Simple moment-based tests for value-at-risk models and discrete distributions*

Christian Bontemps†

October 15, 2014

Abstract

In this paper, we develop moment-based tests for parametric discrete distributions. Moment-based test techniques are attractive as they provide easy-to-implement test statistics. We propose a general transformation that makes the moments of interest insensitive to the parameter estimation uncertainty. This transformation is valid for some extended families of non-differentiable moments that are of great interest in the case of discrete distributions. Considering the power function under local alternatives, we compare this strategy with the

*I would like to thank Torben Andersen, Nour Meddahi, and Enrique Sentana for helpful comments, and participants at various conferences and seminars for helpful discussions. I especially thank David Kaplan for his help in correcting the language. I acknowledge financial support from ANR-05-BLAN-0024-01 and from the chair "Marché des risques et création de valeur Fondation du risque/SCOR.

†Toulouse School of Economics (GREMAQ, IDEI), 21 Allée de Brienne, 31000 Toulouse, France; phone: + 33 5 61 12 85 95, email: christian.bontemps@tse-fr.eu

one in which parameter uncertainty is corrected. The special example of backtesting of value-at-risk (VaR) forecasts is treated in detail, and we provide simple moments that have good size and power properties in Monte Carlo experiments. Additional examples considered are discrete counting processes and the geometric distribution. We finally apply our method to backtesting of VaR forecasts derived from a T-GARCH(1,1) model estimated using foreign exchange-rate data.

Keywords: moment-based tests; parameter uncertainty; discrete distributions; value-at-risk; backtesting.

JEL codes: C12, C18.

1 Introduction

Moment-based tests for testing distributions or particular features of distributions (tail properties, kurtosis) are particularly attractive because of the simplicity of their implementation. These tests are universal because they can consider univariate or multivariate parametric distributions, discrete or continuous distributions, and independent or serially correlated data in the same setting. Moment-based tests have therefore been extensively used in recent contributions related to financial econometrics (Amengual and Sentana, 2011; Amengual *et al.*, 2013; Bai and Ng, 2005; Bontemps and Meddahi, 2005, 2012; Candelon *et al.*, 2011; Chen, 2012; Duan, 2004; Dufour *et al.*, 2003; Fiorentini *et al.*, 2004; Mencia and Sentana, 2012).

In this paper we develop and apply moment-based tests for discrete distributions in an i.i.d. or a serial correlation setting. Particular examples of interest are value-at-risk (VaR) models, discrete counting processes, and discrete choice models. Our framework can be used for both conditional and marginal distributions. We derive a general class of moment conditions that are satisfied under the null hypothesis, and we pick particular moments in this class. There are various guidelines for choosing moments in this class. It is possible to focus on tractability, optimality (against a given alternative), or testing of test-specific features, as in structural modeling. Finally, whatever the reason for the choice, we have a set of moments and the resulting test statistic has an asymptotic chi-square distribution under some usual regularity conditions that ensure the validity of the Central Limit Theorem.

We allow for the presence of estimation uncertainty generated by parameter estimation for residuals or the distribution of interest. For example, in a Poisson counting process, the rate depends on explanatory variables and parameters that are estimated within the sample; in

VaR models, the VaR is not observed but estimated through a (continuous) model of financial returns.

Parameter estimation uncertainty has an impact on the asymptotic distribution of tests and has to be taken into account to obtain accurate size properties. We adopt the approach of Bontemps and Meddahi (2012) and Wooldridge (1990), among others, which involves transforming moments into ones that are orthogonal to the score function. According to the generalized information matrix equality, such moments are robust to parameter estimation uncertainty; in other words, the test statistic does not depend on whether the parameters are estimated or known. Working with robust moments is very attractive for time series data or/and when derivation of the asymptotic distribution of the estimator in closed form is difficult. In particular, in a time series context, using robust moments can drastically simplify the calculations.

Here we generalize the Bontemps–Meddahi (BM) transformation by considering alternatives to the orthogonal projection. We can indeed project our moment along an estimating equation other than the BM score function; we call this an oblique projection. We use the fact that this orthogonalization is still valid for non-differentiable moments of a particular type. This type includes the VaR example and Pearson type-tests that compare empirical and theoretical frequencies of cells. It appears that the generalized information matrix equality can indeed be used for this class of non-differentiable moments. Although the result has been known for some time, there has been no systematic use of this equality in the literature on moment-based tests.

Consideration of non-orthogonal projections can be very attractive as it might lead to quantities that are easily estimated within the data. A key aspect for the small-sample properties of the tests proposed is to have closed forms and to avoid within-sample estimations. The choice of the direction is therefore guided by the problem of deriving closed forms. We illustrate this issue by investigating backtests of VaR measures derived from an underlying T-GARCH process for returns.

We also study the power implications of this orthogonalization strategy. First, there is no loss in working with robust moments. Second, there is no optimal choice of the direction without precise knowledge of the alternative. A particular choice of direction can always be dominated by (or dominate) another choice for specific choices of the alternative. The tractability of the test procedure is ultimately the major guideline for the choice of this direction.

We also prove an appealing and important aspect of working with robust moments in the context of out-of-sample tests. It is known that the central limit theorem for out-of-sample cases does not have the usual expression and depends on the estimation scheme (West, 1996; McCracken, 2000). This is not the case for robust moments, for which the expression appears to be the same and independent of the estimation scheme. This means that the "usual" formula can be applied for any out-of-sample test statistic, which is a noteworthy simplification.

We apply the results to some classical examples of interest. We first study in detail backtests of VaR models. In particular, we derive easily computed procedures that can test the accuracy of VaR forecasts in a GARCH model. These tests are valid regardless of the true conditional mean and variance used to generate the GARCH. We focus in particular on two popular models, the normal GARCH and the T-GARCH models. A Monte Carlo simulation is run to show the performance of the proposed tests. The results suggest that the tests perform well in the set-ups traditionally considered in the literature.

We consider three additional examples. First, we derive tests for Poisson counting models based on the family of Charlier polynomials, which has the nice feature that any polynomial of order higher than two is robust to the parameter uncertainty. Second, we test the geometric distribution in an i.i.d. context. This distribution has been used to model the duration between two consecutive hits in backtesting of VaR models (Christoffersen and Pelletier, 2004; Candelon *et al.*, 2011). In particular, we evaluate the impact of testing its continuous approximation, the exponential family, on the power properties. Simulations suggest that when the data exhibit serial correlation, the power deteriorates, and therefore tests for the true discrete distribution should be used. Finally, we present a slight modification of the well-known Pearson chi-square test that can be used to take parameter uncertainty into account. The difference between the observed and theoretical frequency of the cells considered should be translated by a quantity proportional to the score function. When the parameters are estimated by maximum likelihood estimation (MLE), this modification vanishes, and we recover the usual formula for the Pearson chi-square test.

The remainder of the paper is organized as follows. Section 2 develops the general framework, including the general orthogonalization method, and presents examples that are of particular interest. In Section 3 we construct the class of moments that could be used for testing purposes. We also present particular orthonormal families of polynomials that can be used to test some standard discrete distributions. Section 4 focuses on backtesting VaR models. A few tests are proposed and studied in a Monte Carlo experiment presented in Section 5. Section 6 considers additional examples such as Poisson counting processes and discrete duration models. Finally, Section 7 considers an empirical application that tests VaR forecasts derived from a T-GARCH(1,1) model for daily exchange-rate data. Section 8 concludes the paper. The proofs, details of the calculations, and additional analysis are provided in the Appendix.

2 General results

2.1 Set-up and examples

Let Y be a univariate discrete random variable whose support S is countable. Without loss of generality, we assume that S can be set to $\{0, 1, 2, \ldots, N\}$, where N is either finite or infinite.[1] Let P_θ be a parametric family of distributions for Y indexed by $\theta \in \Theta \subset \mathbb{R}^r$. We assume that Θ is compact. The true distribution of Y is P_{θ^0}, where the true value θ^0 belongs to the interior of Θ. $\mathbb{E}$ denotes the expectation with respect to P_θ, and $\mathbb{V}$ is the variance. $p_i(\theta)$ denotes the probability of observing $Y = i$. $\mathbb{E}_0$, $\mathbb{V}_0$, and $p_i(\theta^0)$ are the same quantities when we consider the true distribution, $\mathcal{P}_{\theta^0}$. The symbol $\top$ denotes the transpose operator. We also adopt the following notation hereafter in this section. For two functions $h_1(y, \theta)$ and $h_2(y, \theta)$, we denote by $\mathbb{E}_0\left[h_1 h_2^\top\right]$ the matrix $\mathbb{E}_0\left[h_1(y, \theta^0) h_2^\top(y, \theta^0)\right]$.

Our framework is adapted to the conditioning case in which X are explanatory variables that may or may not contain past values of Y in the time series case. In this case, P_θ would become $P_{\theta,x}$ and we would test the conditional distribution of $Y \mid X = x$. Here we focus on the marginal case.

Consider a sample of T observations $(y_1, \ldots, y_T)$ that are independent or serially correlated and for which stationarity is assumed. Our goal is to test the null hypothesis that the p.d.f. of y_t is in P_θ with or without specifying the value of θ. Either θ is left undefined and will be

[1]It is indeed possible to map a countable support with a subset of $\mathbb{N}$.

estimated using $y_1, \ldots y_T$, or θ is set equal to some prespecified value. To do this, we pick particular moments $m(\cdot)$ whose expectation is equal to zero under the null hypothesis. Our procedure consists of testing whether the empirical average of these moments is close to zero.

We need regularity conditions to ensure the validity of the Central Limit Theorem and the regularity of the estimator used for θ_0 and the function $m(\cdot)$. These regularity assumptions are introduced because we need them later.

We now provide examples that are of interest in applied economics. Some of them are considered in the Monte Carlo experiment.

Example 1 *VaR models (VaR)*

VaR forecasts are used by financial institutions as a measure of risk exposure. Backtesting procedures are needed to assess the reliability of the models used by these institutions to compute their VaR forecasts.

Let r_t be the daily log return of some given portfolio or equity, and let VaR_t^α be the 1-day-ahead VaR forecast (computed at time $t-1$) for a given level of risk α (value known by the econometrician, generally 5% or 1%). Most of the leading tests are based on the sequence of hits I_t, $I_t = \mathbf{1}\{r_t \leq -\mathrm{VaR}_t^\alpha\}$. Under perfect accuracy, I_t is i.i.d. Bernoulli distributed with parameter α. Christoffersen (1998) considered an LR test in a Markov framework. Christoffersen and Pelletier (2004) and Candelon *et al.* (2011) considered tests based on the distribution of the duration between two consecutive hits.

Example 2 *Counting processes*

Counting processes are used in a wide range of fields (Cameron and Trivedi, 2010). The Poisson distribution is the analog of the normal distribution in the discrete case. This is one leading model in i.i.d. count data. In this model, $p_i(\theta) = e^{-\theta}\frac{\theta^i}{i!}$, but in general there are explanatory variables X and $\theta \equiv \beta^\top X$. This model can be extended to a serially correlated one. The particular case of the Poisson INAR(1) model is considered in Section 6.2.

Example 3 *Discrete choice models*

Discrete choice models describe choices made among a finite set of alternatives. They have played an important role in many subfields: participation in the labor force, urban transport mode choice, and analysis of demand for differentiated product are particular examples among many others. Here, $p_i(\theta, x) = P(Y = i|X = x) = F(a_{i+1} - \beta^\top x; \nu) - F(a_i - \beta^\top x; \nu)$, where $a_0, a_1, \ldots, a_K$ are some threshold values (with the convention $a_0 = -\infty$, $a_K = \infty$), K is the number of choices faced by the decision maker, β is a vector of parameters, and $F(.; \nu)$ is the cumulative distribution function for the error term. Hamilton and Jorda (2002), for example, considered an ordered probit to model the size of the change in the federal funds rate.

A few moment-based tests have been proposed in the literature, including the probit model (Skeels and Vella, 1999), bivariate ordered probit (Butler and Chatterjee, 1997), and ordered probit (Mora and Moro-Egido, 2008).

Example 4 *Pearson χ^2-type tests*

Let $C_1, \ldots, C_K$ be K cells that cover the support of Y. The well-known Pearson χ^2 goodness-of-fit test is based on the set of moments $m_i(y, \theta) = 1_{y \in C_i} - q_i(\theta)$, where $q_i(\cdot)$ is the probability that Y belongs to C_i. Boero *et al.* (2004) studied this test and the sensitivity of its power to the definition of the cells.

2.2 Test statistic

We start with the case in which the true value θ^0 for the parameter θ is known.

Let $m(\cdot)$ be a k-dimensional moment[2] whose expectation under the null hypothesis is equal to 0. Discussion on how to derive and choose the set of moments is postponed to Section 3. Under the null hypothesis that the true distribution of y_t is P_{θ^0},

$$\mathbb{E}_0 m(y, \theta^0) = 0.$$

Assumption (CLT): Central Limit Theorem. We assume that the long-run covariance matrix of $m(\cdot)$, Σ, is finite and positive definite and that the Central Limit Theorem applies. Lower-level assumptions that ensure (CLT) for $m(\cdot)$ can be found in Corollary 5.3 of Hall and Heyde (1980).

Under (CLT), a test statistic ξ_m can be constructed from any consistent estimator $\hat{\Sigma}$ of Σ:

$$\xi_m = \left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} m(y_t, \theta^0) \right)^{\top} \hat{\Sigma}^{-1} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} m(y_t, \theta^0) \right). \tag{1}$$

Under the null hypothesis, this statistic follows an asymptotic chi-square distribution with k degrees of freedom.

It should be noted that picking a finite set of moments does not lead to an omnibus test. Most of the leading tests in the literature are also not omnibus. In a VaR context, the Christoffersen backtesting procedure cannot detect alternatives for which the hits have nonzero autocorrelation of order two or higher. In a continuous context, tests based on skewness and kurtosis measures cannot detect deviation from moments greater than five. However, all these tests are frequently used because they are intuitive, easy to implement, and sufficiently

[2]The k components of m are assumed to be free, that is, the variance of $m(\cdot)$ under the null hypothesis is of full rank.

powerful for the standard alternatives of interest. One of the advantages of using moments is to be able to adapt to the case we consider. This means that we can change the moments of interest depending on the alternative of interest.

2.3 Parameter estimation uncertainty

We consider the case in which there are parameters estimated in the procedure. With abuse of the notation, θ now includes all the parameters estimated from the data, and excludes those that are known. $s_\theta(\cdot)$ is the score function of the model. In the VaR model (Example 1), the hits follow a Bernoulli distribution with parameter α, which is known, and θ is therefore the estimated parameters of the underlying model for the returns. The hit sequence $\{I_t\}_{t=1}^T$ is a function of θ and is a non-differentiable function of the observed returns and the estimated VaR forecasts, $I_t(\theta) = \mathbf{1}\{r_t \leq -\mathrm{VaR}_t^\alpha(\theta)\}$. In the discrete choice models (Example 3), θ includes the parameter of the error term distribution ν, the thresholds a_i, and the parameter β.

2.3.1 Asymptotic expansion

We now impose some regularity assumptions that are necessary to write our first-order expansion.

Assumption (R): Regular estimator. We assume that $\hat{\theta}$, an estimator of θ^0 based on $y_1, \dots, y_T$, satisfies the following expansion:

$$\sqrt{T}(\hat{\theta} - \theta^0) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} w(y_t, \theta^0) + o_P(1),$$

where $w(\cdot)$ is some estimating equation that satisfies the regularity conditions (CLT) and therefore ensures the asymptotic normality of $\hat{\theta}$. $w(\cdot)$, the influence function, can come from an ML estimation or a GMM estimation.

Assumption (GIM): generalized information matrix equality. The generalized information matrix equality

$$\left(\frac{\partial \mathbb{E}_0\left[m(y_t,\theta)\right]}{\partial \theta^\top}\right)_{\theta=\theta_0} + \mathbb{E}_0\left[ms_\theta^\top\right] = 0. \tag{2}$$

is satisfied.

Conditions for the generalized information equality can be found in Tauchen (1985) (e.g., Assumption 2 and 4). It requires, in particular, the continuous differentiability of $\mathbb{E}_0\left[m(y,\theta)\right]$ with respect to θ in some open neighborhood of θ^0. Any differentiable moment of the form

$$m(y,\theta) = \mathbf{1}\{y \in [l(\theta), u(\theta)]\} - p(\theta), \tag{3}$$

where l, u, and p are differentiable functions of θ, satisfies (GIM).

Proposition 1 *Let $m(\cdot,\theta^0)$ be a moment with zero expectation under P_{θ_0} and satisfying Assumption (CLT), where $\hat\theta$ a square-root-consistent estimator of θ^0 that satisfies Assumption (R). Under (GIM), the sequence $m(y_1,\hat\theta),\ldots,m(y_T,\hat\theta)$ satisfies the following expansion:*

$$\sqrt{T}\frac{1}{T}\sum_{t=1}^{T} m(y_t,\hat\theta) = \sqrt{T}\frac{1}{T}\sum_{t=1}^{T} m(y_t,\theta^0) - \mathbb{E}_0\left[ms_\theta^\top\right]\sqrt{T}(\hat\theta-\theta^0) + o_P(1). \tag{4}$$

The proof comes from a combination of the usual Taylor expansion around θ^0

$$\sqrt{T}\frac{1}{T}\sum_{t=1}^{T} m(y_t,\hat\theta) = \sqrt{T}\frac{1}{T}\sum_{t=1}^{T} m(y_t,\theta^0) + \left.\frac{\partial \mathbb{E}\left[m(y_t,\theta)\right]}{\partial \theta^\top}\right|_{\theta=\theta_0}\sqrt{T}(\hat\theta-\theta^0) + o_P(1) \tag{5}$$

combined with the generalized information matrix equality (2).

We are obviously not the first to characterize the asymptotic expansion (4), but there is no systematic use of this alternative expression to treat the parameter uncertainty issue. It has been used in the BM approach for continuous moments but is also valid for the (non-differentiable) class of moments written above.

When θ is estimated, we know that plugging the estimator $\hat{\theta}$ into Equation (1) generally modifies the asymptotic distribution of the test. If we ignore the estimation impact, the size of the test is no longer controlled. We can indeed derive from Equation (4) that the long-run variance of $m(y_t, \hat{\theta})$ is the long-run variance of $m(y_t, \theta^0) - \mathbb{E}_0 \left[m s_\theta^\top \right] w(y_t, \theta^0)$. A notable exception is when $\hat{\theta}$ is the MLE. In this case, the long-run variance is dominated by the long-run variance of $m(y_t, \theta^0)$.[3] Consequently, a test based on (1) uses a higher variance and leads to a conservative test.

More importantly, it is clear from (4) that there are two strategies to deal with the impact of parameter uncertainty. The first consists of correcting for the impact by deriving the joint asymptotic distribution of the two terms on the right-hand side of (4) (Newey, 1985; Mora and Moro-Egido, 2008; Escanciano and Olmo, 2010). This is possible but can be very cumbersome in the time series case and/or for estimators for which the influence function $w(\cdot)$ is not very easy to derive (such as two-steps estimators). The second strategy involves working with moments $m(\cdot)$ that are orthogonal to the true score function (here we call such moments robust moments). In this case, parameter uncertainty has no impact (at the first-order level) on the asymptotic distribution of $\sqrt{T} \frac{1}{T} \sum_{t=1}^{T} m(y_t, \hat{\theta})$ and we can proceed as if θ were known.[4] To do this we transform any moment $m(\cdot)$ into a moment that is orthogonal to the score

[3] When $\hat{\theta}$ is the MLE, $w(y_t, \theta^0) = V_0 \left[s_\theta \right]^{-1} s_\theta(y_t, \theta^0)$. Consequently, $V_0 \left[m(y_t, \theta^0) - \mathbb{E}_0 \left[m s_\theta^\top \right] w(y_t, \theta^0) \right] = V_0 \left[m(y_t, \theta^0) \right] - \mathbb{E}_0 \left[m s_\theta^\top \right] V_0 \left[s_\theta \right]^{-1} \mathbb{E}_0 \left[m s_\theta^\top \right]^\top << V_0 \left[m(y_t, \theta^0) \right]$.

[4] For robust moments, we can use either θ^0 or $\hat{\theta}$ and the test statistic ξ_m in (1) follows the same asymptotic distribution. It is worth noting that, for example, the Jarque–Bera test is not valid when the mean and variance of the normal distribution to be tested are known.

function $s_\theta(\cdot)$. This transformation strategy has been followed in earlier work by Wooldridge (1990) and Chen (2012) and by Bontemps and Meddahi (2012) in a continuous context. We now explore this path in more depth in the next section.

2.3.2 Orthogonalization method

A robust moment is orthogonal to the score function, so all transformations used in the literature implicitly have in common the transformation of a moment into one that is orthogonal to the score function (Wooldridge, 1990; Chen, 2012; Bontemps and Meddahi, 2012). However, the solutions that have been proposed are different. Wooldridge (1990) considered moment-based tests for conditional distributions. In his framework, the matrix involved is the full expectation with respect to the joint distribution of Y **and** X. He proposed a transformation of the instruments $h(X)$ to have orthogonality with respect to the joint distribution of Y and X, and does not refer to the score function. Bontemps and Meddahi (2012) proposed an orthogonal projection of the moment onto the orthogonal of the conditional score. Chen (2012) considered parameter uncertainty in a generalized GARCH model and used the fact that a moment orthogonal to y and y^2 is robust. In a continuous context, which differs from ours, the Khmaladze transformation used by Bai (2003) and Khmaladze and Koul (2004) are similar strategies, as reviewed by Li (2009). Finally, in an earlier study, Bontemps and Meddahi (2005) noted that Hermite polynomials of order greater than three are robust to parameter uncertainty in a general location scale model. Mencia and Sentana (2012) derived an LM test of normality in a multivariate GARCH context. They also obtained a linear combination of Hermite polynomials that are robust to parameter uncertainty. Amengual and Sentana (2011) described a score test for multivariate normality that is also robust to parameter uncertainty.

However, working with the real score function might be very difficult or cumbersome in some specific cases, which is why we propose a generic extension of the BM transformation that does not involve the score function. We first consider an estimating equation $g(\cdot)$ that identifies the parameter θ and satisfies Assumption (CLT). We assume that $g(\cdot)$ is of the same dimension as θ, like the identifying restrictions of a GMM procedure. It is worth highlighting that the estimating function used to estimate θ in the sample can be any estimating equation, such as $g(\cdot)$ itself, the score function, or any other estimating equation. The next proposition proposes a transformation for building a robust moment.

Proposition 2 *Let $g(\cdot)$ be some estimating equation that satisfies (CLT) with the same dimension as θ that identifies θ^0 and provides a regular square-root-consistent estimator of θ^0 (in other words, $\hat{\theta}$ follows Assumption (R)). The moment*

$$\tilde{m}_g(y,\theta) = m(y,\theta) - \left(\frac{\partial \mathbb{E}\left[m(y_t,\theta)\right]}{\partial \theta^\top}\right)_{\theta=\theta^0} \left(\frac{\partial \mathbb{E}\left[g(y_t,\theta)\right]}{\partial \theta^\top}\right)^{-1}_{\theta=\theta^0} g(y,\theta) \tag{6}$$

is a moment robust to parameter estimation uncertainty.

Observe that following the generalized information matrix equality, $\left(\frac{\partial \mathbb{E}[m(y_t,\theta)]}{\partial \theta^\top}\right)_{\theta=\theta^0} = -\mathbb{E}_0[m.s_\theta^\top]$, and similarly for $g(\cdot)$. Consequently, $\mathbb{E}_0[\tilde{m}_g.s_\theta^\top] = 0$ and application of (4) ensures the result.

The two matrices involved in Equation (6) can be computed under the null hypothesis or estimated within the sample. In the latter case, Amengual *et al.* (2013) interpreted the transformation as an IV regression. Bontemps and Meddahi (2012) considered the score function as the estimating equation leading to an orthogonal projection of $m(\cdot)$ on the orthogonal space of the score function (moment denoted by $m^\perp$). Proposition 2 extends their transformation to non-orthogonal projections, which is attractive when it is not practical to deal with the true score.

In terms of the power properties of the test, there are no arguments in favor or against use of an oblique rather than an orthogonal projection (see the discussion in Appendix A.3). The choice of $g(\cdot)$ is guided, for example, by the requirement of having closed-form expressions for the test statistics. We consider this oblique projection in Section 4 in the context of backtesting of VaR forecasts derived from a T-GARCH model.

2.3.3 A simplified procedure: projection on the orthogonal of an auxiliary score

Despite the fact that Proposition 2 provides a strategy for building a robust moment when use of the score is impractical, its attractiveness depends on the choice of $g(\cdot)$. It may happen that the model is so complex that a closed form for (6) is not possible. We now go further by proposing a simplified procedure to build robust moments.

Consider a simple model $\tilde{\mathcal{M}}$ (the auxiliary model) defined by the parametric family of distribution $\tilde{P}(y_t;\beta)$ and let $\tilde{s}_\beta(\cdot)$ be the score for this auxiliary model. Assume that our true model can be linked to this auxiliary model by $\beta = h(X_{t-1},\theta)$, where $h(X_{t-1},\cdot)$ is a differentiable function and X_{t-1} is a collection of variables such that, conditional on X_{t-1}, the distribution of y_t is in $\tilde{P}(y_t;\beta)$.

If a moment $m(y_t;\beta^0)$ is orthogonal to the true score for the auxiliary model, it is orthogonal to the score for the true model. The score for the true model is indeed a linear vcombination of the components of the score in the auxiliary model.[5] Let $\tilde{g}(\cdot)$ be some differ-

[5] $s_\theta(y_t;\theta) = \sum_{j=1}^{\dim\theta} \tilde{s}_{\beta_j}(y_t;\beta)\frac{\partial h_j}{\partial\theta}(X_{t-1},\theta)$. If $m(y_t;\beta^0)$ is orthogonal to the auxiliary score, $m(y_t;h(X_{t-1},\theta^0))$ is orthogonal to the true score by the law of iterated expectations.

entiable estimating equation that identifies the parameters of the auxiliary model and let $\tilde{\mathbb{E}}$ be the expectation for the auxiliary model. The moment

$$\tilde{m}_{\tilde{g}}(y, \beta) = m(y, \beta) - \tilde{\mathbb{E}}[m\tilde{s}_{\beta}^{\top}]\tilde{\mathbb{E}}[\tilde{g}\tilde{s}_{\beta}^{\top}]^{-1}\tilde{g}(y, \beta) \tag{7}$$

is a moment robust to the parameter estimation uncertainty in the true model.

Observe that we can also use the generalized information matrix equality in (7) to derive alternative expressions for the matrices involved in 7.[6]

This approach is particularly appealing because in some cases, building a moment orthogonal to the score for the auxiliary model is much easier than building one for the true model. Interestingly, such a moment remains robust regardless of the functional form $h(\cdot)$. We can therefore derive a class of robust moments once for all cases.

Example 1 In a VaR model, it is relatively easy to derive moments that are robust in the auxiliary model $r_t = \mu + \sigma\varepsilon_t$, where μ and σ are constant. These moments are also robust regardless of the specification of μ and σ, in particular in the class of conditional location-scale models:

$$r_t = \mu(J_{t-1}, \theta^0) + \sigma(J_{t-1}, \theta^0)\varepsilon_t,$$

where J_{t-1} is the information set at time $t-1$ and θ^0 is a vector of parameters to be estimated. This is particularly important for practitioners because robust moments that do not depend on the exact conditional form of the conditional means and variances can be derived (see Section 4 for more details). For example, in the normal GARCH model, the score

[6] For example, $\tilde{\mathbb{E}}[\tilde{g}\tilde{s}_{\beta}^{\top}] = -\tilde{\mathbb{E}}[\frac{\partial \tilde{g}}{\partial \beta}]$.

for the auxiliary model is, up to a scale parameter, the vector whose components are ε_t and $\varepsilon_t^2 - 1$. If we project our moment onto the orthogonal of the space spanned by these two functions, that is, we apply (7), we obtain a moment robust to the parameter uncertainty for **any type** of normal GARCH model. This particular case was considered by Chen (2012).

Example 2 Another interesting example is the Poisson counting process. In Section 6.2, we showed that any Charlier polynomial of order greater than two is robust to the parameter estimation uncertainty. If we assume that the parameter of the Poisson process is linked to some exogenous variables X, $h(X, \theta)$, where θ is a parameter vector to be estimated, the same Charlier polynomials are still robust when θ is estimated within the data.

Illustration using the T-GARCH model Assume that we would like to test whether a parametric T-GARCH (1,1) model without conditional mean is a good model for computing VaR forecasts for a given series of financial returns. The returns r_t are assumed to follow the model

$$r_t = \sigma_t(\theta)\varepsilon_t,$$

where ε_t is an i.i.d. sequence from the standardized Student distribution with ν degrees of freedom. As θ is estimated within the data, we consider a robust moment by projecting onto the orthogonal of the true score function. However, if we have to consider the true score, the covariances do not have closed forms and involve infinite sums. We can therefore apply our last result using a model with constant variance $r_t = \sigma\varepsilon_t$ as the auxiliary model. Orthogonality to the score in this auxiliary model ensures orthogonality to the score of the true T-GARCH model (regardless of the parametric specification of $\sigma_t(\theta)$). We have two parameters to estimate, σ^2 and ν. We can use as the estimating equation $\tilde{g}(\cdot)$ the second and fourth moments in this auxiliary model,

$$\tilde{g}(r_t, \theta) = \begin{bmatrix} r_t^2 - \sigma^2 \\ (r_t^4 - 3\sigma^4)(\nu - 4) - 6\sigma^4 \end{bmatrix}. \tag{8}$$

The correction in Equation (7) now involves the covariance between $\tilde{g}(\cdot)$ and the score function in the auxiliary model with constant variance. It does not depend on the parametric specification of the conditional variance of the T-GARCH model and involves quantities that are only simple functions of the data. For example, after projection the hit sequence $I_t - \alpha$ becomes

$$\tilde{e}_t = I_t - \alpha + \frac{q_\alpha^\nu f_\nu(q_\alpha^\nu)}{2}(\varepsilon_t^2 - 1) + \frac{\partial F_\nu}{\partial \nu}(q_\alpha^\nu)\left(\frac{(\nu-4)^2}{6}(\varepsilon_t^4 - K_\varepsilon) - (\nu-2)(\nu-4)(\varepsilon_t^2 - 1)\right),$$

where $K_\varepsilon = 3 + \frac{6}{\nu-4}$ is the kurtosis of ε_t, q_α^ν is the α quantile of the standardized Student distribution with ν degrees of freedom, and $f_\nu(\cdot)$ ($F_\nu(\cdot)$) is its p.d.f (c.d.f.). Further details are given in Equation (26). Additional calculations are provided in Section 4.2. Simulations in Section 5 highlight the attractive properties of this procedure in terms of both size and power.

2.4 Working with robust moments

It is important to characterize the attractiveness of working with robust moments. First, when a non-robust moment is used, the asymptotic distribution of the test statistic depends on the quality of the estimates. This is not the case for a robust moment as long as $\hat{\theta}$ is a square-root-consistent estimator. Consequently, the test statistic depends only on the choice of the moment, and the critical values of the test statistic can therefore be tabulated using either the asymptotic distributional approximation or simulation techniques (bootstrap techniques can therefore be used to improve the small-sample properties).

Second, a robust moment is robust whether the data are i.i.d. or serially correlated. The alternative of correcting the statistic[7] can require many calculations to compute the covariance between the first and second terms in (4), which are avoided here. It is therefore much more convenient to work with a robust moment in a time series case.

In the Appendix, we compare the power properties between robust moments and correction for parameter uncertainty using Equation (4). It is evident that correcting is equivalent, in a i.i.d. context, to transforming the moment when using for $g(\cdot)$ in (6) the equation used to estimate θ. These two strategies are therefore equivalent. The key issue is the choice of $g(\cdot)$. The second part of our result shows that no particular choice of $g(\cdot)$ dominates the other choices. In other words, there is always some local alternative for which a given estimating equation is better than the others. To conclude, in the absence of optimality, the estimating equation $g(\cdot)$ that appears the more tractable should be selected.

Out-of-sample properties We now prove that working with robust moments is also particularly attractive in a forecasting context. It is known from earlier work (West, 1996; West and McCraken, 1998; McCracken, 2000) that the statistical properties of out-of-sample moments depend on the estimation scheme, that is, whether a recursive, rolling, or fixed scheme is used.

Proposition 3 *Let $R < T$ and $P = T - R$, and let $\hat{\theta}_t$ $(t > R)$ be a sequence of square-*

[7] Correcting means that we compute the joint distribution of the two terms in (4).

root-consistent GMM-type estimators of θ_0 using the data $y_{t-R}, \ldots, y_{t-1}$ (rolling estimator), $y_1, \ldots, y_{t-1}$ (recursive estimator), or $y_1, \ldots, y_R$ (fixed estimator). We assume that $\hat{\theta}_t$ satisfies the expansion (R) for the corresponding values of the time index. We assume that R and P tend to ∞ as T tends to ∞ and that $m(\cdot)$ satisfies the regularity conditions (CLT) and (GIM). If $m(\cdot)$ is a robust moment,

$$\frac{1}{\sqrt{P}} \sum_{t=R+1}^{T} m(y_t, \hat{\theta}_t) = \frac{1}{\sqrt{P}} \sum_{t=R+1}^{T} m(y_t, \theta^0) + o_P(1). \tag{9}$$

The proof is a direct consequence of the fact that the second term in the asymptotic expansion vanishes because of the orthogonality of m to the score function. See, for example, Theorem 4.1 of West (1996), in which the matrix F is the null matrix in this case. The intuition is the same as for the in-sample properties. A robust moment is orthogonal to the score and therefore uncorrelated to local deviations of $\hat{\theta}$ around θ^0.

Therefore, when the moments are robust, the asymptotic variance of out-of-sample moments is essentially the standard long-run variance. We do not have to correct for the estimation scheme. Simulations in Section 5 demonstrated the attractiveness of working with robust moments. They behave better and their size and power properties are similar to those in the in-sample case.

3 Choice of the moments

One appealing property of moment-based tests is the possibility of choosing the appropriate moment. There are many potential guidelines for choosing the moments of interests. We can be interested in tractability and ease of implementation in some cases, and in power against specific alternatives in others. This section provides a guideline about the choice of the moment that is used to test our discrete distributions.

3.1 Adhoc choices

Adhoc choices of moments are always possible. For well-known distributions, one generally knows the first few moments (mean, variance, skewness, and kurtosis) as functions of the parameters. For discrete distributions, one can also simply count the number of realizations of a particular value and compare the expected number of counts with the actual ones (this is the rationale of the standard Pearson chi-squared test).

For the Poisson distribution, we know that it has the property of equidispersion, i.e. the mean and the variance are equal. This gives us the opportunity to test H_0 from the first and second moments together. We could alternatively use the sequence of moments $m_i(y,\theta) = \mathbf{1}\{Y = i\} - p_i(\theta)$ for different i.

3.2 Orthogonal polynomials and Ord's family of discrete distributions

The Ord's family is a well-known extension of the famous Pearson's family[8] to the case of discrete distributions. This family includes the Poisson, binomial, Pascal (or negative binomial), and hypergeometric distributions, as particular examples.

A discrete distribution belongs to the Ord's family if the ratio (we omit the dependence of p_i in θ) $\frac{p_{y+1}-p_y}{p_y}$ equals the ratio of two polynomials $A(.)$ and $B(.)$, where $A(.)$ is affine and $B(.)$ is quadratic.

[8]See Table 1 of BM for orthonormal polynomials related to the Pearson family.

$$\frac{\Delta p_y}{p_y} = \frac{p_{y+1} - p_y}{p_y} = \frac{A(y)}{B(y)} = \frac{a_0 + a_1 y}{b_0 + b_1 y + b_2 y^2}, \tag{10}$$

where Δ is the forward difference operator: $\Delta p_y = p_{y+1} - p_y$.

We can build the associated orthonormal polynomial family Q_j, $j \in \mathbb{N}$, where each polynomial is derived using an analogue of the Rodrigues' formula on finite difference (see Weber and Erdelyi, 1952 or Szegö, 1967):

$$Q_j(y) = \lambda_j \frac{1}{p_y} \Delta^j \left[p_{y-j} B(y) B(y-1) ... B(y-j+1) \right],$$

where λ_j is a constant which ensures that the variance of Q_j is equal to 1.

These orthonormal polynomials are particular moments that can be used for our testing procedure. They are not necessarily the best in terms of power or robust to the parameter estimation uncertainty problem (except for some special cases). However, one advantage is that the variance is known, equal to one. In an i.i.d. context with known parameters, these moments are asymptotically independent with unit variance. It follows that the test statistics based on Q_j are asymptotically $\chi^2(1)$ and independent,

$$\xi_j = \left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} Q_j(y_t) \right)^2 \xrightarrow[T \to \infty]{d} \chi^2(1),$$

$$\xi = \sum_{j=1}^{r} \xi_j \xrightarrow[T \to \infty]{d} \chi^2(r).$$

Another advantage is that the family of orthogonal polynomials is complete in L^2 (see, for example, Gallant, 1980, in a continuous case). Testing the distribution or testing the full sequence of polynomials is therefore equivalent. Appendix C presents some particular examples of Ord's distributions and related polynomial families of interest. Candelon *et* *al.* (2011) used, for example, the Meixner polynomials to test the geometric distributional assumption in a VaR framework.

3.3 A general class of moments

The two previous sections present some particular moments that can be used for testing purposes. There are however some cases where such moments are not so easy to derive. We derive here a general rule for constructing any moment for which the expectation under the null is equal to zero. Let ψ be a function defined on $S \times \Theta$ and such that the expectation under P_θ is finite.

Assumption LB (Lower Bound) $\psi(0, \theta) = 0$.

Assumption LB is just a normalization of the function $\psi(\cdot)$.

Proposition 4 *Let $m(y, \theta)$ be the function defined by*

$$m(y, \theta) = \left[\psi(y+1, \theta) - \psi(y, \theta) + \frac{p_{y+1}(\theta) - p_y(\theta)}{p_y(\theta)} \psi(y+1, \theta) \right]. \tag{11}$$

Under assumption LB,

$$\mathbb{E}_0 m(y, \theta^0) = 0. \tag{12}$$

The proof is given in Appendix A. It is worth noting that the moment built in Proposition 4 is the discrete analogue of the one used in BM (Equation (8), page 983). One could argue that focusing on this class could restrict the range of the tests derived from these moment conditions. It might be the case that the set of moments generated by Eq. (11) could be a small subset in the set of any moments for which we know that the expectation under the null is equal to zero. The next proposition shows in fact that any moment of interest can be generated by the construction presented above.

Proposition 5 *Let $m(y,\theta)$ be a moment such that*

$$\mathbb{E}_0 m(y,\theta^0) = 0. \tag{13}$$

Let $\psi(y,\theta)$ be a function defined on S by:

$$\begin{aligned} &\psi(0,\theta) = 0, \\ &\psi(y,\theta) = \frac{1}{p_y(\theta)} \sum_{k=0}^{y-1} m(k,\theta) p_k(\theta) \quad \text{for } y \geq 1 \end{aligned} \tag{14}$$

*Then, ψ satisfies **LB** and $m(\cdot)$ satisfies the equality in Eq. (11).*

See Appendix A for the proof.

We illustrate the usefulness of Proposition 4 previous results by considering the geometric distribution with parameter θ. In this case, $p_y(\theta) = (1-\theta)^y\theta$ and $\frac{p_{y+1}(\theta)-p_y(\theta)}{p_y(\theta)} = -\theta$. When $\psi(y,\theta) = y$, we obtain the first Meixner polynomial, up to some scale factor, $1-\theta-\theta y$. When $\psi(y,\theta) = y^2$, the moment derived from (11) is a linear combination of the first two Meixner polynomials. The family of functions y^k generates the first k terms of the Meixner family. More generally, Proposition 4 generates a set of moments when one does not have any obvious moment to use.

3.4 Optimal choice of the moments

Moment tests can be interpreted as optimal LM tests against some given models. Let $m(\cdot)$ be a p-order moment used to test our discrete distributional assumption. Let $h(\nu)$ be a function from $\mathbb{R}^p$ to $\mathbb{R}^p$, where ν is a p-dimensional parameter. Assume that $h(0_p) = 0_p$ and that $\nabla h(0_p) = I_p$. Then, $m(\cdot)$ can be interpreted as the LM test of testing the distribution with probability density function $f_0(y)$ against the alternative distribution with pdf $f_a(y) = f_0(y)(1 + h(\nu)^\top m(y))$. One particular choice for h is the identity function $h(\nu) = \nu$. Chesher and Smith (1997) also characterized the family of alternatives such that the moment tests can be interpreted as LR tests in this augmented family.

We derive in Proposition 10 the power under local alternatives for a choice of a robust moment $m(\cdot)$. Without parameter uncertainty, the noncentrality parameter $a(g)$ becomes $a = \frac{(\mathbb{E}_1 m(x_t))^2}{\mathbb{V}_0 m}$. The following inequalities (assuming working with i.i.d. data) provide an upper bound for a (under standard regularity assumptions):

$$\begin{aligned} \frac{(\mathbb{E}_1 m(x_t))^2}{\mathbb{V}_0 m} &= \frac{\left(\int m \frac{q_1}{q_0} q_0\right)^2}{\int m^2 q_0} \\ &= \frac{\left(\int m (\frac{q_1}{q_0} - 1) q_0\right)^2}{\int m^2 q_0} \\ &\leq \frac{\left(\int (\frac{q_1}{q_0} - 1)^2 q_0\right)^2}{\int (\frac{q_1}{q_0} - 1)^2 q_0} = \int (\frac{q_1}{q_0} - 1)^2 q_0. \end{aligned} \tag{15}$$

The last inequality comes from the usual Cauchy-Schwarz inequality. The moment $m(\cdot)$ which reaches the upper bound, i.e. which maximizes the noncentrality parameter, is $\frac{q_1}{q_0} - 1$. This result[9] provides a guideline for the practitioner to choose the moment. If one prefers to manipulate standard moments, like the polynomials related to the Ord's distributions, one can use the polynomials that are the most correlated with this optimal moment.

[9]Bontemps *et al.* (2013) provide a complete discussion of point optimal moment-based tests under a more general framework.

4 Application to the backtesting of VaR models

The Basel Committee on Banking Supervision proposed in 1996 the use of Value-at-Risk models as one possibility for risk management. There is a debate on what is a good measure of risk and whether VaR is adequate (see for example Artzner *et al.*, 1999). However, this is the one that is the most commonly used by financial institutions.

Let r_t be the return at date t for a given financial asset. The Value-at-Risk VaR_t^α is the negative[10] of the α-quantile of the conditional distribution of r_t given J_{t-1}, the information set at date $t-1$:

$$P\left(r_t \leq -VaR_t^\alpha | J_{t-1}\right) = \alpha. \tag{16}$$

The goal of backtesting techniques is to check the accuracy of the model used by a given institution, observing in most of the cases only the VaR forecasts and the returns.

Let I_t be the Hit, i.e. the indicator of bad extreme event:

$$I_t = \begin{cases} 1 & \text{if } r_t \leq -VaR_t^\alpha \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

Under H_0, i.e. the VaR parametric model used by the practitioner is the true model, I_t is i.i.d. Bernoulli distributed with parameter α. I_t is therefore a discrete variable (though built from a continuous model for the returns). Our methodology applies, and this section presents feasible tests that are easy to derive and robust to the parameter uncertainty introduced by the estimation of the model for the returns.

The parameter estimation uncertainty has rarely been taken into account in this literature.

[10] A VaR is positive.

Escanciano and Olmo (2010) characterized the potential size distortion that could arise from ignoring its impact (empirical rejection rates raise to 10% for a 95%-level test) and corrected for it.

We apply our method that considers an auxiliary model. We first assume that the DGP for the return is

$$r_t = \mu + \sigma\varepsilon_t, \tag{18}$$

where $\varepsilon_t \sim i.i.d.\ D(0,1)$. $D(0,1)$ is any continuous distribution of mean 0 and variance 1. The next proposition builds some moments which are robust to the parameter uncertainty.

Proposition 6 *Let $\tilde{s}_\theta(\varepsilon_t)$ be the score function in the model (18), $P = \mathbb{E}\left[I_t.\tilde{s}_\theta^\top\right]$, $V_s = \mathbb{V}(\tilde{s}_\theta)$ and $e_t = I_t - \alpha - PV_s^{-1}\tilde{s}_\theta(\varepsilon_t)$. Let Z_{t-1} be any squared-integrable random variable that belongs to the information set at date $t-1$. The orthogonalized moment*

$$m_t^\perp(\theta) = Z_{t-1}e_t \tag{19}$$

satisfies $\mathbb{E}_0 m_t^\perp(\theta^0) = 0$ and is robust to the parameter uncertainty.

This is a direct application of the results derived in Section 2. A general expression for the matrix P is given in Appendix A.4, see Equation (A.15). Following our earlier results on moments robust in some auxiliary model, the moment in (19) is also robust in the model

$$r_t = \mu_{t-1}(\theta) + \sigma_{t-1}(\theta)\varepsilon_t. \tag{20}$$

In the Monte Carlo section, we study different choices for the past instruments. $Z_{t-1} = 1$ corresponds to the unconditional test (i.e. we test that the frequency of hits is the expected one, α); Z_{t-1} could also be past values or linear combinations of past values of e_t. We summarize the last result in the following corollary.

Corollary 7 *In the model (20), the test statistic*

$$\xi = T\left(\frac{1}{T}\sum_{t=1}^{T} Z_{t-1}e_t\right)^{\top}\left[\mathbb{E}_0[Z_{t-1}Z_{t-1}^{\top}](\alpha(1-\alpha) - PV_s^{-1}P^{\top})\right]^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} Z_{t-1}e_t\right) \tag{21}$$

is asymptotically distributed as a $\chi^2(k)$, where k is the dimension of Z_{t-1}, whether the parameters are estimated or known.

We now detail the expression of the robust moments for two particular GARCH processes, the Normal GARCH and the T-GARCH. The details of the calculations are provided in Appendix B.1 and B.2.

4.1 The Normal GARCH model

In the Normal GARCH model, e_t in Proposition 6 simplifies to

$$e_t = I_t - \alpha + \varphi(n_\alpha)\varepsilon_t + \frac{n_\alpha \varphi(n_\alpha)}{2}\left(\varepsilon_t^2 - 1\right), \tag{22}$$

and its variance is equal to $\alpha(1-\alpha) - \varphi(n_\alpha)^2 - \frac{n_\alpha^2 \varphi(n_\alpha)^2}{2}$, where n_α is the α-quantile of the standard normal distribution and $\varphi(\cdot)$ its pdf.

Assume just here that we have a Normal GARCH model without drift, i.e. $\mu \equiv 0$. The projection onto the orthogonal space of the true score would have given the following quantity $e_t^\star$ in replacement of e_t above:

$$e_t^\star = I_t - \alpha + \frac{n_\alpha \varphi(n_\alpha)}{2}\mathbb{E}\left[\frac{\partial \ln \sigma_t^2(\theta)}{\partial \theta^\top}\right]\mathbb{V}\left[\frac{\partial \ln \sigma_t^2(\theta)}{\partial \theta}\right]^{-1}\frac{\partial \ln \sigma_t^2(\theta)}{\partial \theta^\top}\left(\varepsilon_t^2 - 1\right). \tag{23}$$

The variance of $e_t^\star$ is $\alpha(1-\alpha) - \frac{n_\alpha^2 \varphi(n_\alpha)^2}{2}\mathbb{E}\left[\frac{\partial \ln \sigma_t^2(\theta)}{\partial \theta^\top}\right]\mathbb{V}\left[\frac{\partial \ln \sigma_t^2(\theta)}{\partial \theta}\right]^{-1}\mathbb{E}\left[\frac{\partial \ln \sigma_t^2(\theta)}{\partial \theta^\top}\right]^T$. These quantities, however, involve infinite series which should be estimated within the data.[11] Simulations in the Monte Carlo section, Section 5, suggest that working with e_t instead of working

[11] See Appendix B.1.

with $e_t^\star$ in these examples does not change the power properties but notably simplifies the test procedure.

4.2 The T-GARCH model[12]

Assume for simplicity that the daily returns can be modeled by:[13]

$$r_t = \sigma_t(\theta)\varepsilon_t,$$

where ε_t is an i.i.d. sequence from the standardized Student distribution with ν degrees of freedom. $F_\nu(\cdot)$, $f_\nu(\cdot)$ and q_α^ν denote the cdf, the pdf and the α-quantile of this distribution and we assume that $\nu > 4$. θ and ν can be consistently estimated by Gaussian QMLE combined with the fourth moment of ε_t (see Bollerslev and Wooldridge, 1992) or by MLE. In BM, this distributional assumption is not rejected for most of the daily exchange rate returns.

Contrary to the normal case we do not have closed forms for e_t as we have to consider quantities that involve the score component related to the number of degrees of freedom, ν. We can estimate these quantities within the data or by simulation. If one really wants an explicit form without estimating any matrix, there are two simplifications that can help in deriving explicit test statistics.

First, we can simplify the problem when one assumes that the estimated degrees of freedom of the student distribution are forced to be an integer value (see the simulation exercise of Escanciano and Olmo, 2010, page 41). The parameter estimation uncertainty related to the

[12]The calculations are provided in Appendix B.2.

[13]In Appendix B.2, we present the case with a conditional mean.

estimation of ν vanishes. In this case, e_t in Proposition 6 is equal to

$$e_t = I_t - \alpha - q_\alpha^\nu f_\nu(q_\alpha^\nu)\frac{\nu+3}{2\nu}\left(1 - \frac{(\nu+1)\varepsilon_t^2}{\nu-2+\varepsilon_t^2}\right). \tag{24}$$

A second strategy consists in orthogonalizing the moment from some estimating equation in a constant scale model as discussed in Section 2.3. One estimating equation for $\theta = (\sigma^2, \nu)^\top$ includes both the second and fourth moments of the returns r_t:

$$g(r_t, \theta) = \begin{bmatrix} r_t^2 - \sigma^2 \\ (r_t^4 - 3\sigma^4)(\nu - 4) - 6\sigma^4 \end{bmatrix}. \tag{25}$$

Following the expression of the quantities involved in Equation (6), the orthogonalized version of $I_t - \alpha$ is now

$$\tilde{e}_t = I_t - \alpha + \frac{q_\alpha^\nu f_\nu(q_\alpha^\nu)}{2}(\varepsilon_t^2 - 1) + \frac{\partial F_\nu}{\partial \nu}(q_\alpha^\nu)\left(\frac{(\nu-4)^2}{6}(\varepsilon_t^4 - K_\varepsilon) - (\nu-2)(\nu-4)(\varepsilon_t^2 - 1)\right), \tag{26}$$

where $K_\varepsilon = 3 + \frac{6}{\nu-4}$ is the kurtosis of ε_t.

5 Monte Carlo experiment related to the backtesting of VaR models

In this section, we consider backtests of VaR models, defined in Section 4. The returns of a fictive portfolio/asset are assumed to follow a GARCH (1,1) model with i.i.d. innovations:

$$r_t = \sqrt{\sigma_t^2(\theta)}\varepsilon_t,\ \sigma_t^2(\theta) = \omega + \gamma r_{t-1}^2 + \beta\sigma_{t-1}^2, \tag{27}$$

with $\varepsilon_t \sim D(0,1)$, $\omega = 0.2$, $\gamma = 0.1$ and $\beta = 0.8$. The distribution D considered here is the standard Normal distribution and the standardized Student distribution.

We simulate $T = 250$, 500 or 750 observations which corresponds approximately to, respectively, one, two or three years of trading days. All the results displayed are based on 1 000 replications and each table reports the rejection frequencies for a 5% level test.

5.1 The Normal GARCH model

We first consider the case where the innovation process is Gaussian. In Table 2 and Table 3, we display the in-sample size and power of different competing tests. A VaR model is used for forecasting, but checking the in-sample properties serves as a benchmark for the additional tables. We consider two different VaR measures, with α being respectively equal to 1% and 5%. We now detail the tests presented in the tables. It is worth noting that we can implement our test as such even if the number of actual hits is equal to zero. This particularly interesting when one backtests VaR forecasts with low coverage rate, α.

We first display the unconditional test based on the counting of hits, ignoring the parameter uncertainty, $(I_t - \alpha)^0$. It can be used as a benchmark. Three additional unconditional tests, based on the empirical frequency of hits are also displayed. We either correct for the impact of the parameter uncertainty,[14] in $I_t - \alpha$ or we use robust versions. $e_t^{\star}$ is the robust moment derived from the projection of $I_t - \alpha$ onto the orthogonal space of the true score function (see Equation (23)), e_t is the projection of $I_t - \alpha$ onto the orthogonal space of the score function when one assumes that the volatility is constant (see Equation (22)).

We consider also conditional tests, i.e. tests that detect departure from the independence. These tests are based on the product of e_t with past values, i.e. $e_t e_{t-h}$ for different values of h. We also consider weighted moments $m_k^e = e_t e_{t-1} + \frac{k-1}{k} e_t e_{t-2} + ... + \frac{1}{k} e_t e_{t-k}$, for different

[14]It is the unconditional test proposed by Escanciano and Olmo (2010).

values of k. Additional simulations not provided here suggest that the choice of the weights does not change the order of magnitude of the results. We consider weighted moments based also on $e_t^\star$.

Table 2 presents the in-sample size properties. As expected, the non-robust tests have very bad small sample properties. The robust tests have much better performance though not so good for one year of data and small risk values.[15] The conditional tests behave similarly.

[insert Table 2 here]

In Table 3, we study the power properties by considering three alternatives.[16] In the first alternative, the hit series are built from a VaR measure derived from the Historical Simulation, i.e. taking the empirical α-quantile within the data. Unsurprisingly, the unconditional tests do not have any power as the frequency of hits equalizes its theoretical value by construction. Conditional tests, $e_t e_{t-h}$, do have power, and combining different lags into a single moment (i.e. the moment $m_{t,k}^e$) is the most powerful strategy.

The second alternative consists of simulating a T-GARCH model with the same conditional volatility but with innovation terms ε_t that are distributed following a standardized Student distribution with 4 degrees of freedom. When the VaR measure is computed, Gaussianity is (wrongly) assumed. Power essentially comes from the unconditional tests as the expected frequency of hits is lower than the empirical ones. The rejection rates are very close to each other.

In the third alternative, we simulate an EGARCH model[17] with T(4) innovations, estimating the standard normal GARCH(1,1) model to derive the VaR expression. Both the

[15] See also Escanciano and Olmo (2010), who highlighted this in their simulation exercise.

[16] We only report the results for $\alpha = 5\%$, the results being qualitatively the same for $\alpha = 1\%$.

[17] $\sigma_t^2 = \exp(0.0001 + 0.9 \ln \sigma_{t-1}^2 + 0.3(|\varepsilon_{t-1}| - \sqrt{2/\pi}) - 0.8\varepsilon_{t-1})$.

distributional assumption and the volatility model are wrong. Therefore, both conditional and unconditional tests have power. Like before, the tests related to the correlation between e_t and e_{t-h} have better power properties.

[insert Table 3 here]

In both tables, there is no big differences between test procedures based on e_t or similar ones based on $e_t^\star$. Correcting the Hit sequence for the parameter uncertainty is not better and, in the third alternative, is surprisingly much worse than working with a robust moment Finally, combining in a single moment different weighted past robust e_t increases the power substantially.

In Table 4 and 5, we study the out-of-sample properties. The one-day ahead VaR forecasts are computed with a rolling estimator (Table 4) or using a fixed scheme (Table 5) assuming normality for the innovation term. In both cases, we use $R = 500$ values to estimate the parameter.[18]. We test our moments on $P = 125$ or 250 observations. As highlighted before, robust moment tests do not need any additional correction even for studying out-of-sample performance. We use the same moments and the same DGP's as in the last two tables. We compare the perfomances of the robust tests based on e_t with the ones of the tests based on the correction for the parameter uncertainty (the correction depends on the estimation scheme, see West and Mc Cracken, 2000, for details). We have the same qualitative results as before, except a slight overrejection for the size properties. The power properties are also better for tests based on robust moments (conditional or unconditional). Like for the in-sample case, the power of the tests in the T-GARCH alternative is very low when one decides to use the

[18] Additional simulations with $R = 250$ not provided here yield similar conclusions

strategy which consists in correcting for the parameter uncertainty.

[insert Table 4 and Table 5 here]

5.2 The T-GARCH model

We now consider the T-GARCH model and we use the same volatility model than for the Normal GARCH model (27). Here the innovation distribution is the standardized Student with 8 degrees of freedom.

In Table 6 and 7, we report the size properties (in-sample and out-of-sample) and the power properties when we compute the VaR forecast by historical simulation using the first R values (fixed scheme). The same tests as before are presented. We add the ones based on $\tilde{e}_t$ in Equation (26), which is a robust version of $I_t - \alpha$ after having orthogonalized it using the estimating equation $g(\cdot)$ in Eq. (25). The performances are quite comparable to the ones obtained in the Normal GARCH case. Working with $\tilde{e}_t$ does not change the order of magnitude of the results.

[insert Table 6 and 7 here]

6 Additional examples

6.1 Pearson chi-squared test[19]

Assume that $y_1, \ldots, y_T$ are independently distributed. Let $C_1, \ldots, C_K$ be K cells covering the support of the distribution of Y with $K - 1 > r$, the dimension of the parameter θ. These

[19]The details are provided in Appendix A.5.

cells are the unions of potential outcomes of the variable Y, where the definition of any cell does not depend on the parameter itself. $q_i(\theta)$ is the probability that Y belongs to C_i, i.e. $q_i(\theta) = \sum_{j \in C_i} p_j(\theta)$. For example, in the Poisson case, one could consider the five sets $\{Y = 0\}$, $\{Y = 1\}$, $\{Y = 2\}$, $\{Y = 3\}$, and $\{Y \geq 4\}$. In this case, $q_4(\theta) = \sum_{k=4}^{+\infty} p_k(\theta)$. We assume that all q_i's are strictly positive (which avoids empty cells in population). In this section, $q_i^0 \equiv q_i(\theta^0)$ and $\hat{q}_i = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\{y_t \in C_i\}$. We also assume that the parameter θ is estimated using these K cells. The score function is equal to

$$s(y, \theta) = \sum_{i=1}^{K} \mathbf{1}\{y \in C_i\} \frac{\partial \log q_i(\theta)}{\partial \theta}.$$

A Pearson type test is based on the vector of moments $m(y, \theta) = [m_1(y, \theta), \ldots, m_K(y, \theta)]^\top$ with $m_i(y, \theta) = \mathbf{1}\{y \in C_i\} - q_i(\theta)$, $i \in \{1, \ldots, K\}$. Its variance under the null is the matrix $\Sigma = D - QQ^\top$ of rank $K - 1$, where $D = diag(q_1^0, ..., q_K^0)$ and Q is the $K \times 1$ vector of probabilities $[q_1^0, \ldots, q_K^0]^\top$. Using the fact that a generalized inverse of Σ, Σ^-, is $D^{-1} - \frac{ee^\top}{K}$ where e is the $K \times 1$ vector of 1's, Equation (1) yields the well known Pearson chi-squared statistic,

$$\xi_P = T \left(\sum_{i=1}^{K} \frac{(\hat{q}_i - q_i^0)^2}{q_i^0} \right) \xrightarrow[T \to \infty]{d} \chi^2(K - 1). \tag{28}$$

When θ^0 is estimated by a square root T consistent estimator $\hat{\theta}$, we can apply our methodology and project the moment $m(\cdot)$ onto the orthogonal space of the score function.

Proposition 8 *Let $\hat{\theta}$ be an estimator of θ^0. Let $\overline{s}_\theta$ be the empirical score function (a vector in $\mathbb{R}^r$) at the estimated parameter, $\overline{s}_\theta = \frac{1}{T} \sum_{t=1}^{T} s(y_t, \hat{\theta})$. Let U be the $K \times r$ matrix of partial* *derivatives of* $q_i(\theta)$, $U = \left[\frac{\partial q(\theta)}{\partial \theta^\top}\right]_{\theta=\theta^0}$. *Let* Λ *be the* $K \times r$ *matrix of row vectors* λ_i,

$$\Lambda = \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_K \end{bmatrix} = U\left[U^\top D^{-1} U\right]^{-1}.$$

Then

$$\xi_P^\star = T\left(\sum_{i=1}^{K} \frac{(\hat{q}_i - q_i^0 - \lambda_i \overline{s}_\theta)^2}{q_i^0}\right) \xrightarrow[T\to\infty]{d} \chi^2(K-1-r). \tag{29}$$

Like in the VaR example, the moment $m_i(y,\theta) = \mathbf{1}\{y \in C_i\} - q_i(\theta)$ is modified to be robust against parameter uncertainty. The particular structure of the variance of the score gives us the previous results. We do not have particular simplifications for the calculations of the λ_i's. Note, however, that they are derived from the primitives of the distribution.

The rank reduction in the chi-squared asymptotic distribution from Eq. (28) to Eq. (29) comes from the fact that r constraints are added when one estimates θ. The sum of the partial derivatives of the q_i's with respect to any component of θ is equal to zero (and the true value).

It is finally worth noting that the empirical score, $\overline{s}_\theta$, is equal to zero when $\hat{\theta}$ is the MLE. Therefore the expression for $\xi_P^\star$ in (29) simplifies to the one in (28), i.e. the usual expression for the Pearson test but with a different asymptotic distribution.

6.2 Poisson counting processes

The Poisson process can be viewed as the analogue of the Gaussian distribution for a discrete variable. For a Poisson distribution with parameter θ, $p_y = e^{-\theta}\frac{\theta^y}{y!}$. Following Section 3.2, the orthonormal family associated with the Poisson distribution is the family of Charlier polynomials $C_j^\theta(y)$, $j = 1, 2$, *etc.* They are defined by the recurrence formula

$$C_{j+1}^\theta(y) = \frac{\theta + j - y}{\sqrt{\theta(j+1)}} C_j^\theta(y) - \sqrt{\frac{j}{j+1}} C_{j-1}^\theta(y)$$

for $j \geq 0$, with $C_0^\theta(y) = 1$ and $C_{-1}^\theta(y) = 0$.

The score function $s_\theta(y)$ is proportional to the first Charlier polynomial:

$$s_\theta(y) = \frac{\partial \ln p_y}{\partial \theta} = -1 + \frac{y}{\theta} = -\frac{C_1^\theta(y)}{\sqrt{\theta}}.$$

Any Charlier polynomial of degree greater than or equal to 2 is consequently robust to the parameter estimation uncertainty when one estimates the parameter θ. The same result holds when there are explanatory variables x and when the specification for the parameter θ is $\theta = f(X, \beta)$ where $f(\cdot)$ is a parametric function and β a parameter to be estimated.

The i.i.d. Poisson process can be extended to a dependent process in the family of integer valued autoregressive processes (INAR) introduced by Al-Osh and Alzaid (1987) to model correlated time series with integer values. The INAR (1) process is defined as

$$y_t = \alpha \circ y_{t-1} + \varepsilon_t, \tag{30}$$

where (ε_t) is a sequence of i.i.d. non-negative and integer valued random variables and $\circ$ is the thinning operator. $\alpha \circ y$ is defined as $\sum_{i=1}^{y} u_i$ with $u_i \overset{i.i.d.}{\sim} B(\alpha)$. The probability that u_i is equal to 1 is α whereas the probability that u_i is equal to 0 is $1 - \alpha$, $\alpha \in [0, 1)$. Equation (30) constructs y_t from the sum of two components: the survivorship component of y_{t-1} (where α is the probability of surviving) and the arrival component ε_t. When $\alpha = 0$, we have the i.i.d counting model.

Different marginal distributions of y_t can be generated depending on the distributional assumption made for ε_t (see Al-Osh and Alzaid, 1987 and McKenzie, 1986, for more details).

When $\varepsilon_t \sim \mathcal{P}o(\mu)$, the model is the Poisson INAR(1) process. It is the analog of the AR(1) process with Gaussian innovations. In this case, the marginal distribution of y_t is also a Poisson distribution with parameter $\theta = \frac{\mu}{1-\alpha}$ (see McKenzie, 1988).

The Charlier Polynomials can be used and are still robust to the parameter uncertainty. Now the serial correlation among the y_t's makes the variance matrix different from the identity. In the special case of the INAR(1) process with Poisson innovation, we can prove the following property.

Proposition 9 *If $y_t \sim INAR(1)$ with parameter α and $\mathcal{P}o(\mu)$ innovation process:*

$$Cov\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} C_j^{\frac{\mu}{1-\alpha}}(y_t), \frac{1}{\sqrt{T}}\sum_{t=1}^{T} C_k^{\frac{\mu}{1-\alpha}}(y_t)\right) = \frac{1+\alpha^j}{1-\alpha^j}\delta_{jk}$$

where δ_{jk} is the Kronecker symbol.

The proof is given in the appendix. It comes from the fact that if y_t is Poisson INAR(1) then $Z_t = C_j^{\frac{\mu}{1-\alpha}}(y_t)$ is also AR(1). The test statistics based on the Charlier polynomials are still asymptotically independent in this case, so

$$\xi = \sum_{k=2}^{p}\left(\frac{1-\alpha^k}{1+\alpha^k}\xi_k^2\right) \sim \chi^2(p-1)$$

with $\xi_k = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} C_k^{\frac{\mu}{1-\alpha}}(y_t)$.

In a more general case where the y_t's are marginally Poisson but exhibit serial correlation, the individual test statistics ξ_k are no longer independent. The variance matrix of a joint test of different components nevertheless can be estimated using a HAC procedure.

6.3 Monte Carlo experiment for the Poisson counting processes

We now present some Monte Carlo simulations for the Poisson distributional test. Four sample sizes are considered: 100, 200, 500 and 1000. As in Section 5, all the results displayed are based on 1 000 replications and each table reports the rejection frequencies for a 5% level test.

We consider individual moments based on a single Charlier polynomial C_k and also weighted moments based on the first Charlier polynomials. $C^w_{2,j}$ is the weighted moment combining C_2 up to C_j using Bartlett weights like in Section 5. As a benchmark, we also display the results of the Pearson chi-squared test. We split our sample in $K = 5$ cells $\{Y = 0\}$,$\{Y = 1\}$,$\{Y = 2\}$,$\{Y = 3\}$,$\{Y \geq 4\}$.

We first study the size properties of our tests where the DGP is a Poisson distribution with parameter $\mu = 2$.[20] The results are displayed in Table 8. The finite sample properties of these tests are clearly good for the first polynomials. The rejection rates are very close to 5% even for very small sample sizes (100 observations). The size is similar whether μ is known or estimated though there exist some differences for very small sample sizes.

[insert Table 8 here]

In Table 9, we study the power properties by simulating several alternatives. We focus on two distributions with two parameters, which have the Poisson as limit distribution. All the distributions have the same expectation, here 2, like for the size results. We estimate the parameter assuming (wrongly) that the distribution is a Poisson. This estimator is in fact the QMLE and it is known that it consistently estimates the expectation of the true distribution.

[20] The theoretical probabilities of belonging to the cells $\{Y = 0\}$,$\{Y = 1\}$,$\{Y = 2\}$,$\{Y = 3\}$, and $\{Y \geq 4\}$ are respectively equal to 13.5%, 27.1%, 27.1%, 18.0%, 13.8%.

We simulate a binomial $\mathcal{B}(k, \frac{2}{k})$ for three values of k (10, 15, and 20). When k tends to infinity, the binomial distribution tends to the Poisson distribution. We do the same thing for the Pascal distribution with parameters $(2, \delta)$ for three values of δ: 10, 15, 20. As δ increases, the Pascal distribution also gets closer to the Poisson distribution. We present the same tests as in Table 8.

Unsurprisingly the power of the tests decreases when k and δ increase. For small samples ($n = 100$) it is more and more difficult to detect departure from the null as we go closer to the Poisson distribution. The performance is very good for the other sample sizes and for most of the moments used, especially for the second Charlier polynomial, which detects the over-dispersion in the data.

[insert Table 9 here]

6.4 Testing the geometric distribution versus its continuous counterpart

The geometric distribution is a particular case of the Pascal distribution and is of interest for discrete duration models. The continuous approximation of the geometric distribution is the exponential distribution, whose hazard rate is also constant. In a VaR backtesting framework, the duration between two consecutive hits is geometrically distributed. Christoffersen *et al.* (2008) tested its continuous approximation, whereas Candelon *et al.* (2011) tested the original discrete distribution. Both correct their test by exact methods *à la* Dufour as the number of observed durations is very low.

In this section, we test the geometric distribution and evaluate the loss of power when one tests the exponential distribution. We assume that we observe full durations and do not consider the case of truncated durations. Our experience is therefore slightly different than in the references cited above but can serve as a useful benchmark.

Assume that time has been discretized and that at each period an event occurs with probability α, independently of the past. The duration y between two consecutive events is therefore a geometric distribution with parameter α and $P(y = k) = \alpha(1-\alpha)^{k-1}$ for $k \geq 1$. Following Table 1, we can derive the sequence of polynomials that are specific moments and are also orthogonal under the null.[21] These polynomials are the Meixner polynomials, which satisfy the following recurrence formula:[22]

$$M_{j,\alpha}(y) = \frac{(1-\alpha)(2j-1)+\alpha(j-y)}{j\sqrt{1-\alpha}} M_{j-1,\alpha}(y) - \frac{j-1}{j} M_{j-2,\alpha}(y),$$

with the convention $M_{0,\alpha}(y) = 1$ and $M_{-1,\alpha}(y) = 0$. Furthermore, any polynomial of degree greater than or equal to two is robust as the score function is the first Meixner polynomial.

Consider now the exponential distribution with parameter α. We know that the polynomials associated with the exponential distribution are the Laguerre polynomials, $L_{j,\alpha}(y)$, defined by

$$L_{j,\alpha}(y) = \frac{\alpha y - (2j-1)}{j} L_{j-1,\alpha}(y) - \frac{j-1}{j} L_{j-2,\alpha}(y),$$

with the convention $L_{0,\alpha} = 1$ and $L_{-1,\alpha} = 0$. The first terms of the two families, $M_{1,\alpha}(y)$ and $L_{1,\alpha}(y)$, are both of expectation zero under the null but the variance of the former is 1 whereas the variance of the latter is $\frac{1}{1-\alpha}$. For higher order, the expectation of $L_{j,\alpha}(y)$ is not equal to zero when y is discrete, geometrically distributed. The expectation is however $o(\alpha)$.

[21] Here again, this is a complete family in L^2. It is sufficient to focus on these moments.

[22] There are some differences with respect to the formulas in Table 2 because here the support of the distribution does not contain 0.

For small α (typically 1% or 5%), we do not expect too much difference in the rejection rates.

We now present the Monte Carlo experiment. We first run $n_s = 1000$ simulations of various sample sizes ($T = 50, 100$, and 500) of i.i.d. random variables y_t that are geometrically distributed with parameter α. We consider the case $\alpha = 5\%$.[23] Table 10 and Table 11 present the results. The moments displayed are the first Meixner polynomials and weighted combinations of these polynomials (from order two), weighted similarly to the Charlier polynomials in Table 8. We also display the results related to the Laguerre polynomials.

The size properties are presented in the first block of columns of Table 10. The size properties related to the Meixner polynomials are good, as are the ones related to the Laguerre polynomials though there is some under-rejection that is more severe for higher orders. For the power properties, we consider two scenarios. In the first one (second and third block of columns of Table 10), the data are generated with other values for α (we consider 4% and 6%), but we test the i.i.d. geometric distribution with parameter $\alpha = 5\%$ instead. This DGP mimics a case where the hits in a VaR context are i.i.d. but computed with the wrong distribution of the innovation term of the underlying GARCH model. The most powerful moment is unsurprisingly the first polynomial. This moment measures the distance between the average duration and the expected one. $M_{1,0.05}(y)$ and $L_{1,0.05}(y)$ are very close to each other and lead to similar rejection rates.

[insert Table 10 here]

In the second scenario (Table 11), the DGP is a geometric distribution with serial correlation. It corresponds to a VaR exercise where the conditional variance for the return is

[23] The case with $\alpha = 1\%$, not presented here, yields similar conclusions.

misspecified. We first generate a Gaussian AR(1) process u_t with parameter ρ respectively equal to 0.4, 0.6, or 0.8, for which u_t is marginally distributed as a standard normal variable. We then define our process,

$$y_t = H\left(\frac{\ln(1-\Phi(u_t))}{\ln(1-\alpha)}\right),$$

where $H(\cdot)$ is the ceiling function, i.e. $H(x)$ is the integer value such that $H(x)-1 < x \leq H(x)$. y_t is geometrically distributed and serially correlated. The level of serial correlation is a monotone function of ρ.

When ρ increases, the rejection rates also increase as the durations are more correlated. When ρ is not too large, like in the first two sets of columns of the table, there is a big gain in using the Meixner polynomials. If we consider the weighted polynomials there is a substantial improvement of power. For $\rho = 0.6$ and $T = 50$, we obtain a 31.7% rejection frequency against 17.5% for the Laguerre polynomials. For large values of ρ, the gain is small as both families lead to high rejection rates. In many cases not considering the discrete nature of the process can reduce the power substantially for local deviations from the null.

[insert Table 11 here]

7 Empirical Application

We illustrate our approach on one empirical application related to VaR forecasts. We consider the exchange rate data that have been considered previously in Kim, Shephard and Chib (1998) and also in Bontemps and Meddahi (2005, 2012). These data are observations of weekday close exchange rates[24] from 1/10/81 to 28/6/85. Bontemps and Meddahi (2005) strongly rejected

[24] These rates are the U.K. Pound, French Franc, Swiss Franc, and Japanese Yen, all versus the U.S. Dollar.

the normality assumption for a GARCH(1,1), whereas BM did not reject the T-GARCH(1,1) model for all the series but the SF-US$ one.

Estimation of the T-GARCH (1,1) model by MLE provides parameter estimates that allow us to compute the one day ahead α-VaR forecast for any value of α. We now test the accuracy of the VaR forecasts in sample for the four series for the values $\alpha = 0.5\%$ and $\alpha = 1\%$ using the moments used in Section 4.2. The estimated parameters and the p-values of the tests are presented in Table 12.

The T-GARCH model is globally rejected for all the series but the FF-US$ rate for $\alpha = 0.5\%$. In most of the cases, the rejection is driven by the conditional tests. In other words there are too many consecutive hits. The dynamic of the model should be modified though the T distribution seems not being rejected.

Observe that, in many cases, correcting the moments $(I_t - \alpha)(I_{t-p} - \alpha)$ for the parameter uncertainty does not detect the departure from the null.

In Table 13, we do the same exercise but out-of-sample, using a rolling estimator based on a T-GARCH(1,1) model estimated on the last 500 observations. With 945 observations, we then test our model from the 445 out-of-sample one day ahead VaR forecasts. The results are however qualitatively the same as in the previous table.

[**insert Table 12 and Table 13 here**]

8 Conclusion

We introduced moment-based tests for parametric discrete distributions. Our goal was to present techniques that are easy to implement without losing power for detecting departures from the null hypothesis. Moment techniques are indeed quite easy to adapt to the time series case and can take parameter estimation uncertainty into account.

We worked with robust moments. When working with estimated parameters, this avoids calculation of any correction term that otherwise needs additional estimations, and consequently greatly simplifies the test procedure. The transformation proposed to yield robust moments encompasses the orthogonalization method of Bontemps and Meddahi (2012), although it still yields a moment that is orthogonal to the score. Results for Monte Carlo experiments suggest that the tests derived have good size and power properties for both in-sample and out-of-sample cases. We also applied the method to a large variety of cases. Backtesting of VaR models, Poisson counting processes, and the geometric distribution are presented here, but the techniques developed could be applied to interval forecast evaluation and parametric discrete choices models, among other discrete models.

We also worked with a finite number of moments. There are many examples for which we do not need to have omnibus tests but power against specific alternatives. However, there are some particular distributions for which it is sufficient to test a countable series of polynomials for consistency. For these cases, derivation of an omnibus test is feasible at a small additional cost. Dealing with an infinite number of moments is therefore a natural extension of this agenda.

Distributional assumptions are necessary in applied econometrics to compute forecasts and to make results tractable in structural economic models so that quantities can be estimated in small data sets. However, the assumptions to validate the results derived should be tested if possible, as they may be biased if there is misspecification. The tests derived using our method are attractive because of their powerful and simplicity.

REFERENCES

Al-Osh, M.A. and A.A. Alzaid (1987), "First order integer-valued autoregressive (INAR(1)) processes", *Journal of Time Series Analysis*, 8, 261-75.

Amengual D. and E. Sentana (2011), "Inference in Multivariate Dynamic Models with Elliptical Innovations", *mimeo* CEMFI.

Amengual D., Fiorentini G. and E. Sentana (2013), "Sequential estimation of shape parameters in multivariate dynamic models", *Journal of Econometrics*, 177, 233-249.

Artzner, P., F. Delbaen, J.-M. Eber and D. Heath (1999), "Coherent Measures of Risk", *Mathematical Finance*, 9, 203-228.

Bai, J. (2003), "Testing Parametric Conditional Distributions of Dynamic Models", *Review of Economics and Statistics*, 85, 531-549.

Berkowitz, J., Peter F. Christoffersen and Denis Pelletier (2011), "Evaluating Value-at-Risk Models with Desk-Level Data", *Management Science*.

Boero, G. and Smith, J. and K. F. Wallis (2004), "Sensitivity of the Chi-Squared Goodness-of-Fit Test to the Partitioning of Data", mimeo, University of Warwick.

Bollerslev, T. and J.F. Wooldridge (1992), "Quasi Maximum Likelihood Estimation and Inference in Dynamic Models with Time Varying Covariances", *Econometric Reviews*, 11, 143-172.

Bontemps, C. and N. Meddahi (2005), "Testing Normality: A GMM Approach", *Journal of Econometrics*, 124, 149-186.

Bontemps, C. and N. Meddahi (2012), "Testing Distributional assumptions: A GMM Approach", *Journal of Applied Econometrics*, 27, 978-1012.

Bontemps, C., Dufour J.-M. and N. Meddahi (2013), "Optimal moment-based tests", mimeo Toulouse School of Economics.

Butler J. S. and P. Chatterjee (1997), "Tests of the Specification of Univariate and Bivariate Ordered Probit", *The Review of Economics and Statistics*, 79, 343-347.

Cameron, A.C. and P.K. Trivedi (1998), *Regression Analysis of Count Data*, New York, Cambridge University Press.

Candelon B., Colletaz G., Hurlin C. et Tokpavi S. (2011), "Backtesting Value-at-Risk: a GMM duration-based test", *Journal of Financial Econometrics*, 1-30.

Chen, Y.-T. (2012), "A Simple Approach to Standardized-Residuals-based Higher-Moment Tests", *Journal of Empirical Finance*, 19, 427-453.

Chesher, A. and R. J. Smith (1997), "Likelihood Ratio Specification Tests", *Econometrica*, 65, 627-646.

Christoffersen, P. F. (1998), "Evaluating Interval Forecasts", *International Economic Review*, 39, 841-862.

Christoffersen, P. F. and D. Pelletier (2004), "Backtesting Value-at-Risk: A Duration-Based Approach", *Journal of Financial Econometrics*, 2, 84-108.

Duan, J.C. (2004), "A Specification Test for Time Series Models by a Normality Transformation", working paper, University of Toronto.

Dufour, J.-M., L. Khalaf and M.-C. Beaulieu (2003), "Exact Skewness-Kurtosis Tests for Multivariate Normality and Goodness-of-Fit in Multivariate Regressions with Application to Asset Pricing Models", *Oxford Bulletin of Economics and Statistics*, 65(s1), 891-906.

Escanciano, J.C. and J. Olmo (2010), "Estimation Risk Effects on Backtesting For Parametric Value-at-Risk Models", *Journal of Business and Economic Statistics*, 28, 36-51.

Fiorentini, G., E. Sentana and G. Calzolari (2004), "On the validity of the Jarque-Bera normality test in conditionally heteroskedastic dynamic regression models", *Economics Letters*, 83 307-312.

Gallant, A.R. (1980), "Explicit Estimators of Parametric Functions in Nonlinear Regressions", *Journal of the American Statistical Association*, 75, 182-193.

Hall, A. (2005), "Generalized Method of Moments", Advanced Texts in Econometrics Series, Oxford University Press.

Hamilton J.D. and O. Jorda (2002), "A Model of the Federal Funds Rate Target", *Journal of Political Economy*, 110, 1135-1167.

Hansen, L.P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators", *Econometrica*, 50, 1029-1054.

Khmaladze, E.V. and H. L. Koul (1984), "Martingale transforms goodness-of-fit tests in regression models", *Annals of Statistics*, 32, 995-1034.

Kim, S., N. Shephard and S. Chib (1998), "Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models", *Review of Economic Studies*, 65, 361-393.

Li, B. (2009), "Asymptotically Distribution-Free Goodness-of-Fit Testing: A Unifying View", *Econometric Reviews*, 28, 632-657.

McCracken, M. W. (2000), "Robust out-of-sample inference", *Journal of Econometrics*, 99, 195-223.

McKenzie, E. (1986), "Autoregressive moving-average processes with negative binomial and geometric marginal distributions", *Adv. Appl. prob.*, 18, 679-705.

Mencia J. and E. Sentana (2012), "Distributional tests in multivariate dynamic models with normal and Student t innovations", *The Review of Economics and Statistics*, 94, 133-152.

Mora, J. and A. Moro-Egido (2008), "On specification testing of ordered discrete choice models", *Journal of Econometrics*, 143, 191-205.

Newey, W.K. (1985), "Generalized Method of Moments Specification Testing", *Journal of Econometrics*, 29, 229-256.

Newey, W.K. and K.D. West (1987), "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix", *Econometrica*, 55, 703-708.

Ord, J.K. (1972), *Families of Frequency Distributions*, Vol 30, Griffin's Statistical Monographs and Courses, Hafner Publishing Company, New York.

Skeels, C. L. and F. Vella (1999), "A Monte Carlo investigation of the sampling behavior of conditional moment tests in Tobit and Probit models", *Journal of Econometrics*,, 92, 275-294.

Szegö, G. (1967), *Orthogonal polynomials*, Amer. Math. Soc.

Tauchen, G.E. (1985), "Diagnostic Testing and Evaluation of Maximum Likelihood Models", *Journal of Econometrics*, 30, 415-443.

Weber, M. and A. Erdelyi (1952), "On the finite difference analogue of Rodrigues' formula", *Amer. Math. Monthly*, 59, 163-168.

West, K. D. (1996), "Asymptotic Inference about Predictive Ability", *Econometrica*, 64, 1067-1084.

West, K. D. and M. W. McCracken, (1998), "Regression-Based Tests of Predictive Ability", *International Economic Review*, 39, 817-840.

Wooldridge, J. (1990), "A Unified Approach to Robust, Regression-based Specification Tests", *Econometric Theory*, 6, 17-43.

Appendices

A Proof of the propositions

A.1 Proof of Proposition 4.

We first prove the proposition in the case where N is infinite.

$$\mathbb{E}[\Delta\psi(y,\theta)] = \sum_{i=0}^{+\infty} \left(\psi(i+1,\theta) - \psi(i,\theta)\right) p_i(\theta) \tag{A.1}$$

Reordering the second term of the last expression yields

$$\begin{aligned}
\mathbb{E}[\Delta\psi(y,\theta)] &= \sum_{i=0}^{+\infty} \psi(i+1,\theta)p_i(\theta) - \sum_{i=0}^{+\infty} \psi(i,\theta)p_i(\theta) && \text{(A.2)}\\
&= \sum_{i=0}^{+\infty} \psi(i+1,\theta)p_i(\theta) - \sum_{i=1}^{+\infty} \psi(i,\theta)p_i(\theta) \text{ under } \textbf{LB.} && \text{(A.3)}\\
&= \sum_{i=0}^{+\infty} \psi(i+1,\theta)p_i(\theta) - \sum_{i=0}^{+\infty} \psi(i+1,\theta)p_{i+1}(\theta) && \text{(A.4)}\\
&= -\sum_{i=0}^{+\infty} \psi(i+1,\theta)\left(p_{i+1}(\theta) - p_i(\theta)\right) && \text{(A.5)}\\
&= -\mathbb{E}\left(\psi(y+1,\theta)\frac{\Delta p(y,\theta)}{p(y,\theta)}\right) && \text{(A.6)}
\end{aligned}$$

When N is finite, the proof is similar as $p_i(\theta)$ is equal to zero when $i \geq (N+1)$.

A.2 Proof of Proposition 5.

Let $m(y,\theta)$ a moment such that

$$\mathbb{E}_0 m(y,\theta^0) = 0. \tag{A.7}$$

Let $\psi(y,\theta)$, a function, defined on S by:

$$\begin{aligned}
&\psi(0,\theta) = 0,\\
&\psi(y,\theta) = \frac{1}{p_y(\theta)} \sum_{k=0}^{y-1} m(k,\theta)p_k(\theta) \quad \text{for } y \geq 1
\end{aligned} \tag{A.8}$$

Then,

$$\begin{aligned}\Delta\psi(y,\theta)+\psi(y+1,\theta)\frac{\Delta p_y(\theta)}{p_y(\theta)} &= \psi(y+1,\theta)-\psi(y,\theta)+\psi(y+1,xy,\theta)\left(\frac{p_{y+1}(\theta)}{p_y(\theta)}-1\right)\\ &= \psi(y+1,\theta)\frac{p_{y+1}(\theta)}{p_y(\theta)}-\psi(y,\theta)\\ &= \frac{1}{p_y(\theta)}\sum_{k=0}^{y} m(k,\theta)p_k(\theta)-\frac{1}{p_y(\theta)}\sum_{k=0}^{y-1} m(k,\theta)p_k(\theta)\\ &\text{(using the definition in A.8)}\\ &= m(y,\theta).\end{aligned}$$

Observe that the last equality holds without the expectation.

A.3 Comparison of the two strategies

Let $g_1(\cdot)$ be the GMM estimating equation which is used to estimate θ within the data. We assume that θ is estimated consistently by g_1 under both the null and the alternative. Let $m(y,\theta)$ a moment used to test q_0, the p.d.f. under the null, and $g_2(\cdot)$ an estimating equation which is used to construct the robust version $\tilde{m}_{g_2}$ in (6), i.e. the direction of the projection.

$m^{\perp}(y,\theta) = m(y,\theta) - \mathbb{E}_0\left[m.s_\theta^\top\right]\mathbb{V}_0\left[s_\theta\right]^{-1}s_\theta(y_t)$ is the orthogonal projection of $m(\cdot)$ onto the orthogonal of the score function, i.e. $\tilde{m}_{s_\theta}$.

Similarly, $g_2^{\perp}(y,\theta) = g(y,\theta) - \mathbb{E}_0\left[g.s_\theta^\top\right]\mathbb{V}_0\left[s_\theta\right]^{-1}s_\theta(y_t)$ is the orthogonal projection of $g_2(\cdot)$ onto the orthogonal of the score function.

Proposition 10 *Let $q_1(y) = q_0(y)\left(1+h(y)/\sqrt{T}\right)$, a local alternative (denoted H_1), where $h(\cdot)$ is orthogonal to the score function under the null. $\mathbb{E}_1$ denotes the expectation under H_1.*

(i) If one decides to correct for the parameter uncertainty, the test statistic in (1) becomes

$$\xi_m^{g_1} = T\frac{\left(\frac{1}{T}\sum_{t=1}^{T} m(y_t,\hat{\theta})\right)^2}{\mathbb{V}_0(\tilde{m}_{g_1})}. \tag{A.9}$$

Under H_1, its limiting distribution is a non central χ^2 distribution with one degree of freedom and noncentrality parameter $a(g_1) = \frac{(\mathbb{E}_1[\tilde{m}_{g_1}.h])^2}{\mathbb{V}_0[\tilde{m}_{g_1}]}$.

(ii) If one decides to use the projection of $m(\cdot)$ along $g_2(\cdot)$ onto the orthogonal of the score, the test statistic $\xi^{\perp}_{\tilde{m}_{g_2}}$ is equal to

$$\xi^{\perp}_{\tilde{m}_{g_2}} = T\frac{\left(\frac{1}{T}\sum_{t=1}^{T} \tilde{m}_{g_2}(y_t,\hat{\theta})\right)^2}{\mathbb{V}_0(\tilde{m}_{g_2})}. \tag{A.10}$$

Its asymptotic distribution under H_1 is a non central χ^2 distribution with one degree of freedom and noncentrality parameter $a(g_2)$.

(iii) Comparing the power properties under H_1 of the two strategies is comparing the non-centrality parameters $a(g_1)$ and $a(g_2)$. If $g_1 = g_2$, they have similar properties. The optimal estimating equation $g(\cdot)$ is proportional to $m - \lambda h$, where λ is a scalar. Without knowing h, it is often impossible to rank $a(g_1)$ and $a(g_2)$.

The proof is given below. Proposition 10 characterizes the local power properties of the two strategies. Observe that the noncentrality parameter $a(g)$ under both scenario is also known as the slope of the test i.e. the limit of ξ/T when $T \to +\infty$. Maximizing the power under the sequence of local alternative is maximizing this slope.

Proposition 10 also shows that working with robust moments or correcting are two strategies that can be considered as equivalent. They have the same power properties when they consider the same estimating equation $g(\cdot)$. For two different choices of $g(\cdot)$ there is no obvious ranking between $a(g_1)$ and $a(g_2)$. In *(iv)* we characterize a case where it is optimal to consider the true score as the estimating equation, i.e. working with the orthogonal projection onto the orthogonal score as robust moment. However for any other case such a ranking is not straightforward and the question should be tackled case by case to draw conclusions.

Proof of (i) and (ii):

• If we choose to work with $m(\cdot)$ and to correct for the parameter uncertainty, we have the following asymptotic result:

$$\begin{aligned}\sqrt{T}\frac{1}{T}\sum_{t=1}^{T} m(y_t,\hat{\theta}) &= \sqrt{T}\frac{1}{T}\sum_{t=1}^{T} m(y_t,\theta^0) - \mathbb{E}_0\left[m.s_\theta^\top\right]\sqrt{T}(\hat{\theta}-\theta^0) + o_P(1),\\ &= \sqrt{T}\frac{1}{T}\sum_{t=1}^{T}\Big(m(y_t,\theta^0) - \mathbb{E}_0\left[m.s_\theta^\top\right]\mathbb{E}_0\left[g_1.s_\theta^\top\right]^{-1} g_1(y_t,\theta^0)\Big) + o_P(1),\\ &= \sqrt{T}\frac{1}{T}\sum_{t=1}^{T}\tilde{m}_{g_1}(y_t,\theta^0) + o_P(1),\end{aligned} \tag{A.11}$$

where $\tilde{m}_{g_1}(\cdot)$ is defined in (6). The expression of $\xi_m^{g_1}$ in (A.9) follows.

The variances of any moment under H_0 and under H_1 are equal to each other at the first order:

$$\mathbb{V}_1(\tilde{m}_{g_1}(y_t,\theta^0)) = \mathbb{V}_0(\tilde{m}_{g_1}(y_t,\theta^0)) + o_P(1). \tag{A.12}$$

Moreover, the expectation of $\tilde{m}_{g_1}$ under the alternative can be simplified:

$$
\begin{aligned}
\mathbb{E}_1(\tilde{m}_{g_1}) &= \int \tilde{m}_{g_1}(y,\theta^0)(q_0(y)+h(y)q_0(y)/\sqrt{T})dy \\
&= \frac{1}{\sqrt{T}}\int \tilde{m}_{g_1}(y,\theta^0)h(y)q_0(y)dy \\
&= \frac{1}{\sqrt{T}}\int (m^{\perp}(y,\theta^0) - \mathbb{E}_0\left[m.s_\theta^\top\right]\mathbb{E}_0\left[g_1.s_\theta^\top\right]^{-1} {g_1}^{\perp}(y,\theta^0))h(y)q_0(y)dy \\
&= \frac{1}{\sqrt{T}}\mathbb{E}_0\left[\left(m^{\perp} - \mathbb{E}_0\left[m.s_\theta^\top\right]\mathbb{E}_0\left[g_1.s_\theta^\top\right]^{-1} {g_1}^{\perp}\right).h\right],
\end{aligned}
\tag{A.13}
$$

where $m^{\perp}(y,\theta^0) = m(y,\theta^0) - \mathbb{E}_0\left[m.s_\theta^\top\right]\mathbb{V}_0\left[s_\theta\right]^{-1} s_\theta(y_t)$, i.e. the orthogonal projection of $m(\cdot)$ onto the orthogonal of the score function and where ${g_1}^{\perp}(y,\theta^0) = g_1(y,\theta^0) - \mathbb{E}_0\left[g_1.s_\theta^\top\right]\mathbb{V}_0\left[s_\theta\right]^{-1} s_\theta(y_t)$ is defined similarly.

Consequently,

$$
\begin{aligned}
\xi_m^{g_1} &= \frac{\left(\sqrt{T}\frac{1}{T}\sum_{t=1}^T \tilde{m}_{g_1}(y_t,\theta^0) + o_P(1)\right)^2}{\mathbb{V}_0(\tilde{m}_{g_1})}, \\
&= \frac{\left(\sqrt{T}\left(\frac{1}{T}\sum_{t=1}^T \tilde{m}_{g_1}(y_t,\theta^0) - \mathbb{E}_1(\tilde{m}_{g_1})\right) + \sqrt{T}\mathbb{E}_1(\tilde{m}_{g_1}) + o_P(1)\right)^2}{\mathbb{V}_0(\tilde{m}_{g_1})}, \\
&= \left(\frac{\sqrt{T}\left(\frac{1}{T}\sum_{t=1}^T \tilde{m}_{g_1}(y_t,\theta^0) - \mathbb{E}_1(\tilde{m}_{g_1})\right)}{\sqrt{\mathbb{V}_1(\tilde{m}_{g_1})}} + \frac{\mathbb{E}_0\left[\left(m^{\perp} - \mathbb{E}_0\left[m.s_\theta^\top\right]\mathbb{E}_0\left[g_1.s_\theta^\top\right]^{-1} {g_1}^{\perp}\right).h\right]}{\sqrt{\mathbb{V}_0(\tilde{m}_{g_1})}} + o_P(1)\right)^2, \\
&= \left(Z + \frac{\mathbb{E}_0\left[\left(m^{\perp} - \mathbb{E}_0\left[m.s_\theta^\top\right]\mathbb{E}_0\left[g_1.s_\theta^\top\right]^{-1} {g_1}^{\perp}\right).h\right]}{\sqrt{\mathbb{V}_0(\tilde{m}_{g_1})}}\right)^2 + o_P(1),
\end{aligned}
$$

where Z is a standard normal random variable.

• If now we choose to work with a robust version of $m(\cdot)$, $\tilde{m}_{g_2}$, we do not have to correct. A similar expansion leads to the result.

Proof of (iii): We can apply the Cauchy-Schwarz inequality $(\mathbb{E}_1[\tilde{m}_{g_1}.h])^2 \leq \mathbb{E}_1[\tilde{m}_{g_1}^2]\mathbb{E}_1[h^2]$, we can bound $a(g_1)$ by $\mathbb{E}_1[h^2]$. This upper bound is reached when $\tilde{m}_{g_1} = \lambda h$ for a scalar λ. The set of moments m such that $\tilde{m}_{g_1} = \lambda h$ are moments which are proportional to $\lambda h + \kappa g_1$ or equivalently when g_1 is chosen to be proportional to $m - \lambda h$.

In most of the cases there is no systematic ranking of $a(g_1)$ and $a(g_2)$ for any type of estimating equation. In particular working with the orthogonal projection of the moment onto the orthogonal of the score does not systematically gives a higher slope in a general context.

A.4 Proof of Proposition 6

Let $F_\nu(\cdot)$, $f_\nu(\cdot)$, q_α^ν be respectively the cdf, pdf and α-quantile of the distribution of the innovation term, ε_t. In the constant location-scale model (18), the log pdf of the returns, r_t, is equal to

$$\log \varphi(r_t, \theta) = -\frac{1}{2}\log(\sigma^2) + \log f_\nu\left(\frac{r_t - \mu}{\sigma}\right).$$

The score function is consequently equal to

$$\tilde{s}_\theta(\varepsilon_t) = \begin{bmatrix} -\frac{1}{\sigma}\frac{\partial \log f_\nu}{\partial \varepsilon_t}(\varepsilon_t) \\ -\frac{1}{2\sigma^2}\left(1 + \varepsilon_t \frac{\partial \log f_\nu}{\partial \varepsilon_t}(\varepsilon_t)\right) \\ \frac{\partial \log f_\nu}{\partial \nu}(\varepsilon_t) \end{bmatrix}. \tag{A.14}$$

The projection of $I_t - \alpha = \mathbf{1}\{r_t \leq -VaR_t^\alpha\} - \alpha = \mathbf{1}\{\varepsilon_t \leq q_\alpha^\nu\} - \alpha$ on the orthogonal of the score function is

$$e_t = I_t - \alpha - \mathbb{E}\left[(I_t - \alpha)s_\theta(\varepsilon_t)\right] V_s^{-1}\tilde{s}_\theta(\varepsilon_t).$$

Standard calculations simplifies the covariance between the hit and the score function:

$$P = \mathbb{E}\left[(I_t - \alpha)\tilde{s}_\theta(\varepsilon_t)^\top\right] = \left[-\frac{1}{\sigma}f_\nu(q_\alpha^\nu), -\frac{1}{2\sigma^2}q_\alpha^\nu f_\nu(q_\alpha^\nu), \frac{\partial F_\nu}{\partial \nu}(q_\alpha^\nu)\right]^\top. \tag{A.15}$$

For example, the first component of P, P_1, is equal to

$$P_1 = -\frac{1}{\sigma}\int_{-\infty}^{q_\alpha^\nu} \frac{\partial f_\nu}{\partial \varepsilon}(\varepsilon)\, d\varepsilon = -\frac{1}{\sigma}f_\nu(q_\alpha^\nu).$$

The two other components are derived similarly

For any random variable Z_{t-1} in the information set at time $t-1$, $Z_{t-1}e_t$ is orthogonal to the score using the law of iterated expectations. Moreover

$$\mathbb{V}(Z_{t-1}e_t) = \mathbb{E}\left[Z_{t-1}Z_{t-1}^\top \mathbb{E}(e_t^2|I_{t-1})\right] = \mathbb{E}\left[Z_{t-1}Z_{t-1}^\top\right](\alpha(1-\alpha) - PV_s^{-1}P^\top).$$

A.5 Pearson chi-squared test

Using the notations of Section 6.1, we can compute the variance Σ of the moment vector $m(y,\theta) = [m_1(y,\theta), \ldots, m_K(y,\theta)]^\top$, $\Sigma = D - QQ^\top$. Let e be the $K \times 1$ vector of 1's, a generalized inverse of Σ, Σ^-, is $D^{-1} - \frac{ee^\top}{K}$. Note that $e^\top Q = 1$.

Let $\hat{m}_T^0 = \frac{1}{T}\sum_{t=1}^T m(y_t, \theta^0) = [\hat{q}_1 - q_1^0, \ldots, \hat{q}_K - q_K^0]^\top$. Observe that $e^\top(\hat{m}_T^0) = 0$. The test statistic derived from this moment has a chi-squared distribution with $rk(\Sigma) = K - 1$ degrees of freedom and can be simplified as

$$
\begin{aligned}
\xi_P &= T(\hat{m}_T^0)^\top \Sigma^-(\hat{m}_T^0) \\
&= T(\hat{m}_T^0)^\top D^{-1}(\hat{m}_T^0) \\
&= T\left(\sum_{j=1}^{K} \frac{(\hat{q}_j - q_j^0)^2}{q_j^0}\right)
\end{aligned}
$$

Proof of Proposition 4. When θ is estimated, we need first to compute the covariance between $m(\cdot)$ and the score function and the variance of the score. The score function is :

$$
s_\theta(y) = \frac{\partial \log p_y(\theta)}{\partial \theta}.
$$

The two following matrices are derived using standard calculations. Under the null,

$$
\begin{aligned}
P &= \mathbb{E}_0\left[m.s_\theta^\top\right] = \left(\frac{\partial q_i}{\partial \theta^j}\right)_{i=1,\dots,K;j=1,\dots,r} = U, \\
V_s &= \mathbb{E}_0\left[s_\theta.s_\theta^\top\right] = U^\top D^{-1} U.
\end{aligned}
$$

Let $m^\perp(y,\theta) = m(y,\theta) - PV_s^{-1}s_\theta(y)$.

We know prove that the rang of the variance of $m^\perp$ is equal to $K-r-1$. Note first that the sum of the components of any column of U is equal to zero. Note also that $D^{-1}Q = e$. Consequently, $U^\top D^{-1} Q = 0$.

$$
\begin{aligned}
\mathbb{V}_0(m^\perp) &= D - QQ^\top - U\left[U^\top D^{-1} U\right]^{-1} U^\top \\
&= D\left(I_K - D^{-1/2}QQ^\top D^{-1/2} - D^{-1/2}U\left[U^\top D^{-1}U\right]^{-1} U^\top D^{-1/2}\right) \\
&= D\left(I - C(C^\top C)^{-1} C^\top\right),
\end{aligned}
$$

where C is the $K \times (r+1)$ matrix created by the horizontal concatenation of the $K \times 1$ matrix $D^{-1/2}Q$ and the $K \times r$ matrix $D^{-1/2}U$. This matrix is of rank equal to $r+1$ and note that the first column is orthogonal to the last r columns due to the orthogonality property explained above. The variance matrix is the product of an invertible matrix and an orthogonal projector of rank $K-r-1$. Consequently

$$
T\left(\frac{1}{T}\sum_{t=1}^{T} m^\perp(y_t,\hat{\theta})\right)^\top D^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} m^\perp(y_t,\hat{\theta})\right) \xrightarrow[T\to\infty]{d} \chi^2(K-1-r).
$$

A.6 Proof of Proposition 9

Let first consider the generating function of the orthonormalized Charlier polynomials $C_j^\theta(y)$, $j \in \mathbb{N}$:

$$\sum_{j=0}^{+\infty} C_j^\theta(y) \frac{w^j}{\sqrt{j!\theta^j}} = e^w \left(1 - \frac{w}{\theta}\right)^y$$

In the Poisson INAR(1) model, the marginal distribution of y_t is a Poisson with parameter $\theta = \frac{\mu}{1-\alpha}$.

Using the previous expression with $y \equiv y_t$ and assuming that the sum can commute with $\mathbb{E}_{t-1}$ (the conditional expectation at time $t-1$), one obtains:

$$\sum_{j=0}^{+\infty} \mathbb{E}_{t-1} C_j^\theta(y_t) \frac{w^j}{\sqrt{j!\theta^j}} = e^w \mathbb{E}_{t-1} \left(1 - \frac{w}{\theta}\right)^{y_t}. \tag{A.16}$$

the conditional probability $p(y_t|y_{t-1})$ of y_t conditional on y_{t-1} is equal to (Freeland and McCabe, 2004)

$$p(y_t|y_{t-1}) = \sum_{s=0}^{min(y_t,y_{t-1})} C_{y_{t-1}}^s \alpha^s (1-\alpha)^{y_{t-1}-s} \frac{e^{-\mu}\mu^{y_t-s}}{(y_t-s)!}.$$

We use this last expression to calculate the second part of (A.16).

$$\begin{aligned}
\mathbb{E}_{t-1}\left(1-\frac{w}{\theta}\right)^{y_t} &= \sum_{k=0}^{+\infty} p(k|y_{t-1}) \left(1 - \frac{w(1-\alpha)}{\mu}\right)^k \\
&= \sum_{k=0}^{+\infty} \sum_{s=0}^{min(k,y_{t-1})} C_{y_{t-1}}^s \alpha^s (1-\alpha)^{y_{t-1}-s} \frac{e^{-\mu}\mu^{k-s}}{(k-s)!} \left(1 - \frac{w(1-\alpha)}{\mu}\right)^k \\
&= \sum_{s=0}^{y_{t-1}} \sum_{k=s}^{+\infty} C_{y_{t-1}}^s \alpha^s (1-\alpha)^{y_{t-1}-s} \frac{e^{-\mu}\mu^{k-s}}{(k-s)!} \left(1 - \frac{w(1-\alpha)}{\mu}\right)^k \\
&= \sum_{s=0}^{y_{t-1}} C_{y_{t-1}}^s \alpha^s (1-\alpha)^{y_{t-1}-s} e^{-w(1-\alpha)} \left(1 - \frac{w(1-\alpha)}{\mu}\right)^s \\
&= e^{-w(1-\alpha)} \left(1 - \frac{\alpha w(1-\alpha)}{\mu}\right)^{y_t-1}
\end{aligned}$$

We can now plug the last result into (A.16) to get

$$\sum_{j=0}^{+\infty} \mathbb{E}_{t-1} C_j^{\frac{\mu}{1-\alpha}}(y_t) \frac{w^j}{\sqrt{j!\left(\frac{\mu}{1-\alpha}\right)^j}} = e^{w\alpha}\left(1 - \frac{\alpha w(1-\alpha)}{\mu}\right)^{y_t-1}$$
$$= \sum_{j=0}^{+\infty} \alpha^j C_j^{\frac{\mu}{1-\alpha}}(y_{t-1}) \frac{w^j}{\sqrt{j!\left(\frac{\mu}{1-\alpha}\right)^j}}$$

and so, making each term of w^j equal, we obtain

$$\mathbb{E}_{t-1} C_j^{\frac{\mu}{1-\alpha}}(y_t) = \alpha^j C_j^{\frac{\mu}{1-\alpha}}(y_{t-1}).$$

$C_j^{\frac{\mu}{1-\alpha}}(y_t)$ is therefore an AR(1) process with parameter α^j. The expression of the covariance follows immediately.

B Calculations related to Section 4

B.1 The GARCH(1,1) with normal innovations

We consider here a GARCH(1,1) model with independent normal innovations.

Let $\Phi(\cdot)$, $\varphi(\cdot)$, n_α be respectively the cdf, the pdf and the α-quantile of the standard normal distribution.

Following Proposition 6, the covariance P and the variance of the score in the constant location-scale model V_s are equal to, in the particular case of a Normal GARCH model,

$$P = \left[-\frac{\varphi(n_\alpha)}{\sigma}, -\frac{\varphi(n_\alpha)n_\alpha}{2\sigma^2}\right], \text{ and } V_s = diag(\frac{1}{\sigma}, \frac{1}{2\sigma^2}).$$

The robust moment is therefore based on the new robust term

$$e_t = I_t - \alpha + \varphi(n_\alpha)\varepsilon_t + \frac{n_\alpha \varphi(n_\alpha)}{2}\left(\varepsilon_t^2 - 1\right). \tag{B.17}$$

If we consider now the model without drift:

$$r_t = \sqrt{\sigma_t^2(\theta)}\varepsilon_t,\ \sigma_t^2(\theta) = \omega + \gamma r_{t-1}^2 + \beta\sigma_{t-1}^2,$$

The score function, up to a scale factor, is equal to

$$s_\theta(r_t) = \frac{\partial \ln \sigma_t(\theta)}{\partial \theta}\left(\left(\frac{r_t}{\sigma_t(\theta)}\right)^2 - 1\right).$$

Therefore

$$V_s = \mathbb{V}(s_\theta) = 2\mathbb{E}\left[\frac{\partial \ln \sigma_t(\theta)}{\partial \theta}\frac{\partial \ln \sigma_t(\theta)}{\partial \theta^\top}\right]$$

and the covariance between the hit function, I_t and the score function is

$$P = \mathbb{E}\left(\mathbf{1}\{r_t \leq \sigma_t(\theta)n_\alpha\}s_\theta^\top(r_t)\right) = -q_\alpha\varphi(q_\alpha)\mathbb{E}\left[\frac{\partial \ln \sigma_t(\theta)}{\partial \theta^\top}\right].$$

The projection, $e_t^\star$ of $I_t - \alpha$ onto the orthogonal space of the score function is

$$e_t^\star = I_t - \alpha + \frac{q_\alpha\varphi(q_\alpha)}{2}\mathbb{E}\left[\frac{\partial \ln \sigma_t(\theta)}{\partial \theta^\top}\right]\mathbb{E}\left[\frac{\partial \ln \sigma_t(\theta)}{\partial \theta}\frac{\partial \ln \sigma_t(\theta)}{\partial \theta^\top}\right]^{-1} s_\theta(r_t). \tag{B.18}$$

The variance of $e_t^\star$ is equal to:

$$\left(\alpha(1-\alpha) - \frac{\left(q_\alpha\varphi(q_\alpha)\right)^2}{2}\mathbb{E}\left[\frac{\partial \ln \sigma_t(\theta)}{\partial \theta^\top}\right]\mathbb{E}\left[\frac{\partial \ln \sigma_t(\theta)}{\partial \theta}\frac{\partial \ln \sigma_t(\theta)}{\partial \theta^\top}\right]^{-1}\mathbb{E}\left[\frac{\partial \ln \sigma_t(\theta)}{\partial \theta^\top}\right]^\top\right).$$

The last matrices can be estimated in the sample using the following results:

$$\begin{aligned}
\frac{\partial \ln \sigma_t(\theta)}{\partial \omega} &= \frac{1}{2\sigma^2(\theta)}\frac{1}{1-\beta},\\
\frac{\partial \ln \sigma_t(\theta)}{\partial \gamma} &= \frac{1}{2\sigma^2(\theta)}\sum_{k=1}^{+\infty}\beta^{k-1}r_{t-k}^2,\\
\frac{\partial \ln \sigma_t(\theta)}{\partial \beta} &= \frac{1}{2\sigma^2(\theta)}\sum_{k=1}^{+\infty}\beta^{k-1}\sigma_{t-k}^2.
\end{aligned}$$

B.2 The T-GARCH(1,1) model

We now consider the general T-GARCH model

$$r_t = \mu_{t-1}(\theta) + \sigma_{t-1}(\theta)\varepsilon_t,$$

where $\mu_{t-1}(\cdot)$ and $\sigma_{t-1}^2(\cdot)$ are the conditional mean and variance of r_t given the past and where ε_t is a i.i.id sequence from a **standardized** Student distribution with ν degrees of freedom. $F_\nu(\cdot)$, $f_\nu(\cdot)$, q_α^ν are respectively the cdf, the pdf and the α-quantile of the **standardized** Student distribution.

The pdf is equal to

$$f_\nu(\varepsilon_t) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\Gamma(1/2)\sqrt{\nu-2}}\frac{1}{\left(1+\frac{\varepsilon_t^2}{\nu-2}\right)^{(\nu+1)/2}}.$$

Following Proposition 6, a robust version of $I_t - \alpha$ can be derived using a constant location scale-model for the returns, $r_t = \mu + \sigma\varepsilon_t$. The values of P is given in Equation (A.15). The score function in the auxiliary model is equal to

$$\tilde{s}_\theta(\varepsilon_t) = \begin{bmatrix} -\frac{1}{\sigma}\frac{\partial \log f_\nu}{\partial \varepsilon_t}(\varepsilon_t) \\ -\frac{1}{2\sigma^2}\left(1 + \varepsilon_t \frac{\partial \log f_\nu}{\partial \varepsilon_t}(\varepsilon_t)\right) \\ \dfrac{\partial \log f_\nu}{\partial \nu}(\varepsilon_t) \end{bmatrix} = \begin{bmatrix} \frac{\nu+1}{\sigma}\frac{\varepsilon_t}{{\varepsilon_t}^2+\nu-2} \\ -\frac{1}{2\sigma^2}\left(1 - \frac{(\nu+1)\varepsilon_t^2}{{\varepsilon_t}^2+\nu-2}\right) \\ \dfrac{\partial \log f_\nu}{\partial \nu}(\varepsilon_t) \end{bmatrix}.$$

Variance of the score function We now give some details related to the calculation of the variance of the score. The first component is uncorrelated to the two other components by symmetry.

$$\begin{aligned} \mathbb{V}(\frac{\varepsilon_t}{{\varepsilon_t}^2+\nu-2}) &= \int_{-\infty}^{+\infty} \frac{\varepsilon^2}{(\varepsilon^2+\nu-2)^2} f_\nu(\varepsilon) d\varepsilon \\ &= \frac{\nu}{\nu-2}\int_{-\infty}^{+\infty} \frac{z^2}{(z^2+\nu)^2} h_\nu(z) dz \\ &= \frac{\nu}{\nu-2}\left(\mathbb{E}(\frac{1}{z^2+\nu}) - \nu\mathbb{E}(\frac{1}{(z^2+\nu)^2})\right), \end{aligned}$$

where $z = \varepsilon\sqrt{\frac{\nu}{\nu-2}}$, follows a Student distribution with ν degrees of freedom and $h_\nu(\cdot)$ is its pdf (we use the same change of variables in this section, for the other calculations). These expectations are standard (see Appendix C.1 of BM in particular) and

$$\mathbb{V}(\frac{\varepsilon_t}{{\varepsilon_t}^2+\nu-2}) = \frac{\nu}{\nu-2}\frac{1}{(\nu+1)(\nu+3)}.$$

The variance of the second component is computed similarly.

$$\begin{aligned} \mathbb{V}\left(1 - \frac{(\nu+1)\varepsilon_t^2}{{\varepsilon_t}^2+\nu-2}\right) &= \int_{-\infty}^{+\infty} \frac{(-\nu\varepsilon^2+\nu-2)^2}{(\varepsilon^2+\nu-2)^2} f_\nu(\varepsilon) d\varepsilon \\ &= \nu^2 \int_{-\infty}^{+\infty} \frac{(z^2-1)^2}{(z^2+\nu)^2} h_\nu(z) dz \\ &= \nu^2\left(1 - 2(\nu+1)\mathbb{E}(\frac{1}{z^2+\nu}) + (\nu+1)^2\mathbb{E}(\frac{1}{(z^2+\nu)^2})\right) \\ &= \frac{2\nu}{\nu+3}. \end{aligned}$$

We do not have particular closed forms for $G = \mathbb{E}\left(-\frac{1}{2\sigma^2}\left(1 - \frac{(\nu+1)\varepsilon_t^2}{{\varepsilon_t}^2+\nu-2}\right)\frac{\partial \log f_\nu}{\partial \nu}(\varepsilon_t)\right)$ and $H = \mathbb{V}\left(\frac{\partial \log f_\nu}{\partial \nu}(\varepsilon_t)\right)$. However, we can either estimate them within the data or by simulation techniques.

The variance matrix V_s is therefore equal to

$$V_s = \begin{bmatrix} \frac{1}{\sigma^2}\frac{\nu}{\nu-2}\frac{\nu+1}{\nu+3} & 0 & 0 \\ 0 & \frac{1}{2\sigma^4}\frac{\nu}{\nu+3} & G \\ 0 & G & H \end{bmatrix}.$$

Oblique projection for building a robust moment An alternative is to project $I_t - \alpha$ along an estimating equation. In this case we can use the first two moments for μ and σ^2 and the fourth moment to estimate ν. The estimating equation can be

$$g(r_t, \theta) = \begin{bmatrix} r_t - \mu \\ (r_t - \mu)^2 - \sigma^2 \\ ((r_t - \mu)^4 - 3\sigma^4)(\nu - 4) - 6\sigma^4 \end{bmatrix}. \tag{B.19}$$

Using the results of Section 2.3, we can use a new robust version, $\tilde{e}_t$ of $I_t - \alpha$

$$\tilde{e}_t = I_t - \alpha - \mathbb{E}[m.\tilde{s}_\theta^\top]\mathbb{E}[g.\tilde{s}_\theta^\top]^{-1} g(r_t, \theta) \tag{B.20}$$

The first matrix in (B.20) is exactly the matrix P derived in Eq. (A.15) and does not depend on the choice of $g(\cdot)$. The second matrix replaces the matrix V_s^{-1} and is equal to:

$$M = \mathbb{E}\left[g.\tilde{s}_\theta^\top\right]^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 6\sigma^2(\nu-2) & \frac{-6\sigma^4}{\nu-4} \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{(\nu-4)(\nu-2)}{\sigma^2} & \frac{-(\nu-4)}{6\sigma^4} \end{bmatrix}$$

Let us give some details about the calculations. We denote by $g_1(\cdot)$, $g_2(\cdot)$, $g_3(\cdot)$ the three components of g. $g_1(\cdot)$ is uncorrelated with the second and third components, $\tilde{s}_{\theta,2}(\cdot)$ and

$\tilde{s}_{\theta,3}(\cdot)$ of the score function. The covariance with the first component of the score is derived using the same method as for the variance of the score.

$$\mathbb{E}\left[g_1.\tilde{s}_{\theta,1}\right] = \int_{-\infty}^{+\infty} \frac{(\nu+1)\varepsilon^2}{\varepsilon^2+\nu-2} f_\nu(\varepsilon)d\varepsilon = \int_{-\infty}^{+\infty} \frac{(\nu+1)z^2}{z^2+\nu} h_\nu(z)dz = 1.$$

For the second component, $g_2(\cdot)$, it is also, by symmetry, uncorrelated to the first component of the score. the covariance with the second component is equal to

$$\begin{aligned}
\mathbb{E}\left[g_2.\tilde{s}_{\theta,2}\right] &= -\frac{1}{2}\int_{-\infty}^{+\infty} \left(1 - \frac{(\nu+1)\varepsilon^2}{\varepsilon^2+\nu-2}\right)(\varepsilon^2-1) f_\nu(\varepsilon)d\varepsilon \\
&= -\frac{1}{2}\mathbb{E}\left(-\nu(\varepsilon^2-1) + (\nu-2)(\nu+1)\frac{\varepsilon^2-1}{\varepsilon^2+\nu-2}\right) \\
&= -\frac{(\nu-2)(\nu+1)}{2}\mathbb{E}\left(\frac{\varepsilon^2-1}{\varepsilon^2+\nu-2}\right) \\
&= -\frac{(\nu-2)(\nu+1)}{2}\mathbb{E}\left(1-(\nu-1)\frac{1}{\varepsilon^2+\nu-2}\right) \\
&= 1.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{E}\left[g_2.\tilde{s}_{\theta,3}\right] &= -\frac{1}{2}\mathbb{E}\left(\log(1+\frac{\varepsilon^2}{\nu-2})(\varepsilon^2-1)\right) + \frac{\nu+1}{2(\nu-2)}\mathbb{E}\left(\frac{\varepsilon^2(\varepsilon^2-1)}{\varepsilon^2+(\nu-2)}\right) \\
&= -\frac{\nu-2}{2\nu}\mathbb{E}\left(\log(1+\frac{z^2}{\nu})(z^2-\frac{\nu}{\nu-2})\right) + \frac{\nu+1}{2\nu}\mathbb{E}\left(\frac{z^2(z^2-\frac{\nu}{\nu-2})}{z^2+\nu}\right).
\end{aligned}$$

The second term is equal to $\frac{1}{\nu-2}$ using $\mathbb{E}(z^2) = \frac{\nu}{\nu-2}$ and $\mathbb{E}(\frac{1}{z^2+\nu}) = \frac{1}{\nu+1}$. The first can be computed using the continuous analog of (11) (see BM) for the Student distribution with $\psi(z) = z(z^2+\nu)\log(1+\frac{z^2}{\nu})$. This equation yields

$$\mathbb{E}\left(\log(1+\frac{z^2}{\nu})(z^2-\frac{\nu}{\nu-2})\right) = \frac{2}{\nu-2}\mathbb{E}(z^2) = \frac{2\nu}{(\nu-2)^2}.$$

Consequently: $\mathbb{E}\left[g_2.\tilde{s}_{\theta,3}\right] = 0$.

For the covariance of $g_3(\cdot)$ with the score components, the details are provided below (the same type of calculations are used):

$$
\begin{aligned}
\mathbb{E}\left[g_3.\tilde{s}_{\theta,2}\right] &= -\frac{\sigma^2}{2}\mathbb{E}\left(\left(\frac{-\nu\varepsilon^2+\nu-2}{\varepsilon^2+\nu-2}\right)\left((\varepsilon^4-3)(\nu-4)-6\right)\right)\\
&= -\frac{\sigma^2\nu}{2}\mathbb{E}\left(\left(\frac{-z^2+1}{z^2+\nu}\right)\left(\frac{(\nu-2)^2}{\nu^2}(\nu-4)z^4-3(\nu-2)\right)\right)\\
&= -\frac{\sigma^2\nu(\nu+1)}{2}\mathbb{E}\left(\frac{1}{z^2+\nu}\left(\frac{(\nu-2)^2}{\nu^2}(\nu-4)z^4-3(\nu-2)\right)\right)\\
&= 6\sigma^2(\nu-2).
\end{aligned}
$$

$$
\mathbb{E}\left[g_3.\tilde{s}_{\theta,3}\right] = -\frac{\sigma^4}{2}\mathbb{E}\left(\log(1+\frac{\varepsilon^2}{\nu-2})\left((\nu-4)\varepsilon^4-3(\nu-2)\right)\right)+\frac{(\nu+1)\sigma^4}{2(\nu-2)}\mathbb{E}\left(\frac{\varepsilon^2}{\varepsilon^2+(\nu-2)}\left((\nu-4)\varepsilon^4-3(\nu
$$

The first can be computed using the continuous analog of (11) (see BM) for the Student distribution with $\psi(z)=z^3(z^2+\nu)\log(1+\frac{z^2}{\nu})$. This equation yields

$$
\mathbb{E}\left(\log(1+\frac{z^2}{\nu})(-(\nu-4)z^4+3\nu z^2)\right)=\frac{-6\nu^2}{(\nu-2)(\nu-4)}.
$$

Consequently

$$
\mathbb{E}\left(\log(1+\frac{\varepsilon^2}{\nu-2})\left((\nu-4)\varepsilon^4-3(\nu-2)\right)\right)=\frac{12(\nu-3)}{\nu-4}.
$$

The second term follows:

$$
\begin{aligned}
&\mathbb{E}\left(\frac{\varepsilon^2}{\varepsilon^2+(\nu-2)}\left((\nu-4)\varepsilon^4-3(\nu-2)\right)\right)\\
&= \mathbb{E}\frac{z^2}{z^2+\nu}\left(\frac{(\nu-2)^2(\nu-4)}{\nu^2}z^4-3(\nu-2)\right)\\
&= -\nu\mathbb{E}\frac{1}{z^2+\nu}\left(\frac{(\nu-2)^2(\nu-4)}{\nu^2}z^4-3(\nu-2)\right)\\
&= -\nu\mathbb{E}\frac{1}{z^2+\nu}\left(\frac{(\nu-2)^2(\nu-4)}{\nu^2}z^2(z^2+\nu)-\frac{(\nu-2)^2(\nu-4)}{\nu}(z^2+\nu)+(\nu-2)^2(\nu-4)-3(\nu-2)\right)\\
&= \frac{12(\nu-2)}{\nu+1}.
\end{aligned}
$$

Therefore

$$
\mathbb{E}\left[g_3.\tilde{s}_{\theta,3}\right]=\frac{-6\sigma^4}{\nu-4}.
$$

The orthogonalization of $I_t-\alpha$ along the estimating equation, $g(\cdot)$, yields

$$
\tilde{e}_t = I_t-\alpha+f_\nu(q^\nu_\alpha)\varepsilon_t+\frac{q^\nu_\alpha f_\nu(q^\nu_\alpha)}{2}(\varepsilon_t^2-1)+\frac{\partial F_\nu}{\partial\nu}(q^\nu_\alpha)\left(\frac{(\nu-4)^2}{6}(\varepsilon_t^4-K_\varepsilon)-(\nu-2)(\nu-4)(\varepsilon_t^2-1)\right), \tag{B.21}
$$

where $K_\varepsilon=3+\frac{6}{\nu-4}$ is the kurtosis of ε_t.

C Examples of Ord's distributions

We provide here particular examples of discrete distributions. The definition of the orthonormal polynomial family is provided in Table 1.

The Poisson distribution When $Y \sim \mathcal{P}o(\mu)$, the probability distribution function of Y is:

$$p_y = e^{-\mu}\frac{\mu^y}{y!}$$

The orthonormal family associated to the Poisson distribution is the family of Charlier polynomials $C_j(y,\mu)$. As

$$\frac{\partial \ln p_y}{\partial \mu} = -1 + \frac{y}{\mu} = -\frac{C_1(y,\mu)}{\sqrt{\mu}},$$

Charlier polynomials of degree greater or equal to 2 are robust to the parameter estimation uncertainty when one estimates the parameter μ.

The Pascal distribution The Pascal distribution is also known as the negative binomial distribution. It extends the Poisson distribution to some cases where the variance could be greater than the mean of the distribution (the overdispersion that Poisson counting processes fail to fit). The negative binomial distribution is also known as a Poisson-Gamma mixture.

When $Y \sim \mathcal{P}a(\mu,\delta)$,

$$p_y = \left(\frac{\mu}{\mu+\delta}\right)^y \left(\frac{\delta}{\mu+\delta}\right)^\delta \frac{\Gamma(y+\delta)}{\Gamma(\delta)\Gamma(y+1)}$$

When $\delta \to +\infty$, the Pascal distribution tends to the Poisson distribution. The orthonormal polynomials associated to this distribution are the Meixner polynomials $M_j(y,\mu,\delta)$.

When $\delta = 1$, the Pascal distribution is the geometric distribution ($\alpha = \frac{1}{\mu+1}$). Candelon *et al.* (2011) test this discrete distribution in a context of backtesting.

The binomial distribution The probability distribution function of the Binomial distribution is:

$$p_y = \binom{N}{y} p^y (1-p)^{N-y}$$

where $p \leq 1$

In this case, the orthogonal polynomials $K_j(y,N,p)$ are the Krawtchouk polynomials. They can be used for testing probit and logit models.

Figures and Tables

Table 1: Ord's family and orthonormal polynomials.

Name	p_y Recursive relationship	A	B	Q_1
Poisson	$e^{-\mu}\frac{\mu^y}{y!}$	$-(y-\mu+1)$	$y+1$	$\frac{\mu-y}{\sqrt{\mu}}$
	$Q_{j+1}(y) = \frac{\mu+j-y}{\sqrt{\mu(j+1)}}Q_j(y) - \sqrt{\frac{j}{j+1}}Q_{j-1}(y)$			
Pascal	$\left(\frac{\mu}{\mu+\delta}\right)^y\left(\frac{\delta}{\mu+\delta}\right)^\delta \frac{\Gamma(y+\delta)}{\Gamma(\delta)\Gamma(y+1)}$	$\frac{\mu}{\mu+\delta}(y+\delta)-(y+1)$	$y+1$	$\frac{\mu\delta-\delta y}{\sqrt{\mu\delta(\mu+\delta)}}$
	$Q_{j+1}(y) = \frac{\mu(2j+\delta)+\delta(j-y)}{\sqrt{\mu(\mu+\delta)(j+\delta)(j+1)}}Q_j(y) - \sqrt{\frac{j(\delta+j-1)}{(j+1)(\delta+j)}}Q_{j-1}(y)$			
Geometric	$(1-\alpha)^y\alpha$	$-\alpha(y+1)$	$y+1$	$\frac{1-\alpha-\alpha y}{\sqrt{1-\alpha}}$
	$Q_{j+1}(y) = \frac{(1-\alpha)(2j+1)+\alpha(j-y)}{\sqrt{1-\alpha}(j+1)}Q_j(y) - \frac{j}{j+1}Q_{j-1}(y)$			
Binomial	$\binom{N}{y}p^y(1-p)^{N-y}$	$-(y-Np+q)$	$q(y+1)$	$\frac{pN-y}{\sqrt{pqN}}$
	$Q_{j+1}(y) = \frac{p(N-j)+qj-y}{\sqrt{pq(N-j)(j+1)}}Q_j(y) - \sqrt{\frac{j(N-j+1)}{(j+1)(N-j)}}Q_{j-1}(y)$			

$\frac{p_{y+1}-p_y}{p_y} = \frac{A(y)}{B(y)}$. Q_j is the orthogonal polynomial of degree j, normalized.

In sample properties						
	$\alpha = 0.01$			$\alpha = 0.05$		
T	250	500	750	250	500	750
$(I_t - \alpha)^0$	1.90	1.80	3.60	0.90	2.00	2.10
e_t	2.00	7.70	3.60	4.00	5.50	5.70
$e_t^\star$	2.00	7.70	3.60	3.60	5.00	5.70
$I_t - \alpha$	1.90	8.50	3.60	3.40	4.10	5.70
$e_t e_{t-1}$	2.50	5.20	7.40	6.10	4.10	4.90
$e_t e_{t-2}$	2.60	5.10	6.80	4.60	5.40	4.80
$e_t e_{t-3}$	4.10	4.50	6.90	6.00	5.30	4.80
m_3^e	4.20	5.60	6.60	4.80	4.50	4.10
m_5^e	5.00	4.60	4.20	4.90	4.30	4.40
m_{10}^e	3.90	3.80	4.30	5.80	4.90	4.60
$e_t^\star e_{t-1}^\star$	2.20	5.30	7.60	6.20	4.60	5.40
$e_t^\star e_{t-2}^\star$	2.10	5.10	7.00	4.70	5.20	4.30
$e_t^\star e_{t-3}^\star$	3.70	4.40	7.20	6.40	5.10	5.10
$m_3^{e^\star}$	4.10	6.30	6.80	4.60	4.80	4.50
$m_5^{e^\star}$	4.80	5.40	4.10	5.20	4.20	4.50
$m_{10}^{e^\star}$	4.50	4.00	3.90	4.80	5.20	4.30
$(I_t - \alpha)(I_{t-1} - \alpha)$	1.90	5.20	7.20	3.60	2.90	3.30
$(I_t - \alpha)(I_{t-2} - \alpha)$	2.00	4.50	6.90	3.90	3.50	3.90
$(I_t - \alpha)(I_{t-3} - \alpha)$	2.70	4.10	7.60	6.20	4.70	4.20

Note: for each sample size T, we report the rejection frequencies for a 5% significance level test of the accuracy of the one day-ahead VaR forecasts computed from the estimation of a GARCH normal model. The different moments are detailed in Section 5.

Table 2: Size of the Backtest - Normal GARCH model

	Alternatives								
	Hist. Simulation			T-GARCH			EGARCH		
T	250	500	750	250	500	750	250	500	750
$(I_t - \alpha)^0$	0.00	0.00	0.00	2.90	8.80	13.70	6.00	12.10	19.30
e_t	0.00	0.00	0.00	9.30	15.20	23.60	10.00	20.00	29.10
$e_t^\star$	0.00	0.00	0.00	8.90	14.10	22.70	12.20	22.20	28.90
$I_t - \alpha$	0.00	0.00	0.00	7.30	13.30	22.30	9.90	23.50	28.80
$e_t e_{t-1}$	15.40	16.60	15.90	7.90	9.20	11.30	16.90	26.40	36.80
$e_t e_{t-2}$	13.00	12.30	14.80	11.00	11.60	14.10	12.40	17.70	21.20
$e_t e_{t-3}$	13.40	15.50	12.70	10.30	12.80	13.40	10.80	15.20	17.80
m_3^e	15.80	16.80	18.50	10.60	11.80	13.80	21.80	35.10	44.00
m_5^e	16.40	17.80	20.60	12.20	12.80	14.70	21.70	35.70	47.10
m_{10}^e	16.10	19.00	21.50	12.10	14.00	15.40	20.10	36.00	44.20
$e_t^\star e_{t-1}^\star$	17.80	16.80	17.70	5.60	6.90	7.60	16.30	25.40	35.70
$e_t^\star e_{t-2}^\star$	13.20	13.60	16.60	9.20	8.90	10.30	12.30	17.70	20.80
$e_t^\star e_{t-3}^\star$	14.60	17.20	14.50	8.60	9.70	9.00	10.50	15.40	17.70
$m_3^{e^\star}$	16.20	18.20	21.40	7.40	8.80	8.40	20.60	34.60	43.30
$m_5^{e^\star}$	16.70	19.40	23.70	9.00	11.10	10.30	20.60	35.50	46.70
$m_{10}^{e^\star}$	15.80	21.50	24.70	10.20	11.90	12.50	19.00	36.20	44.40
$(I_t - \alpha)(I_{t-1} - \alpha)$	9.50	15.10	20.40	2.80	2.60	2.60	9.30	13.10	14.80
$(I_t - \alpha)(I_{t-2} - \alpha)$	8.10	12.90	18.90	3.20	1.60	2.10	8.20	7.70	9.20
$(I_t - \alpha)(I_{t-3} - \alpha)$	6.70	13.80	15.70	3.80	3.50	3.40	6.60	5.10	6.00

Note: for each sample size T, we report the rejection frequencies for a 5% significance level test of the accuracy of the one day-ahead VaR forecasts computed from the estimation of a GARCH normal model. The different moments are detailed in Section 5.

Table 3: In-sample power properties of the VaR Backtest - Normal GARCH model, $\alpha = 5\%$.

	Size		Power					
			HS		T-GARCH		EGARCH	
	$P = 125$	$P = 250$	$P = 125$	$P = 250$	$P = 125$	$P = 250$	$P = 125$	$P = 250$
$(I_t - \alpha)^0$	32.00	14.70	61.10	51.60	28.90	18.20	35.20	18.00
e_t	5.40	4.10	23.50	31.10	16.50	20.20	7.00	6.70
$e_t^\star$	5.10	4.20	25.40	32.80	15.50	20.80	7.30	6.80
$I_t - \alpha$	4.90	3.90	30.00	36.00	3.80	5.70	4.60	5.60
$e_t e_{t-1}$	6.80	6.70	22.50	26.40	9.80	14.60	8.80	8.80
$e_t e_{t-2}$	6.70	5.20	21.20	25.20	12.10	14.60	7.50	7.90
$e_t e_{t-3}$	6.80	5.70	18.30	25.00	12.40	14.60	7.50	6.90
m_3^e	6.50	5.80	23.80	30.20	13.30	17.70	7.30	8.60
m_5^e	6.20	6.30	23.40	33.10	13.00	18.00	8.40	8.80
m_{10}^e	7.20	5.90	24.00	35.50	13.50	17.10	8.20	9.00
$(I_t - \alpha)(I_{t-1} - \alpha)$	5.00	5.50	19.60	27.10	2.90	3.50	6.50	6.50
$(I_t - \alpha)(I_{t-2} - \alpha)$	4.80	4.50	19.70	27.40	3.70	3.20	5.10	5.70
$(I_t - \alpha)(I_{t-3} - \alpha)$	5.20	4.50	17.20	25.50	4.00	4.70	6.20	6.50

Table 4: Out-of-sample properties Rolling Scheme - Normal GARCH model - $\alpha = 5\%$, $R = 500$ observations.

	Size		Power					
			HS		T-GARCH		EGARCH	
	$P = 125$	$P = 250$	$P = 125$	$P = 250$	$P = 125$	$P = 250$	$P = 125$	$P = 250$
$(I_t - \alpha)^0$	36.00	18.80	38.50	24.30	33.70	23.60	44.70	32.50
e_t	4.90	3.80	9.70	10.30	16.30	22.40	32.10	43.70
$e_t^\star$	5.50	4.30	10.00	10.70	16.60	22.90	31.90	43.40
$I_t - \alpha$	4.90	6.40	7.70	11.10	4.60	8.30	14.10	15.90
$e_t e_{t-1}$	6.30	6.10	7.20	7.50	10.40	12.40	27.30	37.20
$e_t e_{t-2}$	6.60	6.00	7.00	6.70	11.20	13.30	24.40	32.10
$e_t e_{t-3}$	8.70	7.60	8.50	8.70	12.50	15.80	19.90	27.10
m_3^e	6.90	6.90	8.10	7.50	13.10	16.60	29.60	41.40
m_5^e	7.20	7.10	7.90	8.10	14.20	17.90	32.30	43.00
m_{10}^e	7.20	6.30	8.80	9.10	14.60	19.50	32.70	43.30
$(I_t - \alpha)(I_{t-1} - \alpha)$	5.00	5.10	6.00	6.10	3.20	4.30	18.50	25.10
$(I_t - \alpha)(I_{t-2} - \alpha)$	4.80	5.00	6.40	6.00	4.10	3.60	13.40	19.10
$(I_t - \alpha)(I_{t-3} - \alpha)$	6.00	5.70	7.00	6.90	4.00	4.70	12.90	16.20

Table 5: Out-of-sample properties Fixed Scheme - Normal GARCH model - $\alpha = 5\%$, $R = 500$ observations.

	Size			Power HS		
T	250	500	750	250	500	750
e_t	4.80	5.50	3.90	0.00	0.00	0.00
$e_t^\star$	3.70	3.80	3.70	0.00	0.00	0.00
$\tilde{e}_t$	5.00	4.60	4.80	0.00	0.00	0.00
$I_t - \alpha$	3.40	3.40	3.70	0.00	0.00	0.00
$e_t e_{t-1}$	5.00	4.50	5.30	15.20	15.40	18.30
$e_t e_{t-2}$	6.10	5.00	5.30	13.10	12.90	17.50
$e_t e_{t-3}$	6.40	4.60	5.20	12.20	14.40	16.00
m_3^e	5.00	5.40	5.10	13.70	16.60	21.70
m_5^e	4.60	5.50	4.70	12.00	17.90	22.70
m_{10}^e	3.70	5.20	5.20	11.70	18.10	23.20
$e_t^\star e_{t-1}^\star$	4.90	4.70	4.70	15.00	16.60	19.10
$e_t^\star e_{t-2}^\star$	5.50	5.10	5.50	13.10	12.70	20.40
$e_t^\star e_{t-3}^\star$	5.90	4.90	4.70	13.40	14.90	17.00
$m_3^{e^\star}$	4.50	5.30	4.60	13.70	19.30	25.00
$m_5^{e^\star}$	3.80	5.30	4.20	14.40	19.60	26.00
$m_{10}^{e^\star}$	3.60	5.60	5.40	12.80	21.90	28.90
$\tilde{e}_t \tilde{e}_{t-1}$	4.50	4.50	5.30	11.50	14.70	16.40
$\tilde{e}_t \tilde{e}_{t-2}$	5.50	5.20	4.90	10.00	12.00	15.50
$\tilde{e}_t \tilde{e}_{t-3}$	6.20	5.20	5.00	9.10	11.90	13.20
$m_3^{\tilde{e}}$	3.70	4.20	4.60	9.60	14.60	19.20
$m_5^{\tilde{e}}$	3.20	4.50	4.10	9.00	15.30	19.10
$m_{10}^{\tilde{e}}$	3.30	3.90	4.70	9.00	16.80	20.00
$(I_t - \alpha)(I_{t-1} - \alpha)$	5.10	3.50	3.90	8.30	13.80	22.10
$(I_t - \alpha)(I_{t-2} - \alpha)$	4.90	3.30	3.80	8.80	10.80	21.40
$(I_t - \alpha)(I_{t-3} - \alpha)$	5.50	4.30	5.30	6.10	13.30	17.10

Table 6: Size and Power of the Backtest - T-GARCH model - In sample - $\alpha = 5\%$

	Size		Power HS	
	$P = 125$	$P = 250$	$P = 125$	$P = 250$
e_t	6.10	6.60	11.70	14.40
$\tilde{e}_t$	6.10	6.30	11.20	13.00
$I_t - \alpha$	4.00	4.10	8.00	9.60
$e_t e_{t-1}$	7.30	6.40	8.70	9.00
$e_t e_{t-2}$	5.60	5.40	7.30	7.40
$e_t e_{t-3}$	6.40	6.10	8.20	8.80
m_3^e	5.90	5.10	6.60	8.20
m_5^e	5.50	4.90	6.60	7.60
m_{10}^e	6.20	5.60	7.90	8.20
$\tilde{e}_t \tilde{e}_{t-1}$	7.30	4.90	7.70	7.10
$\tilde{e}_t \tilde{e}_{t-2}$	5.10	3.90	5.90	4.80
$\tilde{e}_t \tilde{e}_{t-3}$	4.70	5.10	5.90	6.20
$m_3^{\tilde{e}}$	5.80	4.20	5.80	5.80
$m_5^{\tilde{e}}$	5.20	4.10	5.30	5.80
$m_{10}^{\tilde{e}}$	4.70	4.50	5.60	6.10
$(I_t - \alpha)(I_{t-1} - \alpha)$	5.30	5.00	6.70	7.10
$(I_t - \alpha)(I_{t-2} - \alpha)$	3.70	4.10	5.20	5.60
$(I_t - \alpha)(I_{t-3} - \alpha)$	4.70	5.20	6.20	6.40

Table 7: Out-of-sample properties - Fixed Scheme - T-GARCH model - $\alpha = 5\%$

θ^0 known					θ^0 estimated by MLE				
T	100	200	500	1000	T	100	200	500	1000
C_1	4.94	4.56	4.96	5.24					
C_2	4.92	4.22	5.60	5.10	C_2	4.48	4.44	5.52	4.86
C_3	5.20	4.86	4.64	5.06	C_3	4.74	4.66	4.46	5.00
C_4	2.82	3.74	3.86	4.46	C_4	2.74	3.56	3.62	4.30
$C^w_{2,3}$	5.14	4.86	5.16	5.22	$C^w_{2,3}$	5.26	4.88	5.14	5.24
$C^w_{2,4}$	5.02	4.94	5.24	5.24	$C^w_{2,4}$	5.14	5.02	5.30	5.16
$C^w_{2,5}$	5.44	5.06	5.22	5.00	$C^w_{2,5}$	5.48	5.18	5.22	5.00
χ^2_P	9.06	8.46	8.50	8.24	χ^2_P	5.02	5.04	5.10	4.84

Note: The data are i.i.d. from a $\mathcal{P}o(2)$ distribution. The results are based on 10 000 replications.

Table 8: Size of the Poisson tests

Binomial distribution $\mathcal{B}(k, \frac{2}{k})$														
	k=10					k=15					k=20			
T	100	200	500	1000	T	100	200	500	1000	T	100	200	500	1000
C_2	26.44	56.68	94.04	99.94	C_2	12.24	25.74	58.08	89.04	C_2	7.86	15.30	35.54	64.96
C_3	1.04	1.56	2.06	4.72	C_3	1.70	1.62	1.56	2.22	C_3	2.34	2.02	1.88	1.80
C_4	0.66	0.72	1.44	2.54	C_4	1.00	1.26	1.08	1.58	C_4	1.22	1.24	1.54	1.64
$C^w_{2,3}$	27.74	52.24	89.18	99.60	$C^w_{2,3}$	14.06	24.86	51.36	83.00	$C^w_{2,3}$	9.92	15.54	31.48	57.94
$C^w_{2,4}$	22.70	41.78	78.62	97.80	$C^w_{2,4}$	11.68	20.40	41.92	71.96	$C^w_{2,4}$	9.08	12.74	26.22	46.58
$C^w_{2,5}$	18.32	33.02	68.20	94.18	$C^w_{2,5}$	10.02	16.50	33.96	62.30	$C^w_{2,5}$	8.08	10.92	21.90	39.68
χ^2_P	9.86	24.48	69.68	97.52	χ^2_P	5.86	10.36	27.72	62.10	χ^2_P	4.64	7.14	16.02	34.68

Pascal distribution $\mathcal{P}a(2.\delta)$														
	δ=10					δ=15					δ=20			
T	100	200	500	1000	T	100	200	500	1000	T	100	200	500	1000
C_2	30.84	46.96	80.94	97.86	C_2	18.22	27.04	51.54	79.36	C_2	13.34	18.54	34.24	57.42
C_3	13.36	15.32	19.02	23.30	C_3	9.60	11.32	12.02	15.28	C_3	8.22	8.84	9.70	11.86
C_4	9.68	13.04	17.80	24.50	C_4	6.68	8.66	11.42	15.32	C_4	5.70	7.16	8.98	11.14
$C^w_{2,3}$	22.70	37.72	69.84	93.60	$C^w_{2,3}$	13.82	20.78	41.46	68.18	$C^w_{2,3}$	10.18	14.64	26.48	46.10
$C^w_{2,4}$	20.32	34.56	65.10	90.32	$C^w_{2,4}$	12.34	18.82	35.74	61.82	$C^w_{2,4}$	9.52	12.58	22.52	40.68
$C^w_{2,5}$	18.58	29.66	58.80	85.30	$C^w_{2,5}$	11.52	16.30	31.40	55.14	$C^w_{2,5}$	8.86	11.32	20.10	35.84
χ^2_P	17.26	26.50	56.20	87.24	χ^2_P	11.10	14.62	29.26	53.00	χ^2_P	8.90	10.74	18.42	32.84

Table 9: Power of the Poisson tests

Size				Power with wrong α							
				$\alpha = 4\%$				$\alpha = 6\%$			
T	50	100	250	T	50	100	250	T	50	100	250
M_1	0.00	0.00	0.00	M_1	44.00	67.60	95.24	M_1	18.46	39.26	81.68
M_2	3.14	3.18	4.24	M_2	16.80	20.70	27.86	M_2	2.16	1.94	2.94
M_3	1.52	2.18	3.32	M_3	9.48	13.26	18.28	M_3	0.58	0.54	0.56
M_4	0.86	0.90	1.14	M_4	6.94	8.42	10.54	M_4	0.22	0.16	0.24
$M^w_{2,3}$	4.46	4.12	4.94	$M^w_{2,3}$	13.82	15.92	20.28	$M^w_{2,3}$	2.78	3.18	4.46
$M^w_{2,4}$	4.84	4.44	5.12	$M^w_{2,4}$	10.46	12.02	15.68	$M^w_{2,4}$	3.78	3.98	5.32
$M^w_{2,5}$	4.88	4.60	5.20	$M^w_{2,5}$	8.96	10.16	13.18	$M^w_{2,5}$	4.26	4.48	5.86
L_1	0.00	0.00	0.00	L_1	42.18	66.12	94.92	L_1	16.82	37.00	80.34
L_2	2.44	2.60	4.90	L_2	14.52	17.06	20.34	L_2	1.40	0.86	1.06
L_3	1.30	2.18	3.72	L_3	8.12	11.42	15.66	L_3	0.46	0.44	0.88
L_4	1.02	1.30	1.56	L_4	6.84	8.56	11.48	L_4	0.26	0.44	0.60
$L^w_{2,3}$	2.22	2.48	2.60	$L^w_{2,3}$	12.34	15.56	22.14	$L^w_{2,3}$	1.10	0.76	0.78
$L^w_{2,4}$	1.56	1.72	1.98	$L^w_{2,4}$	8.94	11.90	16.94	$L^w_{2,4}$	0.60	0.42	0.52
$L^w_{2,5}$	1.14	1.22	1.56	$L^w_{2,5}$	7.24	9.76	14.10	$L^w_{2,5}$	0.36	0.28	0.38

Table 10: Size and Power of the geometric distributional test

$\rho = 0.4$				$\rho = 0.6$				$\rho = 0.8$			
T	50	100	250	T	50	100	250	T	50	100	250
M_1	0.00	0.00	0.00	M_1	0.00	0.00	0.00	M_1	0.00	0.00	0.00
M_2	8.48	11.58	19.96	M_2	19.24	34.06	61.04	M_2	41.58	66.52	93.64
M_3	3.62	6.00	11.70	M_3	12.70	23.40	46.04	M_3	37.30	57.68	77.94
M_4	1.68	2.00	4.16	M_4	5.94	8.80	21.38	M_4	20.64	34.04	63.52
$M^w_{2,3}$	11.10	15.18	25.54	$M^w_{2,3}$	26.34	43.76	73.06	$M^w_{2,3}$	53.74	77.22	97.42
$M^w_{2,4}$	12.20	17.04	28.92	$M^w_{2,4}$	29.48	48.08	78.76	$M^w_{2,4}$	59.90	82.38	98.68
$M^w_{2,5}$	13.00	17.46	30.42	$M^w_{2,5}$	31.70	50.92	81.66	$M^w_{2,5}$	62.84	84.80	98.96
L_1	0.00	0.00	0.00	L_1	0.00	0.00	0.00	L_1	0.00	0.00	0.00
L_2	6.36	8.70	12.80	L_2	15.74	27.32	50.96	L_2	37.18	62.04	91.82
L_3	2.32	3.50	5.92	L_3	8.98	15.82	32.70	L_3	31.22	52.48	74.84
L_4	1.32	1.14	2.06	L_4	3.74	5.28	11.14	L_4	15.62	26.38	53.00
$L^w_{2,3}$	4.52	6.22	9.14	$L^w_{2,3}$	9.08	15.48	29.36	$L^w_{2,3}$	20.52	39.90	71.24
$L^w_{2,4}$	2.84	4.58	7.14	$L^w_{2,4}$	6.70	11.88	23.82	$L^w_{2,4}$	16.62	35.92	68.96
$L^w_{2,5}$	1.94	3.14	5.22	$L^w_{2,5}$	5.26	9.16	18.58	$L^w_{2,5}$	13.70	30.22	58.98

Table 11: Power of the geometric distributional test with non i.i.d. geometric variables

	UK-US$	FF-US$	SF-US$	Yen-US$
$\hat{\omega}$	7.2e-07	1.5e-06	9.2e-08	2.9e-06
$\hat{\alpha}$	0.07	0.08	0.04	0.09
$\hat{\beta}$	0.92	0.89	0.96	0.86
$\hat{\nu}$	8.81	12.26	6.73	7.41

Note: MLE of the T-GARCH(1,1) model for daily exchange rates.

	$\alpha = 0.5\%$				$\alpha = 1\%$			
	UK-US$	FF-US$	SF-US$	Yen-US$	UK-US$	FF-US$	SF-US$	Yen-US$
e_t	0.64	0.28	0.97	0.56	0.36	0.78	0.00	0.89
$\tilde{e}_t$	0.66	0.07	0.54	0.83	0.43	0.52	0.37	0.69
$I_t - \alpha$	0.73	0.40	0.90	0.54	0.37	0.87	0.01	0.85
$e_t e_{t-1}$	0.84	0.84	0.43	0.00	0.94	0.00	0.64	0.00
$e_t e_{t-2}$	0.31	0.70	0.99	0.34	0.00	0.91	0.76	0.30
$e_t e_{t-3}$	0.93	0.99	0.31	0.81	0.76	0.80	0.00	0.73
m_3^e	0.69	0.63	0.37	0.00	0.02	0.82	0.18	0.00
m_5^e	0.60	0.71	0.42	0.00	0.02	0.74	0.06	0.00
m_{10}^e	0.59	0.87	0.24	0.00	0.11	0.85	0.03	0.00
$\tilde{e}_t \tilde{e}_{t-1}$	0.92	0.54	0.84	0.00	0.85	0.00	0.86	0.00
$\tilde{e}_t \tilde{e}_{t-2}$	0.50	0.47	1.00	0.98	0.00	0.50	0.99	0.94
$\tilde{e}_t \tilde{e}_{t-3}$	0.85	0.77	0.91	0.94	0.93	0.98	0.14	0.91
$m_3^{\tilde{e}}$	0.74	0.39	0.89	0.00	0.04	0.51	0.60	0.00
$m_5^{\tilde{e}}$	0.71	0.39	0.93	0.00	0.03	0.60	0.53	0.00
$m_{10}^{\tilde{e}}$	0.72	0.43	1.00	0.00	0.10	0.60	0.47	0.01
$(I_t - \alpha)(I_{t-1} - \alpha)$	0.91	0.97	0.82	0.81	0.63	0.00	0.36	0.75
$(I_t - \alpha)(I_{t-2} - \alpha)$	0.91	0.97	0.86	0.80	0.63	0.83	0.39	0.74
$(I_t - \alpha)(I_{t-3} - \alpha)$	0.91	0.97	0.86	0.81	0.62	0.83	0.00	0.75

Note: we test the accuracy of the one day ahead VaR forecast computed from a T-GARCH(1,1) model for different level of risk α for the four daily exchanges rates. The p-value of the test statistics are reported. The notations are defined in Section 5.

Table 12: Backtesting of VaR forecasts for the T-GARCH(1,1) model

	$\alpha = 0.5\%$				$\alpha = 1\%$			
	UK-US\$	FF-US\$	SF-US\$	Yen-US\$	UK-US\$	FF-US\$	SF-US\$	Yen-US\$
e_t	0.95	0.65	0.04	0.68	0.91	0.95	0.27	0.25
$\tilde{e}_t$	0.83	0.96	0.00	0.31	0.70	0.47	0.01	0.35
$I_t - \alpha$	0.06	0.22	0.88	0.59	0.08	0.44	0.03	0.21
$e_t e_{t-1}$	0.44	0.87	0.09	0.99	0.45	0.89	0.02	0.90
$e_t e_{t-2}$	0.00	0.37	0.91	0.22	0.00	0.85	0.89	0.02
$e_t e_{t-3}$	0.37	0.99	0.43	0.51	0.45	0.60	0.36	0.62
m_3^e	0.00	0.72	0.13	0.41	0.00	0.90	0.03	0.21
m_5^e	0.00	0.68	0.12	0.36	0.00	0.70	0.10	0.12
m_{10}^e	0.00	0.43	0.38	0.25	0.00	0.11	0.59	0.66
$\tilde{e}_t \tilde{e}_{t-1}$	0.84	0.86	0.89	0.77	0.71	0.95	0.27	0.96
$\tilde{e}_t \tilde{e}_{t-2}$	0.00	0.70	0.94	0.40	0.00	0.97	0.86	0.22
$\tilde{e}_t \tilde{e}_{t-3}$	0.93	0.86	0.66	0.45	0.83	0.80	0.64	0.78
$m_3^{\tilde{e}}$	0.00	0.98	0.79	0.37	0.00	0.92	0.50	0.49
$m_5^{\tilde{e}}$	0.00	0.95	0.70	0.26	0.00	0.85	0.79	0.38
$m_{10}^{\tilde{e}}$	0.00	0.62	0.56	0.15	0.00	0.35	0.94	0.79
$(I_t - \alpha)(I_{t-1} - \alpha)$	0.70	0.77	0.95	0.86	0.58	0.72	0.45	0.64
$(I_t - \alpha)(I_{t-2} - \alpha)$	0.70	0.78	0.93	0.86	0.57	0.71	0.50	0.59
$(I_t - \alpha)(I_{t-3} - \alpha)$	0.70	0.78	0.93	0.85	0.57	0.71	0.44	0.64

Note: we test the accuracy of the one day ahead VaR forecast computed for different level of risk α for the four daily exchanges rates.

Table 13: Backtesting of VaR forecasts, out-of-sample