APT with Idiosyncratic Variance Factors

PRELIMINARY AND INCOMPLETE: PLEASE DO NOT CITE WITHOUT PERMISSION

Eric Renault*, Thijs van der Heijden†, and Bas J.M. Werker‡

January 9, 2016

Abstract

Recent research has documented the existence of common factors in individual asset's idiosyncratic variances or squared idiosyncratic returns. We provide an Arbitrage Pricing Theory that leads to a linear factor structure for prices of squared excess returns. This pricing representation allows us to study the interplay of factors at the return level with those in idiosyncratic variances. We document the presence of a common volatility factors. Linear returns do not have exposure to this factor when using at least five principal components as linear factors. The price of the common volatility factor is zero.

JEL codes : C58, G12.

Key words and phrases: Common volatility factors, Option prices.

*Department of Economics, Brown University, 64 Waterman Street, Providence RI 02912, USA, email: eric_renault@brown.edu

†Department of Finance, The University of Melbourne, 198 Berkeley Street, Carlton VIC 3010, Australia, email: thijsv@unimelb.edu.au

‡Econometrics and Finance Group, Netspar, Tilburg University, P.O. Box 90153, 5000 LE, Tilburg, The Netherlands. E-mail: Werker@TilburgUniversity.edu. We thank conference participants at the Econometrics of High-Dimensional Risk Networks at the Stefanovich Center and seminar participants at Tilburg University for helpful comments and suggestions.

1 Introduction

Recently, several papers have documented the presence of a common factor in idiosyncratic volatilities from a linear return factor model, arguing that this factor is priced which would be at odds with standard theory. This line of research started with Ang, Hodrick, Xing, and Zhang (2006) who coined their result the "idiosyncratic volatility puzzle". Recent contributions are Duarte, Kamara, Siegel, and Sun (2014) and Herskovic, Kelly, Lustig, and Van Nieuwerburgh (Forthcoming). In the present paper, we revisit this puzzle within the Ross (1976) Arbitrage Pricing Theory (APT) framework. Specifically, we propose a different formulation of the classic, single-period, APT in terms of (cumulative) portfolios of assets in the economy. Intuitively, our formulation can be viewed as a transposed version of standard continuous-time finance theory, where the index of the stochastic process refers to an asset index rather than time. The advantage of our approach is that it easily extends the APT for linear returns to squared returns. In contrast to Ang, Hodrick, Xing, and Zhang (2006), Duarte, Kamara, Siegel, and Sun (2014) and Herskovic, Kelly, Lustig, and Van Nieuwerburgh (Forthcoming), we do not study idiosyncratic volatility as a possible missing factor in linear returns, but instead consider the factor structure of squared returns directly. Our model predicts the presence of a set of risk prices related to the squared linear factors as well as to any additional factors driving the idiosyncratic variances. This allows us to study the effect of possibly omitted factors at the linear return level and to obtain more precise estimates of the price of common volatility factors. We thereby shed new light on the mixed evidence on the presence and price of common volatility factors.

We focus on excess squared excess returns, i.e., squared excess returns minus their price. Thereby, these squared excess returns have zero price as well and, thus, can be interpreted as excess returns themselves. In order to construct these excess squared excess returns, we compute the price of squared excess returns using the spanning results of Bakshi and Madan (2000). They show that the price of any payoff that is a twice-differentiable function of the underlying security is given by a combination of a position in a risk-free asset, a forward contract and a suitable portfolio of put and call options.

In our empirical analysis of S&P500 index firms over the period 1996-2013, we document the presence of a common factor in idiosyncratic variances in addition to (the squares of) the factors in the linear excess returns. We extract up to ten factors of the linear return model using principal components and analyse the factor structure of the squared residual (idiosyncratic) returns. We then include both the linear factors and the squared return factor in a Fama and MacBeth (1973) analysis. The squared return factor has some incremental explanatory power in the linear return model even with ten principal components included. However, the loadings on the squared return factor are insignificantly different from zero when ten principal components are included. In addition, irrespective of the number of principal components that are included, the price of risk of the squared return factor is insignificantly different from zero. This is in contrast to the same analysis that uses the five Fama and French (2015) factors, where both the average loading and the price of risk of the squared return factor are significantly different from zero. The squared return factor also has a substantially higher explanatory power for linear returns in the Fama and French (2015) case than when using principal components.

In order to understand our results, it is useful to distinguish the concepts of *statistical* and *financial* factor models. In a statistical factor model one extracts (e.g., using principal components) factors such that the residuals become cross-sectionally uncorrelated, i.e., diversifiable. In a financial factor model, one extracts factors (e.g., the Fama-French factors) such that the residuals become idiosyncratic in the sense that they do not command a risk premium, i.e., have zero price. The Arbitrage Pricing Theory states that, under an additional no-arbitrage assumption, a statistical factor model implies a financial factor model. The converse, however, does not hold. That is, there may exist non-diversifiable risks that do not command a risk premium, i.e., have zero price. Thus, using Fama-French factors at the linear return level, may leave a common (non-diversifiable) factor in the "idiosyncratic" residuals. The square of this factor will show as a common factor in the "idiosyncratic" variances and it may or may not be priced. The contribution of the present paper is to show that, in line with the intuition that diversifiable risk cannot command any risk premium, the use of a statistical factor model at the linear return level, still leads to a common factor in idiosyncratic variances, but we empirically find this common factor to be idiosyncratic in the sense that it has zero price.

Our paper is also related to the literature on skewness in asset pricing, which started with Kraus and Litzenberger (1976) showing that investors exhibiting non-increasing absolute risk aversion is equivalent with an extension of the CAPM that incorporates skewness as the covariance between the asset return and the squared market return. Harvey and Siddique (2000) focus on the cross-section of expected returns and use conditional rather than unconditional skewness. They write down a model in which the pricing kernel is linear in the market return and market return squared. 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 risk is linked to the portfolio that optimally hedges the squared market return. Since volatility is an important determinant of option prices, our paper is also related to the literature on the factor structure in option prices, e.g., Christoffersen, Fournier, and Jacobs (2015). Instead of studying factor structures in option prices, we use these prices to obtain a of squared excess returns. As follows from our theory, this price of a quadratic transformation is much more easily studied in an Arbitrage Pricing Theory framework than the more complicated non-linearities in option prices.

The remainder of the paper is structured as follows. In Section 2, we propose a new formulation of the APT model for linear returns. We use this new formulation, in Section 3, to study an approximate factor structure in excess squared excess returns (i.e., idiosyncratic variances) and derive testable implications. In Section 4 we describe the sample and the variables we construct. Section 5 contains the empirical results and Section 6 concludes.

2 The APT revisited

We start our theoretical analysis by providing a new proof of the classical Arbitrage Pricing Theory (APT). Instead of, e.g., Al-Najjar (1998) and Gagliardini, Ossola, and Scaillet (2014), we consider cumulative portfolios of assets to obtain the APT. A precise link with existing APT results is provided in Remark 1 below. The advantage of our approach is that it readily extends to squared returns, the subject of Section 3; at the level of the linear returns there is not much new.

Consider n traded assets with (arithmetically compounded) excess returns $R_i^{(n)}$, $i = 1, \ldots, n$. Recall that excess returns have price zero, i.e., they refer to zero-investment opportunities. In this paper we call any investment with zero price an excess return.

In order to formalize our formulation of the assumption of an approximate factor structure, we construct portfolios, for given $u \in [0, 1]$, consisting of $1/n$ exposures in the first u fraction of the assets. Such a portfolio thus has excess return

$$R^{(n)}(u) = \frac{1}{n} \sum_{i=1}^{\lfloor un \rfloor} R_i^{(n)}. \tag{1}$$

Note that $R^{(n)}(u)$ is an alternative representation of the available assets $R_i^{(n)}$ in the market[1]. We have $R^{(n)}(0) = 0$ and $R^{(n)}(1)$ represents an equally weighted portfolio in all available assets in the economy.

The rewrite from original assets with excess returns $R_i^{(n)}$ to portfolios indexed by $u \in [0, 1]$ facilitates a formal analysis of factor models. Observe that in the definition below, no moment restrictions are imposed on the excess returns, factors, or idiosyncratic errors.

Definition 1 *The (sequence of) excess return process(es) $R^{(n)}$ is said to satisfy an* approximate factor structure *if there exists a K-dimensional (random) factor F and deterministic finite-variation functions α and β such that we may write*

$$R^{(n)}(u) = \int_0^u \alpha(v)\,dv + \int_0^u \beta^{\intercal}(v)\,dv F + Z^{(n)}(u), \tag{2}$$

where $Z^{(n)}$ converges to zero.

Definition 1 formalizes our assumption of a factor structure. In order to illustrate the more abstract we results, we introduce an example that will also form the basis of our empirical analysis later.

Example For simplicity we consider the case of a single factor only, i.e., $K = 1$. Let the excess returns be given by

$$R_i^{(n)} = \alpha_i + \beta_i F + (\varphi_{0i} + \varphi_{1i} G)^{1/2}\, \nu_i, \tag{3}$$

for constants α_i, β_i, φ_{0i}, and φ_{1i} and where G is a common positive volatility factor. We assume that the ν_i's are i.i.d. zero-mean random variables, independent of both F and G, whose variances are normalized to unity. Moreover, we will assume the φ_{0i} and φ_{1i} to be bounded away from zero and infinity. Consequently,

$$R^{(n)}(u) \quad = \quad \frac{1}{n}\sum_{i=1}^{\lfloor un \rfloor} \alpha_i + \left(\frac{1}{n}\sum_{i=1}^{\lfloor un \rfloor} \beta_i\right) F + \frac{1}{n}\sum_{i=1}^{\lfloor un \rfloor} (\varphi_{0i} + \varphi_{1i} G)^{1/2}\, \nu_i. \tag{4}$$

In order to satisfy Definition 1 we assume the existence of finite-variation functions α and β

[1] Formally, $R^{(n)}$ is considered a stochastic process in $D[0, 1]$, the set of cadlag functions on $[0, 1]$, equipped with the supremum norm $\|\cdot\|$. All convergences of stochastic processes in this paper are weak convergence in $(D[0, 1], \|\cdot\|)$.

such that[2]

$$\frac{1}{n}\sum_{i=1}^{\lfloor un\rfloor}\alpha_i - \int_0^u \alpha(v)\mathrm{d}v = o\left(\frac{1}{\sqrt{n}}\right), \tag{5}$$

$$\frac{1}{n}\sum_{i=1}^{\lfloor un\rfloor}\beta_i - \int_0^u \beta(v)\mathrm{d}v = o\left(\frac{1}{\sqrt{n}}\right). \tag{6}$$

These conditions impose some stability on the intercepts and the factor loadings. Intuitively, the functions α and β are approximately given by $\alpha(u) \approx \alpha_{\lfloor un\rfloor}$ and $\beta(u) \approx \beta_{\lfloor un\rfloor}$, $0 \le u \le 1$.

Concerning the process $Z^{(n)}$, we now may write

$$Z^{(n)}(u) = \frac{1}{n}\sum_{i=1}^{\lfloor un\rfloor}\left(\varphi_{0i} + \varphi_{1i}G\right)^{1/2}\nu_i + o\left(\frac{1}{\sqrt{n}}\right). \tag{7}$$

The functional law of large numbers now gives convergence of $Z^{(n)}$ to zero. We will not provide details here as this result is an immediate consequence of the functional central limit theorem we apply to verify the conditions of Definition 2 below. □

Remark 1 *Several other formalization of the classical APT result exist in the literature, e.g., Gagliardini, Ossola, and Scaillet (2014) and Al-Najjar (1998). Those papers often start from excess returns written as* $R_i^{(n)} = \alpha_i + \beta_i^\intercal F + u_i^{(n)}$, *so that*

$$\begin{aligned} R^{(n)}(u) &= \frac{1}{n}\sum_{i=1}^{\lfloor un\rfloor}\left(\alpha_i + \beta_i^\intercal F + u_i^{(n)}\right) \\ &= \frac{1}{n}\sum_{i=1}^{\lfloor un\rfloor}\alpha_i + \left(\frac{1}{n}\sum_{i=1}^{\lfloor un\rfloor}\beta_i\right)^\intercal F + \frac{1}{n}\sum_{i=1}^{\lfloor un\rfloor}u_i^{(n)}. \end{aligned} \tag{8}$$

Assuming that the partial sums of α_i *and* β_i *converge as in (5) and (6), we have* $Z^{(n)}(u) = \frac{1}{n}\sum_{i=1}^{\lfloor un\rfloor} u_i^{(n)} + o(1)$. *Gagliardini, Ossola, and Scaillet (2014) impose* $\rho\left(Var\left\{u^{(n)}\right\}\right)/n \to 0$ *where* $u^{(n)} = \left(u_i^{(n)}\right)_{i=1}^n$ *and* ρ *denotes the maximum eigenvalue. Let* $\iota_n(u)$ *be an* n*-vector whose*

[2] The o-term here again is in terms of uniform convergence in $D[0,1]$. Moreover, for the results in this section it would be enough to require $o(1)$ convergence, but $o(n^{-1/2})$ will be needed when revisiting this example in Section 3.

first $\lfloor un \rfloor$ elements are one and the others are zero. Then we have

$$\begin{aligned} Var\left\{ Z^{(n)}(u) \right\} &= Var\left\{ \frac{1}{n} \sum_{i=1}^{\lfloor un \rfloor} u_i^{(n)} \right\} \\ &= \frac{1}{n^2} \iota_n(u)^\intercal Var\left\{ u^{(n)} \right\} \iota_n(u) \\ &\leq \frac{1}{n^2} \rho\left(Var\left\{ u^{(n)} \right\} \right) \iota_n(u)^\intercal \iota_n(u) \\ &= \frac{\lfloor un \rfloor}{n^2} \rho\left(Var\left\{ u^{(n)} \right\} \right). \end{aligned} \tag{9}$$

Assumption 4 in Gagliardini, Ossola, and Scaillet (2014) implies that the latter term converges to zero and, in particular, thus implies our Definition 1.

In order to derive the APT pricing implications, we consider portfolios of the base assets $R_i^{(n)}$, $i = 1, \dots, n$. Formally, we identify such a portfolio by a finite-variation function h. This portfolio's excess return is then, by definition,

$$\int_{u=0}^{1} h(u) \mathrm{d} R^{(n)}(u). \tag{10}$$

Taking $h(u) = 1$, we would find the excess return of an equally weighted portfolio with exposures $1/n$ to all assets. A value-weighted portfolio can be obtained by choosing $h(u)$ proportional to the relative market share of the u-th asset in the economy. As we work with excess returns, note in particular that increments in $R^{(n)}$ are also excess returns of portfolios consisting of a subset of the entire asset universe.

We can now state our version of the APT.

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

$$\alpha(u) = -\beta(u)^\intercal \lambda. \tag{11}$$

Proof: The proof is classical, but we provide a version that is convenient for our setup. Let h denote any portfolio without exposure to the factors, i.e., with $\int_{u=0}^{1} h(u)\beta(u)\mathrm{d}u = 0$. Then, from (2), the induced portfolio returns are

$$\int_{u=0}^{1} h(u)\mathrm{d}R^{(n)}(u) = \int_{u=0}^{1} h(u)\alpha(u)\mathrm{d}u + \int_{u=0}^{1} h(u)\mathrm{d}Z^{(n)}(u). \tag{12}$$

As both h and $Z^{(n)}$ are of finite variation, we find by partial integration

$$\int_{u=0}^{1} h(u)\mathrm{d}Z^{(n)}(u) = Z^{(n)}(1)h(1) - \int_{u=0}^{1} Z^{(n)}(u)\mathrm{d}h(u) - \sum_{0\leq u\leq 1} \Delta Z^{(n)}(u)\Delta h(u),$$

which converges to zero since $Z^{(n)}$ does. Consequently,

$$\int_{u=0}^{1} h(u)\mathrm{d}R^{(n)}(u) \to \int_{u=0}^{1} h(u)\alpha(u)\mathrm{d}u. \tag{13}$$

In the absence of arbitrage, we must have $\int_{u=0}^{1} h(u)\alpha(u)\mathrm{d}u = 0$. As this must hold for any finite-variation function h orthogonal to all components of β, we have $\alpha(u) = -\beta(u)^{\intercal}\lambda$ for some vector λ. □

Remark 2 - Repackaging *An important point in theoretical foundations of the APT is that its assumptions should be invariant under so-called "repackaging", see, e.g., Al-Najjar (1999). Loosely speaking this means that the assumptions should be invariant with respect to reordering the assets and forming portfolios. It's easy to see that our Definition 1 indeed obeys to this invariance.*

Consider first a reordering of the assets. First observe that, from (2), we have, for given $w \in [0, 1]$,

$$R^{(n)}(u) - R^{(n)}(w) = \int_{w}^{u} \alpha(v)\,dv + \int_{w}^{u} \beta^{\intercal}(v)\,dv + Z^{(n)}(u) - Z^{(n)}(w). \tag{14}$$

Now consider $p+1$ *fixed constants* $0 = u_0 < u_1 < \ldots < u_p = 1$. *A reordering of assets can now be obtained by permuting the* p *intervals* $[u_{j-1}, u_j]$, $j = 1, \ldots, p$. *It's clear that 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.*

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* $\int_0^u h^*(v)\,dR^{(n)}(v)$. *Such* *process obviously satisfies Definition 1 as soon as $R^{(n)}$ does. Indeed, we have, in view of (2),*

$$\int_0^u h^*(v)\,dR^{(n)}(v) = \int_0^u h^*(v)\alpha(v)\,dv + \int_0^u h^*(v)\beta^\intercal(v)\,dv + \int_0^u h^*(v)\,dZ^{(n)}(v),$$

using the same arguments as in the proof of Proposition 1, we find that the last term converges to zero. Moreover, $h^*\alpha$ and $h^*\beta$ are of finite variation (as the product of finite variation functions). As a result, $\int_0^u h^*(v)\,dR^{(n)}(v)$ satisfies an approximate factor structure as well.

Remark 3 - Factor-mimicking portfolios *If the excess return processes $R^{(n)}$ satisfy an approximate factor structure, Proposition 1 implies that we may write*

$$R^{(n)}(u) = \int_0^u \beta^\intercal(v)\,dv\,(F - \lambda) + Z^{(n)}(u). \tag{15}$$

This allows us to use the idea of factor mimicking portfolios. Choose a K-dimensional function H of finite variation on $[0,1]$ such that

$$\int_{u=0}^1 H(u)\beta^\intercal(u)\,du = I_K, \tag{16}$$

the $K \times K$ identity matrix. This is possible as long as the components of β are linearly independent.

Then the K portfolios induced by H satisfy

$$\int_{u=0}^1 H(u)\,dR^{(n)}(u) = (F - \lambda) + \int_{u=0}^1 H(u)\,dZ^{(n)}(u). \tag{17}$$

This implies in turn that these K portfolios could be used as factors, i.e., we have

$$R^{(n)}(u) = \int_0^u \beta^\intercal(v)\,dv \int_{u=0}^1 H(u)\,dR^{(n)}(u) + \left[Z^{(n)}(u) - \int_0^u \beta^\intercal(v)\,dv \int_{u=0}^1 H(u)\,dZ^{(n)}(u)\right]. \tag{18}$$

Remark 4 - Omitted factors *It is useful to consider the situation of possibly omitted factors. So suppose that the excess return process satisfies an approximate factor structure with factors (F, F_o), i.e.,*

$$R^{(n)}(u) = \int_0^u \alpha(v)\,dv + \int_0^u \beta^\intercal(v)\,dv F + \int_0^u \beta_o^\intercal(v)\,dv F_o + Z^{(n)}(u), \tag{19}$$

where $Z^{(n)}$ converges to zero. Assume now that the researcher omits the factors F_o from the *analysis. This researcher effectively considers the "idiosyncratic" errors* $\int_0^u \beta_o^\intercal(v)\,dv F_o + Z^{(n)}(u)$*. This will only converge to zero if* $\beta_o = 0$*. Consequently, Definition 1 precisely identifies the correct number of factors and formalizes the necessary (asymptotic) orthogonality of* F *and* $Z^{(n)}$*.*

Some papers have documented within-industry correlation patterns that point to industry-specific factors, compare, e.g., Ait-Sahalia and Xiu (2015). From a statistical point of view, such industry factors present themselves in the form of a block-diagonal covariance structure (in case assets are sorted by industry). The question whether such industry factors should be included as market-wide factors is essentially an empirical one. From a theoretical point of view, they should be included in case the size of the industry relative to the total market does not vanish asymptotically. Indeed, in that case the industry risk cannot be diversified.

3 The APT for squared returns

The main theoretical contribution of the present paper is to provide Arbitrage Pricing Theory implications for *squared* excess returns. We will, therefore, in this section provide an additional assumption on the factor structure in Definition 1, such that we can *deduce* a factor structure for the squared excess returns, and, thus, consider their pricing.

We reinforce Definition 1 to the following, stronger, condition.

Definition 2 *The excess return process* $R^{(n)}$ *is said to satisfy a* second-order approximate factor structure *if, additionally to the conditions in Definition 1, we have*

$$\sqrt{n}Z^{(n)}(u) \quad \overset{\mathcal{L}}{\to} \quad Z(u), \tag{20}$$

$$\left[\sqrt{n}Z^{(n)}, \sqrt{n}Z^{(n)}\right](u) \quad \overset{\mathcal{L}}{\to} \quad [Z, Z]\,(u), \tag{21}$$

for some stochastic process Z *whose quadratic variation satisfies the factor structure*

$$[Z, Z]\,(u) = \int_0^u \varphi_0(v)\,dv + \int_0^u \varphi_1^\intercal(v)\,dv G, \tag{22}$$

for deterministic finite-variation functions φ_0 *and* φ_1 *and a* K_S*-dimensional factor* G*.*

The above definition imposes directly weak convergence on both the idiosyncratic errors $Z^{(n)}$ as well as its quadratic variation. A sufficient condition for (20) to imply (21) is the so-called P-UT condition which is sometimes more easily checked, see Jacod and Shiryaev (2003) Section VI.6a for more details.

Example continued In order to verify the conditions in Definition 2 for our example, we need to study weak convergence of $\sqrt{n}Z^{(n)}(u)$ from (7), i.e., the convergence of

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor un \rfloor} (\varphi_{0i} + \varphi_{1i} G)^{1/2} \nu_i, \tag{23}$$

and its limiting quadratic variation. Note that conditionally on the value of G, we can apply the functional central limit theorem for independent, but not necessarily identically distributed, random variables. Under the additional conditions

$$\frac{1}{n} \sum_{i=1}^{\lfloor un \rfloor} \varphi_{0i} \quad \rightarrow \quad \int_0^u \varphi_0(v) \mathrm{d}v, \tag{24}$$

$$\frac{1}{n} \sum_{i=1}^{\lfloor un \rfloor} \varphi_{1i} \quad \rightarrow \quad \int_0^u \varphi_1(v) \mathrm{d}v, \tag{25}$$

for finite-variation functions φ_0 and φ_1, we find

$$[Z, Z](u) = \int_0^u \varphi_0(v) \mathrm{d}v + G \int_0^u \varphi_1(v) \mathrm{d}v, \tag{26}$$

As a result, the conditions in Definition 2 are satisfied. □

Now consider the situation where squared excess returns are traded assets as well. In Section 5 we will use plain vanilla options traded on individual assets to reconstruct these asset payoffs using a well-known technique going back to Breeden and Litzenberger (1978) and Bakshi and Madan (2000). We denote the market prices of these squared excess return by $p_i^{(n)}$. Then, we can define the *excess squared excess returns*[3] as

$$S_i^{(n)} = \left(R_i^{(n)}\right)^2 - p_i^{(n)}, \tag{27}$$

In line with the notations above, we assume that we may write, for a deterministic finite-

[3]Recall that we use the term excess return for any asset that has zero price.

variation function p,

$$\frac{1}{n}\sum_{i=1}^{\lfloor un \rfloor} p_i^{(n)} \to \int_0^u p(v)\mathrm{d}v, \tag{28}$$

again uniformly in $u \in [0,1]$. Subsequently, we define the *squared return* process[4] $S^{(n)}$ in a conformable way as

$$S^{(n)}(u) = \frac{1}{n}\sum_{i=1}^{\lfloor un \rfloor} S_i^{(n)}. \tag{29}$$

The main theoretical result of our paper is now the following.

Proposition 2 *If the excess return process $R^{(n)}$ satisfies a second-order approximate factor structure and (28) holds, then the squared return process $S^{(n)}$ satisfies an approximate factor structure with α given by*

$$\alpha_S(u) = \varphi_0(u) - p(u), \tag{30}$$

factors[5] $\mathit{vech}\left([F-\lambda]\,[F-\lambda]^\intercal\right)$, *with loadings*

$$\beta_{SF}(u) = \mathit{vech}\left(\beta(u)\beta^\intercal(u)\right), \tag{31}$$

and additional factors G with loadings

$$\beta_{SG}(u) = \varphi_1(u). \tag{32}$$

Proof: As the return process $R^{(n)}$ satisfies the conditions of Proposition 1, we may rewrite (1) as

$$\begin{aligned} R_i^{(n)} &= n\left[R^{(n)}\left(\frac{i}{n}\right) - R^{(n)}\left(\frac{i-1}{n}\right)\right] \qquad (33)\\ &= n\int_{(i-1)/n}^{i/n} \beta^\intercal(v)\mathrm{d}v\,(F-\lambda) + n\left[Z^{(n)}\left(\frac{i}{n}\right) - Z^{(n)}\left(\frac{i-1}{n}\right)\right]. \end{aligned}$$

[4]A more precise name would be the excess squared excess return process, but we use the term squared return process for convenience.

[5]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.

This implies

$$
\begin{aligned}
S^{(n)}(u) &= \frac{1}{n}\sum_{i=1}^{\lfloor un\rfloor}\left(R_i^{(n)}\right)^2 - p_i^{(n)} \\
&= n\sum_{i=1}^{\lfloor un\rfloor}\left[\int_{(i-1)/n}^{i/n}\beta^{\mathsf{T}}(v)\mathrm{d}v\,(F-\lambda)\right]^2 - \frac{1}{n}\sum_{i=1}^{\lfloor un\rfloor}p_i^{(n)} \\
&\quad + n\sum_{i=1}^{\lfloor un\rfloor}\left[Z^{(n)}\left(\frac{i}{n}\right) - Z^{(n)}\left(\frac{i-1}{n}\right)\right]^2 \\
&\quad + 2n\sum_{i=1}^{\lfloor un\rfloor}\int_{(i-1)/n}^{i/n}\beta^{\mathsf{T}}(v)\mathrm{d}v\,[F-\lambda]\left[Z^{(n)}\left(\frac{i}{n}\right) - Z^{(n)}\left(\frac{i-1}{n}\right)\right].
\end{aligned} \tag{34}
$$

We consider the convergence of the above four terms separately. For simplicity, we only give the proof for $K = 1$. With respect to the first term, we know that β is bounded, say by M. We establish that this first term essentially is a Riemann sum. Indeed, we have

$$
\begin{aligned}
&\left| n\sum_{i=1}^{\lfloor un\rfloor}\left(\int_{(i-1)/n}^{i/n}\beta(v)\mathrm{d}v\right)^2 - \sum_{i=1}^{\lfloor un\rfloor}\beta^2\left(\frac{i-1}{n}\right)\frac{1}{n}\right| \\
&= \left|\frac{1}{n}\sum_{i=1}^{\lfloor un\rfloor}\left[\left(n\int_{(i-1)/n}^{i/n}\beta(v)\mathrm{d}v\right)^2 - \beta^2\left(\frac{i-1}{n}\right)\right]\right| \\
&\leq \frac{2M}{n}\sum_{i=1}^{\lfloor un\rfloor}\left|n\int_{(i-1)/n}^{i/n}\beta(v)\mathrm{d}v - \beta\left(\frac{i-1}{n}\right)\right| \\
&\leq 2M\sum_{i=1}^{\lfloor un\rfloor}\int_{(i-1)/n}^{i/n}\left|\beta(v) - \beta\left(\frac{i-1}{n}\right)\right|\mathrm{d}v
\end{aligned}
$$

As β is of bounded variation, we may write $\beta = \beta_+ - \beta_-$ where both β_+ and β_- are increasing. We thus find

$$
\begin{aligned}
&\left| n\sum_{i=1}^{\lfloor un\rfloor}\left(\int_{(i-1)/n}^{i/n}\beta(v)\mathrm{d}v\right)^2 - \sum_{i=1}^{\lfloor un\rfloor}\beta^2\left(\frac{i-1}{n}\right)\frac{1}{n}\right| \\
&\leq 2M\sum_{i=1}^{\lfloor un\rfloor}\int_{(i-1)/n}^{i/n}\left|\beta_+(v) - \beta_+\left(\frac{i-1}{n}\right) - \left\{\beta_-(v) - \beta_-\left(\frac{i-1}{n}\right)\right\}\right|\mathrm{d}v \\
&\leq 2M\left[\sum_{i=1}^{\lfloor un\rfloor}\int_{(i-1)/n}^{i/n}\left|\beta_+\left(\frac{i}{n}\right) - \beta_+\left(\frac{i-1}{n}\right)\right|\mathrm{d}v + \sum_{i=1}^{\lfloor un\rfloor}\int_{(i-1)/n}^{i/n}\left|\beta_-\left(\frac{i}{n}\right) - \beta_-\left(\frac{i-1}{n}\right)\right|\mathrm{d}v\right] \\
&\leq \frac{2M}{n}\left[\beta_+(1) - \beta_+(0) + \beta_-(1) - \beta_-(0)\right],
\end{aligned}
$$

which converges to zero. Consequently, the first term in (34) converges to the limit of the

Riemann sums $\sum_{i=1}^{\lfloor un \rfloor} \left(\beta\left(\frac{i-1}{n}\right)(F-\lambda)\right)^2 \frac{1}{n}$, i.e., to $\int_{v=0}^{u} (\beta(v)(F-\lambda))^2 \, \mathrm{d}v$.

The second term in (34) converges given (28) and the third one in view of (21).

Finally, consider the last term in (34). By Cauchy-Schwarz and the previous results, we find

$$
\begin{aligned}
& n \sum_{i=1}^{\lfloor un \rfloor} \int_{(i-1)/n}^{i/n} \left[\beta(v) - \beta\left(\frac{i-1}{n}\right) \mathrm{d}v\right] \left[Z^{(n)}\left(\frac{i}{n}\right) - Z^{(n)}\left(\frac{i-1}{n}\right)\right] \\
\leq & \sqrt{n \sum_{i=1}^{\lfloor un \rfloor} \left(\int_{(i-1)/n}^{i/n} \beta(v) - \beta\left(\frac{i-1}{n}\right) \mathrm{d}v\right)^2} \\
& \times \sqrt{n \sum_{i=1}^{\lfloor un \rfloor} \left[Z^{(n)}\left(\frac{i}{n}\right) - Z^{(n)}\left(\frac{i-1}{n}\right)\right]^2}
\end{aligned}
$$

For increasing β, we may bound the first square-root further by

$$
\begin{aligned}
& \sqrt{n \sum_{i=1}^{\lfloor un \rfloor} \left(\int_{(i-1)/n}^{i/n} \beta\left(\frac{i}{n}\right) - \beta\left(\frac{i-1}{n}\right) \mathrm{d}v\right)^2} \\
\leq & \sqrt{\frac{1}{n} \sum_{i=1}^{\lfloor un \rfloor} \left(\beta\left(\frac{i}{n}\right) - \beta\left(\frac{i-1}{n}\right)\right)^2},
\end{aligned}
$$

which converges to zero. For general finite-variation β the same result again follows from writing it as the difference of two increasing functions. Consequently, the limit of the fourth term in (34) equals that of

$$
[F-\lambda] \sum_{i=1}^{\lfloor un \rfloor} \beta\left(\frac{i-1}{n}\right) \left[Z^{(n)}\left(\frac{i}{n}\right) - Z^{(n)}\left(\frac{i-1}{n}\right)\right] \tag{35}
$$

As $\sqrt{n} Z^{(n)}$ converges in law, the above expression converges to zero.

Taking these claims together, we find that

$$
S^{(n)}(u) - \int_{v=0}^{u} (\beta(v)(F-\lambda))^2 \, \mathrm{d}v + \int_{0}^{u} p(v) \mathrm{d}v - [Z, Z](u), \tag{36}
$$

converges to zero. In view of (22), this concludes the proof. □

As the squared return process $S^{(n)}$ satisfies an approximate factor structure, Proposition 1 immediately gives the following corollary.

Corollary 1 *If the excess return process $R^{(n)}$ satisfies a second-order approximate factor structure and (28) holds, then there exists a $K(K+1)/2$-dimensional vector of prices of risk δ and* *a K_S-dimensional vector of prices of risk η such that*

$$\varphi_0(u) - p(u) = vech\,(\beta(u)\beta^\intercal(u))^\intercal\,\delta + \varphi_1^\intercal(u)\eta. \tag{37}$$

Corollary 1 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 (37) gives the effect of the linear return factors F on 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 (37) also gives the consequences for pricing "idiosyncratic" variances in case some factors have been omitted in the linear return factor model. Clearly, in such 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 to the squared factor loadings $\text{vech}\,(\beta(u)\beta^\intercal(u))$ are given by a parameter δ that is unrelated to the prices of risk at the linear return factor model λ. An empirical advantage of this finding is that inference about the price of squared returns/idiosyncratic variances is not hampered by possibly weak identification of the price of risk λ.

The second term in (37), $\varphi_1^\intercal(u)\eta$, gives the pricing effect of common factors in truly idiosyncratic variances. Quadratic returns command a linear 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, i.e., linear in φ_1. Notice that the idiosyncratic variance factor G may be correlated with the linear return factors F or their squares. The no-arbitrage condition does neither impose nor exclude this.

In Section 5 we use standard Fama and MacBeth (1973) regressions to identify, in particular, the prices of risk η for the idiosyncratic variance factors using S&P500 stocks as described in Section 4.

4 Data

4.1 Sample construction

Each last trading day of the month for the period between January 1996 and December 2013, we extract the index constituents of the S&P500 index from Compustat (using ticker "I0003"). We merge the stock indentifying information with daily stock returns from CRSP using the WRDS linking table and compute cumulative 30-calendar day returns from month end to match the maturity of the OptionMetrics implied volatility surface detailed below. We merge the stock data from CRSP with the OptionMetrics standardized implied volatility surface for a maturity of 30 calendar days. The implied volatility surface contains smoothed implied volatilities for a standardized set of deltas ranging from -0.8 to -0.2 for puts and 0.2 to 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 the smoothed implied volatility is finite. We compute the stock's forward price using realised dividends over the life of the option from the OptionMetrics dividend file, discounted using the interpolated risk-free rates in the OptionMetrics zero-coupon yield file.

Call and put implied volatility smiles are not always identical, and the standardized call and put deltas yield slightly different implied strike prices, i.e., moneyness defined as the strike price over the forward price. 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 use the vector of average smoothed implied volatilities to compute implied volatilities for non-observed moneyness levels using linear interpolation. Outside the observed range of moneyness levels, we assume the implied volatility is constant at the endpoints in the observed data. We compute Black-Scholes option prices from the interpolated implied volatility curve for each stock-date combination.

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[6], a bond and a forward contract,

$$H(S) = H(K_0) + (S - K_0)H_S(K_0) + \mathrm{e}^{r\tau} \int_0^{K_0} H_{SS}(K)P(K)\mathrm{d}K \\ + \mathrm{e}^{r\tau} \int_{K_0}^{\infty} H_{SS}(K)C(K)\mathrm{d}K, \tag{38}$$

with K_0 a predetermined cut-off level separating the strike space into put and call options, r the continuously-compounded risk-free rate and τ the relevant maturity. We seek to compute the price of the discretely compounded squared excess return,

$$p_{it} = E^Q \left\{ R_{iT}^2 \right\} = E^Q \left\{ H(S) \right\},$$

with

$$H(S) = \left(\frac{S}{S_0} - \mathrm{e}^{r\tau} \right)^2, \tag{39}$$

so that

$$H_S(S) = \frac{2}{S_0} \left(\frac{S}{S_0} - \mathrm{e}^{r\tau} \right), \tag{40}$$

$$H_{SS}(S) = \frac{2}{S_0^2}. \tag{41}$$

Plugging (39)-(41) into (38), setting $K_0 = S_0$ and taking risk-neutral expectations, we obtain

$$p_{it} = (1 - \mathrm{e}^{r\tau})^2 + \frac{2}{S_0}(1 - \mathrm{e}^{r\tau}) + \frac{2}{S_0^2} \int_0^{S_0} P(K)\mathrm{d}K + \frac{2}{S_0^2} \int_{S_0}^{\infty} C(K)\mathrm{d}K, \tag{42}$$

which shows that for the squared excess return, each of the options in the replicating portfolio will be given the same weight. 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, which ensures that the integral is defined properly. Figure 1 plots the time series of the equal-weighted cross-sectional average price of squared excess returns

[6] Individual equity options are American rather than European. Since we use only out-of-the-money options, the early exercise premium will be small. Ofek, Richardson, and Whitelaw (2004) report a median early exercise premium equal to 70 bps for at-the-money put options. The bid-ask spread of those options is an order of magnitude larger than that.

of S&P500 stocks. Our final sample contains 894 different firms over the 216 months between January 1996 and December 2013.

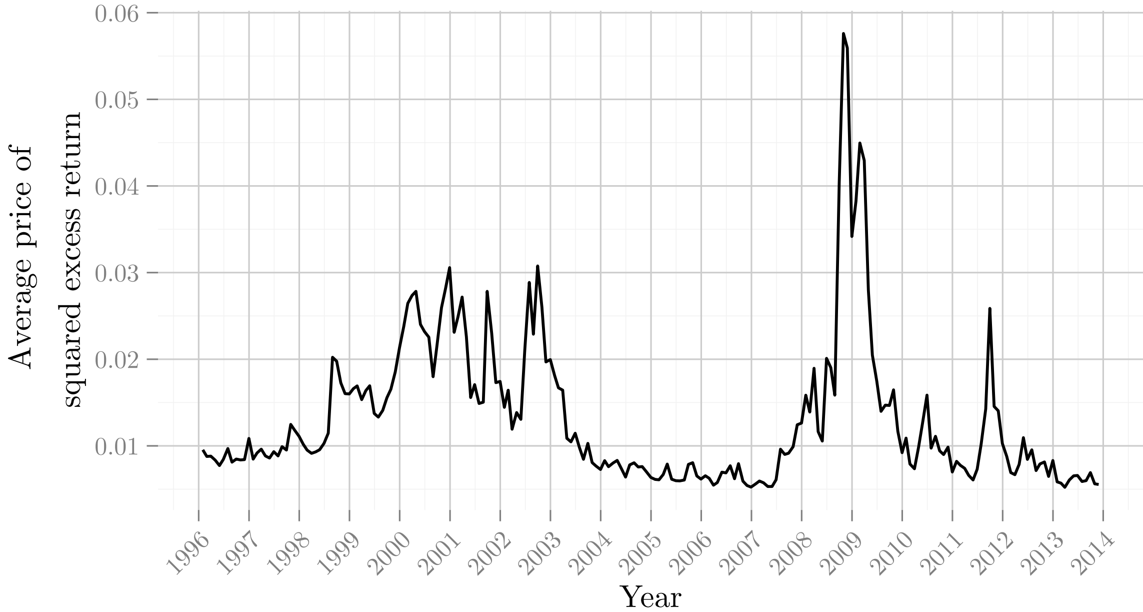

Figure 1: **Time series of cross-sectional average price of squared excess return of S&P500 stocks**

This figure plots the time series of the equal-weighted cross-sectional average price of the squared excess return of S&P500 stocks. The price is constructed each last trading day of the month using (42) for a standardized maturity of 30 calendar days. The sample period covers January 1996 to December 2013.

5 Squared return factors in S&P500 stocks

We extract statistical factors (F) using principal components on the panel of 30-calendar day excess returns on the 121 firms with a complete time series, and estimate the loadings (β) of each firm to each factor in a firm-level time series regression, assuming the factor loadings remain constant over the sample period. Figure 2 plots the cumulative fraction of the total variance of monthly linear excess returns explained as a function of the number of included principal components. A second principal components analysis on the squared residuals (ϵ^2) of the time-series regressions of all firms with a complete return time series identifies any additional squared return factors (G). A firm-level time-series regression of ϵ^2 on the factors G yields the factor loadings φ.

We fit an AR(1) model to the G time series, and retain the innovations. In the remainder, G refers to these innovations. We examine the correlation between G and F, as well as the coefficient estimate on G in a time-series regression of linear returns on F and G. If our model is correctly specified, the coefficient estimate on G should be zero for linear returns. However, as observed in Section 3, F and G may be correlated so the R^2 of a regression of linear returns on G does not have to be zero.

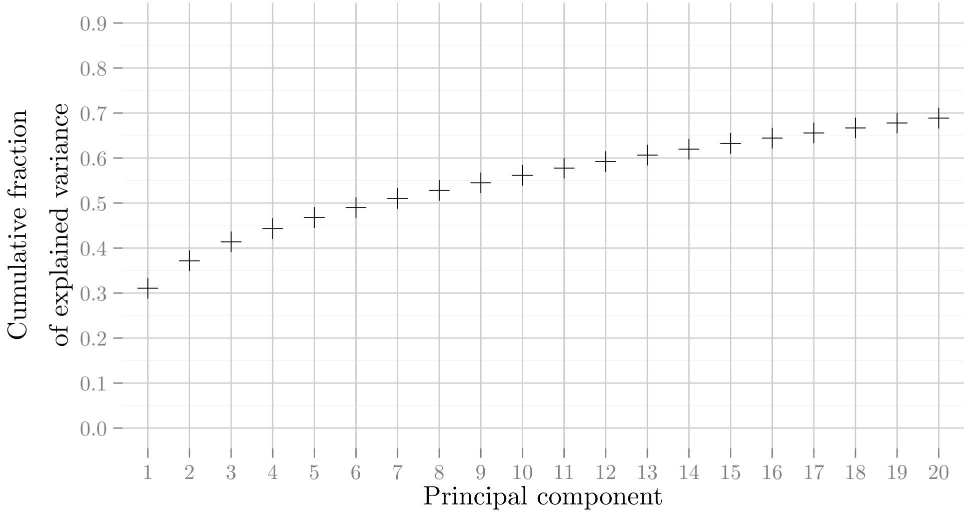

Figure 2: **Cumulative Fraction of Excess Return Variance Explained by Principal Components**

Cumulative fraction of variance explained by principal components of the stock-level excess returns R_{it}. Returns are measured over a 30-day horizon starting at the close of the last trading day of each month. Data is from CRSP, covering the period January 1996 to December 2013 (216 months) for the firms in the S&P500 index, 894 firms in total. The principal components are extracted from the time series of the 121 firms with a complete history over the sample period.

Recalling Section 3, the main prediction of our theoretical model is

$$\varphi_0(u) - p(u) = \text{vec}\left(\beta(u)\beta^\intercal(u)\right)^\intercal \delta + \varphi_1^\intercal(u)\eta. \tag{43}$$

We estimate this cross-sectional regression for each of the 216 months in our sample and repeat the above analyses for a different number (two, five or ten) F factors and a single G factor. To relate to the existing literature, we also run the same regressions using the five Fama and French (2015) factors in the linear returns model. Table 1 contains the results. The first line shows that there is a common factor in squared residual returns; the first principal component of the squared residual returns explains between 10 and 15% of the total variation, confirming the large body of literature on common factors in idiosyncratic volatility, e.g., Ang, Hodrick, Xing, and Zhang (2006) and Chen and Petkova (2012). The next part of the table examines Fama and MacBeth (1973) regressions. We report statistics on the R^2 of the cross-sectional regressions of the second-stage. The results confirm that the squared return factor G has some explanatory power for the linear returns, but that the additional explanatory power on top of the linear return factors F is small. Notably, the explanatory power of the squared return factor for linear returns is substantially higher when using residual returns from the Fama and French (2015) model to extract G. If only a small number of principal components are included, then the average loading of linear returns on G is significantly negative (the principal components are standardised to have zero mean and unit variance). This suggests the presence of an omitted factor in F. Including 10 principal components leaves the average loading of linear returns on G insignificantly different from zero.

Regardless of the number of principal components included in the first step, the last two rows of Table 1 show that the hypothesis that the risk premium on G equals zero cannot be rejected. The exception again is the model that uses the Fama and French (2015) factors rather than principal components. In this case, both the loadings of linear returns on G as well as the risk premium η are significantly different from zero, confirming the results in Herskovic, Kelly, Lustig, and Van Nieuwerburgh (Forthcoming).

Table 1: **Factor Models estimated on S&P500 Firm Returns**

This table reports model estimates from the analysis of monthly S&P500 firm returns over the period January 1996-December 2013 (216 months). Firstly, a varying number of principal components (F) is extracted from the 121 complete time series of linear returns. The column headers in the table refer to the number of principal components used. The last column uses the Fama and French (2015) factors rather than principal components. Secondly, the first principal component of the residual squared returns is extracted. The first row in the table shows the fraction of the total variance of the squared residual returns explained by this principal component. An AR(1)-model is then fitted to the principal component and the innovations retained as the factor (G). A Fama and MacBeth (1973) regression using the components (F) and/or (G) is then fitted to the return series of all 894 firms in the sample assuming a constant loading. Rows 2-6 in the table contain the average R^2 and its standard error of the second-stage cross-sectional regressions. Rows 7-8 contain the average loading to (G) across the 894 firms and its standard error. Finally, rows 10-11 contain the estimate of the price-of-risk η from (43). The statistical significance is represented by asterisks, where ***, **, and * represent significance at the 1%, 5%, and 10% levels, respectively. Standard errors are reported in parenthesis.

	2-PC	5-PC	10-PC	5-FF
Fraction explained by first PCA	0.151	0.124	0.095	0.112
Average R^2 with both F & G	0.208***	0.328***	0.504***	0.413***
(s.e.)	(0.010)	(0.011)	(0.010)	(0.013)
Average R^2 with only G	0.062***	0.038***	0.045***	0.174***
(s.e.)	(0.004)	(0.003)	(0.003)	(0.015)
Average R^2 with only F	0.173***	0.315***	0.485***	0.362***
(s.e.)	(0.011)	(0.011)	(0.010)	(0.013)
Average loading on G	−0.009***	−0.008***	0.001	0.009**
(s.e.)	(0.001)	(0.001)	(0.002)	(0.004)
η	−0.085	−0.030	−0.007	0.116***
(s.e.)	(0.054)	(0.048)	(0.034)	(0.009)

6 Summary and conclusions

We propose a new formulation of the classic Ross (1976) Arbitrage Pricing Theory which allows an extension to squared excess returns. For the set of S&P500 stocks over the period 1999-

2013, we document the presence of a common factor in residual volatilities of a linear factor model. However, while this factor appears to be priced when using the Fama and French (2015) factors in the linear return model, it is not priced when using principal components as factors. In a future version of the paper, we plan to include several extensions to the current analysis. Firstly, the one-period formulation presented in this paper can be easily extended to a multi-period setting. Unlike multi-period equilibrium-based asset pricing models, in which agents' demand for assets will generally be a combination of a speculative demand and a hedge demand, the multi-period APT only requires a period-by-period no-arbitrage condition as long as all our assets trade each period.

Secondly, the pricing kernel in our model for squared returns will be quadratic in the linear factors F and linear in the quadratic factors G. One of the criticisms of standard (linear) factor models is that the linear version of the pricing kernel can take on negative values. A linear-quadratic formulation opens up the possibility that the pricing kernel will always be positive, but it is an empirical question whether that indeed is true.

References

Ait-Sahalia, Y., and D. Xiu (2015): "Using Principal Component Analysis to Estimate a High Dimensional Factor Model with High-Frequency Data," .

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

——— (1999): "On the Robustness of Factor Structures to Asset Repackaging," *Journal of Mathematical Economics*, 31, 309–320.

Ang, A., R. J. Hodrick, Y. Xing, and X. Zhang (2006): "The Cross-Section of Volatility," *Journal of Finance*, 61(1), 259–299.

Bakshi, G., and D. Madan (2000): "Spanning and derivative-security valuation," *Journal of Financial Economics*, 55(2), 205–238.

Boloorforoosh, A. (2014): "Is Idiosyncratic Volatility Risk Priced ? Evidence from the Physical and Risk-Neutral Distributions," .

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

Chabi-Yo, F., D. P. Leisen, and E. Renault (2014): "Aggregation of preferences for skewed asset returns," *Journal of Economic Theory*, 154, 453–489.

Chen, Z., and R. Petkova (2012): "Does Idiosyncratic Volatility Proxy for Risk Exposure?," *Review of Financial Studies*, 25(9), 2745–2787.

Christoffersen, P., M. Fournier, and K. Jacobs (2015): "The Factor Structure in Equity Options," .

Dobi, D., and M. Avellaneda (2014): "Modeling Volatility Risk in Equity Options Market : A Statistical Approach," .

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 (2015): "A five-factor asset pricing model," *Journal of Financial Economics*, 116(1), 1–22.

Fama, E. F., and J. D. MacBeth (1973): "Risk, Return, and Equilibrium: Empirical Tests," *Journal of Political Economy*, 81(3), 607–636.

Gagliardini, P., E. Ossola, and O. Scaillet (2014): "Time-Varying Risk Premiums in Large Cross-Sectional Equity Datasets," Working Paper.

Harvey, C. R., and A. Siddique (2000): "Conditional Skewness in Asset Pricing Tests," *The Journal of Finance*, 55(3), 1263–1295.

Herskovic, B., B. Kelly, H. Lustig, and S. Van Nieuwerburgh (Forthcoming): "The Common Factor in Idiosyncratic volatility: quantitative asset pricing implications," *Journal of Financial Economics*.

Jacod, J., and A. N. Shiryaev (2003): *Limit Theorems for Stochastic Processes*, vol. 288 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Kraus, A., and R. H. Litzenberger (1976): "Skewness Preference and the Valuation of Risk Assets," *The Journal of Finance*, 31(4), 1085–1100.

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.

Ross, S. A. (1976): "The arbitrage theory of capital asset pricing," *Journal of Economic Theory*, 13(3), 341–360.

Stambaugh, R. F., J. Yu, and Y. Yuan (2015): "Arbitrage Asymmetry and the Idiosyncratic Volatility Puzzle," *Journal of Finance*.