Arbitrage Pricing Theory for Idiosyncratic Variance Factors

PRELIMINARY AND INCOMPLETE: PLEASE DO NOT CITE WITHOUT PERMISSION

Eric RENAULT*, Thijs VAN DER HEIJDEN†, and Bas J.M. WERKER‡

February 7, 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 (2016). 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 classical 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 theoretical advantage of our approach is twofold. On the one hand, by considering economies with a continuum of assets, we share, with Al-Najjar (1998)'s seminal work, 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 will be shown to encompass the concept of approximate factor structure recently put forward by Gagliardini, Ossola, and Scaillet (2014). On the other hand, our approach in terms of stochastic processes (where the continuous "time" index is actually an index for a specific portfolio) allows us to resort to the theory of quadratic variations and related tools. In short, our new formulation of an approximate factor structure easily extends the APT for linear returns to squared returns and thus to (idiosyncratic) volatilities.

In contrast to Ang, Hodrick, Xing, and Zhang (2006), Duarte, Kamara, Siegel, and Sun (2014) and Herskovic, Kelly, Lustig, and Van Nieuwerburgh (2016), 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 disentangle the effect of possibly omitted factors at the linear return level from possible factors in idiosyncratic variances. On the one hand, if a factor is forgotten in the factor pricing of linear returns, obviously its squared value will show up as a common factor of idiosyncratic volatilities. However, there is no argument, either theoretical or empirical, that prevents new factors (i.e., unrelated to the factors at the linear return level) to show up in idiosyncratic variances.

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 analyze 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. However, the focus of our paper to document whether the price of risk of the squared return factor is, economically and statistically, different from zero. We therefore focus on excess squared excess returns, i.e., squared excess returns minus their price. 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 value 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. By using this set of excess squared excess returns, we are able to check that, irrespective of the number of principal components that are included (namely 5 or 10), 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 empirical results shed new light on the idiosyncratic volatility puzzle in at least three respects. First, in contrast with several papers in the extant literature, our focus of interest is beyond the role of possible forgotten factors in the linear factor model of standard asset returns. By using up to ten principal components, we make sure that no common statistical factor is missing and we still find a common factor in squared idiosyncratic residuals. Obviously, this common factor captures less (about ten percent) of the total variance of squared residuals than when using Fama-French type factors. In the latter case, people find numbers up to 30 percent or more. See, e.g., Duarte, Kamara, Siegel, and Sun (2014) and Herskovic, Kelly, Lustig, and Van Nieuwerburgh (2016). Second, 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 option prices to obtain a price 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. Third, 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 to an extension of the standard Capital Asset Pricing Model that incorporates skewness as the covariance between asset returns 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 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. In other words, 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 one due to the tracking error on the squared market return.

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 common factors in idiosyncratic variances, the main topic of this paper. At the level of 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 actually call any investment with zero price an excess return.

In order to formalize the assumption of an approximate factor structure, we construct cumulative portfolios. That is, for given $u \in [0, 1]$, we construct an equally weighted portfolio consisting of $1/n$ exposures in the first u fraction of the assets.[1] Such a cumulative 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 simply an alternative representation of the available assets $R_i^{(n)}$ in the market. We have $R^{(n)}(0) = 0$ and $R^{(n)}(1)$ represents an equally weighted portfolio in all available assets in the economy. Formally, $R^{(n)}$ is 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 $(D[0, 1], \|\cdot\|)$.

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, the factors, or the idiosyncratic errors.[2]

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.

Our formulation of a factor model is, technically, of a different nature than existing results in the literature. The following remark shows that our setup encompasses recently proposed alternatives.

[1] In Remark 1 we extend this to non-equally weighted portfolios.

[2] In particular, it is not even assumed at this stage that the errors $Z^{(n)}$ and the factor F are uncorrelated; see Remark 5.

Remark 1 *Several other formalizations 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 sequences of excess returns written as*

$$R_i^{(n)} = \alpha_i^{(n)} + \beta_i^{(n)\intercal} F + \varepsilon_i^{(n)}, \tag{3}$$

with $E\varepsilon_i^{(n)} = 0$. *Now, portfolios with excess returns* $R^{(n)}(u)$ *can be defined as above. The setting in Gagliardini, Ossola, and Scaillet (2014) is arguably more general in the sense that the portfolios considered are not necessarily equally weighted. This can be included in our setting by considering weights* $\delta_i^{(n)} > 0$ *satisfying*

$$\sum_{i=1}^{n} \delta_i^{(n)} = 1, \tag{4}$$

$$n \sum_{i=1}^{n} \left(\delta_i^{(n)}\right)^2 = O(1). \tag{5}$$

The condition that the weights $\delta_i^{(n)}$ *sum to one is made for convenience only and immaterial as we consider excess returns. The second condition is reminiscent of the concept of bounded Asymptotic Quadratic Variation in Time (AQVT) introduced in Mykland and Zhang (2006). To see that, note that the weights* $\delta_i^{(n)}$ *may be seen as the lengths of consecutive intervals*[3] $(\Delta_{j-1}^{(n)}, \Delta_j^{(n)}]$ *in* $[0, 1]$, *i.e.,*

$$\Delta_j^{(n)} = \sum_{i=1}^{j} \delta_i^{(n)}, \quad j = 1, \ldots, n,$$

with $\Delta_0^{(n)} = 0$. *These weights can be used to construct a continuum of portfolios much like (1). More precisely we define*

$$R_\Delta^{(n)}(u) = \sum_{i=1}^{n} R_i^{(n)} \delta_i^{(n)} 1_{[\Delta_i^{(n)} \le u]}. \tag{6}$$

Note that conditions (4) and (5) are in particular fulfilled for the weights $\delta_i^{(n)} = 1/n$, *leading to the equally weighted portfolio* $R_\Delta^{(n)}(u) = R^{(n)}(u)$. *In general,* $R_\delta^{(n)}(u)$ *is a stochastic process (with sample paths in* $D[0, 1]$*), decomposes into*

$$R_\Delta^{(n)}(u) = \sum_{i=1}^{n} \alpha_i^{(n)} \delta_i^{(n)} 1_{[\Delta_i^{(n)} \le u]} + \sum_{i=1}^{n} \beta_i^{(n)\intercal} F \delta_i^{(n)} 1_{[\Delta_i^{(n)} \le u]} + \sum_{i=1}^{n} \varepsilon_i^{(n)} \delta_i^{(n)} 1_{[\Delta_i^{(n)} \le u]}.$$

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

[3] It is tempting to call the "time intervals". However, note that the index u does not indicate time but the fraction of assets included in a portfolio.

and β defined by

$$\sum_{i=1}^{n} \alpha_i^{(n)} \delta_i^{(n)} 1_{[\Delta_i^{(n)} \leq u]} \rightarrow \int_0^u \alpha(v) dv \tag{7}$$
$$\sum_{i=1}^{n} \beta_i^{(n)} \delta_i^{(n)} 1_{[\Delta_i^{(n)} \leq u]} \rightarrow \int_0^u \beta(v) dv \tag{8}$$

These conditions impose sufficient stability on the intercepts α and the factor loadings β. Intuitively, these functions α and β are approximatively given by

$$u \in (\Delta_{i-1}^{(n)}, \Delta_i^{(n)}] \Rightarrow \alpha(u) \approx \alpha_i^{(n)} \text{ and } \beta(u) \approx \beta_i^{(n)}. \tag{9}$$

The additional generality obtained by allowing non-equal weights may be, from a practical point of view, limited. Indeed, Theorem 1 below shows that they are not needed to obtain the classical Arbitrage Pricing Theory. In order to keep in line with the existing literature, we now show that they do play a role in the interpretation of assumptions needed to get convergence to zero of the residual process

$$Z_\Delta^{(n)}(u) = \sum_{i=1}^{n} \varepsilon_i^{(n)} \delta_i^{(n)} 1_{[\Delta_i^{(n)} \leq u]}.$$

One easily verifies

$$Var\left\{Z_\Delta^{(n)}(u)\right\} \leq n \left(\sum_{i=1}^{n} \left(\delta_i^{(n)}\right)^2\right) \frac{\rho\left(Var\left\{\left[\varepsilon_i^{(n)}\right]_{i=1}^{n}\right\}\right)}{n},$$

where $\rho(A)$ stands for the maximum eigenvalue of a symmetric matrix A. Now, Assumption APR3 in Gagliardini, Ossola, and Scaillet (2014) is akin to assuming

$$\frac{\rho\left(Var\left\{\left[\varepsilon_i^{(n)}\right]_{i=1}^{n}\right\}\right)}{n} \rightarrow 0,$$

so that the AQVT assumption (5) implies

$$\sup_{u \in [0,1]} Var\left\{Z_\Delta^{(n)}(u)\right\} \rightarrow 0.$$

As a result, the conditions in Definition 1 are satisfied.

In other words, up to reweighing that is immaterial in our setting as explained above, our definition of approximate factor structure encompasses the setting of Gagliardini, Ossola, and Scaillet (2014) as a particular case. The convergence derived above corresponds to Lemma 13 in Appendix 3 of Gagliardini, Ossola, and Scaillet (2014). They stress *that reweighing matters for them because it allows them to formalize the concept of a block dependence structure. Such a structure may be empirically relevant, for instance, in the case of unobserved industry specific factors that are independent among industries . Since reweighing is mathematically immaterial in our setting, we will throughout focus on equally weighted portfolios in our theoretical derivations.*

It is worth acknowledging that 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.

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

Example As we are particularly interested in the pricing of idiosyncratic variance factors, we consider a standard stochastic volatility model. For simplicity, we focus on a single factor ($K = 1$). Consider

$$R_i = \alpha_i + \beta_i F + (\omega_i + \varphi_i G)^{1/2} \nu_i, \tag{10}$$

for constants α_i, β_i, ω_i, and φ_i 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 assume the ω_i and φ_i to be bounded away from zero and infinity.

Under the regularity conditions (7) and (8) on the α and β (in the context of equally weighted portfolios, i.e., $\delta_i^{(n)} = 1/n$), we get an approximate factor structure with

$$Z^{(n)}(u) = \frac{1}{n} \sum_{i=1}^{\lfloor un \rfloor} (\omega_i + \varphi_i G)^{1/2} \nu_i + o(1). \tag{11}$$

Then, the functional law of large numbers gives the required convergence of the process $Z^{(n)}$ to zero. Clearly, this result relies on the assumed cross-sectional independence of the idiosyncratic errors ν_i. We will not provide details here as this law of large numbers is an immediate consequence of the (functional) central limit theorem we apply to verify the conditions of Definition 2 below. $\square$

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

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

Taking $h(u) = 1$, we would find the excess return of an equally weighted portfolio with exposures $1/n$ to all n 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.

Remark 2 - Factor-mimicking portfolios *If the excess returns $R^{(n)}$ satisfy an approximate factor structure, we can define a K-dimensional function H of finite variation on $[0, 1]$ such that*

$$\int_{u=0}^{1} H(u)\beta^{\intercal}(u)\,du = I_K, \tag{13}$$

the $K \times K$ identity matrix. This is possible as long as the components of β are linearly independent.[4] *Then the K portfolios induced by H, i.e.,*

$$\tilde{F} = \int_{u=0}^{1} H(u)\,dR^{(n)}(u), \tag{14}$$

can also be used as factors. To see that, note that (2) implies

$$\tilde{F} = \int_{0}^{1} \alpha(u)H(u)\,du + F + \int_{0}^{1} H(u)\,dZ^{(n)}(u).$$

Hence, we may write, again using (2),

$$\begin{aligned}
R^{(n)}(u) &= \int_0^u \alpha(v)\,dv \\
&\quad + \int_0^u \beta^{\intercal}(v)\,dv \left[\tilde{F} - \int_0^1 \alpha(w)H(w)\,dw - \int_0^1 H(w)\,dZ^{(n)}(w)\right] + Z^{(n)}(u) \\
&= \int_0^u \left[\alpha(v) - \beta^{\intercal}(v)\int_0^1 \alpha(w)H(w)\,dw\right] dv + \int_0^u \beta^{\intercal}(v)\,dv\tilde{F} \\
&\quad + \left[Z^{(n)} - \int_0^u \beta^{\intercal}(v)\,dv \int_0^1 H(w)\,dZ^{(n)}(w)\right],
\end{aligned}$$

where the last term indeed converges to zero. Observe that switching to factor mimicking

[4]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 $\langle f, g\rangle = \int_0^1 f(u)g(u)\mathrm{d}u$.

portfolios in this way does not affect the factor loadings β, but the intercept α is affected.

Remark 3 - 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 with respect to forming portfolios. It's easy to see that our Definition 1 indeed obeys to this invariance.*

Consider first a reordering of the assets. Note that (2) implies, for given $w \in [0, 1]$,

$$R^{(n)}(u) - R^{(n)}(w) = \int_w^u \alpha(v)\,dv + \int_w^u \beta^\intercal(v)\,dv + Z^{(n)}(u) - Z^{(n)}(w). \tag{15}$$

Now consider $p+1$ fixed constants $0 = u_0 < u_1 < \ldots < u_p = 1$. A reordering of assets can 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 Theorem 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.

We can now state our version of the APT; its proof can be found in the appendix.

Theorem 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*

$$\alpha(u) = -\beta(u)^\intercal\lambda, \tag{16}$$

up to a set of Lebesgue measure zero.

We end this section's recollection of the Arbitrage Pricing Theory with a few remarks. These are not unique to our setting, but revisit some discussions in the vast literature on the APT.

Remark 4 - Factor 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. We return to it, from an empirical point of view, in Section 5.

First note that 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, i.e., such that the variance matrix of F is non-singular.[5] *However, this does not exclude that a strictly smaller factor space would not also be valid. Indeed, assume that one of the components of the factor loadings β is a linear combination of the other components; for instance suppose that, for almost every $u \in [0,1]$, we have*

$$\beta_1(u) = \sum_{k=2}^{K} \zeta_k \beta_k(u).$$

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, we will maintain throughout 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 is zero (a.e.). Under this maintained assumption, it is not possible to write an approximate factor structure with less than K factors.

Conversely, 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.,

$$R^{(n)}(u) = \int_0^u \alpha(v)\,dv + \int_0^u \beta^{\intercal}(v)\,dvF + \int_0^u \beta_o^{\intercal}(v)\,dvF_o + Z^{(n)}(u), \tag{17}$$

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 $\int_0^u \beta_o^{\intercal}(v)\,dvF_o + Z^{(n)}(u)$. This will only converge to zero if $\beta_o = 0$. Consequently, Definition 1 precisely identifies the correct number of factors.

Remark 5 - Identification of factors

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 of F and $Z^{(n)}$, but merely that the process $Z^{(n)}$ vanish asymptotically. Actually, observe that orthogonality of factors and idiosyncratic terms is assumed in Gagliardini, Ossola, and Scaillet (2014), though never used in their proofs. However, in case the convergence

[5] For sake of expositional simplicity we assume here that this variance matrix exists.

to zero of the process $Z^{(n)}$ is also uniform in L^2, i.e.,

$$\sup_{u\in[0,1]} Var\left\{Z^{(n)}(u)\right\} \to 0,$$

we obtain by Cauchy-Schwartz

$$\sup_{u\in[0,1]} \left|Cov\left\{F, Z^{(n)}(u)\right\}\right| \to 0. \tag{18}$$

As, by definition,

$$Z^{(n)}(u) \approx \frac{1}{n}\sum_{i=1}^{[un]}\left\{R_i^{(n)} - \alpha_i^{(n)} - \beta_i^{(n)\intercal}F\right\} = \frac{1}{n}\sum_{i=1}^{[un]}\varepsilon_i^{(n)},$$

it is natural to impose that $\varepsilon_i^{(n)}$ is not correlated with F. In that case, we can interpret (2) as a regression equation. Of course, (18) does not imply that $\varepsilon_i^{(n)}$ is uncorrelated with F for every individual asset i. It is however standard practice in the empirical literature to assume the idiosyncratic errors to be uncorrelated with the factors for each individual asset. We also impose this identifying condition in Section 5. However, note that this practice has been criticized, going as far as questioning the testability of Ross (1976)'s APT. Actually, the approach of characterizing the APT in an economy with a continuum of assets, as in Al-Najjar (1998) and Gagliardini, Ossola, and Scaillet (2014), is precisely motivated by this criticism. Our paper addresses the issue by looking at cumulative portfolios of assets.

Besides its conceptual relevance, Remark 5 also refers to the estimation of factor models using time-series regressions. A criticism of single-period equilibrium models is that they do generally not readily extend to multiple periods as multiperiod asset demands will generally contain hedge demands as well. It's useful to observe that such criticism does not hold for a multi-period application of the APT. As the APT is based on a no-arbitrage assumption, its conclusions extend to multi-period settings as long as the no-arbitrage assumption is imposed each period. Once the factor loadings β have been identified by a time-series regression, Theorem 1 can be applied period-by-period. Clearly, both factor loadings β and prices of risk λ may become time-varying in that case.

For future reference, we recall the construction of a pricing kernel in the APT setting. We formulate, as usual, the kernel in a setting where (2) has a regression interpretation.

Remark 6 - Pricing kernel
If asset i's idiosyncratic errors $\varepsilon_i^{(n)}$ are uncorrelated with the factors F, Theorem 1 implies the regression relationship

$$R_i^{(n)} = \beta_i^{(n)\intercal}(F-\lambda) + \varepsilon_i^{(n)}, \tag{19}$$

with $E\varepsilon_i^{(n)} = 0$ *and* $Cov\left\{F, \varepsilon_i^{(n)}\right\} = 0$.

Then, a pricing kernel $M^{(n)}$ *satisfies, for all* $i = 1, \ldots, n$,

$$E\left\{R_i^{(n)} M^{(n)}\right\} = 0. \tag{20}$$

Relation (20) holds if and only if, for all $i = 1, \ldots, n$,

$$\beta_i^{(n)\intercal} E\left\{(F - \lambda) M^{(n)}\right\} + Cov\left\{\varepsilon_i^{(n)}, M^{(n)}\right\} = 0. \tag{21}$$

Following standard arguments we find that the APT pricing relation (19) is tantamount to the existence of a pricing kernel $M^{(n)}$ *that is affine in the factor* F. *Indeed, with* $M^{(n)} = a + b^\intercal F$, *we find that the (gross) risk-free rate* R_F *satisfies*

$$\frac{1}{R_F} = E\left\{M^{(n)}\right\} = a + b^\intercal E\left\{F\right\}.$$

Moreover, from (21), we find that a *and* b *solve*

$$Var\{F\}\, b + (E\{F\} - \lambda)\left(a + b^\intercal E\{F\}\right) = Var\{F\}\, b + \frac{E\{F\} - \lambda}{R_F} = 0.$$

The price of the factor F *itself is thus given by*

$$E\left\{F M^{(n)}\right\} = Var\{F\}\, b + \frac{E\{F\}}{R_F} = \frac{\lambda}{R_F}. \tag{22}$$

In other words, the parameter λ *can always be interpreted as the price of the future payoff* $R_F F$. *In case* F *is itself an excess return, we thus must have* $\lambda = 0$. *When* F *is extracted using statistical methods,* F *will generally be an affine transformation of excess returns, usually standardized to have zero mean and unit variance. In that case* λ *is not necessarily zero and* $F - \lambda$ *has zero price, i.e., is an excess return.*

3 The APT for variance factors

The main theoretical contribution of the present paper is to provide Arbitrage Pricing Theory implications for *squared* excess returns. These will subsequently be used to derive pricing implications for idiosyncratic volatility factors . The strategy we follow is to 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 using Theorem 1.

Consider as before excess returns $R_i^{(n)}$, for $i = 1, \ldots, n$, that induce the cumulative portfolios returns $R^{(n)}(u)$ as in (1). As $R^{(n)}$ is piecewise constant, its quadratic variation

$\left[R^{(n)}, R^{(n)}\right]$ satisfies

$$n\left[R^{(n)}, R^{(n)}\right](u) = \frac{1}{n}\sum_{i=1}^{\lfloor un \rfloor}\left(R_i^{(n)}\right)^2. \tag{23}$$

Obviously, squared excess returns are not excess returns themselves. However, they can be seen as payoffs on traded assets. In our empirical analysis we will use plain vanilla options traded on individual assets to reconstruct the no-arbitrage price of squared excess returns using a well-known technique going back to Breeden and Litzenberger (1978) and Bakshi and Madan (2000). We denote the option-induced no-arbitrage market price of the squared excess return $\left(R_i^{(n)}\right)^2$ by $p_i^{(n)}$. Then, we can define the *excess squared excess returns*[6] as

$$S_i^{(n)} = \left(R_i^{(n)}\right)^2 - p_i^{(n)}, \tag{24}$$

Applying the notation (1) to these excess squared excess returns leads to what we call the *squared return* process[7] as

$$S^{(n)}(u) = \frac{1}{n}\sum_{i=1}^{\lfloor un \rfloor} S_i^{(n)} = \left[\sqrt{n}R^{(n)}, \sqrt{n}R^{(n)}\right](u) - \frac{1}{n}\sum_{i=1}^{\lfloor un \rfloor} p_i^{(n)}. \tag{25}$$

As

$$\sqrt{n}R^{(n)}(u) = \sqrt{n}\int_0^u \alpha(v)\mathrm{d}v + \sqrt{n}\int_0^u \beta^\intercal(v)\mathrm{d}vF + \sqrt{n}Z^{(n)}(u)$$

one already gets the intuition that our strengthening of Definition 1 lies in assuming a central limit type of convergence. To make this precise, we assume, in line with the notations above, that we may write

$$\frac{1}{n}\sum_{i=1}^{\lfloor un \rfloor} p_i^{(n)} \to \int_0^u p(v)\mathrm{d}v, \tag{26}$$

for a deterministic finite-variation function p.

This discussion leads to the following strengthening of Definition 1 which will allow us to analyze the pricing of factor structures in (idiosyncratic) variances.

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*

$$\begin{aligned} \sqrt{n}Z^{(n)}(u) &\overset{\mathcal{L}}{\to} Z(u), & (27)\\ \left[\sqrt{n}Z^{(n)}, \sqrt{n}Z^{(n)}\right](u) &\overset{\mathcal{L}}{\to} [Z,Z](u), & (28) \end{aligned}$$

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

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

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

$$[Z, Z](u) = \int_0^u \omega(v)\, dv + \int_0^u \varphi^{\intercal}(v)\, dv G, \tag{29}$$

for deterministic finite-variation functions ω and φ 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 (27) to imply (28) 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 (11), i.e., the convergence of

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor un \rfloor} (\omega_i + \varphi_i G)^{1/2} \nu_i, \tag{30}$$

and its limiting quadratic variation. Moreover, we need a reinforcement of (7)–(8) to[8]

$$\frac{1}{n} \sum_{i=1}^{\lfloor un \rfloor} \alpha_i^{(n)} - \int_0^u \alpha(v) \mathrm{d}v = o\left(\frac{1}{\sqrt{n}}\right), \tag{31}$$

$$\frac{1}{n} \sum_{i=1}^{\lfloor un \rfloor} \beta_i^{(n)} - \int_0^u \beta(v) \mathrm{d}v = o\left(\frac{1}{\sqrt{n}}\right). \tag{32}$$

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

$$\frac{1}{n} \sum_{i=1}^{\lfloor un \rfloor} \omega_i \rightarrow \int_0^u \omega(v) \mathrm{d}v, \tag{33}$$

$$\frac{1}{n} \sum_{i=1}^{\lfloor un \rfloor} \varphi_i \rightarrow \int_0^u \varphi(v) \mathrm{d}v, \tag{34}$$

for finite-variation functions ω and φ, we find

$$[Z, Z](u) = \int_0^u \omega(v) \mathrm{d}v + G \int_0^u \varphi(v) \mathrm{d}v, \tag{35}$$

As a result, the conditions in Definition 2 are satisfied. $\square$

[8] With the interpretation (9), these assumptions amount to saying that the intercept α and the factor loadings β are Hölder-continuous of a degree strictly larger than 1/2.

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

Theorem 2 *If the excess return process $R^{(n)}$ satisfies a second-order approximate factor structure and (26) holds, then the squared return process $S^{(n)}$ satisfies an approximate factor structure with α given by*

$$\alpha_S(u) = \omega(u) - p(u), \tag{36}$$

factors[9] $vech\left([F-\lambda]\,[F-\lambda]^\intercal\right)$*, with loadings*

$$\beta_{SF}(u) = vech\left(\beta(u)\beta^\intercal(u)\right), \tag{37}$$

and additional factors G with loadings

$$\beta_{SG}(u) = \varphi(u). \tag{38}$$

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

Corollary 1 *If the excess return process $R^{(n)}$ satisfies a second-order approximate factor structure and (26) 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*

$$\omega(u) - p(u) = -vech\left(\beta(u)\beta^\intercal(u)\right)^\intercal \delta - \varphi^\intercal(u)\eta. \tag{39}$$

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 (39) 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 (39) 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 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.

[9] 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.

The second term in (39), $\varphi^\intercal(u)\eta$, 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. 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.

Our main theoretical result in Theorem 2 can also be written in the form of a beta-pricing relationship for return variance, much akin the standard beta-pricing relation for expected returns. We explain this is the setting of our example that also forms the basis of our empirical study in Section 5.

Example continued Using Theorem 1, we again find

$$R_i = \beta_i^\intercal (F - \lambda) + \varepsilon_i,$$

with

$$\varepsilon_i = (\omega_i + \varphi_i G)^{1/2} \nu_i.$$

This immediately implies

$$\operatorname{Var}\{R_i\} = \beta_i^\intercal \operatorname{Var}\{F\} \beta_i + \omega_i + \varphi_i^\intercal E\{G\}.$$

Assume, similarly to the assumption discussed in Remark 5, that the pricing relationship (39) is valid for each asset individually, i.e.,

$$\omega_i - p_i = -\operatorname{vech}(\beta_i \beta_i^\intercal)^\intercal \delta - \varphi_i^\intercal \eta.$$

Then, we have for each asset,

$$\begin{aligned} E\{R_i\} &= \beta_i^\intercal E\{F - \lambda\}, \\ \operatorname{Var}\{R_i\} - p_i &= \operatorname{vech}(\beta_i \beta_i^\intercal)^\intercal [\operatorname{vech}(\operatorname{Var}\{F\}) - \delta] + \varphi_i^\intercal E\{G - \eta\}. \end{aligned}$$

The first equation is the standard beta pricing relation for expected excess returns. Similarly, the second equation gives a beta-type pricing relation for the variance of excess returns. When seen as an expected excess return, after subtracting the price p_i of the squared return, the excess return variance displays a beta-type pricing relation with coefficients φ_i vis-à-vis the variance factors G. Note however the correction with the premium on the variance of the factors. The various prices of risk can also be understood through the study of the induced pricing kernel . □

Remark 7 - Dynamic Factor Models *As mentioned before, our pricing results can be used as well in a dynamic setting in which case the moments of interest would be conditional on the relevant past information. While conditional beta pricing has a long history , conditional factor models have also been used to get more parsimonious models of conditional variances of returns . This paper is one of the first to bridge the gap between these two strands of literature. Typically, a model like (10) provides a parsimonious model for the conditional variance of returns. In a more general formulation, we may write the conditional variance of excess returns as*

$$Var_{t-1}\left\{(R_{it})_{i=1}^{n}\right\} = \Omega + \Phi^{\intercal} E_{t-1}\left\{GG^{\intercal}\right\}\Phi.$$

where Ω and Φ are coefficient matrices describing the (joint) volatility dynamics of the excess returns. As a result, the idiosyncratic variance factors G, besides their role in pricing nonlinear derivatives, also provide information on the volatility dynamics that are not already captured by the factors F at the linear excess return level.

Before turning to standard Fama and MacBeth (1973) regressions and GMM methods to identify, in particular, the prices of risk η for the idiosyncratic variance factors, we conclude this section with a remark concerning the price of skewness and kurtosis.

Remark 8 - Pricing skewness and kurtosis *The pricing kernel M has so far been characterized using no arbitrage conditions only. However, it is possible to come up with economic interpretations of M based on the characterization of an equilibrium in an economy with an exogenous supply of risky assets. It is well known that if investors only care about the mean and the variance of returns, the only relevant pricing factor is the excess market return, say $F = R_M - R_F$. However, if agents also care about skewness, it is known that the squared excess market return becomes a relevant factor as well . This would imply a pricing kernel of the form*

$$M = a + b\left(R_M - R_F\right) + c\left[\left(R_M - R_F\right)^2 - \Pi\left(R_M - R_F\right)^2\right],$$

where Π denotes the pricing operator $\Pi X = E\{XM\}$.

Based on small noise expansions, Chabi-Yo, Leisen, and Renault (2014) have shown that the value of b and c would be

$$b = -\frac{1}{R_F}\frac{1}{\tau}, \quad c = \frac{1}{R_F}\frac{\rho}{\tau^2},$$

where τ and ρ stand for an average (across investors) of risk tolerances and skewness tolerances, respectively. However, Chabi-Yo, Leisen, and Renault (2014) also point out that when investors have heterogeneous preferences for skewness, an additional factor G

should be introduced

$$M = a + b\left(R_M - R_F\right) + c\left[\left(R_M - R_F\right)^2 - \Pi\left(R_M - R_F\right)^2\right] + dG.$$

In terms of excess return, we have

$$G = (R_M - R_F)(R_{sk} - R_F) - \Pi(R_m R_{sk}) + 2R_F - R_F^2$$

where R_{sk} *is a portfolio return defined from the affine regression of* R_M^2 *on the "linear" returns* R_i*,* $i = 1, ...n$*, i.e.,*

$$Cov\left[R_M^2, (R_i)_{1\leq i\leq n}\right]\left[Var\left\{(R_i)_{1\leq i\leq n}\right\}\right]^{-1}(R_i)_{1\leq i\leq n} = A + BR_{sk}.$$

In other words, it is precisely because the squared market return cannot be perfectly traded with linear portfolios that an additional factor shows up, that does not coincide with the cubic market return R_M^3*, usually introduced to price kurtosis. Note that the coefficient of* G *in the pricing kernel* M *is precisely non zero because skewness tolerances are heterogeneous across investors*

$$d = \frac{2}{R_F}\frac{Var(\rho)}{\tau^3},$$

where $Var(\rho)$ *stands for the cross-sectional variance (across investors) of skewness tolerances. Note that this remark does not preclude the introduction within* G *(and thus also in the pricing kernel) an additional factor* R_M^3 *precisely focused on pricing kurtosis. Chabi-Yo, Leisen, and Renault (2014) show that its coefficient in the pricing kernel will be proportional to an average kurtosis tolerance. This is in line with pricing the risk encapsulated in squared returns, that is the variance of squared returns.*

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[10], a bond and a forward contract,

$$\begin{aligned} H(S) = H(K_0) + (S - K_0)H_S(K_0) + \mathrm{e}^{r\tau} \int_0^{K_0} H_{SS}(K)P(K)\mathrm{d}K \\ + \mathrm{e}^{r\tau} \int_{K_0}^{\infty} H_{SS}(K)C(K)\mathrm{d}K, \end{aligned} \tag{40}$$

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

$$H(S) = \left(\frac{S}{S_0} - \mathrm{e}^{r\tau} \right)^2, \tag{41}$$

[10]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.

so that

$$H_S(S) = \frac{2}{S_0}\left(\frac{S}{S_0} - \mathrm{e}^{r\tau}\right), \tag{42}$$

$$H_{SS}(S) = \frac{2}{S_0^2}. \tag{43}$$

Plugging (41)-(43) into (40), setting $K_0 = S_0$ and taking risk-neutral expectations, we obtain

$$p_{it} = (1-\mathrm{e}^{r\tau})^2 + \frac{2}{S_0}(1-\mathrm{e}^{r\tau}) + \frac{2}{S_0^2}\int_0^{S_0} P(K)\mathrm{d}K + \frac{2}{S_0^2}\int_{S_0}^{\infty} C(K)\mathrm{d}K, \tag{44}$$

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

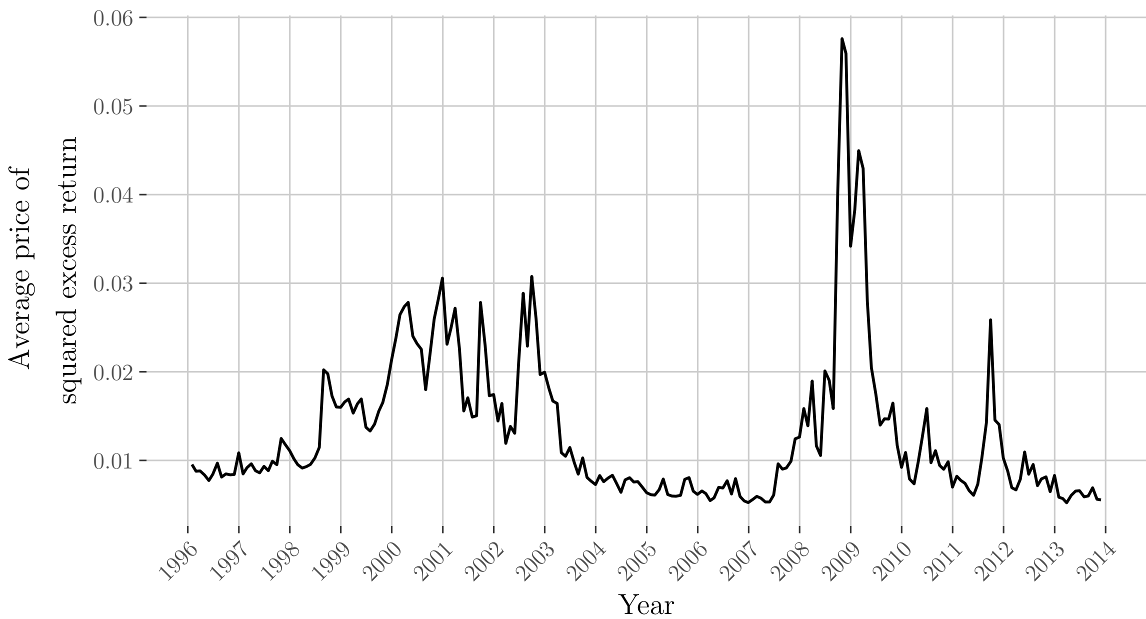

Figure 1: **Time series of cross-sectional average price of squared excess return of S&P500 stocks**

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 (44) 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.

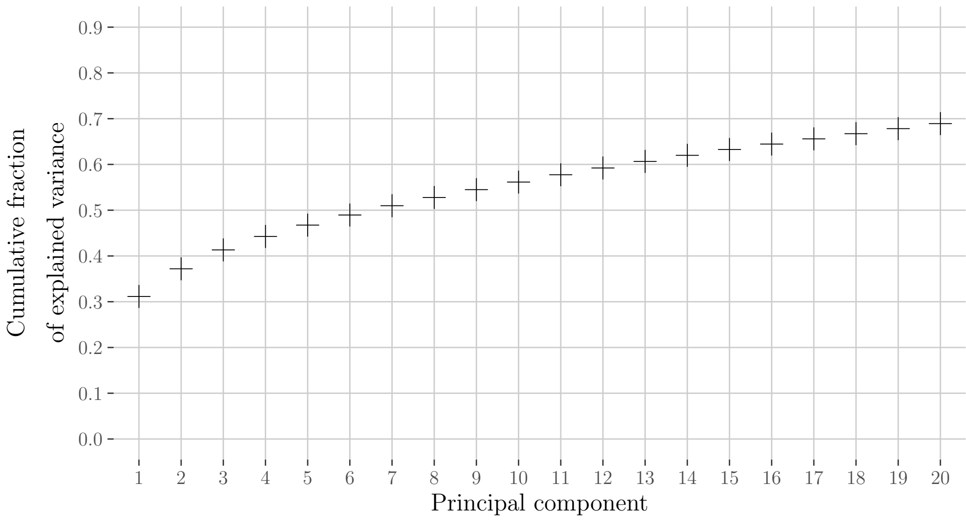

Figure 2: **Cumulative Fraction of Excess Return Variance Explained by Principal Components**

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

$$\omega(u) - p(u) = \mathrm{vec}\left(\beta(u)\beta^{\intercal}(u)\right)^{\intercal}\delta + \varphi_1{}^{\intercal}(u)\eta. \tag{45}$$

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 **?**.

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.

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 (45). 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	0.127	0.096	0.116
Average R^2 with both F & G	0.208***	0.323***	0.506***	0.396***
(s.e.)	(0.010)	(0.012)	(0.010)	(0.013)
Average R^2 with only G	0.064***	0.047***	0.072***	0.163***
(s.e.)	(0.005)	(0.005)	(0.009)	(0.015)
Average R^2 with only F	0.177***	0.314***	0.503***	0.356***
(s.e.)	(0.011)	(0.012)	(0.010)	(0.013)
Average loading on G	−0.010***	−0.008***	−0.002	0.008*
(s.e.)	(0.001)	(0.001)	(0.001)	(0.004)
η	−0.141**	−0.085	−0.023	0.115***
(s.e.)	(0.060)	(0.054)	(0.027)	(0.009)

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.

A Proofs

PROOF: of Theorem 1 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

$$\int_{u=0}^{1} h(u)\mathrm{d}R^{(n)}(u) = \int_{u=0}^{1} h(u)\alpha(u)\mathrm{d}u + \int_{u=0}^{1} h(u)\mathrm{d}Z^{(n)}(u). \tag{46}$$

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,

$$\int_{u=0}^{1} h(u)\mathrm{d}R^{(n)}(u) \to \int_{u=0}^{1} h(u)\alpha(u)\mathrm{d}u. \tag{47}$$

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. □

PROOF: of Theorem 2 As the return process $R^{(n)}$ satisfies the conditions of Theorem 1, we may rewrite (1) as

$$\begin{aligned} R_i^{(n)} &= n\left[R^{(n)}\left(\frac{i}{n}\right) - R^{(n)}\left(\frac{i-1}{n}\right)\right] \\ &= n\int_{(i-1)/n}^{i/n} \beta^{\intercal}(v)\mathrm{d}v\,(F-\lambda) + n\left[Z^{(n)}\left(\frac{i}{n}\right) - Z^{(n)}\left(\frac{i-1}{n}\right)\right]. \end{aligned} \tag{48}$$

This implies

$$
\begin{aligned}
S^{(n)}(u) &= \frac{1}{n}\sum_{i=1}^{\lfloor un \rfloor}\left(R_i^{(n)}\right)^2 - p_i^{(n)} \\
&= n\sum_{i=1}^{\lfloor un \rfloor}\left[\int_{(i-1)/n}^{i/n}\beta^{\intercal}(v)\mathrm{d}v\,(F-\lambda)\right]^2 - \frac{1}{n}\sum_{i=1}^{\lfloor un \rfloor}p_i^{(n)} \\
&\quad + n\sum_{i=1}^{\lfloor un \rfloor}\left[Z^{(n)}\left(\frac{i}{n}\right) - Z^{(n)}\left(\frac{i-1}{n}\right)\right]^2 \\
&\quad + 2n\sum_{i=1}^{\lfloor un \rfloor}\int_{(i-1)/n}^{i/n}\beta^{\intercal}(v)\mathrm{d}v\,[F-\lambda]\left[Z^{(n)}\left(\frac{i}{n}\right) - Z^{(n)}\left(\frac{i-1}{n}\right)\right].
\end{aligned}
\tag{49}
$$

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 (49) 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 $\int_{v=0}^{u} \left(\beta(v)\,(F-\lambda)\right)^2 \mathrm{d}v$.

The second term in (49) converges given (26) and the third one in view of (28).

Finally, consider the last term in (49). 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 (49) equals that of

$$[F-\lambda] \sum_{i=1}^{\lfloor un \rfloor} \beta\left(\frac{i-1}{n}\right) \left[Z^{(n)}\left(\frac{i}{n}\right) - Z^{(n)}\left(\frac{i-1}{n}\right)\right] \tag{50}$$

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

Taking these claims together, we find that

$$S^{(n)}(u) - \int_{v=0}^{u} \left(\beta(v)\,(F-\lambda)\right)^2 \mathrm{d}v + \int_0^u p(v)\mathrm{d}v - [Z, Z](u), \tag{51}$$

converges to zero. In view of (29), this concludes the proof. □

References

Aït-Sahalia, Y., and D. Xiu (2015): "Using Principal Component Analysis to Estimate a High Dimensional Factor Model with High-Frequency Data," http://ssrn.com/abstract=2669506.

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.

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," .

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 (2016): "The Common Factor in Idiosyncratic volatility: quantitative asset pricing implications," *Journal of Financial Economics*, 119(2), 249–283.

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.

Mykland, P. A., and L. Zhang (2006): "ANOVA for diffusions and Itô processes," *The Annals of Statistics*, 48(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.

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