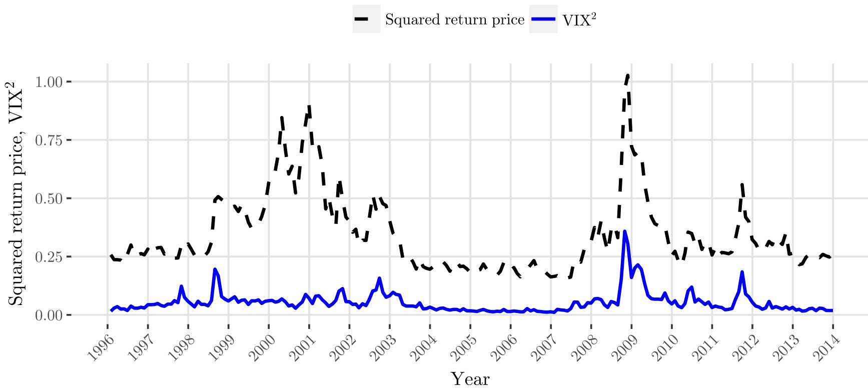

Figure 2 Time series of cross-sectional average price of squared excess return and VIX^2 index. This figure plots the time series of the annualized equal-weighted cross-sectional average price of the squared excess return of all optionable stocks in the OptionMetrics database satisfying the filters of Section 3.1. The price is constructed each last trading day of the month using Equation (52) for a standardized maturity of 30 calendar days. The sample period covers January 1996 to December 2013. Also plotted is the VIX^2 from the CBOE over the same period. The correlation between the two series is 0.72.

risk premium for G should be zero for linear returns.[24] Using the standard delta method, this test amounts to testing the hypothesis $H_0 : \beta_G \lambda_G = 0$, where λ_G is the price-of-risk of G. The hypothesis H_0 is not rejected for any of the 34 test assets at the 5% level for any model. However, this test may be conservative due to the singularity of the Jacobian under the null. Sorting portfolios by their β_G, the 80th percentile portfolio earns on average 6–7 bps less per month than the 20th percentile portfolio in any of the models we estimate below. In contrast, for portfolios sorted by the loading on the Herskovic et al.'s (2016) CIV factor, they report an average return spread of -45 bps per month between the top and bottom quintile portfolios.

3.2 Variance Factors and Industry Portfolios

The pricing restrictions of Corollary 1 present a set of moment conditions, which we estimate using GMM (details are in Appendix B). The parameter vector Θ to be estimated contains the components $\Theta = (\beta, \pi, \delta, \eta)$. The β parameters represent the factor loadings of the test assets for the different factors. Since all our factors are (scaled) excess returns, they have a zero price π, which is confirmed by our point estimates for π. The δ parameters represent the prices of risk of the implied factors at the squared return level and finally the η parameters represent the price of risk of the variance factor G.

For the 34 industry test portfolios, we estimate five different models, varying in the number and composition of factors for linear returns: (1) the first three principal components (henceforth PCA3), (2) the first four principal components (PCA4), (3) the first four principal components and a volatility factor (PCA4 + G), (4) three Fama–French factors

24 As observed in Section 2.3, this does not mean that G should not have any explanatory power for linear returns. The factors F and G may be correlated so the R^2 of a regression of linear returns on just G does not have to be zero.

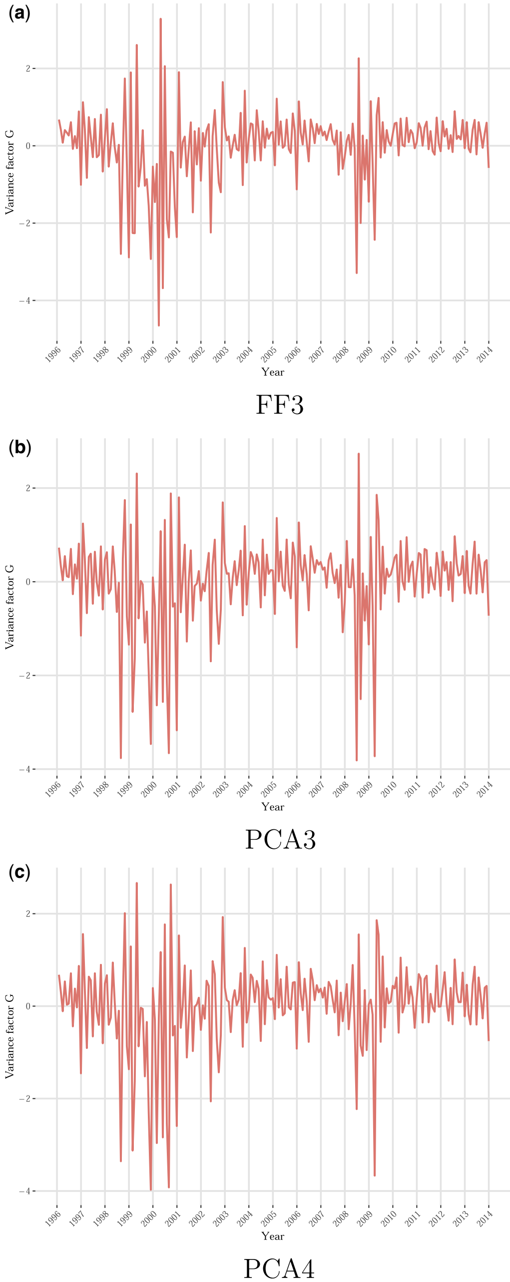

Figure 3 Time series of variance factor *G*: (a) FF3, (b) PCA3, and (c) PCA4. Time series of variance factor realizations. The variance factor is computed using residual returns of three different factor models for linear returns. The Fama and French (1993) three-factor model and PCA using the first three or four principal components. Returns are measured over a period of 30 calendar days from the last trading day of each month. The sample from which the principal components are extracted contains all 291 optionable stocks with a complete return history over the sample period 1996–2013.

(FF3), and (5) three Fama–French factors and a volatility factor following Herskovic et al. (2016) (FF3 + G).[25,26]

The main parameter of interest is the price of risk associated with G in the squared return equation, η. Table 1 reports the point estimate for η and its GMM heteroskedasticity and autocorrelation-adjusted standard error. For all models, the point estimate and associated standard error for η support the conclusion that idiosyncratic variance is a priced risk factor when considering squared returns. The average monthly excess squared excess return is −21 bps. The spread in the φ loadings between the 20th and 80th percentile of our test assets varies between 18 and 21 bps for the different models. Combining that with the η point estimates leads to an average squared return that is 5 (for the FF3 and FF3 + G models) to 7 bps (for the PCA3 and PCA4 models) per month higher for the 20th percentile portfolio than for the 80th percentile φ portfolio.

The results in Table 1, together with the lack of evidence for the risk premium on G in linear returns being non-zero as reported in Section 3.1.2, suggest that idiosyncratic variance risk is priced for squared returns only and is not a missing risk factor in the linear return model. Our integrated approach to pricing both linear and squared returns allows us to pin down the way in which idiosyncratic variance risk is priced in the cross-section. Our methodology differs substantially from existing papers, who generally take the following approach to study the price of the variance factor G in linear returns. First, they compute a panel of idiosyncratic returns using stock-level Fama–French regressions or PCA analysis. Second, they extract a principal component from the panel of squared residual returns (Duarte et al., 2014) or use the (change in the) average idiosyncratic variance directly as a factor (Herskovic et al., 2016). Third, they form quintile portfolios by sorting stocks on the loading with respect to the volatility factor and examine the return differences between the extreme quintiles as a proxy for the price of risk of idiosyncratic variance. We address the pricing of the idiosyncratic variance factor G in an asset pricing model that directly explains squared, rather than linear, returns, and in which all parameters are estimated jointly in a single step.

3.2.1 Prices of risk of implied variance factors

As part of the GMM estimation of the squared returns model, we obtain estimates for δ, the prices of risk of the factors $(F-\pi)(F-\pi)^\top$ implied by the linear returns model. Tables 2–7 contain the point estimates and standard errors of the δ parameters for the FF3, FF3 + G, PCA3, PCA3 + G, PCA4, and PCA4 + G models.

The first principal component in the PCA3/PCA4 model can be viewed as a scaled (with unit variance) proxy of the market return, having a correlation of 0.972 with the Fama–French market factor. The estimated price of risk of this component, δ_{11}, equals 1.9 for both PCA3 and PCA4 (2.0 for PCA4 + G). When scaling the principal component such that its variance is equal to the variance of the market return in the Fama–French model, the δ estimate corresponds to 0.0045 per month, or 5.4% per year. In the FF3 model, the price of risk of the squared market return is estimated at 0.0038 per month or 4.6% per year. These

25 Using the first five principal components yields similar results as the PCA4 model, but makes identification in the squared returns model harder due to the increased dimensionality of the implied factors. The same is true when using the five Fama–French factors.

26 The Hansen (1982) J-test does not reject any of the specifications.

Table 1 FF34 industry portfolios—η estimates

Model	η point estimate	η Std error	λ_G point estimate	λ_G Std error
FF3	0.269	(0.128)		
FF3 + G	0.262	(0.297)	0.088	(0.164)
PCA3	0.378	(0.136)		
PCA3 + G	0.407	(0.139)	0.122	(0.163)
PCA4	0.320	(0.135)		
PCA4 + G	0.298	(0.136)	0.094	(0.158)

Notes: For the 34 Fama–French industry portfolios with a full history of data, this table reports η point estimates and standard errors. The price-of-risk of the variance risk factor G for squared returns, η, is estimated using GMM on the moment conditions outlined in Equations (60) and (61). The different models use either the first three (PCA3) or four (PCA4) principal components or the three Fama–French factors (FF3) as the linear factors F. For each model, we also estimate a version in which the set of linear return factors F is augmented with the idiosyncratic variance factor G (FF3 + G, PCA3 + G, and PCA4 + G). In these "+ G" model variants, the price-of-risk of G for linear returns, λ_G, is defined as $\lambda_G = E(G) - \pi_G$. In the "+ G" models, G shows up in two different places in the squared return equation: once as itself and also as G^2 and $F^\top G$ in the set of implied factors. Irrespective of whether G is added to F or not, residual returns used to compute G are only based on the original F factors. The principal components are extracted from the panel of 291 firms with a complete history of both return and option data over the sample period. All returns are measured over 30 calendar days from the last trading day of a month.

estimates are of the same order of magnitude as other estimates of the market variance risk premium in the literature, using quite different methodologies.[27] It is also important to stress that since our model estimates risk premia for all factors jointly, the point estimates of η and δ suggest that market variance risk and idiosyncratic variance risk are both priced risk factors.

Estimates of δ for the squared SMB and HML factors and any cross-products in the FF3 model are not significantly different from zero. For the PCA3/PCA4 models, there is some evidence of the squared second principal component being priced, just like the cross-product between the first and second component. The sign of the coefficients points to offsetting exposures between the second PC squared and the cross-product of the first and second PC. The second and further PCs do not have as clear an interpretation as the first PC, so the economic interpretation of the δ estimates for the second PC and the cross-product of the first and second PC is not clear. We note that the second PC has a correlation of 0.74 with the HML factor and †0.25 with the SMB factor but zero correlation with the MKT factor.

3.2.2 The shape of the pricing kernel

We now turn to the shape of the pricing kernel implied by GMM parameter estimates. Recent literature argues that, as a function of the market return, the empirical pricing

27 Bollerslev, Tauchen, and Zhou (2009) and Drechsler and Yaron (2011), for example, find an annualized variance risk premium of 2.2%. The authors confirmed to us that the numbers reported in Bollerslev, Tauchen, and Zhou (2009) are monthly variance premiums in basis points, rather than annualized percentages as suggested in the caption of their Table 1. Londono (2015) reports a variance risk premium of around 2% in his sample of country indices.

Table 2 FF3/δ estimates

	F_1	F_2	F_3
F_1	0.004***		
	(0.000)		
F_2	–0.001	0.010	
	(0.001)	(0.006)	
F_3	0.000	0.005	0.003
	(0.001)	(0.004)	(0.003)

Notes: For the Fama–French three-factor model, this table reports the estimated prices-of-risk, δ, corresponding to the implied variance factors $FF^{\top}$. Each number indicates the δ corresponding to the product of the factors in the corresponding row and column. F1 equals the Fama–French market factor, F2 the SMB factor, and F3 the HML factor. The model is estimated using GMM on the moment conditions outlined in Equations (60) and (61) with the 34 of the 49 Fama–French industry portfolios with a full return history over the sample period as test assets. The sample period covers January 1996 to December 2013. ***, **, and * denote point estimates statistically different from zero at the 1%, 5%, and 10% level, respectively.

Table 3 FF3 + G/δ estimates

	F_1	F_2	F_3	F_4
F_1	0.003***			
	(0.001)			
F_2	0.000	0.015		
	(0.002)	(0.018)		
F_3	0.001	0.003	0.001	
	(0.002)	(0.011)	(0.008)	
F_4	–0.010	0.707*	0.403	–6.052
	(0.089)	(0.428)	(0.320)	(20.909)

Notes: For the Fama–French three-factor model augmented with the variance factor G in the linear return equation, this table reports the estimated prices-of-risk, δ, corresponding to the implied variance factors $FF^{\top}$. Each number indicates the δ corresponding to the product of the factors in the corresponding row and column. F1 equals the Fama–French market factor, F2 the SMB factor, F3 the HML factor, and F4 the variance factor G which is now included in the set of linear factors again. The model is estimated using GMM on the moment conditions outlined in Equations (60) and (61) with the 34 of the 49 Fama–French industry portfolios with a full return history over the sample period as test assets. The sample period covers January 1996 to December 2013. ***, **, and * denote point estimates statistically different from zero at the 1%, 5%, and 10% level, respectively.

kernel implied by S&P500 index option prices is non-monotonic. In particular, the pricing kernel is generally found to display an upward sloping region as a function of the market return for moderately positive market returns.[28]

Chabi-Yo, Garcia, and Renault (2008) provide a unifying explanation in terms of state dependence in case states are observed by the investors but not by the researcher. They show that models with regime shifts in fundamentals, preferences, or beliefs can rationalize

28 See, for example, Aït-Sahalia and Lo (2000); Jackwerth (2000); Bakshi, Madan, and Panayotov (2010); and Christoffersen, Heston, and Jacobs (2013).

Table 4 PCA3/δ estimates

	F_1	F_2	F_3
F_1	1.916***		
	(0.159)		
F_2	−0.632	0.876***	
	(0.097)	(0.338)	
F_3	−0.304	0.360	0.479
	(0.210)	(0.428)	(1.206)

Notes: For the three-principal component factor model, this table reports the estimated prices-of-risk, δ, corresponding to the implied variance factors $FF^{\top}$. Each number indicates the δ corresponding to the product of the factors in the corresponding row and column. The model is estimated using GMM on the moment conditions outlined in Equations (60) and (61) with the 34 of the 49 Fama–French industry portfolios with a full return history over the sample period as test assets. The sample period covers January 1996 to December 2013. ***, **, and * denote point estimates statistically different from zero at the 1%, 5%, and 10% level, respectively.

Table 5 PCA3 + G/δ estimates

	F_1	F_2	F_3	F_4
F_1	1.880***			
	(0.221)			
F_2	−0.477***	0.907**		
	(0.134)	(0.394)		
F_3	−0.223	−0.338	0.011	
	(0.254)	(0.877)	(1.341)	
F_4	−1.077	−0.440	6.003	8.773
	(0.920)	(1.280)	(5.454)	(13.555)

Notes: For the three-principal component factor model augmented with the variance factor G in the linear return equation, this table reports the estimated prices-of-risk, δ, corresponding to the implied variance factors $FF^{\top}$. Each number indicates the δ corresponding to the product of the factors in the corresponding row and column. F4 refers to the variance factor G which is now included in the set of linear factors again. The model is estimated using GMM on the moment conditions outlined in Equations (60) and (61) with the 34 of the 49 Fama–French industry portfolios with a full return history over the sample period as test assets. The sample period covers January 1996 to December 2013. ***, **, and * denote point estimates statistically different from zero at the 1%, 5%, and 10% level, respectively.

the puzzles put forward in Aït-Sahalia and Lo (2000) and Jackwerth (2000). The absolute risk aversion and pricing kernel extracted from calibrated prices in these economies exhibit the same puzzling features as in the original papers and are inconsistent with the usual assumptions of decreasing marginal utility and positive risk aversion. However, investor utility is well-behaved and her risk aversion remains positive. The bottom line is that, insofar as investors' behavior depends on factors the statistician does not observe, the pricing kernel puzzle can be steeply U-shaped while individual behaviors are not at odds with economic theory. This argument is arguably model specific but able to unify many popular models. When one wants to use a model-free approach, for instance as in Schneider and

Table 6 PCA4/δ estimates

	F_1	F_2	F_3	F_4
F_1	1.987***			
	(0.192)			
F_2	−0.629***	0.781*		
	(0.142)	(0.461)		
F_3	−0.296	0.840	−0.413	
	(0.222)	(0.548)	(1.203)	
F_4	0.410	−1.227*	−2.193	2.459
	(0.445)	(0.642)	(1.909)	(1.529)

Notes: For the four-principal component factor model, this table reports the estimated prices-of-risk, δ, corresponding to the implied variance factors $FF^\top$. Each number indicates the δ corresponding to the product of the factors in the corresponding row and column. The model is estimated using GMM on the moment conditions outlined in Equations (60) and (61) with the 34 of the 49 Fama–French industry portfolios with a full return history over the sample period as test assets. The sample period covers January 1996 to December 2013. ***, **, and * denote point estimates statistically different from zero at the 1%, 5%, and 10% level, respectively.

Table 7 PCA4 + G/δ estimates

	F_1	F_2	F_3	F_4	F_5
F_1	2.005***				
	(0.242)				
F_2	−0.580**	0.741			
	(0.242)	(0.570)			
F_3	−0.390	0.785	−0.544		
	(0.331)	(0.663)	(1.074)		
F_4	0.376	−2.057*	−0.735	2.247	
	(0.568)	(1.127)	(2.250)	(2.792)	
F_5	1.054	0.035	−1.822	−6.345	10.324
	(1.235)	(1.708)	(5.978)	(4.661)	(15.647)

Notes: For the four-principal component factor model augmented with the variance factor G in the linear return equation, this table reports the estimated prices-of-risk, δ, corresponding to the implied variance factors $FF^\top$. Each number indicates the δ corresponding to the product of the factors in the corresponding row and column. F5 refers to the variance factor G which is now included in the set of linear factors again. The model is estimated using GMM on the moment conditions outlined in Equations (60) and (61) with the 34 of the 49 Fama–French industry portfolios with a full return history over the sample period as test assets. The sample period covers January 1996 to December 2013. ***, **, and * denote point estimates statistically different from zero at the 1%, 5%, and 10% level, respectively.

Trojani (2019), it may not allow for such extreme U-shaped SDFs, while confirming the puzzle.

As discussed in Section 2.6, our model gives rise to a quadratic pricing kernel, which obviously tightly constrains and possibly overly magnifies the departure from monotonicity of implied pricing kernels. The system (48)–(50) allows us to pin down the shape of the pricing kernel as a function of the factors.[29] To compare the behavior of the pricing kernel as a function of the market return in our model, with the well-documented evidence described above, we proceed as follows. We assume that the market return and the factors F are jointly normally distributed over our sample period. So that also the conditional distribution of the Fs given the market return is normal. We compute the conditional average of the Fs as we vary the 30-day market return between its 5th and 95th percentile in the sample. These conditional averages are then used as inputs to compute the pricing kernel as a function of the market return only from Equation (43). Figures 4–6 contain the results. The PCA3 and PCA4 pricing kernels align with the recent results in the literature about non-monotonic or U-shaped pricing kernels. Compared with a log-normal pricing kernel with a Sharpe ratio of 0.5,[30] the pricing kernels we estimate here put a substantially higher price on low returns. For a market return of –8% on a 30-day horizon, the log-normal pricing kernel takes a value of 1.28 versus 1.8 (1.6) for the quadratic pricing kernel using the PCA3 (PCA4) model.[31]

The pricing kernel of the FF3 model shows an (economically counterintuitive) inverted U-shape. It predicts investors are willing to pay to take on both extreme downside and extreme upside risk, which is particularly problematic for the downside risk case. It is important to note that in the GMM estimation we do not force the pricing kernel to have a U-shape, but our model does imply a quadratic pricing kernel. We do not include the pricing kernel plot for the FF3 + G model as there was no convergence for the system (48)–(50) in that case.

3.3 Individual Stocks as Test Assets

In this section, we repeat the empirical analysis of Section 3 using individual stocks rather than portfolios as test assets. Specifically, we use all 100 stocks that were part of the S&P500 index, had options traded on them, and had a complete time series of monthly returns for the full sample period. This test is motivated by results in Ang, Liu, and Schwarz (2017) who argue that using individual stocks as test assets improves the identification of risk premiums, even though the factor loadings are estimated with more noise than for portfolios.

Table 8 confirms the results in Table 1. The point estimates of η are all significantly different from zero irrespective of the specific factor model used and the standard errors of η are about 40–60% lower for the individual stocks which is to be expected given the increased number of test assets by a factor 3.

3.4 Size-B/M Portfolios as Test Assets

Our third set of test assets (see Table 9) are 25 Fama–French size-B/M-sorted portfolios.[32] With the exception of the Fama–French three-factor model, we find no evidence of the

29 We numerically solve the system of equations and note that the optimization routine did not converge for the model that includes both the three Fama–French factors and the variance factor in the linear returns model.

30 The Sharpe ratio of the CRSP total market index over the period 2007–2016 was 0.53.

31 The other parameters we used are $r = 0.05$, $\sigma = 0.16$, and $T = 30/365$.

32 The portfolios are constructed as an independent sort each year in June by market cap and the ratio of book equity to market equity at the end of the prior financial year.

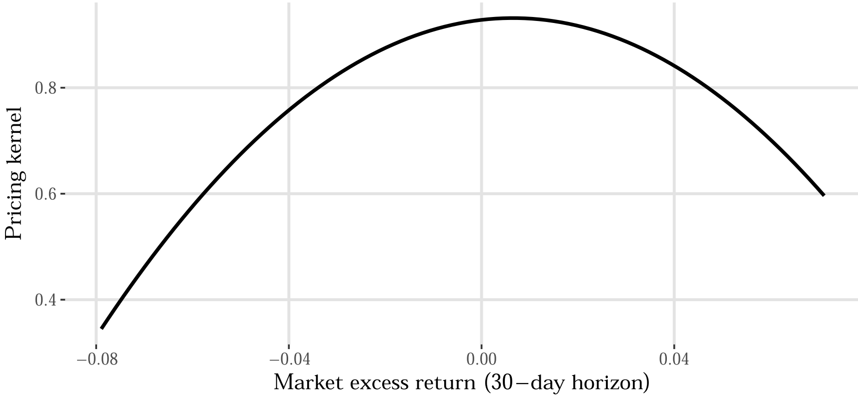

Figure 4 Pricing kernel as function of market return—FF3. This figure plots the quadratic pricing kernel of Section 2.6 as a function of the market excess return only. The parameter estimates of the FF3 model estimated on 34 of the Fama–French 49 industry portfolios with a complete return history over the period 1996–2013 are used. In generating this figure, the factors are assumed to be distributed jointly normal, and the HML and SMB factors as well as the variance factor *G* are fixed at their conditional expected value given the return of the market factor indicated on the horizontal axis. To do this, we first run a regression of each of the Fama–French factors and *G* on the market factor. The return on the horizontal axis corresponds to a maturity of 30 calendar days.

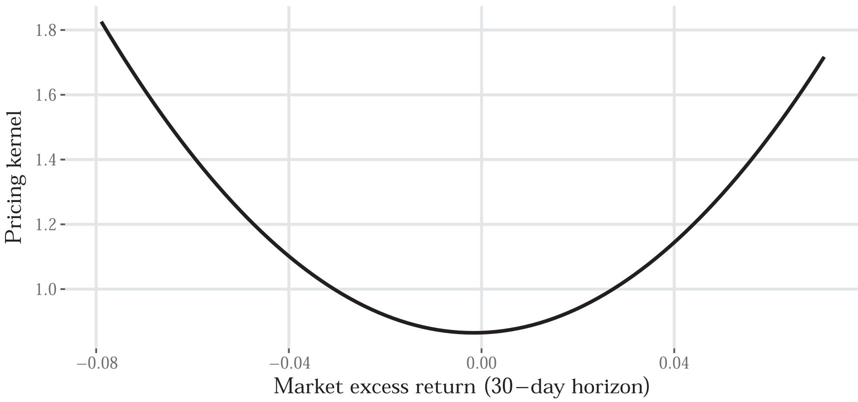

Figure 5 Pricing kernel as function of market return—PCA3. This figure plots the quadratic pricing kernel of Section 2.6 as a function of the rescaled first principal component only. The parameter estimates of the PCA3 model estimated on 34 of the Fama–French 49 industry portfolios with a complete return history over the period 1996–2013 are used. In generating this figure, we first regress each of the factors (the set of linear return factors *F* and the variance factor *G*) on the Fama–French market return factor. The 30-day return on the Fama–French market factor is plotted on the horizontal axis. The values of the *F* and *G* factors are fixed at their conditional mean given the Fama–French market return realization.

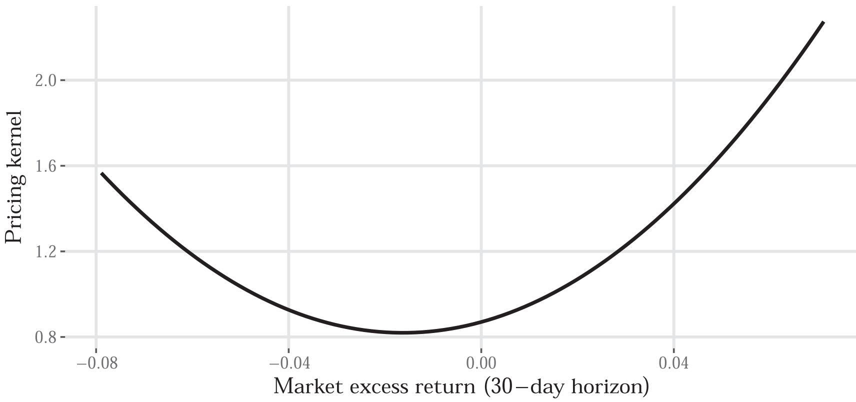

Figure 6 Pricing kernel as function of market return—PCA4. This figure plots the quadratic pricing kernel of Section 2.6 as a function of the rescaled first principal component only. The parameter estimates of the PCA4 model estimated on 34 of the Fama–French 49 industry portfolios with a complete return history over the period 1996–2013 are used. In generating this figure, we first regress each of the factors (the set of linear return factors *F* and the variance factor *G*) on the Fama–French market return factor. The 30-day return on the Fama–French market factor is plotted on the horizontal axis. The values of the *F* and *G* factors are fixed at their conditional mean given the Fama–French market return realization.

Table 8 S&P500 stocks as test assets—η estimates

Model	η point estimate	η Std error	λ_G Point estimate	λ_G Std error
FF3	0.428	(0.089)		
FF3 + G	0.490	(0.119)	0.011	(0.104)
PCA3	0.335	(0.072)		
PCA3 + G	0.438	(0.139)	0.026	(0.104)
PCA4	0.343	(0.081)		
PCA4 + G	0.407	(0.102)	−0.007	(0.101)

Notes: For the test assets comprised of individual S&P500 stocks, this table reports η point estimates and standard errors. The price-of-risk of the variance risk factor G for squared returns, η, is estimated using GMM on the moment conditions outlined in Equations (60) and (61). The different models use either the first three (PCA3) or four (PCA4) principal components or the three Fama–French factors (FF3) as the linear factors F. For each model, we also estimate a version in which the set of linear return factors F is augmented with the idiosyncratic variance factor G (FF3 + G, PCA3 + G, and PCA4 + G). In these "+ G" model variants, the price-of-risk of G for linear returns, λ_G, is defined as $\lambda_G = E(G) - \pi_G$. In the "+ G" models, G shows up in two different places in the squared return equation: once as itself and also as G^2 and $F^\top G$ in the set of implied factors. Irrespective of whether G is added to F or not, residual returns used to compute G are only based on the original F factors. The principal components are extracted from the panel of 291 firms with a complete history of both return and option data over the sample period. All returns are measured over 30 calendar days from the last trading day of a month.

Table 9 Twenty-five size-B/M portfolios—η estimates

Model	η point estimate	η Std error	λ_G point estimate	λ_G Std error
FF3	0.367	(0.135)		
FF3 + G	0.644	(0.378)	–0.239	(0.215)
PCA3	0.240	(0.190)		
PCA3 + G	0.289	(0.292)	–0.261	(0.214)
PCA4	0.273	(0.246)		
PCA4 + G	0.356	(0.236)	–0.387	(0.213)

Notes: For the 25 size-B/M-sorted portfolios, this table reports η point estimates and standard errors. The price-of-risk of the variance risk factor G for squared returns, η, is estimated using GMM on the moment conditions outlined in Equations (60) and (61). The different models use either the first three (PCA3) or four (PCA4) principal components or the three Fama–French factors (FF3) as the linear factors F. For each model, we also estimate a version in which the set of linear return factors F is augmented with the idiosyncratic variance factor G (FF3 + G, PCA3 + G, and PCA4 + G). In these "+ G" model variants, the price-of-risk of G for linear returns, λ_G, is defined as $\lambda_G = E(G) - \pi_G$. In the "+ G" models, G shows up in two different places in the squared return equation: once as itself and also as G^2 and $F^\top G$ in the set of implied factors. Irrespective of whether G is added to F or not, residual returns used to compute G are only based on the original F factors. The principal components are extracted from the panel of 291 firms with a complete history of both return and option data over the sample period. All returns are measured over 30 calendar days from the last trading day of a month.

variance factor to be priced for these test assets. The point estimate of the η parameter is similar to those of the industry portfolios, but the standard errors are quite large, which could be caused by the low beta dispersion for the Fama–French size-B/M portfolios. The delta estimates corresponding to the squared factors for these test assets are all statistically significantly different from zero for the squared FF3 factors. That is, the squared market excess return, the squared SMB, and squared HML factors are all priced when considering the size-B/M-sorted portfolios of squared returns as test assets. This result could be driven by the correlation between the MKT, SMB, and HML factors themselves. The coefficients on the implied risk factors (β^2) are always positive; coupled with $E(F^2) - \delta < 0$ for all three Fama–French factors, this means that the risk premiums on these factors are negative. A one-standard deviation increase in β^2 leads to a decrease in annualized expected squared return of 0.28%, 2.97%, and 0.32% for the market, SMB, and HML factors, respectively.

3.5 Variance-Sorted Portfolios as Test Assets

Our forth and final set of test assets are portfolios sorted by IV. Specifically, at the end of each month for each firm, the time series of daily stock returns within that month is regressed on the five Fama and French (2015) factors. The standard deviation of the residuals from that regression defines the IV for that firm-month. We sort the cross-section of firms into 25 portfolios by firm-level IV.[33] The IVOLs of the stocks in our sample are highly correlated, similar to the results reported by Herskovic et al. (2016).

33 The Ang et al. (2006b) puzzle is not present in the universe of optionable stocks, with the monthly Fama–French five-factor alpha of a portfolio long the lowest IVOL stocks and short the highest IVOL stocks is an insignificant –1.5 bps.

Table 10 Twenty-five IVOL portfolios—η estimates

Model	η point estimate	η Std error	λ_G point estimate	λ_G Std error
FF3	0.640	(29.153)		
FF3 + G	0.325	(19.543)	5.261	(525.343)
PCA3	0.314	(0.228)		
PCA3 + G	0.303	(0.150)	0.116	(0.192)
PCA4	0.437	(0.499)		
PCA4 + G	0.448	(0.419)	0.007	(0.200)

Notes: For the 25 IVOL-sorted portfolios, this table reports η point estimates and standard errors. The price-of-risk of the variance risk factor G for squared returns, η, is estimated using GMM on the moment conditions outlined in Equations (60) and (61). The different models use either the first three (PCA3) or four (PCA4) principal components or the three Fama–French factors (FF3) as the linear factors F. For each model, we also estimate a version in which the set of linear return factors F is augmented with the idiosyncratic variance factor G (FF3 + G, PCA3 + G, and PCA4 + G). In these "+ G" model variants, the price-of-risk of G for linear returns, λ_G, is defined as $\lambda_G = E(G) - \pi_G$. In the "+ G" models, G shows up in two different places in the squared return equation: once as itself and also as G^2 and $F^\top G$ in the set of implied factors. Irrespective of whether G is added to F or not, residual returns used to compute G are only based on the original F factors. The principal components are extracted from the panel of 291 firms with a complete history of both return and option data over the sample period. All returns are measured over 30 calendar days from the last trading day of a month.

The first principal component explains 35% of the variation in a panel of monthly IVOL for the 291 stocks with a complete time series in our data set. However, this comovement does not appear priced, as Table 10 reports. With the exception of the "PCA3 + G" model, none of the η coefficient estimates is statistically significant, which may be caused by the weak identification of the parameter due to the relatively small cross-sectional variation in the factor loadings φ of the IVOL portfolios for the set of optionable stocks we consider.

Overall we conclude that our analyses confirm the presence of a priced idiosyncratic variance risk factor affecting average squared returns for industry portfolios and individual stocks as test assets. Among the sets of test assets we consider, these two sets have got the highest cross-sectional dispersion in their factor loadings and therefore are expected to provide the best identification. For the variance-sorted portfolios and size-b/m-sorted portfolios, we find very limited evidence of a priced risk factor.

In contrast with the existing literature, we find evidence only for the idiosyncratic variance factor affecting average squared returns. The risk premium on this factor for linear returns is not different from zero. We also confirm the pricing of market variance as a risk factor. Across the various sets of test assets, the risk premium on market variance for the average test asset is substantially higher (45 bps/month) than the risk premium on the idiosyncratic variance factor (ca. 5 bps/month).

5 Summary and Conclusions

We propose a new formulation of the classic Ross (1976) APT and link this formulation to work by Al-Najjar (1998) and Gagliardini et al. (2016). Our formulation implies an APT for *squared* excess returns. We show how a factor structure for linear returns gives rise to implied risk factors for the squared returns as well as allowing for (a) separate idiosyncratic variance factor(s). The implied factors carry their own prices of risk which are distinct from the prices of risk of the original factors at the linear return level. Possibly, also the idiosyncratic variance factor is priced.

In our empirical analysis using Fama–French industry portfolios as test assets, we confirm the presence of a priced risk factor in idiosyncratic volatilities. However, in contrast to recent papers such as Duarte et al. (2014) and Herskovic et al. (2016), we find no evidence of this factor being priced for linear returns, but only for squared returns. This suggests that the IV factor does not represent a missing risk factor in the linear returns model.

Our conclusions are the same whether we consider a model for linear returns with the Fama–French factors or with principal components as factors. Using the 100 stocks that were part of the S&P500 during our sample period (1996–2013) and have a full history of returns over that period as test assets again leads us to conclude that there is a factor in idiosyncratic volatilities but that it is only priced for squared returns. Comparing the point estimates of the IV factor's price of risk for the four sets of test assets, we see that the estimate for the price of risk from the squared returns increases in magnitude, while the estimate for the linear returns decreases in magnitude, providing further support for our conclusion.

The risk premium estimates for the implied risk factors in the squared returns model confirm that idiosyncratic variance and market variance are priced separately. For the average test asset, the magnitude of the risk premium for market variance risk is considerably larger (45 bps/month) than that of idiosyncratic variance risk (5 bps/month).

We also estimate the induced pricing kernel of the model as a function of the market excess return only (using the conditional joint distribution of the factors to evaluate the other factors at their conditional mean). Other than the functional form of the pricing kernel being quadratic in the factors, we do not impose any restrictions on its shape. The induced pricing kernel takes a plausible convex form with positive values throughout when using principal components as factors, but displays a concave shape when using Fama–French factors.

A natural question is how our APT for squared returns extends of other nonlinear transformations of excess returns. One important example would be the factor pricing of put and call options, as an alternative to the analyses in Kadan, Liu, and Tang (2017) and Christoffersen, Fournier, and Jacobs (2018). For these securities, our framework would link the factor structure in option returns to non-linear transformations of the factors that explain the cross section of returns of the underlying assets. We leave this for future research.

Appendix A: Proofs

Proof of Proposition 1: Combining Equations (3) and (7) we find

$$
\begin{aligned}
S^{(n)}(u) &= \int_0^u([\nabla R^{(n)}(v)]^2 - p^{(n)}(v))\mathrm{d}v \\
&= \int_0^u([\alpha_R(v) + \beta_R(v)^\top F + \nabla Z^{(n)}(v)]^2 - p^{(n)}(v))\mathrm{d}v \\
&= \int_0^u [\alpha_R(v) + \beta_R(v)^\top F]^2\mathrm{d}v + \int_0^u([\nabla Z^{(n)}(v)]^2 - p^{(n)}(v))\mathrm{d}v \\
&+2\int_0^u[\alpha_R(v) + \beta_R(v)^\top F][\nabla Z^{(n)}(v)]\mathrm{d}v \\
&= \int_0^u [\alpha_R(v) + \beta_R(v)^\top F]^2\mathrm{d}v + \int_0^u [\nabla Z^{(n)}(v)]^2 - p^{(n)}(v)\mathrm{d}v \\
&+2\int_0^u[\alpha_R(v) + \beta_R(v)^\top F]\mathrm{d}Z^{(n)}(v) \\
\underset{\rightarrow}{w^*} &\int_0^u [\alpha_R(v) + \beta_R(v)^\top F]^2\mathrm{d}v + \int_0^u[\omega_S(v) - p(v)]\mathrm{d}v + \int_0^u \varphi_S^\top(v)G\mathrm{d}v,
\end{aligned}
\tag{53}
$$

where the last equality follows in view of Equations (10), (1), and (2). □

Proof of Proposition 2: The conditions imposed immediately imply that $R_\Delta^{(n)}$ satisfies Equation (3). □

Proof of Proposition 3: In order to verify Equation (10), observe

$$
\begin{aligned}
&\int_0^u [\nabla Z_\Delta^{(n)}(v)]^2\mathrm{d}v = \sum_{i=1}^n [\varepsilon_i^{(n)}]^2 \int_0^u 1_{\{v\in U_i^{(n)}\}}\mathrm{d}v \\
&= \sum_{i=1}^n (\omega_i^{(n)} + \varphi_i^{(n)\top}G + \nu_i^{(n)}) \int_0^u 1_{\{v\in U_i^{(n)}\}}\mathrm{d}v \\
&\overset{w^*}{\rightarrow} \int_0^u \omega_S(v)\mathrm{d}v + \int_0^u \varphi_S(v)^\top G\mathrm{d}v,
\end{aligned}
\tag{54}
$$

as $n \to \infty$. □

Proof of Proposition 4: The unknown coefficients a, b, c, and d in the candidate pricing kernel M in Equation (43) are determined by the pricing relations

$$
0 = ER_i^{(n)}M = \frac{ER_i^{(n)}}{R_f} + \mathrm{Cov}\left\{R_i^{(n)}, M\right\}, \tag{55}
$$

$$
0 = E\left(\left(R_i^{(n)}\right)^2 - p_i^{(n)}\right)M = \frac{E\left(R_i^{(n)}\right)^2 - p_i^{(n)}}{R_f} + \mathrm{Cov}\left\{\left(R_i^{(n)}\right)^2, M\right\}. \tag{56}
$$

Under the Conditions (44)–(47), we obtain a valid pricing kernel if, for all β_i and all φ_i, we have

$$\begin{aligned}0 = \beta_i^{(n)\top}\frac{EF-\pi}{R_f} + \beta_i^{(n)\top}\mathrm{Var}\{F-\pi\}b \\ + \beta_i^{(n)\top}\mathrm{Cov}\Big\{F-\pi,(F-\pi)^\top c(F-\pi)\Big\} \\ + \beta_i^{(n)\top}\mathrm{Cov}\Big\{F-\pi,(G-\eta)^\top\Big\}d\end{aligned} \tag{57}$$

$$\begin{aligned}0 = &\frac{\beta_i^{(n)\top}E\Big[(F-\pi)(F-\pi)^\top-\delta\Big]\beta_i^{(n)}+\varphi_i^{(n)\top}E\{G-\eta\}}{R_f} \\ &+\mathrm{Cov}\Big\{\beta_i^{(n)\top}\Big[(F-\pi)(F-\pi)^\top-\delta\Big]\beta_i^{(n)},(F-\pi)^\top\Big\}b \\ &+\varphi_i^{(n)\top}\mathrm{Cov}\Big\{(G-\eta),(F-\pi)^\top\Big\}b \\ &+\mathrm{Cov}\Big\{\beta_i^{(n)\top}\Big[(F-\pi)(F-\pi)^\top-\delta\Big]\beta_i^{(n)},(F-\pi)^\top c(F-\pi)\Big\} \\ &+\varphi_i^{(n)\top}\mathrm{Cov}\Big\{(G-\eta),(G-\eta)^\top\Big\}d.\end{aligned} \tag{58}$$

As these relations have to hold for all $\beta_i^{(n)}$ and $\varphi_i^{(n)}$, we obtain Equations (48)–(50). □

Appendix B: GMM Estimation

We estimate our APT model, including the idiosyncratic variance factors, by GMM. We detail the exact moment conditions we use in this Appendix.

At the linear return level, we use the moment condition implied by Equation (41), that is,

$$R_{it} = \beta_i^\top(F_t-\pi)+\varepsilon_i, \tag{59}$$

where we omit the number of available assets n from the notation and add time indexes. In line with Section 2.5, we interpret this as an unconditional regression, that is, the return process is assumed to be stationary and induced moment conditions are

$$E\left[\begin{pmatrix}1\\F_t\end{pmatrix}\otimes\big(R_{it}-\beta_i^\top(F_t-\pi)\big)_{1\le i\le n}\right]=0, \tag{60}$$

so there are $(K+1)n$ moment conditions for $Kn+K$ parameters.

The moment conditions for the squared returns are based on Equation (42). This leads to the following moment conditions[34]

34 For a symmetric $K\times K$ matrix A, vech(A) equals the $K(K+1)/2$ column vector obtained by vectorizing the lower triangular part of A. We introduce it here to avoid using multi-collinear instruments in $(F_t-\pi)(F_t-\pi)^\top$.

$$E\left[\begin{pmatrix} 1 \\ \mathrm{vech}\left((F_t-\pi)(F_t-\pi)^\top\right) \\ G_t \end{pmatrix} \otimes \left(R_{it}^2 - p_{it} - \beta_i^\top\left[(F_t-\pi)(F_t-\pi)^\top - \delta\right]\beta_i - \varphi_i^\top(G_t-\eta)\right)_{1\le i\le n}\right] = 0. \qquad (61)$$

The linear return parameters $(\beta_i)_{1\le i\le n}$ and π are present in both Equations (60) and (61). The moment conditions for squared returns (61) feature the additional parameters $(\varphi_i)_{1\le i\le n}$, δ, and η. We estimate the parameters by a single-step GMM procedure based on Equations (60) and (61).

Putting everything together, we see that the GMM system is over-identified with a total number of $(K+1)n + \left(1+\frac{K(K+1)}{2}+K_G\right)n$ equations, while $Kn+K+\frac{K(K+1)}{2}+K_G n + K_G$ parameters need to be estimated. A standard test for the over-identified moments equations would thus feature, under the null of correct specification, $n-K+n-K_G+(n-1)K(K+1)/2$ degrees of freedom. The HAC estimator uses a Bartlett kernal with Newey–West optimal bandwidth values.

B.1 Initial Values

We initialize the GMM estimation as follows. Starting values for the β parameters follow from time-series regressions of test asset returns on the F factors. The π parameters are set to zero, reflecting the fact that the F factors are (scaled) excess returns. The λ parameters (from $\lambda = E(F)-\pi$) are set equal to the unconditional average of the factors. The φ parameters follow from time-series regressions of the excess squared excess returns on the G variance factor. The δ and η parameters are initialized by first running cross-sectional regressions of the average excess squared excess returns on $\mathrm{vech}\beta\beta^\top$ and φ, and then computing the difference between the average of $\mathrm{vech}(F-\pi)(F-\pi)^\top$ and the coefficients on $\mathrm{vech}\beta\beta^\top$ (for the δ parameters) or the difference between G and the coefficient on φ (for the η parameter).

References

AïT-Sahalia, Y., and A. W. Lo. 2000. Nonparametric Risk Management and Implied Risk Aversion. *Journal of Econometrics* 94: 9–51.

Al-Najjar, N. I. 1995. Decomposition and Characterization of Risk with a Continuum of Random Variables. *Econometrica* 63: 1195–1224.

Al-Najjar, N. I. 1998. Factor Analysis and Arbitrage Pricing in Largest Asset Economies. *Journal of Economic Theory* 78: 231–262.

Al-Najjar, N. I. 1999. On the Robustness of Factor Structures to Asset Repackaging. *Journal of Mathematical Economics* 31: 309–320.

Ang, A., J. Chen, and Y. Xing. 2006a. Downside Risk. *Review of Financial Studies* 19: 1191–1239.

Ang, A., R. J. Hodrick, Y. Xing, and X. Zhang. 2006b. The Cross-Section of Volatility and Expected Returns. *The Journal of Finance* 61: 259–299.

Ang, A, J. Liu, and K. Schwarz.2020. Using Stocks or Portfolios in Tests of Factor models. Journal of Financial and Quantitative Analysis 55: 709–750.

Avramov, D., T. Chordia, G. Jostova, and A. Philipov. 2013. Anomalies and Financial Distress. *Journal of Financial Economics* 108: 139–159.

Bakshi, G., and D. Madan. 2000. Spanning and Derivative-Security Valuation. *Journal of Financial Economics* 55: 205–238.

Bakshi, G., D. Madan, and G. Panayotov. 2010. Returns of Claims on the Upside and the Viability of U-Shaped Pricing Kernels. *Journal of Financial Economics* 97: 130–154.

Bali, T. G., and N. Cakici. 2008. Idiosyncratic Volatility and the Cross Section of Expected Returns. *Journal of Financial and Quantitative Analysis* 43: 29–58.

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

Breeden, D., and R. Litzenberger. 1978. Prices of State-Contingent Claims Implicit in Option Prices. *The Journal of Business* 51: 621–651.

Carr, P., and R. Lee. 2009. Volatility Derivatives. *Annual Review of Financial Economics* 1: 319–339.

Carr, P., and L. Wu. 2009. Variance Risk Premiums. *Review of Financial Studies* 22: 1311–1341.

Chabi-Yo, F., R. Garcia, and E. Renault. 2008. State Dependence Can Explain the Risk Aversion Puzzle. *Review of Financial Studies* 21: 973–1011.

Chabi-Yo, F., D. P. J. Leisen, and E. Renault. 2014. Aggregation of Preferences for Skewed Asset Returns. *Journal of Economic Theory* 154: 453–489.

Chamberlain, G., and M. Rothschild. 1983. Factor Structure, and Mean–Variance Analysis on Large Asset Markets. *Econometrica* 51: 1281–1304.

Chen, Z., and R. Petkova. 2012. Does Idiosyncratic Volatility Proxy for Risk Exposure? *Review of Financial Studies* 25: 2745–2787.

Christoffersen, P., M. Fournier, and K. Jacobs. 2018. The Factor Structure in Equity Options. *The Review of Financial Studies* 31: 595–637.

Christoffersen, P., S. Heston, and K. Jacobs. 2013. Capturing Option Anomalies with a Variance-Dependent Pricing Kernel. *Review of Financial Studies* 26: 1963–2006.

Connor, G., R. A. Korajczyk, and O. Linton. 2006. The Common and Specific Components of Dynamic Volatility. *Journal of Econometrics* 132: 231–255.

Daigler, R. T., and L. Rossi. 2006. A Portfolio of Stocks and Volatility. *The Journal of Investing* 15: 99–106.

Dash, S., and M. T. Moran. 2005. VIX as a Companion for Hedge Fund Portfolios. *The Journal of Alternative Investments* 8: 75–80.

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

Driessen, J., P. J. Maenhout, and G. Vilkov. 2009. The Price of Correlation Risk: Evidence from Equity Options. *The Journal of Finance* 64: 1377–1406.

Duarte, J., A. Kamara, S. Siegel, and C. Sun. 2014. "The Systematic Risk of Idiosyncratic Volatility." University of Washington Working Paper.

Fama, E. F., and K. R. French. 1993. Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics* 33: 3–56.

Fama, E. F., and K. R. French. 2015. A Five-Factor Asset Pricing Model. *Journal of Financial Economics* 116: 1–22.

Gagliardini, P., E. Ossola, and O. Scaillet. 2016. Time-Varying Risk Premiums in Large Cross-Sectional Equity Datasets. *Econometrica* 84: 985–1046.

Goncalves-Pinto, L., B. D. Grundy, A. Hameed, T. van der Heijden, and Y. Zhu. 2020. Why Do Option Prices Predict Stock Returns? The Role of Price Pressure in the Stock Market. *Management Science* 66: 3903–3926. http://papers.ssrn.com/abstract=2695145

Hansen, L. P. 1982. Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica* 50: 1029–1054.

Harvey, C. R., and A. Siddique. 2000. Conditional Skewness in Asset Pricing Tests. *The Journal of Finance* 55: 1263–1295.

Herskovic, B., B. Kelly, H. Lustig, and S. Van Nieuwerburgh. 2016. The Common Factor in Idiosyncratic Volatility: Quantitative Asset Pricing Implications. *Journal of Financial Economics* 119: 249–283.

Jackwerth, J. C. 2000. Recovering Risk Aversion from Option Prices and Realized Returns. *Review of Financial Studies* 13: 433–451.

Kadan, O., F. Liu, and X. Tang. 2017. A Factor Structure for Implied Volatility. *SSRN Electronic Journal*. http://dx.doi.org/10.2139/ssrn.2969089.

Korsaye, S. A., A. Quaini, and F. Trojani. 2020. "Smart Stochastic Discount Factors." Working Paper.

Kraus, A., and R. H. Litzenberger. 1976. Skewness Preference and the Valuation of Risk Assets. *The Journal of Finance* 31: 1085–1100.

Lehmann, B. N. 1990. Residual Risk Revisited. *Journal of Econometrics* 45: 71–97.

Londono, J.-M. 2015. The Variance Risk Premium Around the World. Available at SSRN: https://ssrn.com/abstract=2517020 or http://dx.doi.org/10.2139/ssrn.2517020.

Mykland, P. A., and L. Zhang. 2006. ANOVA for Diffusions and Itô Processes. *The Annals of Statistics* 34: 1931–1963.

Ofek, E., M. Richardson, and R. F. Whitelaw. 2004. Limited Arbitrage and Short Sales Restrictions: Evidence from the Options Markets. *Journal of Financial Economics* 74: 305–342.

Rehman, Z., and G. Vilkov. 2012. "Risk-Neutral Skewness: Return Predictability and Its Sources." Working Paper.

Ross, S. A. 1976. The Arbitrage Theory of Capital Asset Pricing. *Journal of Economic Theory* 13: 341–360.

Schneider, P., and F. Trojani. 2019. (Almost) Model-Free Recovery. *The Journal of Finance* 74: 323–370.

Signori, O., M. Briere, and A. Burgues. 2010. Volatility Exposure for Strategic Asset Allocation. *The Journal of Portfolio Management* 36: 105–116.

Uppal, R., P. Zaffaroni, and I. Zviadadze. 2018. "Correcting Misspecified Stochastic Discount Factors." Working Paper.

van der Heijden, T., Q. Zeng, and Y. Zhu. 2019. "A Multifactor Model of Idiosyncratic Volatility." Working Paper, The University of Melbourne.