

Moment-based tests for discrete distributions

Christian Bontemps*

September 30, 2008

Abstract

In this paper, we develop moment-based tests for parametric discrete distributions. We characterize the class of moments for which the expectation is equal to zero under the null hypothesis. Moment-based test techniques are attractive as they provide easy-to-implement test statistics. Moreover, we can handle both the serial correlation and the parameter estimation uncertainty using traditional approaches that have been used in the GMM literature. There are various potential applications but we pay particularly attention to discrete counting processes, discrete choice models and the backtests of VaR models. We also assess the finite sample size properties of our tests through a Monte Carlo exercise.

Keywords: Moment, Test, Discrete Distribution, Poisson, Value-at-Risk, Backtesting, Probit.

JEL codes: C12, C15.

*Toulouse School of Economics & IDEI, 21 allée de Brienne, 31000 Toulouse, FRANCE. E-mail: bontemps@cict.fr

1 Introduction

Distributional tests have been extensively studied in the literature since Pearson's seminal paper of 1900 which introduced the well-known chi-squared goodness of fit test. This formally includes tests based on the cumulative distribution function (Kolmogorov, 1933, Smirnov, 1939), the characteristic function (Epps and Pulley, 1983, Koutrouvelis and Kellermeier, 1981) and on particular moments of the data (the Jarque-Bera test for normality in a regression context). The literature has focused mainly on serial correlation after the introduction of the generalized method of moments (see Hansen, 1982) and the development of financial econometrics which provides many empirical applications (see, for example, Bontemps and Meddahi (2005), Chen (2007b), Duan (2004), Dufour and al. (2003), Fiorentini and al. (2004), among many others).

Moment-based techniques can be very powerful in estimating parameters but can also be used for diagnostic purposes like the test for over-identifying restrictions derived in the GMM literature (Hansen, 1982). The goal of this paper is to use moment-based tests for testing discrete distributions. Particular examples of interest are discrete choice models, Value-at-Risk models and discrete counting processes.

There are several advantages to using moment-based tests. The first one is universality. Many tests have been developed for testing either the Normal or the Uniform distributions. One of the strategy is to first transform the variable of interest into a variable which follows one of these two distributions using (at least in a first step) the Probability Integral Transform (PIT) of Rosenblatt (1952). It is worth noting that it also transforms a serially correlated series into a i.i.d. one. It has been used for example in the context of density forecasts by Diebold, Gunther and Tay (1998) (see also Diebold, Hahn and Tay (1999) in the multivariate case, Bai (2003), Bai and Chen (2008) and Kalliovirta (2006)). This solution does not work in the case of discrete data or, at least, is not very efficient as there are different ways to define a continuous c.d.f. in this case. Moments are however well defined in many contexts. They are suitable for any type of parametric distribution, univariate or multivariate, discrete or continuous and can deal with independent or serially correlated data.

Moment estimation techniques have also been a big success as they can estimate some parameters involved in a subpart of an economic model without estimating the whole model. We can transpose this to our testing purpose. There are some cases (as with the VaR models) where we do not want to focus on the entire distribution but on one specific part. Moment techniques do not need to consider the entire distribution but can test a particular feature of the data (the variance, the tail, the kurtosis, etc.). It has therefore the appealing property of being flexible. This flexibility can help us to derive a moment-based test to any question under concern but also can help us to have some gain in statistical power. In fact, our goal is not to develop an omnibus test. We are sometimes more concerned with detecting departure in some given direction and omnibus tests, in this context, are not very powerful. As for the moment-based tests, we can choose the particular moments which

will be able to detect more precisely the departure from the null.

Testing procedure concerns observable or estimated variables. Many often, there are parameters involved in a model for which we do not want to test particular values. We can for example postulate in a structural economic model that a given counting process is a Poisson process and tests this hypothesis whatever the true rate. But we would also like to test that the Value-at-Risk model of a financial institution performs well. In this case, the VaR is not observed but estimated through a model for the financial returns. The accuracy of the test will also depend on the precision of the estimator. This parameter estimation uncertainty has to be taken into account when one goes from a world with all parameters known to a world where some parameters of interest are estimated. If one does not take it into account, it can deteriorate the size and power properties of the test (see, for example, Escanciano-Olmo (2007) in a VaR framework). On some occasions the distortion is not very important and one could ignore it, on others it is not.

A lot of past articles have solved this problem in very specific cases. The Jarque-Bera test for normality (1980) treats the problem for residuals in homoskedastic regressions (but is not adapted to observed variables). More recently, Bai (2003) used a Kolmogorov-Smirnov type approach with the combination of the Khmaladze transformation (1981) to deal with the parameter estimation uncertainty. Duan (2004) first transforms the variable into a normal one before treating the parameter estimation uncertainty by using special linear combinations of the moments tested. Both solutions used the c.d.f. and are therefore not implementable for discrete cases. The framework of M-tests developed by Newey (1985) and Tauchen (1985) allows us to treat this problem in a relatively simple way.

In this paper we focus on tests for discrete parametric distributions. We treat alternatively conditional and marginal distributions. We derive a general class of moment conditions which are satisfied under the null hypothesis. We pick particular moments in this class and derive a test statistic which is asymptotically chi-squared distributed. There are various guidelines for the choice of the moments in this class: one could be tractability where we simply count the number of occurrences of some given cells, one could be optimality where we use a moment which is optimal against a given alternative, one could be ease of interpretation where we want to give some meaning about the reasons of a potential rejection. At the end, whatever the choice of the moment, we work with a computationally tractable and easy to implement statistic.

Simplicity is one of our major concerns. We do not want to use simulation techniques for calculating the critical values or to estimate the distribution nonparametrically and compare this estimate with the parametric one (which provides generally non standard distributions in presence of parameter estimation uncertainty, see, for example, Delgado and Stute, 2008). Our tests are standard and the asymptotic critical values could be used even for (relatively) small sample size.

We treat the problem of the parameter estimation uncertainty by focusing on moments which are robust to

this problem. It is very rare to work in a case where there is no estimation. Either the variables are coming from a first step estimation or we have parameters involved. In most of the cases, we are indifferent to the true value of these parameters. There are various way to deal with that problem. In the classic treatment, one can compute the covariance between the moment of interest and the local deviation around the true value of the estimated parameter (see, for example, Newey, 1985). Recent approaches have dealt with finding procedures which could get rid of this additional term in a more general way. Duan (2004) used special combinations, Chen (2007a) used another transformation in a general regressional context, Bontemps and Meddahi (2007) used a moment which is modified by a projection as in Woolridge (1990) who, however, used a different one. We follow this approach in this paper. One potential risk is to loose a large part of the power properties of the moment used for the test. It appears that this technique is as efficient as it could be in the MLE framework.

We allow for the presence of serial correlation of unknown form assuming that the Central Limit Theorem still applies. Under some regularity conditions, a given moment can be used indifferently in a i.i.d. context or in a time series context. What matters at the end is the estimation of the variance matrix. A lot of past works on GMM have focused on the estimation of this matrix with the presence of serial correlation (Newey and West, 1987 and Andrews, 1991, among others). This gave rise to the heteroskedastic and autocorrelation consistent estimator of the variance matrix, which was developed for estimating purpose but which can be used for our testing procedure. When one ignores the form of the time dependency, the HAC framework allows researchers to use moment-based tests by computing the variance in a non-parametric way. Particular cases could be counting processes where the variance matrix is difficult to compute.

A Monte Carlo simulation experiment is run in different contexts and shows that the empirical performances of the tests are quite good, even in small samples.

Section 2 develops the general framework and some examples that are treated in the simulation experiment. Section 3 constructs the class of moments which could be used for the testing purpose. It characterizes all potential choices with respect to some predefined criteria. Section 4 develops some classical examples under the light of this approach. Particular examples are discrete choice models, Poisson counting processes and VaR models. Section 5 is devoted to Monte Carlo simulations about the testing procedure on small sample sizes (100 to 1000). Section 7 presents some empirical application on financial data. Section 8 concludes the paper.

2 General results

2.1 The Setup

Let Y be a univariate discrete random variable whose support S is discrete and countable. We denote by X the vector of explanatory variables (including potentially lagged values of Y_t). The number of coordinates of X is equal to p . We denote by P_x^0 the true conditional distribution of $Y|X = x$.

The support S could be indexed by some subset of \mathbb{Z} , I possibly \mathbb{Z} itself:

$$S = \bigcup_{i \in I \subset \mathbb{Z}} a_i.$$

Let $l = \inf\{i, i \in I\}$ and $r = \sup\{i, i \in I\}$ the infimum and supremum of the support S . l and r can be finite or infinite¹.

Let $P_{\theta,x}$ be a parametric family of conditional distributions indexed by $\theta \in \Theta \subset \mathbb{R}^s$. We will denote by E_θ the expectation with respect to this conditional distribution, by E_X the expectation with respect to the distribution of X and by E_0 the expectation with respect to the true joint distribution of (Y, X) . The notations for the variances are defined similarly. \top denotes the transpose operator.

The conditional probability of observing a_i as outcome is

$$p_i(x, \theta) = p(a_i, x, \theta) = P_{\theta,x}(Y = a_i | X = x).$$

For a given moment $m(y, x, \theta)$, its conditional expectation (assuming it is finite) is equal to:

$$E_\theta m(y, x, \theta) = \sum_{i=l}^r p_i(x, \theta) m(a_i, x, \theta).$$

Our goal is to test a parametric distributional assumption on the conditional distribution of $Y|X$:

$$\exists \theta^0 \in \Theta \subset \mathbb{R}^s : P(P_x^0 = P_{\theta^0,x}) = 1. \quad (1)$$

In the paper, we will focus on the case where $\Theta = \{\theta_0\}$, where we postulate the value of θ_0 , and on the case $\Theta = \mathbb{R}^s$. A typical example of the former is Example 3 below whereas Example 1 and 2 are examples of the latter. We first exhibit examples which are of particular interest in applied econometrics. They will be discussed in Section 4.

Example 1 *Discrete choice models*

Discrete choices models describe choices made among a given set of alternatives. They have played an important role in many subfields; participation to the labor force, urban transport mode choice, analysis of

¹The points of the support S are not necessarily equidistant.

demand for differentiated product are particular examples among many others. A few tests have been proposed in the literature. Skeels and Vella (1999) for the probit model, Butler and Chatterjee (1997) for bivariate ordered probit and Mora and Moro-Egido (2008) for ordered probit have also developed moment-based tests. For a discrete choice model:

$$p_i(x, \theta) = \Phi(a_{i+1} - \beta x) - \Phi(a_i - \beta x), \quad (2)$$

where a_0, a_1, \dots, a_{K+1} are some threshold values (with the convention $a_0 = -\infty, a_{K+1} = \infty$), K the number of choices faced by the decision maker and $\theta = (a_1, \dots, a_K, \beta)$.

Example 2 *Counting processes*

One of the leading model in i.i.d. count data is the Poisson model (see for example Cameron and Trivedi, 1998):

$$p_i(x, \theta) = P(Y = i | X = x; \theta) = e^{-\theta^\top x} \frac{(\theta^\top x)^i}{i!}.$$

We do not generally postulate any particular value for θ which is therefore estimated in the data. It is worth noting that this model could be easily extended to a serially correlated model. The particular case of INAR(1) model will be considered in Section 4.2.

Example 3 *Value-at-Risk (VaR)*

Value-at-Risk (VaR) are derived by financial institutions as a measure for risk exposition. As they are often internally developed, it is crucial to check the accuracy of the model used. Basically, the data are often composed only by a two-dimensional vector (return, VaR) which is observed across time.

Let r_t be the daily log-return of some given portfolio, equity, etc. and VaR_t the one day-ahead VaR forecast 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 where

$$I_t = \mathbf{1}\{r_t \leq VaR_t\}.$$

Under perfect accuracy I_t is i.i.d Binomial distributed with parameter α . Kupiec (1995) has considered a test based on the first-order moment of I_t , Christoffersen (1998) has considered a LR test in a Markov framework for testing both the conditional and unconditional distribution.

2.2 Test statistics

Consider a sample of T observations, independent or serially correlated, $(y_1, x_1), \dots, (y_T, x_T)$. We assume in a first step that the variables are observed and that we know the true value for the parameter θ , *i.e.* θ_0 . The case where there is a first step estimation involved in the testing procedure is treated in the following subsection on parameter estimation uncertainty.

Let m be a k -dimensional moment whose expectation under the null is equal to 0. We will focus in Section 3 on the moments we can use. Under H_0 :

$$E_0 m(y, x, \theta^0) = 0.$$

Throughout the paper, we assume that the matrix Σ defined by

$$\Sigma \equiv \lim_{T \rightarrow +\infty} \text{Var}_0 \left[\frac{1}{\sqrt{T}} \sum_{t=1}^T m(y_t, x_t, \theta^0) \right] = \sum_{h=-\infty}^{+\infty} E_0 [m(y_t, x_t, \theta^0) m^\top(y_{t-h}, x_{t-h}, \theta^0)], \quad (3)$$

is finite and positive definite. Under some regularity conditions (see for example White, 1984), we know that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T m(y_t, x_t, \theta^0) \longrightarrow \mathcal{N}(0, \Sigma)$$

which leads to the test statistic of interest ξ_m :

$$\xi_m = \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T m(y_t, x_t, \theta^0) \right)^\top \Sigma^{-1} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T m(y_t, x_t, \theta^0) \right), \quad (4)$$

which is asymptotically chi-squared distributed with k degrees of freedom². For the feasibility of the test procedure, one needs the matrix Σ or one consistent estimator. One candidate $\hat{\Sigma}$ could be the HAC estimator *à la* Newey-West (1987):

$$\hat{\Sigma} = \sum_{j=-M}^M w(j, M) \left(\frac{1}{T-|j|} \sum_{t=1}^{T-|j|} m(y_t, x_t, \theta^0) m^\top(y_{t+|j|}, x_{t+|j|}, \theta^0) \right),$$

for some appropriate choices of the weight functions $w(j, M)$ and of the number of lags M (see for example Andrews, 1991). The HAC estimator is known to provide good estimators when the persistence of the process is not too high.

In the context of cross-sectional observations, we have

$$\Sigma = \text{Var}[m(y, x)] = E_0 [m(y, x, \theta^0) m^\top(y, x, \theta^0)], \quad (5)$$

and

$$\hat{\Sigma} = \frac{1}{T} \sum_{t=1}^T m(y_t, x_t, \theta^0) m^\top(y_t, x_t, \theta^0).$$

It is worth noting that one key of good small sample size properties is an accurate estimate of the matrix Σ .

²The k components of m are assumed to be free.

2.3 The parameter estimation uncertainty problem

In most of the applications, the testing framework involves some parameters which have to be estimated. A given parametric distribution is generally known up to some finite dimensional parameters and it happens that the variable of interest could be the result of a first step estimation (like the residuals in a regression). In Example 1, we pay attention to the assumption that $Y|X$ is Poisson distributed regardless of the true value for θ . In VaR models (Example 3), the hit sequence is derived from the comparison between the raw returns and the VaR forecasts. This VaR forecast is generally computed through a model for the tail or for the entire conditional distribution of the returns. This model often involves some parameters which need to be estimated from the data.

It is well known that the test statistics have generally different asymptotic distributions when a consistent estimator is plugged in place of the true value. The Box-Pierce Test is a famous example. When one wants to test that ε_t is a White-Noise process one can compute the empirical autocorrelations \hat{r}_h for different lags $h = 1 \dots K$. Under the null,

$$\xi_{BP} = T \sum_{h=1}^K \hat{r}_h^2 \sim \chi^2(K).$$

However when ε_t is estimated, like in an $ARMA(p, q)$ model, the same statistic is still asymptotically chi-squared distributed but the degrees of freedom are now $K - p - q$ instead of K .

Let first assume for simplicity that the variable of interest is observable, that the parameters which are estimated are the parameters of the distribution of interest and that the moment is derivable with respect to this parameter θ . We denote by $m(y, x, \theta)$ such a moment.

A standard Taylor expansion around the true value of the parameter θ^0 provides the asymptotic difference between the process based on the (unknown) true value $\sqrt{T} \frac{1}{T} \sum_{t=1}^T m(y_t, x_t, \theta^0)$ and the process based on the estimated parameter, $\hat{\theta}$, $\sqrt{T} \frac{1}{T} \sum_{t=1}^T m(y_t, x_t, \hat{\theta})$:

$$\sqrt{T} \frac{1}{T} \sum_{t=1}^T m(y_t, x_t, \hat{\theta}) = \sqrt{T} \frac{1}{T} \sum_{t=1}^T m(y_t, x_t, \hat{\theta}) + \underbrace{E_0 \left[\frac{\partial m}{\partial \theta}(y, x, \theta) \right]_{\theta=\theta^0}}_{=Q} \sqrt{T}(\hat{\theta} - \theta^0) + o(1) \quad (6)$$

It can be shown (see for example Newey, 1985) that the variance of the left part of Equation (6), when $\hat{\theta}$ is the MLE, is the variance of

$$m^\perp(y, x, \theta) = m(y, x, \theta) - E_0 [m(y, x, \theta) s_\theta^\top(y, x)] \left[V_0 [s_\theta(y, x)]^{-1} \right] s_\theta(y, x) \quad (7)$$

at $\theta = \theta^0$ ($s_\theta(y, x)$ is the conditional score function). As $E_0 [m^\perp(y, x, \theta) s_\theta^\top(y, x)] = 0$, we have the

following variance equality:

$$\begin{aligned} V_0 m(y, x, \theta) &= V_0 m^\perp(y, x, \theta) + V_0 \left(E_0 [m(y, x, \theta) s_\theta^\top(y, x)] [V_0 [s_\theta(y, x)]^{-1}] s_\theta(y, x) \right) \\ &= V_0 m^\perp(y, x, \theta) + E_0 [m(y, x, \theta) s_\theta^\top(y, x)] [V_0 s_\theta(y, x)]^{-1} E_0 [m(y, x, \theta) s_\theta^\top(y, x)]^\top. \end{aligned} \quad (8)$$

Using the previous equality, we deduce that $(Vm^\perp)^{-1} \gg (Vm)^{-1}$. Plugging $\hat{\theta}$ in the place of θ^0 in (4) provides therefore a lower random variable than the expected chi-squared one. Ignoring the parameter estimation uncertainty yields underrejection of the null hypothesis. The bias depends on the ratio between these two variance matrices and so, looking at (8), on the covariance between m and s_θ .

The problem of parameter estimation uncertainty is usually fixed by computing the covariance between the first two terms of the right hand side of Equation (6); Newey (1985), Skeels and Vella (1999), Mora and Moro-Egido (2008) are some particular examples.

Another approach followed by Bontemps and Meddahi (2007) and others (see below) consists in deriving moment conditions for which the expectation of the derivative with respect to the parameter (the matrix Q defined in (6)) is equal to zero. We follow here the same approach.

In this case, we can plug the root-T consistent estimator of the true parameter in (4) without any change in the asymptotic expansion as it can be checked in Equation (6). The test statistic used when one knows the true value of the parameter could therefore be used with the estimated one without any additional precaution.

Using some adaptation of the generalized Information Matrix equality (see Tauchen, 1985, or Newey and McFadden, 1994, page 2163), we can reexpress Q in Equation (6) as $-E_0 m(y_t, x_t, \theta) s_\theta(y_t, x_t)$. Finding a moment such that $Q = 0$ is equivalent to find a moment orthogonal to the score function³. Such moments are not so easy to find. A particular case in our framework is the family of Charlier polynomials which are specific moments for testing the Poisson model. The score function is indeed proportional to the first Charlier polynomial. So any higher order polynomial is orthogonal to the score function.

When a given m is not orthogonal, we can still construct another moment by projecting this original moment m onto the space spanned by the score function. A robust moment can be written in the following form:

$$m^\perp(y, x, \theta) = m(y, x, \theta) - E_0 [m(y, x, \theta) s_\theta^\top(y, x)] [V_0 s_\theta(y, x)]^{-1} s_\theta(y, x). \quad (9)$$

A few remarks can be made about this strategy. First we do not oblige the square root consistent estimator of θ^0 to be estimated in the same step though we do not forbid this case. In particular, the J-test for overidentifying restrictions (see Hansen, 1982) is implicitly involved in this framework (like in Butler and Chatterjee, 1997).

³Is it obvious while looking at Equation (8) because, in this case, $Vm = Vm^\perp$.

Second, the estimator is not necessary the MLE. However, in this particular case, the test statistic constructed from (4) with m^\perp is exactly the same than the original one constructed from m after correcting for the parameter estimation uncertainty⁴.

Last, there are many other strategies to make a statistic robust to the parameter estimation uncertainty. Other particular approaches are among others Wooldridge (1990), Duan (2003), Chen (2007). They all share the explicit objective to work with one or several moments such that the matrix Q defined in Equation (6) is equal to zero. One potential drawback of these approaches is to destroy some power by projecting "too much". This is not the case, at least, for ours as we have just noticed that in the case of the MLE, all the information that could be used is still in the test statistic.

2.4 Non differentiable moments

The results exposed above is valid under some conditions of smoothness for m . There are cases of particular interest for which this assumption is no longer valid. In a VaR framework (see Section 4.4), we focus uniquely on the left tail of the distribution of the returns. The moment is based on estimating the frequency of the data under a given quantile (generally 1 or 5%) and to compare it with the theoretical one:

$$m(y) = \mathbf{1}\{y \in]-\infty, q_\alpha(\theta)]\} - \alpha.$$

The cell under concern therefore depends on the parameter through the estimation of this quantile and the resulting moment is non differentiable with respect to θ . Another example is derived from the Pearson chi-squared type test. We want to test the whole distribution of a continuous random variable and we divide the set of potential outcomes into equally probable cells. The moments of interest are no longer differentiable.

We treat here the case of a moment defined by:

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

where l, u, p are smooth functions of the parameter θ . In the Taylor expansion (6), Q is no longer the expectation of the derivative of the moment as the moment itself is not differentiable. However we can prove (see the appendix) that:

$$\begin{aligned} Q &= E_0 \left(\frac{\partial}{\partial \theta} u(x_t, \theta) \frac{\partial}{\partial y} F(u(x_t, \theta); \theta) - \frac{\partial}{\partial \theta} l(x_t, \theta) \frac{\partial}{\partial y} F(l(x_t, \theta); \theta) - \frac{\partial}{\partial \theta} p(x_t, \theta) \right)_{\theta=\theta^0} \\ &= -E_0 (m(y, x, \theta) s_\theta(y, x)), \end{aligned} \quad (10)$$

⁴The empirical mean of the score function is exactly zero at the ML estimator $\hat{\theta}$ which means that $\frac{1}{T} \sum_{t=1}^T m(y, x, \hat{\theta}) = \frac{1}{T} \sum_{t=1}^T m^\perp(y, x, \hat{\theta})$. The variance of each term is Vm^\perp as explained previously.

where $F(\cdot, \theta)$ is the conditional c.d.f of $Y|X = x$. The generalized Information Matrix equality is still valid. The parameter estimation uncertainty could still be treated by computing the asymptotic expansion of the right part of (6) with the help of Equation (10) or with the projection method detailed above.

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 moment of interests. We can be interested in tractability and ease of implementation in some cases, in power against specific alternatives in others. This section describes how to derive the moments which will be used to test our discrete distributions.

3.1 Ad-hoc choices

Ad-hoc choices of moments are always possible. For well-known distributions, one generally knows the first 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.

Let take the example of the Poisson distribution and imagine that we want to test that Y is marginally Poisson distributed with parameter λ_0 . We know that, in this case, the mean and the variance are equal to λ_0 . This gives us the opportunity to test H_0 from the first and second moments together or separately. We could alternatively use the sequence of moments:

$$m_i(y, \lambda_0) = \mathbf{1}\{Y = i\} - p_i(\lambda_0)$$

for different i . This is the basis of the well-known Pearson chi-squared test.

3.2 Orthogonal polynomials and Ord's family of discrete distributions

There are specific distribution families for which we can easily construct a sequence of moments for which the expectation is equal to zero. null expectation. The Ord's family is a well-known extension of the famous Pearson's family to the case of discrete distributions. This family includes, as particular examples, the Poisson, the Binomial, the Pascal (or negative binomial) and the hypergeometric distributions. For the clarity of the exposition, we will omit the dependence in X but this can be adapted to the conditional case quite easily.

A discrete distribution with p.d.f. p_y ⁵ belongs to the Ord's family if the ratio $\frac{p_{y+1} - p_y}{p_y}$ equals the ratio of

⁵ p_y denotes $P(Y = y)$.

two polynomials $A(\cdot)$ and $B(\cdot)$, where $A(\cdot)$ is affine and $B(\cdot)$ is quadratic.

$$\frac{\Delta p_y}{p_y} = \frac{p_{y+1} - p_y}{p_y} = \frac{A(y)}{B(y)} = \frac{a_0 + a_1 y}{b_0 + b_1 y + b_2 y^2}, \quad (11)$$

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

Particular moments for which the expectation is equal to zero under the null are coming from the associated orthonormal polynomial family Q_j , $j \in \mathbb{N}$. They are defined using an analogue of the Rodrigues' formula on finite difference (see Weber and Erdelyi, 1952 or Szegő, 1967):

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

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

These orthonormal polynomials are particular moments which can be used for our testing procedure. They are not necessarily neither the best in term of power nor robust to the parameter estimation uncertainty problem (except for some special cases). However, one advantage is that the variance is known, equal to one. The knowledge of the variance increases the small sample size properties of the test (see for example Mora and Moro-Egido, 2008). In a 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.

$$\xi_j = \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T Q_j(y_t) \right)^2 \sim \chi^2(1)$$

$$\xi = \sum_{k=1}^r \xi_k \sim \chi^2(r)$$

3.2.1 Examples of Ord's distributions

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

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 $C_j^\mu(y)$. They are defined in the appendix. As

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

it is therefore straightforward to prove that Charlier polynomials of degree greater or equal to 2 are robust to the parameter estimation uncertainty when one estimates the parameter μ^6 .

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 $Y \sim \mathcal{Pa}(\mu, \delta)$,

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

When $\delta \rightarrow +\infty$, the Pascal distribution tends to the Poisson distribution. The orthonormal polynomials associated to this distribution are the Meixner polynomials $M_j^{(\mu, \delta)}(y, \mu, \delta)$. The recursion relationship is defined in the appendix.

When $\delta = 1$, the Pascal distribution is the geometric distribution ($\alpha = \frac{1}{\mu+1}$). This distribution is of interest in discrete duration data. For example, Christoffersen and Pelletier (2004) prove that, in a VaR framework, the duration between two consecutive hits is geometrically distributed. They however use the continuous approximation (*i.e.* the exponential distribution). Candelon and al. (2008) test the discrete distribution.

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 $K_j^{(N,p)}(y)$ are the Krawtchouk polynomials. They will be used for testing probit and logit models.

3.3 A general class of moments

The two previous subsections expose some particular moments which can be used for testing purpose. We derive here a general rule for constructing any moment for which the expectation under the null is equal to zero.

For a function f defined on $S \times \mathbb{R}^p \times \Theta$, we define three operators: Δ the forward difference operator, ∇ the backward difference operator and $_{+1}$ the forward operator.

$\forall i \in I:$

⁶We recall that we omit here the conditioning variable X for simplicity. The same result holds for $\mu \equiv \lambda^\top X$ when the parameter of the Poisson distribution depends on some explanatory variables.

$$\begin{aligned}\Delta f(a_i, x, \theta) &= f(a_{i+1}, x, \theta) - f(a_i, x, \theta), \\ \nabla f(a_i, x, \theta) &= f(a_i, x, \theta) - f(a_{i-1}, x, \theta), \\ f_{+1}(a_i, x, \theta) &= f(a_{i+1}, x, \theta),\end{aligned}$$

with some conventions when the upper bound is finite ($f(a_{r+1}, x, \theta) = K$ a given normalization constant) or when the lower bound is finite ($f(a_{l-1}, x, \theta) = K'$).

When a_i are equidistant and belongs to \mathbb{Z} , Δ and ∇ are the standard backward and forward difference operators.

Let ψ a test function defined on $S \times \mathbb{R}^p \times \Theta$, and such that the expectation under $P_{\theta, x}$ is finite.

Assumption LB (Lower Bound) If l is finite, $\psi(a_l, x, \theta) = 0 \forall x, \theta \in \mathbb{R}^p \times \Theta$.

Proposition 1 *Under Assumption LB:*

$$E_{\theta} \left[\Delta \psi(y, x, \theta) + \frac{\Delta p(y, x, \theta)}{p(y, x, \theta)} \psi_{+1}(y, x, \theta) \right] = 0 \quad (12)$$

Let $m(y, x, \theta)$ be a moment for which the conditional expectation is equal to 0. Then let $\psi(y, x, \theta)$ defined on S by:

$$\begin{aligned}\psi(y, x, \theta) &= \frac{1}{p_y(x, \theta)} \sum_{k=l}^{y-1} m(a_k, x, \theta) p_k(x, \theta) \text{ for } y > a_l \\ \psi(a_l, x, \theta) &= 0 \text{ if } l \text{ is finite.}\end{aligned} \quad (13)$$

Then, ψ satisfies **LB** and

$$\Delta \psi(y, x, \theta) + \frac{\Delta p(y, x, \theta)}{p(y, x, \theta)} \psi_{+1}(y, x, \theta) = m(y, x, \theta)$$

The proof is given in the appendix. One can observe that assumption LB vanishes when $l = -\infty$.

The first part of the proposition shows how we can construct a moment such that its conditional expectation is equal to 0. One could argue that focusing on this class could restrict the range of the tests derived from these moment conditions. The second part of the proposition shows that this is however not the case and that there is a one-to-one relationship between m and ψ . We therefore do not loose anything by focusing on moment of the form (12).

Remark 1 One can notice that a second moment condition could be derived for some test function $\tilde{\psi}(y, x, \theta)$.

This moment conditions is:

$$E_{\theta}[\Delta\tilde{\psi}(y, x, \theta) + \frac{\nabla p_y(x, \theta)}{p_y(x, \theta)}\tilde{\psi}(y, x, \theta)] = 0 \quad (14)$$

under the assumption that $\tilde{\psi}(a_{r+1}, x, \theta) = 0$ when r is finite. However, Equation (14) is simply Equation (12) with $\tilde{\psi}(y, x, \theta) = \frac{p_y(x, \theta)}{p_{y-1}(x, \theta)}\psi(y, x, \theta)$ for $y > a_l$. We will therefore only use Equation (12).

In the sequel, we will speak in term of moments. We know now how we can generate any moment for testing a given discrete distribution. As mentioned before, one of the appealing properties of our moment-based tests is that we can adapt our testing procedure. Some particular choices of m (or ψ) will be guided by tractability, other choices by optimality.

3.4 Optimal moments

An omnibus test has power against any alternative. However it is sometimes preferable to use admissible tests in order to have more power against specific alternatives. One advantage of the moment-based tests is that we can choose one or several moments to have power against one or several alternatives.

Let assume that we want to test in a i.i.d. context:

$$H_0 \exists \theta^0 \in \Theta \subset \mathbb{R}^s : P(P_x^0 = P_{\theta^0, x}) = 1$$

against:

$$H_a : \exists \lambda^0 \in \Lambda \subset \mathbb{R}^s : P(P_x^0 = Q_{\lambda^0, x}) = 1,$$

where $Q_{\lambda, x} \cap P_{\theta, x} = \emptyset$.

Focusing on the cases where moments are square-integrable with respect to the null and the alternative assumptions, and assuming that the variance matrix in (4) is estimated under the null hypothesis (theoretically or by simulations) but not in the sample, we can expand the test-statistics ξ_m :

$$\xi_m = T \frac{\left[\frac{1}{T} \sum_{t=1}^T m(y_t, x_t, \theta^0) \right]^2}{\frac{1}{T} \sum_{t=1}^T m^2(y_t, x_t, \theta^0)},$$

which is equivalent when the expectation under the alternative is different from zero to (E_a denotes the expectation under H_a):

$$\xi_m \sim T \frac{[E_a m(y_t, x_t, \theta^0)]^2}{E_0 m^2(y_t, x_t, \theta^0)} \quad (15)$$

when T goes to infinity.

The optimal moment which maximizes the approximate slope of ξ_m is:

$$m_{opt}(y, x) = \frac{dP_{\theta^0, x}}{dP_{\lambda^0, x}} - 1$$

when θ^0 and λ^0 are known (see Bontemps and Meddahi, 2008). When θ has to be estimated, the optimal moment in a MLE context is the optimal one under full knowledge of the parameters projected, like previously, on the orthogonal of the score function. What matters in this paper is that we are able to derive a moment which have optimal properties when one wants to distinguish the null from a well specified parametric family.

3.5 The conditional case

We have detailed different guidelines for choosing moments such that the conditional expectation is equal to zero. However, in the case where there are some explanatory variables X involved, the tests used will be based on an unconditional moment. The true distribution of the explanatory variable is in most of the cases beyond the scope of the econometrician (though it can have some impact on the power properties of the tests). The unconditional moment will be therefore constructed as the product of some instrument (any function of X) with the conditional moment under concern. If $E_\theta m(y, x, \theta) = 0$, then:

$$E_0 h(x) m(y, x, \theta) = 0.$$

There are many choices possible for h ($1, x, \cos(kx), \sin(kx), \mathbf{1}\{x \leq A\}$, etc.). We will use particular choices in our Monte Carlo section.

4 Examples

In this section we come back to our leading examples which were presented in Section 2. and are motivations for our test procedures.

4.1 Discrete choice models

Discrete choice models have been widely used in applied economics. Though there exists now techniques to estimate nonparametrically or semiparametrically discrete models (Manski, 1975, Horowitz, 1992, Matzkin, 1992), parametric models are still often used in applied subfields. Moro and Mora-Egido (2008) compare the

performances of some nonparametric tests (based on the comparison between the parametric and nonparametric estimators) and some parametric ones. It appears that moment-based tests can still be competitive w.r.t. nonparametric ones even for big sample size and few conditioning variables. In this subsection, we first focus on the univariate probit model before considering the ordered ones. We also assume that the data are i.i.d. It is therefore straightforward to consider the dependent case. When we work with dependent variables, there is nothing different but another estimator of the variance matrix in Equation (4).

4.1.1 Binary models

In this case, Y have two possible outcomes standardized to 0 and 1. The conditional probability $p_y(x, \theta)$ is equal to

$$p_1(x, \theta) = P(Y = 1|X = x, \theta) = F(x^\top \beta; \nu), \quad p_0(x, \theta) = 1 - F(x^\top \beta; \nu),$$

where F is a c.d.f.⁷ with parameter ν (possibly a vector) and $\theta^\top = (\beta^\top, \nu^\top)$. F can be the standard normal distribution in the case of the probit model, the logistic distribution for the logit model (ν does not appear in these two cases) or any other parametric distribution. We denote by p_x the conditional probability of observing $Y = 1$, $F(x^\top \beta; \nu)$ for notational convenience.

Lemma 2 *In a binary choice model, any moment for which the conditional expectation w.r.t. the true distribution is equal to zero can be written as:*

$$m(y, x, \theta) = h(x)K_1^{p_x}(y), \tag{16}$$

where $K_1^{p_x}$ is the first Krawtchouk polynomial⁸ (see Szegö, 1967).

From Proposition 1 we know how to construct any moment for which conditional expectation w.r.t the true distribution is equal to zero. The proof in the appendix uses the fact that this moment only depends on two functions $m(0, x, \theta)$ and $m(1, x, \theta)$.

Assume, from now, for avoiding too many notations that ν the parameters of the c.d.f. F are known or does not exist like in the probit or logit case⁹ (hence $\theta = \beta$). The conditional score function s_β of a discrete choice model is trivially known to have conditional expectation equal to zero:

$$s_\beta(y, x) = -x \frac{f(x^\top \beta)}{1 - p_x} \left(1 - \frac{y}{p_x} \right),$$

⁷We denote by $f(z)$ the derivative of $F(z; \nu)$ w.r.t z .

⁸ $K_1^p(y) = \frac{1}{\sqrt{p(1-p)}} (p - y)$.

⁹There is no specific difficulty associated to this parameter.

and $h(x) = -x \frac{f(x^\top \beta)}{\sqrt{p_x(1-p_x)}}$ in this case. The conditional moment test for testing omitted variable w is also expressed like (16) with

$$h(x) = -w \frac{f(x^\top \beta)}{\sqrt{p_x(1-p_x)}},$$

(see Skeels and Vella, 1999, for other moment-based tests). One can also stretch the fact that this first Krawtchouk polynomial is also known (up to some scale parameter) as the generalized residual (see Gourieroux et al., 1987).

The moment defined above are generally not robust to the parameter estimation uncertainty because β is generally estimated. The correction needed is here linked to the choice of $h(x)$ as any moment which will be used for the test only differentiate with another one through this weight function. Following the strategy exposed in Section 2.3,

$$m^\perp(y, x) = h^\perp(x) K_1^{p_x}(y)$$

is robust to the estimation uncertainty. The expression for h^\perp is:

$$h^\perp(x) = h(x) - \lambda_h (V_0 s_\beta(y, x))^{-1} x \frac{f(x^\top \beta)}{\sqrt{p_x(1-p_x)}},$$

where $\lambda_h = E_X \left(h(x) x^\top \frac{f(x^\top \beta)}{\sqrt{p_x(1-p_x)}} \right)$. Moreover the variance of the score function could be simplified, using the fact that $V_\theta K_1^{p_x}(y) = 1$:

$$V_0 s_\beta(y, x) = E_X \left(x x^\top \frac{f^2(x^\top \beta)}{(1-p_x)p_x} \right).$$

We can now derive the test-statistics ξ_h to test our distributional assumption:

$$\xi_h = T \left(\frac{1}{T} \sum_{t=1}^T h^\perp(x_t) K_1^{F(x_t^\top \hat{\beta})}(y_t) \right)^\top V_h^{-1} \left(\frac{1}{T} \sum_{t=1}^T h^\perp(x_t) K_1^{F(x_t^\top \hat{\beta})}(y_t) \right) \sim \chi^2(\dim h)$$

which is asymptotically chi-squared distributed with $\dim h$ degrees of freedom. The variance matrix V_h can be simplified using the conditional expectation of Y under the null:

$$V_h = V_0 (m^\perp(y, x, \theta)) \tag{17}$$

$$= V_0 (m(y, x, \theta)) - \lambda_h V_s^{-1} \lambda_h' \tag{18}$$

$$= E_X (h(x) h^\top(x)) - \lambda_h V_s^{-1} \lambda_h' \tag{19}$$

There are various ways to estimate the previous variance matrix. One can compute the empirical variance of m^\perp in the sample using $\frac{1}{T} \sum_{t=1}^T (m^\perp(y_t, x_t, \hat{\theta}))^2$ or using the empirical counterpart of the expression (19). Though they are asymptotically equivalent, it appears that the small sample size properties dramatically depends on the way the variance is estimated. The sensitivity of tests to alternative expressions of the variance

matrix was first underlined by Orme (1990) in the context of the Information Matrix test applied to discrete choice models (see also Skeels and Vella, 1999). Mora and Moro-Egido (2008), in a similar context, shows that (19) performs better.

4.1.2 Ordered discrete choice models

We just focus here on the ordered models with three outputs ($y = 0, 1$ or 2) to prove the adaptability of our procedure to any case. Some helpful computations are exposed in the appendix. The ordered models have also been considered in Butler and Chatterjee (1997) as well as in Mora and Moro-Egido (2008). The first authors used a J-test in a GMM framework where as the last ones used Pearson' tests *à la* Andrews (1988). All the results derived in this subsection can be generalized to more than three outcomes. The ordered discrete choice model assumes that there is a latent variable Y^* , a threshold value μ and a p-dimensional parameter β such that:

$$Y^* = X^\top \beta + U,$$

with U distributed according to a distribution with c.d.f. F . F could depend on some extra parameter, ν . Y the observed outcome is 0 if the latent variable is negative, equal to 1 if it stands between 0 and μ and otherwise equal to 2. Using the previous notations, we have:

$$p_0(x, \theta) = F(-x^\top \beta; \nu), \quad p_1(x, \theta) = F(\mu - x^\top \beta; \nu) - F(-x^\top \beta; \nu), \quad \text{and} \quad p_2(x, \theta) = 1 - F(\mu - x^\top \beta; \nu),$$

where θ summarizes in a single notation μ, β and ν .

Using the result of Proposition 1, any potential moment which could be used for testing purpose could be expressed from two test functions $\psi_1(x)$ and $\psi_2(x)$ (which can also depend on θ):

$$\begin{aligned} m(0, x, \theta) &= \psi_1(x) + \frac{p_1(x, \theta) - p_0(x, \theta)}{p_0(x, \theta)} \psi_1(x) = \frac{p_1(x, \theta)}{p_0(x, \theta)} \psi_1(x), \\ m(1, x, \theta) &= \psi_2(x) - \psi_1(x) + \frac{p_2(x, \theta) - p_1(x, \theta)}{p_1(x, \theta)} \psi_2(x) = \frac{p_2(x, \theta)}{p_1(x, \theta)} \psi_2(x) - \psi_1(x), \\ m(2, x, \theta) &= -\psi_2(x), \end{aligned}$$

which can be summarized in a single expression:

$$\begin{aligned} m(y, x, \theta) &= -\psi_2(x) \frac{y(y-1)}{2} - \left(\frac{p_2(x, \theta)}{p_1(x, \theta)} \psi_2(x) - \psi_1(x) \right) y(y-2) + \frac{p_1(x, \theta)}{p_0(x, \theta)} \psi_1(x) \frac{(y-1)(y-2)}{2} \\ &= \lambda_2(x)(y^2 - E_\theta y^2) + \lambda_1(x)(y - E_\theta y). \end{aligned}$$

(20)

In the binary choice model, any moment was proportional to the first Krawtchouk polynomial up to the choice of a weight function. In a model with K outcomes, any moment is the linear combination of the first $K - 1$ moments. We can base our choice either on $\psi_1(x)$ and $\psi_2(x)$ or on $\lambda_1(x)$ and $\lambda_2(x)$. The appendix provides the one-to-one relationships between $\psi_1(x)$, $\psi_2(x)$ and $\lambda_1(x)$, $\lambda_2(x)$.

Let denote by $f(x)$ the p.d.f of $U(\frac{d}{dx}F(x; \nu))$ and $f_\nu(x)$ the partial derivative of F with respect to ν ($\frac{d}{d\nu}F(x; \nu)$).

The component of the score function which corresponds to the derivative w.r.t. β is derived from $\psi_1(x) = -x \frac{f(-x^\top \beta)}{p_1(x, \theta)}$ and $\psi_2(x) = -x \frac{f(\mu - x^\top \beta)}{p_2(x, \theta)}$. The expression of the other components are given in the appendix.

We can define similarly a moment which is robust to the parameter estimation uncertainty using the same argument than for the probit model. The score function and the moment are both functions of $y^2 - E_\theta y^2$ and $y - E_\theta y$. We can express the covariance between two moments as functions of x and first four conditional moments of Y . The appendix provides all the technical details to construct the robust moment m^\perp defined in Equation (9).

4.2 The Poisson INAR(1) process

The general integer valued autoregressive process (INAR) was introduced by Al-Osh and Alzaid (1987) to model correlated time series with integer values. It is the extension of the AR process to count data. The INAR (1) process is defined as

$$y_t = \alpha \circ y_{t-1} + \epsilon_t, \quad (21)$$

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 $u_i \sim B(1 - \alpha)$, i.i.d. . 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 (21) 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 . This model trivially nests the i.i.d case with $\alpha = 0$.

Different marginal distributions of y_t can be generated, and it depends on the distributional assumption made for (ϵ_t) (see Al-Osh and Alzaid, 1987, McKenzie, 1986, for more details). When $\epsilon_t \sim \mathcal{Po}(\mu)$, we have 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 $\frac{\mu}{1-\alpha}$ (see McKenzie, 1988).

As mentioned previously, the Charlier polynomials can be used as particular moments for testing Poisson distributions. However, due to the serial correlation among y_t the test statistics are *a priori* no longer

independent. However, in the special case of the INAR(1) process with Poisson innovation, we can prove that:

Proposition 3 *If $y_t \sim INAR(1)$ with parameter α and $\mathcal{P}o(\mu)$ innovation process:*

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

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 $Z_t = C_j^{1-\frac{\mu}{\alpha}}(y_t)$ is also AR(1). The test statistics based on the Charlier polynomials are still asymptotically independent in this case.

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

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

It is worth noting that except in the i.i.d. case, the Charlier polynomials are no longer robust to the parameter estimation uncertainty. We can construct moment robust in this case using the results derived in Freeland and McCabe (2004) who characterized the conditional score function.

4.3 Pearson chi-squared test

Let us first assume that y_1, \dots, y_n are i.i.d. (we change our notations from y_1, \dots, y_T to y_1, \dots, y_n to stress the independence) and that there are no explanatory variables. Let C_1, \dots, C_K , K cells such that $S = \bigcup_{i=1..K} C_i$. Let $p_i(\theta)$ be the theoretical probability of belonging to cell i . We assume that all p_i 's are strictly positive (which avoids to consider empty cells). We assume in a first step that the definition of the cells itself does not depend on the parameter.

Let us first consider the case of θ known equal to θ^0 . We will use p_i^0 for $p_i(\theta^0)$ and q_i for the empirical frequency of cell i :

$$q_i = \frac{1}{n} \sum_{j=1}^n \mathbf{1}\{y_j \in C_i\}.$$

The Pearson chi-squared test is a moment-based test using the $K-1$ moments:

$$m_i(y, \theta) = \mathbf{1}\{y \in C_i\} - p_i(\theta)$$

for $i = 1..K - 1$ (the last moment being 1 minus the sum of the first $K - 1$ moments).

The variance matrix of this $K-1$ dimensional moment is standard but derived in the appendix. Plugging everything in (4), we recover the expression of the well-known Pearson chi-squared test statistic:

$$\xi_P = n \left(\sum_{j=1}^K \frac{(q_j - p_j^0)^2}{p_j^0} \right) \sim \chi^2(K-1). \quad (22)$$

When θ^0 is unknown but estimated by a square root n consistent estimator $\hat{\theta}$, we could make these moments robust by projecting them onto the orthogonal of the space spanned by the score function. Let θ_i denote the i th-component of θ and $\bar{s}_\theta = \frac{1}{n} \sum_{i=1}^n s(y_i, \hat{\theta})$ be the averaged score function at the estimator.

Proposition 4

$$\xi = n \left(\sum_{j=1}^{K-1} \frac{(q_j - \hat{p}_j - \lambda \bar{s}_\theta)^2}{\hat{p}_j} + \frac{(q_K - \hat{p}_K + (K-1)\lambda \bar{s}_\theta)^2}{\hat{p}_K} \right) \sim \chi^2(K-1-s), \quad (23)$$

where the exact expression is omitted here but given in the appendix (λ is a non linear function of the p_i 's and the partial derivatives of the p_i 's).

This proposition shows how the Pearson chi-squared test can be modified when θ is estimated. The rank reduction comes from the fact that s constraints are added reducing the degrees of freedom (the sum of each partial derivative of the log probability with respect to each component of θ is equal to zero).

One particular case is the MLE. In this case, \bar{s}_θ is therefore equal to zero and the expression of ξ in (23) simplifies to

$$\xi = n \left(\sum_{j=1}^K \frac{(q_j - p_j)^2}{p_j} \right).$$

This is the usual Pearson chi-squared statistic. This particular result can be found in MaCurdy and Ryu (2003).

When y_1, \dots, y_n are no longer i.i.d, Proposition 4 is no longer valid. However, we can base our test procedure on the robust moments $m_j^\top(y) = 1_{y \in C_j} - p_j - \lambda \bar{s}$ and estimate the variance matrix with a HAC procedure *à la* Newey-West (1987).

When the cells depend on θ , the result of Section 2.4 shows that Q is still equal to the opposite of the covariance between the moment and the score function. One particular example concerns the backtest of VaR models which is considered in the next section.

4.4 Backtesting Value-at-Risk 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 and al., 1999). However this is the one which is the most commonly used by financial institutions.

Let r_t be the return at date t for a given portfolio. The Value-at-Risk VaR_α^t is the opposite¹⁰ of the percentage of loss that the portfolio return¹¹ will exceed with probability $1 - \alpha\%$:

$$P(r_t \leq -VaR_\alpha^t) = \alpha. \quad (24)$$

The practitioners compute one period ahead VaR forecasts using numerous models which are often kept unknown from the econometrician. 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 $-VaR_{t+1|t}^\alpha$ be the one period ahead forecast for the α VaR made at time t for time $t + 1$.

Let I_t be the indicator of bad extreme event:

$$I_t = \begin{cases} 1 & \text{if } r_t \leq -VaR_{t|t-1}^\alpha \\ 0 & \text{otherwise} \end{cases} \quad (25)$$

Under H_0 , *i.e.* the VaR model is the true model, I_t is i.i.d Bernoulli distributed with parameter α .

Let $m_T = \frac{1}{T} \sum_{t=1}^T I_t - \alpha$. Under H_0 and without parameter estimation uncertainty, the test statistic is¹²:

$$\xi_{UC} = T \frac{m_T^2}{\alpha(1-\alpha)} \sim \chi^2(1). \quad (26)$$

ξ_{UC} is the Kupiec test which, in this case, is equivalent to the unconditional test of Christoffersen (1998).

There is no other possibility for testing the marginal distribution of I_t . One can only improve the power of the test by testing the independence. Christoffersen (1998) augments his test by doing a LR test in a first order Markov chain framework. Basically, it tests the independence between I_t and I_{t-1} . We can therefore add some additional moments and test that

$$E[g(z_{t-1})(I_t - \alpha)] = 0,$$

for any regular function g , and any random variable z_{t-1} which belongs to the information set at time $t - 1$.

In the Monte Carlo experiment, we will use $\cos(r_{t-h})$ and $\cos(2r_{t-h})$ for $h = 1$ to 4.

Parameter estimation uncertainty has rarely been taken into account in this literature. An exception is a recent contribution from Escanciano and Olmo (2007). They showed in a particular simulation setting the potential bias that could arise. It should be fairly noticed that one can correct for the uncertainty only when one knows the underlying model that have been used for estimating the Value-at-Risk. In many cases,

¹⁰a VaR is positive.

¹¹Theoretically, the VaR is the value of the loss itself and should be referred to the amount invested initially. However, without loss of generality, we can speak in term of returns.

¹² $Vm_T = \alpha(1 - \alpha)$.

this is not true and the econometrician only knows the values predicted one day before and the raw returns. Without further information, it seems vain to take the estimation uncertainty into account. We know that if the moment-based test is rejected when one ignores the parameter estimation uncertainty, it will be rejected when it is considered (ignoring the parameter estimation uncertainty leads to a conservative test).

Using Equations (10) and (9), one can construct a moment robust to the parameter estimation uncertainty when we know the model.

Assume, here, that the model for the returns is a GARCH(1,1) with normal innovations¹³:

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

with $\varepsilon_t \sim \mathcal{N}(0, 1)$ and $\theta^\top = (\omega, \gamma, \beta)^\top$. Then, we can prove that the test statistic ξ_{UC} in (26) should be replaced by:

$$\xi_{UC}^\perp = T \frac{m_T^2}{V^\perp},$$

where $V^\perp = \alpha(1-\alpha) - \frac{(q_\alpha \phi(q_\alpha))^2}{2} E_0 \left[\frac{\partial \ln \sigma_t(\theta)}{\partial \theta^\top} \right] E_0 \left[\frac{\partial \ln \sigma_t(\theta)}{\partial \theta} \frac{\partial \ln \sigma_t(\theta)}{\partial \theta^\top} \right]^{-1} E_0 \left[\frac{\partial \ln \sigma_t(\theta)}{\partial \theta^\top} \right]^\top$, ϕ is the p.d.f. of the standard normal distribution and q_α its α -quantile.

The difference with (26) comes from the expression of the variance in the denominator which is smaller than the one without parameter estimation uncertainty. It is worth noting that the distortion is not too high when α tends to zero as in this case $q_\alpha \phi(q_\alpha)$ also tends to zero.

We can adapt the framework to the moments $g(z_{-1})(I_t - \alpha)$. Escanciano and Olmo (2007) have proved that there was no distortion when $g(z_{t-1}) = I_{t-h} - \alpha$, $h \geq 1$.

5 Monte-Carlo simulations

In this section, we provide Monte Carlo simulations to assess the finite sample properties of our test procedures. We consider alternatively the three examples exposed in the motivation of the paper. All the simulations are based on 10 000 replications of the samples.

5.1 The binary probit model

We consider here the binary probit model. In the simple parametric model, the underlined latent variable is assumed to be linear in the observables and the independent additive error term is assumed to have a normal distribution. Any departure from these hypothesis leads to biased estimators if one uses the ML estimator.

¹³The construction is similar for a GARCH-Student process.

Following the notations of Section 4, the null hypothesis is:

$$H_0 : P(Y = 1|X) = \phi(x'\beta),$$

for some $\beta \in \mathbb{R}^d$ where ϕ is the p.d.f. of the standard normal variable.

Any moment of interest can be indexed by the instrument function $h(x)$ because, as explained previously, we can write

$$m(y, x) = h(x) \left(\frac{y - \Phi(x^T \beta)}{\sqrt{\Phi(x^T \beta)(1 - \Phi(x^T \beta))}} \right). \quad (27)$$

Like in Mora and Moro-Egido (2008), the model tested is a probit model with one explanatory variable X which is normally distributed:

$$Y = \mathbf{1}\{\beta_1 + \beta_2 X + U \geq 0\}, U \sim \mathcal{N}(0, 1).$$

$\beta^\top = (\beta_1, \beta_2) = (0, 1)$ is estimated using MLE. For the power properties, we use three directions of deviations, coming from the simulation exercise of Mora and Moro-Egido (2008):

- For the distribution of the error term we consider the logistic distribution and the Student distributions with 5, 10 and 20 degrees of freedom as alternatives.
- We consider some non linearities of the latent variable with respect to the observable X by adding $c(X^2 - 1)$ to the regular expression of the latent variable for $c = 0.2, 0.4, 0.6$ (denoted $O - 0.2, O - 0.4$ and $O - 0.6$ in the Table).
- We consider heteroskedasticity of the error term. The variance of U given $X = x$ is equal to $\exp(dX^2 - d^2/2)$, for $d = 0.2, 0.4, 0.6$ (denoted $H - 0.2, H - 0.4$ and $H - 0.6$ in the Table). The unconditional distribution of U is still a standard normal distribution.

For the size properties as well as for the power properties, we use several choices for the instruments: $h(x) = 1, x, x^2, \cos(x), \cos(2x)$ are ad-hoc choices¹⁴, $h(x) = \frac{\phi(x^T \beta)}{\sqrt{\Phi(x^T \beta)(1 - \Phi(x^T \beta))}}$ yields to the classical moment $\mathbf{1}\{Y = 1\} - \Phi(x^T \beta)$ (denoted Std in the Table), $h(x) = (x^2 - 1) \frac{\phi(x^T \beta)}{\sqrt{\Phi(x^T \beta)(1 - \Phi(x^T \beta))}}$ is the instrument proposed by Skeels and Vella (1999) for omitted variable (here $x^2 - 1$, noted Omitted in the Table), $h(x) = (x^T \beta)x^2 \frac{\phi(x^T \beta)}{\sqrt{\Phi(x^T \beta)(1 - \Phi(x^T \beta))}}$ is the instrument proposed by Skeels and Vella for testing heteroskedasticity (noted Hetero in the Table), $h_{opt}(x) = \frac{F_a(x) - F_0(x)}{\sqrt{\Phi(x^T \beta)(1 - \Phi(x^T \beta))}}$ is the instrument which gives the highest approximated slope for distinguishing F_0 the c.d.f. of U under the null and F_a the c.d.f. under the alternative

¹⁴The cos and sin are the first members of the family $\exp(ikx)$, $k \in \mathbb{N}$. We could have considered all the family to construct a real specification test.

(labeled "Optilog" and "Opti-t5" for power against the logistic and the Student(5) distributions). The row labeled "All" consists in the joined test with $h(x)$ being $\sin(kx)$ and $\cos(kx)$ for $k = 1$ to 4. Table 2 displays the rejection frequencies for a 5%-level test for the sample sizes $n = 500$ and $n = 1000$ (there is not very much power for $n = 100$). All variances are computing using the formula (19) except for the second column "Size*" for which every projection involved in the construction is estimated in the data. Like in Orme (1990), Skeels and Vella (1999) and Mora and Moro-Egido (2008) the over-rejection rate is quite severe for many instruments in this case even for $n = 1000$.

The simple moment "Std" which compares the conditional probability of $Y = 1$ in the sample with the theoretical performs well. It comes from the fact that all departure tested here leads to biased estimators for β . The test "Omitted" is also very powerful. This test has been derived by Skeels and Vella to detect omitted variables. This test, however, performs better than the one derived for detecting the heteroskedasticity of the error term. In a different experiment, Skeels and Vella found relatively low power for the test "Hetero". The gaussian test seems to detect departure from normality quite efficiently as well as the optimal tests. These optimal tests (in the sense of the approximated slope) are quite close to the more powerful tests. The test "All" is quite powerful (though never the first) in many directions.

5.2 Poisson counting process

In this subsection, we focus on testing the Poisson distribution. We focus on testing a marginal distribution in a i.i.d. context. In this case, the score function is proportional to the first Charlier polynomial. The Charlier polynomials of higher degrees are therefore robust to the problem of the parameter estimation uncertainty. We will use these moments to test our distributional assumption.

We consider $K=7$ cells defined by $\{Y = 0\}, \{Y = 1\}, \{Y = 2\}, \{Y = 3\}, \{Y = 4\}, \{Y = 5\}, \{Y \geq 6\}$. Like before, $\hat{p}_j = p_j(\hat{\theta})$ is the theoretical probability of $Y \in C_j$ estimated at $\hat{\theta}$ and q_j the empirical one. We compare our tests with two well-known tests derived from the Cressie-Read divergence family. The Cressie-Read statistic with parameter α , CR^α , is a pseudo-distance measure between the theoretical and the empirical distributions:

$$CR^\alpha = \frac{2n}{\alpha(\alpha + 1)} \sum_{i=1}^K q_i \left(\left(\frac{q_i}{p_i} \right)^\alpha - 1 \right),$$

and is asymptotically $\chi^2(K - 2)$ distributed under the null that Y is Poisson distributed (one degree less because of the estimation of the parameter of the Poisson distribution). When $\alpha = 1$ the Cressie-Read statistic is the Pearson chi-squared statistic. When α tends to 0 this is the empirical log-likelihood ratio.

Four sample sizes are considered: 100, 200, 500 and 1000. We first study the size properties of our tests: we start by simulating i.i.d. samples of Poisson distributions with parameter $\theta^0 = 2$ (mean and variance are

equal to θ^0)¹⁵

In Table 3, we assume that θ is known or estimated. We provide the rejection frequencies at a 5% level test. Two sets of moments are considered: the individual moments based on a single Charlier polynomial $C_k^{\theta^0}$ and the joint moments based on the first Charlier polynomials: $C_{2-3}^{\theta^0}$ is the joined moment combining $C_2^{\theta^0}$ and $C_3^{\theta^0}$, $C_{2-4}^{\theta^0}$ is the joined moment combining $C_2^{\theta^0}$ to $C_4^{\theta^0}$. We also display the results for the two Cressie-Read statistics.

The finite sample properties of these tests are clearly good for all polynomials. The rejection rates are very close to 5% even for very small sample sizes (100 observations). When θ is estimated (we use the MLE), the size performance differences between knowing θ or estimating it are very low but exist in very small sample sizes. Even in the “worst” cases, the results are quite good. The CR statistics have also good sample size properties.

In Table 4 we study the power properties by simulating several alternatives. We focus on two distributions with two parameters and which have the Poisson as limit distribution: a binomial distribution and a negative binomial distribution. They all have the same expectation θ^0 . The parameter is estimated by a QMLE procedure using the Poisson model ($\hat{\theta}$ is therefore the mean of the data).

We simulate a binomial $\mathcal{B}(k, \frac{\theta^0}{k})$ for three values of k (10, 15 and 20). We know that, 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 than in the previous Table.

We first observe that the power of the tests decreases when k and δ increase. For the 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 performances are very good for the other sample sizes and for most of the moments used, more especially for the second Charlier polynomial which detects the over-dispersion in the data.

5.3 Backtesting VaR model

We consider in this Monte Carlo experiment the backtests of VaR models, defined in Section 4.4. The returns of a fictive portfolio are assumed to follow a GARCH (1,1) model with normal innovations:

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

with $\varepsilon_t \sim \mathcal{N}(0, 1)$, $\omega = 0.2$, $\gamma = 0.1$ and $\beta = 0.9$.

¹⁵The theoretical probabilities of belonging to cell C_1 to C_7 are respectively equal to 13.5%, 27.1%, 27.1%, 18.0%, 9.0%, 3.6%, 1.2%.

We simulate $T = 1500$ observations which corresponds approximately to six years of trading days. From $t = 500$ to 1495 and every 5 observations, we estimate a GARCH(1,1)-Normal model on the last 500 observations. We use these estimations to compute the 1-day ahead forecasts of the 1% and 5% VaR for the five following days (before reestimating the model, etc.).

We finally have 1000 daily VaR forecasts (for each value of α) which have been estimated at day -1 (the VaR is estimated out of sample like in the real life). We construct the hit sequence by comparing the raw returns and the VaR forecasts. We consider four sample sizes by focusing on the first year than on the first two years, etc. (sample sizes from 250 to 1000).

For the size properties, we consider as a benchmark the artificial case where the parameters are assumed known by the econometrician, than we use the MLE to estimate them (we therefore ran 200 estimations for the four years).

For the power properties, we consider 3 different series. The first two VaR series are constructed from a GARCH (1,1) with standardized Student innovations, where the degrees of freedom are set to 5 and 9. We assume in this case that we estimate blindly a GARCH(1,1) with normal innovations. The estimator is therefore the QMLE. We expect the VaR forecasts to be lower than the accurate ones, especially for $\alpha = 0.01\%$ ¹⁶.

All these simulations are drawn 5000 times. In Tables 5 through 7 we report the rejection frequencies at a 5% level test for several moments and tests. C_{uc} is the unconditional version of the LR test derived in Christoffersen (1998), ξ_{uc} is the unconditional moment-based test. Both are asymptotically equivalent. C_c is the conditional test of Christoffersen (1998) which tests the independence between I_t and I_{t-1} . "Markov" sums the two Christoffersen' tests, C_{uc} and C_c . We also use lagged values of returns to test the independence between I_t and the past information. We use $\cos(kr_{t-h})$ for $k = 1$ and 2 , $h = 1, \dots, 4$. We do not consider the parameter estimation uncertainty. As mentioned above, the econometrician rarely observes the model used by the practitioner and consequently, can not correct for this uncertainty. Moreover, the distortion vanishes when α becomes smaller. We expect a small distortion for $\alpha = 1\%$.

We do not use the tests based on the duration between two consecutive hits. In some cases, the test could not be run and it would have biased the comparisons to drop these particular samples (more specially when $\alpha = 1\%$ for 250 and 500 observations). More importantly, for the feasibility of the test procedure, we choose to display the rejection rates using the asymptotic critical values. Our goal is to provide tests for which the asymptotic approximation works quite well. We do not want to use simulation techniques to compute the critical values. The number of durations observed is very small even for a sample size equal to 1000. It would have been unfair to display the raw rejection frequencies using the asymptotic critical values.

¹⁶The 1% and 5% quantile are respectively equal to -2.326 and -1.645 for the standard normal distribution, -2.508 and -1.610 for the standardized Student with 8 degrees of freedom and -2.606 and -1.560 for the standardized Student with 5 degrees of freedom.

Table 5 displays the size properties. When θ is known, each test but the one based on the independence between $\cos(r_{t-1})$ and I_t has an accurate size. The conditional test C_c is also upward biased which is compatible with the last result. This bias increases with the sample size. Though asymptotically equivalent, C_{uc} and ξ_{uc} behaves differently for small sample and the latter performs better. When θ is estimated, the rejection frequencies are quite similar. The distortion due to the estimation of the parameters (at least for this GARCH (1,1) model) is negligible.

Table 6 presents the results when the innovations are standardized Student. As expected, the power comes from the unconditional tests as, at least for $\alpha = 1\%$, the quantiles of the normal and of the standardized Student are quite different. We have obviously more power for $\nu = 5$. Like before, ξ_{uc} performs better than C_{uc} . When $\alpha = 5\%$, the tests do not have very much power because of the proximity of the two quantiles.

The results for VaR computed from the Historical Simulation are presented in Table 7. The power comes from the correlation among the hits. The corresponding tests have more power when $\alpha = 5\%$ because we have more observations to detect this correlation. It is worth noting that the instruments $\cos 2r_{t-h}$ do not have any power and so, are not considered in the empirical application.

6 Empirical Application

We consider in this section the backtests of VaR models on some US index. We use the S&P 500 daily index for the period 2003-2007. The first two years only help to provide past data for estimating the different competing models. Figure 1 displays the price and estimated volatility for the period 2005-2007 (we only use the first 750 trading days). The activity is quite homogeneous in the first two years while the volatility increases dramatically during the last year. We therefore expect to have more hits corresponding to the last 250 observations.

Like in the Monte Carlo experiment, we estimate each competing VaR model on a weekly basis using the last 500 observations. We then forecast the 1% and 5% daily VaR for the following week using these estimates. We consider 5 competing models for the VaR:

- a GARCH(1,1)-Normal model with constant mean,
- a GARCH(1,1)-Student model with constant mean
- a GARCH(1,1) model with constant mean estimated by QMLE where the quantile of the innovation process is estimated with the Cornish-Fisher approximation,
- the Historical Simulation VaR model (HS) using the past 500 observations,

- the Filtered Historical Simulation VaR model (FHS). For this model, we estimate the quantile of the distribution of the innovations using the past 500 observations.

Figure 2 displays the 1% VaR forecasts for each model. As expected, the model HS provides VaR forecasts which do not vary too much.

Figure 3 displays the volatility path with respect to the hit sequences. There is a hit for each "jump" in the volatility path. Surprisingly the Cornish Fisher approximation seems to fit better than the other models especially for the last period. Table 8 presents the time index for the hits. We have respectively 20, 17, 9, 15 and 16 hits for all the models. Under H_0 , the average number should be 7. Most of the hits occur in the last year when the volatility increases.

Table 9 presents the test-statistics for testing the accuracy of the 1% VaR model. The tests are defined in the Monte Carlo Section. We also present their corresponding p-value in parentheses. For the first year, no model can be rejected. With two years of observations, the GARCH-Normal is rejected. As usual, the rejection comes from the unconditional test. We know that a GARCH-Normal does not fit the left part of the tail of the innovations of the process (the degrees of freedom for the Student distribution are around 10). When we includes the last year, all the models but the Cornish-Fisher are rejected. We have indeed more than 7 hits (though we should expect 2.5) for these 4 models.

Table 10 presents the same results for the 5% VaR forecasts. Only HS is rejected or close to rejection due to the failure in fitting the dynamic of the process (HS gives too flat VaR forecasts).

7 Conclusion

We introduce in this paper moment-based tests for parametric discrete distributions. Our goal is to present techniques that are easy to implement but without losing power to detect departure from the null hypothesis. As already mentioned, moment techniques are quite easy to adapt to the time series case and to take into account the parameter estimation uncertainty.

Our tests have good size and good power properties even for small samples. This power is due to the fact that we consider particular moments instead of many moments together. There are many examples where we do not want to have omnibus tests but power against specific alternatives. We show while focusing on several different examples, that the construction of the tests do not require any sophisticated technique and that we can adapt the testing framework to many cases.

It should make these tests attractive for applied econometricians for which the scope is the application and who want to avoid too many technicalities. Distributional assumptions are necessary in applied econometrics to compute forecasts, to make results tractable in structural economic models, to estimate quantities in small data

set. However, one should (if possible) test the assumptions made to emphasize the results derived. Otherwise, one can not avoid the risk to have biased estimations due to misspecification.

One direction for future research is the extension to the multivariate case. There are many potential applications. Moreover, in a recent contribution, Bontemps and Meddahi (2008) have characterized optimal moments for testing a given null against a given alternative. Another direction for future research is therefore to combine this optimal approach to the many moment case. As a matter of fact, when one considers all the class of moments characterized in the paper, one has many strategies to join them in a single test. One could characterize the one which gives optimal power in some given directions.

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.
- Andrews, D.W.K. (1988), "Chi-square diagnostic tests for econometric models: theory", *Econometrica*, 56, 1419-53.
- Andrews, D.W.K. (1991), "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation", *Econometrica*, 60, 953-966.
- 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.
- Bai, J. and Z. Chen (2008), "Testing multivariate distributions in GARCH models", *Journal of Econometrics*, 143, 19-36.
- Ben-Akiva, M. E. and S. R. Lerman (1985), *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press.
- Bontemps, C. and N. Meddahi (2005), "Testing Normality: A GMM Approach," *Journal of Econometrics*, 124, 149-186.
- Bontemps, C. and N. Meddahi (2007), "Testing Distributional assumptions: A GMM Approach," mimeo University of Toulouse.
- Bontemps, C. and N. Meddahi (2008), "Optimal moment-based tests," mimeo University of Toulouse.
- 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., G. Colletaz, C. Hurlin and S. Tokpavi (2008), "Backtesting Value-at-Risk: a GMM duration-based test", Working Paper.
- Chen, Y.T. (2007a), "Moment Tests for Standardized Error Distributions: A Simple Robust Approach", mimeo.
- Chen, Y.T. (2007b), "Moment-Based Copula Tests for Financial Returns", *Journal of Business & Economic Statistics*, 25, 377-397.
- Chihara, T.S. (1978), *An Introduction to Orthogonal Polynomials*, Gordon and Breach Science Publishers.
- Christoffersen, P. F. (1998), "Evaluating Interval Forecasts," *International Economic Review*, 39, pp. 841-862.
- Christoffersen, P. F. and D. Pelletier (2004), "Backtesting Value-at-Risk: A Duration-Based Approach", *Journal of Financial Econometrics*, 2, 84-108.

- Delgado, M.A. and W. Stute (2008), "Distribution-free specification tests of conditional models", *Journal of Econometrics*, 143, 37-55.
- Diebold, F.X., T.A. Gunther and A.S. Tay (1998), "Evaluating Density Forecasts, with Applications to Financial Risk Management," *International Economic Review*, 39, 863-883.
- Diebold, F.X., J. Hahn, and A.S. Tay (1999), "Multivariate Density Forecast Evaluation and Calibration in Financial Risk Management: High-Frequency Returns on Foreign Exchange," *Review of Economics and Statistics*, 81, 661-673.
- Diebold, F.X., A.S. Tay, and K.F. Wallis (1999), "Evaluating Density Forecasts of Inflation: The Survey of Professional Forecasters," in *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive W.J. Granger*, R.F. Engle and H. Whited, Eds., pp. 76-90, Oxford University Press.
- 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.
- Epps, T.W and L.B. Pulley (1983), "A Test for Normality Based on the Empirical Characteristic Function," *Biometrika*, 70, 723-726.
- Escanciano, J.C. and J. Olmo (2007), "Estimation Risk Effects on Backtesting For Parametric Value-at-Risk Models", mimeo.
- 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
- Freeland, R.K. and B.P.M. McCabe (2004), "Analysis of low count time series data by Poisson autoregression", *Journal of Time Series Analysis*, 25, 701-722.
- Hansen, L.P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029-1054.
- Hansen, L.P. (2001), "Generalized Method of Moments Estimation: A Time Series Perspective," *International Encyclopedia of Social and Behavioral Sciences*.
- Horowitz, J.L. (1992), "A Smoothed Maximum Score Estimator for the Binary Response Model", *Econometrica*, 60, 505-551.
- Jarque, C.M. and A.K. Bera (1980), "Efficient Tests for Normality, Homoscedasticity and serial Independence of Regression Residuals," *Economics Letters*, 6, 255-259.
- Johnson, N.L, S. Kotz and N. Balakrishnan (1994), *Discrete Continuous Univariate Distributions*, Wiley.
- Kalliovirta, L. (2006), "Misspecification Tests Based on Quantile Residuals", Discussion Paper No 124, Helsinki Center of Economic Research.

- Khmaladze, E.V. (1981), "Martingale Approach in the Theory of Goodness-of-Tests," *Theory of Probability and its Application*, 26, 240-257.
- Kolmogorov, A.N. (1933), "Sulla Determinazione Empirica di una Legge di Distribuzione," *Giornale Ist. Attuari.*, 4, 83-91.
- Koutrouvelis, I.A. and J. Kellermeyer (1981), "A Goodness-of-Fit Test Based on the Empirical Characteristic Function when Parameters must be Estimated," *J. R. Statist. Soc. B*, 43, 173-176.
- Kupiec, P. H. (1995), "Techniques for Verifying the Accuracy of Risk Measurement Models," *Journal of Derivatives*, winter, 73-84.
- MaCurdy, T.E. and K. Ryu (2003), "Equivalence results in chi-square tests", *Economics Letters*, 80, 329-36.
- Manski, C. F. (1975), "Maximum Score Estimation of the Stochastic Utility Model of Choice", *Journal of Econometrics*, 3, 205-228.
- Matzkin, R. (1992), "Nonparametric and Distribution-Free Estimation of the Binary Choice and the Threshold-Crossing Models", *Econometrica*, 60, 239-270.
- McKenzie, E. (1986), "Autoregressive moving-average processes with negative binomial and geometric marginal distributions", *Adv. Appl. prob.*, 18, 679-705.
- McKenzie, E. (1988), "Some ARMA models of dependent sequence of Poisson counts", *Adv. Appl. prob.*, 20, 822-35.
- 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.
- Orme, C.D. (1990), "The Small Sample Performance of the Information Matrix Test", *Journal of Econometrics*, 46, 309-331.
- Rao, C. R. (1965), *Linear statistical inference and its applications*, Wiley.
- Rosenblatt, M. (1952), "Remarks on a multivariate transformation", *Annals of Mathematical Statistics*, 23, 470-472.
- Schoutens, W. (2000), *Stochastic Processes and Orthogonal Polynomials*, Lecture Notes in Statistics 146, Springer-Verlag. G.
- 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.
- Smirnov, N.V. (1939), "Sur les Écarts de la Courbe de Distribution Empirique," *French/Russian Su Matematiceskii Sbornik N. S.*, 6, 3-26.

- Szegő, 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.
- White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1-25.
- White, H. (1984), *Asymptotic Theory For Econometricians*, New York Academic Press.
- Wooldridge, J. (1990), "A Unified Approach to Robust, Regression-based Specification Tests," *Econometric Theory*, 6, 17-43.

Appendix

Parameter estimation uncertainty for non-smooth moments.

We focus here on a special case when the moment-based test is related to a cell for which the bounds are functions of the parameter θ . One example treated in this paper is the backtest of VaR models (Example 3); another is the Pearson chi-squared test where the cells are defined by empirical quantiles. We assume here that (y_t, x_t) , $t = 1, \dots, T$ are i.i.d.

The moment of interest is defined as:

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

where l, u, p are continuously differentiable with respect to θ locally around θ^0 . The difference between the moment taken at the estimated parameter $\hat{\theta}$ and the true value θ^0 can be decomposed in the following manner:

$$\begin{aligned} \hat{z}_t^T &= m(y_t, x_t, \hat{\theta}) - m(y_t, x_t, \theta^0) \\ &= \mathbf{1}\{y_t \in [l(x_t, \hat{\theta}), u(x_t, \hat{\theta})]\} - \mathbf{1}\{y_t \in [l(x_t, \theta^0), u(x_t, \theta^0)]\} - (p(x_t, \hat{\theta}) - p(x_t, \theta^0)) \\ &= \underbrace{(\mathbf{1}\{y_t \leq u(x_t, \hat{\theta})\} - \mathbf{1}\{y_t \leq u(x_t, \theta^0)\})\mathbf{1}\{y_t \geq l(x_t, \theta^0)\}}_{=W_{t,1}^T} \\ &\quad + \underbrace{(\mathbf{1}\{y_t \geq l(x_t, \hat{\theta})\} - \mathbf{1}\{y_t \geq l(x_t, \theta^0)\})\mathbf{1}\{y_t \leq u(x_t, \theta^0)\}}_{=W_{t,2}^T} \\ &\quad + \underbrace{(\mathbf{1}\{y_t \geq l(x_t, \hat{\theta})\} - \mathbf{1}\{y_t \geq l(x_t, \theta^0)\})(\mathbf{1}\{y_t \leq u(x_t, \hat{\theta})\} - \mathbf{1}\{y_t \leq u(x_t, \theta^0)\})}_{=W_{t,3}^T} \\ &\quad - \underbrace{(p(x_t, \hat{\theta}) - p(x_t, \theta^0))}_{=W_{t,4}^T} \end{aligned}$$

Let us first focus on the first term $W_{t,1}^T$. This term depends on the sample size through the estimated parameter $\hat{\theta}$. We denote by $F(\cdot; \theta^0)$ the conditional c.d.f. of Y . We can compute the conditional expectation and variance of $W_{t,1}^T$:

$$E_{\theta} W_{t,1}^T = F(u(x_t, \hat{\theta}); \theta^0) - F(u(x_t, \theta^0); \theta^0) \tag{28}$$

$$= \frac{\partial u(x, \theta)}{\partial \theta^\top} \frac{\partial F}{\partial y}(u(x, \theta); \theta^0) \Big|_{\theta=\theta^0} (\hat{\theta} - \theta^0) + o(\hat{\theta} - \theta^0) \tag{29}$$

$$V_{\theta} W_{t,1}^T = E(W_{t,1}^T)^2 - (F(u(x_t, \hat{\theta}); \theta^0) - F(u(x_t, \theta^0); \theta^0))^2 \tag{30}$$

$$= F(u(x_t, \hat{\theta}); \theta^0) + F(u(x_t, \theta^0); \theta^0) - 2F(\min(u(x_t, \hat{\theta}), u(x_t, \theta^0)); \theta^0) \tag{31}$$

$$- (F(u(x_t, \hat{\theta}); \theta^0) - F(u(x_t, \theta^0); \theta^0))^2 \tag{32}$$

$$= o(\hat{\theta} - \theta^0) \tag{33}$$

A standard Taylor expansion for the empirical mean of $W_{t,1}^T$ is:

$$\frac{1}{T} \sum_{t=1}^T W_{t,1}^T = E_0 W_{t,1}^T + \frac{1}{\sqrt{T}} \sqrt{V_0 W_{t,1}^T} Z + o\left(\frac{1}{T}\right), \quad (34)$$

where Z is a standard normal variable.

Incorporating (29) and (33) into (34), gives the expansion of the empirical process $\sqrt{T} \frac{1}{T} \sum_{t=1}^T W_{t,1}^T$ around θ^0 :

$$\begin{aligned} \sqrt{T} \frac{1}{T} \sum_{t=1}^T W_{t,1}^T &= \sqrt{T} E_0 W_{t,1}^T + \sqrt{V_0 W_{t,1}^T} Z + o\left(\frac{1}{\sqrt{T}}\right) \\ &= E_X \left(\frac{\partial u(x, \theta)}{\partial \theta^T} \frac{\partial F}{\partial y} (u(x, \theta); \theta^0) \Big|_{\theta=\theta^0} \right) \sqrt{T} (\hat{\theta} - \theta^0) + o(1). \end{aligned} \quad (35)$$

We can similarly consider the other terms to prove that:

$$\sqrt{T} \frac{1}{T} \sum_{t=1}^T \hat{z}_t^T = Q \sqrt{T} (\hat{\theta} - \theta^0) + o(1), \quad (36)$$

where

$$Q = E_X \left(\frac{\partial}{\partial \theta} u(x_t, \theta) \frac{\partial}{\partial y} F(u(x_t, \theta); \theta) - \frac{\partial}{\partial \theta} l(x_t, \theta) \frac{\partial}{\partial y} F(l(x_t, \theta); \theta) - \frac{\partial}{\partial \theta} p(x_t, \theta) \right)_{\theta=\theta^0}.$$

We will now derive the generalized Information Matrix equality in this particular case. From:

$$\int_{l(x_t, \theta)}^{u(x_t, \theta)} f(y; \theta) dy = p(x, \theta),$$

we can differentiate this equality with respect to θ . The bounds depend on θ contrary to the usual case. The equality follows immediately.

Proof of Proposition 1.

$$E_\theta [\Delta \psi(y, x, \theta)] = \sum_{i=l}^r \Delta \psi(a_i, x, \theta) p_i(x, \theta) \quad (37)$$

We first prove the proposition in the case where l and r are finite. Reordering the second term of the last

expression yields to:

$$E[\Delta\psi(y, x, \theta)] = \sum_{i=l}^r \psi(a_{i+1}, x, \theta)p_i(x, \theta) - \sum_{i=l}^r \psi(a_i, x, \theta)p_i(x, \theta) \quad (38)$$

$$= \sum_{i=l}^{r-1} \psi(a_{i+1}, x, \theta)p_i(x, \theta) - \sum_{i=l-1}^{r-1} \psi(a_{i+1}, x, \theta)p_{i+1}(x, \theta) + Kp_r(x, \theta) \quad (39)$$

$$= \sum_{i=l}^{r-1} \psi(a_{i+1}, x, \theta) (p_i(x, \theta) - p_{i+1}(x, \theta)) - \psi(a_l, x, \theta)p_l(x, \theta) + Kp_r(x, \theta) \quad (40)$$

$$= \sum_{i=l}^r \psi(a_{i+1}, x, \theta) (p_i(x, \theta) - p_{i+1}(x, \theta)) \text{ under LB.} \quad (41)$$

$$= -E_\theta \left(\psi_{+1}(y, x, \theta) \frac{\Delta p(y, x, \theta)}{p(y, x, \theta)} \right) \quad (42)$$

When r is infinite, we know that each term of Equation (39) converges due to the assumption of finite expectation of $\psi(y, x, \theta)$ with respect to P_x^0 . Thus, Equation (39) becomes

$$E[\Delta\psi(y, x, \theta)] = \sum_{i=l}^{+\infty} \psi(a_{i+1}, x, \theta)p_i(x, \theta) - \sum_{i=l-1}^{+\infty} \psi(a_{i+1}, x, \theta)p_{i+1}(x, \theta) \quad (43)$$

which gives the same result.

When l is infinite, the second term of Equation (40) vanishes due to the finite conditional expectation of ψ : $\psi(a_i, x, \theta)p_i(x, \theta) \xrightarrow{i \rightarrow -\infty} 0$. This Equation becomes

$$E[\Delta\psi(y)] = \sum_{y=-\infty}^{r-1} \psi(a_{i+1}, x, \theta)p_i(x, \theta) - \sum_{y=-\infty}^{r-1} \psi(a_{i+1}, x, \theta)p_{i+1}(x, \theta) + Kp_r(x, \theta) \quad (44)$$

when r is finite and

$$E[\Delta\psi(y)] = \sum_{y=-\infty}^{+\infty} \psi(y+1)p_y - \sum_{y=-\infty}^{+\infty} \psi(y+1)p_{y+1} \quad (45)$$

when r is infinite. In both cases, we can write;

$$E[\Delta\psi(y, x, \theta)] = -E_\theta \left(\psi_{+1}(y, x, \theta) \frac{\Delta p(y, x, \theta)}{p(y, x, \theta)} \right)$$

Proof of Lemma 2

Any function $f(y, x)$ defined on the support $S = \{0, 1\} \times \mathbb{R}^p$ can be expressed as:

$$\begin{aligned} f(y, x) &= f(0, x)(1 - y) + f(1, x)y \\ &= f(0, x) + p_x (f(1, x) - f(0, x)) + (f(1, x) - f(0, x)) (y - p_x) \end{aligned}$$

Any real valued function on S is a linear combination of the first Krawtchouk polynomials $K_0^{p_x}$ (i.e. the constant 1) and $K_1^{p_x}$. If the conditional expectation of $f(y, x)$ is equal to zero, the component associated to the constant $f(0, x) + p_x(f(1, x) - f(0, x)) = 0$ is equal to zero and

$$f(y, x) = (f(1, x) - f(0, x))(y - p_x) = \left(\sqrt{p_x(1-p_x)}(f(1, x) - f(0, x))\right) K_1^{p_x}(y).$$

Moment robust to parameter the estimation uncertainty in a binary model:

Following the subsection on the parameter estimation uncertainty, we know that a robust moment could be constructed by projecting the original one onto the orthogonal space spanned by the score function:

$$m^\perp(y, x, \theta) = m(y, x, \theta) - [E_0(m(y, x, \theta)s_\beta^\top(y, x))] [V_0s_\beta(y, x)]^{-1} s_\beta(y, x).$$

Using the law of iterated expectation, we can simplify the two matrices involved in the previous equality:

$$\begin{aligned} E_0(m(y, x)s_\beta^\top(y, x)) &= E_X E_\beta \left(-h(x)x^\top \frac{f(x^\top \beta)}{\sqrt{p_x(1-p_x)}} K_1^{p_x}(y)^2 \right) \\ &= -E_X \left(h(x)x^\top \frac{f(x^\top \beta)}{\sqrt{p_x(1-p_x)}} \right) \end{aligned}$$

as $E_\beta K_1^{p_x}(y)^2 = 1$.

Similarly,

$$\begin{aligned} V_0s_\beta(y, x, \beta) &= E_X V_\beta \left(xx^\top \left(\frac{f(x^\top \beta)}{\sqrt{p_x(1-p_x)}} \right)^2 K_1^{p_x}(y)^2 \right) \\ &= E_X \left(xx^\top \frac{f^2(x^\top \beta)}{(1-p_x)p_x} \right) \end{aligned}$$

We now plug the last two terms into the expression of m^\perp :

$$m^\perp(y, x, \theta) = h^\perp(x) K_1^{p_x}(y),$$

where $h^\perp(x) = h(x) - E_X \left(h(x)x^\top \frac{f(x^\top \beta)}{\sqrt{p_x(1-p_x)}} \right) \left[E_X \left(xx^\top \frac{f^2(x^\top \beta)}{(1-p_x)p_x} \right) \right]^{-1} x \frac{f(x^\top \beta)}{\sqrt{p_x(1-p_x)}}$.

Ordered models

We focus here on the ordered probit model with three outcomes, 0, 1 and 2. The conditional expectation of Y^k is equal to $p_1(x, \theta) + 2^k p_2(x, \theta)$. Any moment for which the expectation under the null is equal to zero can be expressed as a linear combination of two polynomials:

$$m(y, x, \theta) = \lambda_1(x)(y - E_\theta y) + \lambda_2(x)(y^2 - E_\theta y^2).$$

The score function with respect to the parameter β can be expressed using this form:

$$\begin{aligned} s_\beta(y, x) &= -\frac{x}{2} (y^2 - E_\theta(y^2)) \left(f(-x^\top \beta) \left(\frac{1}{p_0(x, \theta)} + \frac{2}{p_1(x, \theta)} \right) - f(\mu - x^\top \beta) \left(\frac{2}{p_1(x, \theta)} + \frac{1}{p_2(x, \theta)} \right) \right) \\ &\quad + \frac{x}{2} (y - E_\theta(y)) \left(f(-x^\top \beta) \left(\frac{3}{2p_0(x, \theta)} + \frac{4}{p_1(x, \theta)} \right) - f(\mu - x^\top \beta) \left(\frac{4}{p_1(x, \theta)} + \frac{1}{p_2(x, \theta)} \right) \right) \end{aligned}$$

There is a one-to-one relationship between the test functions $\psi_1(x)$, $\psi_2(x)$ and $\lambda_1(x)$, $\lambda_2(x)$:

$$\begin{aligned}\psi_1(x) &= -\lambda_1(x) \left(p_0(x, \theta) + 2 \frac{p_0(x, \theta)p_2(x, \theta)}{p_1(x, \theta)} \right) - \lambda_2(x) \left(p_0(x, \theta) + 4 \frac{p_0(x, \theta)p_2(x, \theta)}{p_1(x, \theta)} \right) \\ \psi_2(x) &= -\lambda_1(x) (p_1(x, \theta) + 2p_0(x, \theta)) - \lambda_2(x) (p_1(x, \theta) + 4p_0(x, \theta)) \\ \lambda_1(x) &= -\psi_1(x) \left(2 + \frac{3p_1(x, \theta)}{2p_0(x, \theta)} \right) + \psi_2(x) \left(\frac{1}{2} + \frac{2p_2(x, \theta)}{p_1(x, \theta)} \right) \\ \lambda_2(x) &= \psi_1(x) \left(1 + \frac{p_1(x, \theta)}{2p_0(x, \theta)} \right) - \psi_2(x) \left(\frac{1}{2} + \frac{p_2(x, \theta)}{p_1(x, \theta)} \right).\end{aligned}$$

The score component $s_\beta(y, x)$ can be build from the test functions $\psi_1(x) = -x \frac{f(-x^\top \beta)}{p_1(x, \theta)}$ and $\psi_2(x) = -x \frac{f(\mu - x^\top \beta)}{p_2(x, \theta)}$.

We can express similarly the other components of the score function. For the component w.r.t. the threshold value μ , $\psi_1(x) = 0$ and $\psi_2(x) = \frac{f(\mu - x^\top \beta)}{p_2(x, \theta)}$; for the component w.r.t. the parameter ν of the c.d.f. of the error term, $\psi_1(x) = \frac{f_\nu(-x^\top \beta)}{p_1(x, \theta)}$ and $\psi_2(x) = \frac{f_\nu(\mu - x^\top \beta)}{p_2(x, \theta)}$.

In order to compute a moment robust to the parameter estimation uncertainty, one has to compute first the covariance between two moment restrictions, $m(y, x, \theta) = \lambda_1(x)(y - E_\theta y) + \lambda_2(x)(y^2 - E_\theta y^2)$ and $n(y, x, \theta) = \nu_2(x)(y^2 - E_\theta y^2) + \nu_1(x)(y - E_\theta y)$:

$$\begin{aligned}Cov_\theta(m(y, x, \theta), n(y, x, \theta)) &= E_X [\mu_2(x)\lambda_2(x)(E_\theta y^4 - (E_\theta y^2)^2)] + E_X [\nu_1(x)\lambda_1(x)(E_\theta y^2 - (E_\theta y)^2)] \\ &\quad + E_X [(\nu_1(x)\lambda_2(x) + \nu_2(x)\lambda_1(x)) (E_\theta y^3 - (E_\theta y^2)(E_\theta y))].\end{aligned}$$

We can use these results with (9) to achieve the construction.

Proof of Proposition 3

Let first consider the generating function of the orthonormalized Charlier polynomials $C_m^a(y)$, $m \in \mathbb{N}$:

$$\sum_{m=0}^{+\infty} C_m^a(y) \frac{w^m}{\sqrt{m!a^m}} = e^w \left(1 - \frac{w}{a} \right)^y$$

Here the marginal distribution of y_t is a Poisson with parameter $\frac{\mu}{1-\alpha}$.

Using the previous expression with $y \equiv y_t$ and assuming that the sum can commute with E_{t-1} , one obtains:

$$\sum_{m=0}^{+\infty} E_{t-1} C_m^a(y_t) \frac{w^m}{\sqrt{m!a^m}} = e^w E_{t-1} \left(1 - \frac{w}{a} \right)^{y_t}. \quad (46)$$

We know (see, for example, Freeland and McCabe, 2004) that the conditional probability $p(y_t | y_{t-1})$ of y_t conditional on y_{t-1} is:

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

We use this expression to compute the following conditional expectation

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

We can now plug the last result into (46) to get

$$\begin{aligned}
\sum_{m=0}^{+\infty} E_{t-1} C_m^{\frac{\mu}{1-\alpha}}(y_t) \frac{w^m}{\sqrt{m! \left(\frac{\mu}{1-\alpha} \right)^m}} &= e^{w\alpha} \left(1 - \frac{\alpha w(1-\alpha)}{\mu} \right)^{y_{t-1}} \\
&= \sum_{m=0}^{+\infty} \alpha^m C_m^{\frac{\mu}{1-\alpha}}(y_{t-1}) \frac{w^m}{\sqrt{m! \left(\frac{\mu}{1-\alpha} \right)^m}}
\end{aligned}$$

and so, making each term of w^m equal, we get:

$$E_{t-1} C_m^{\frac{\mu}{1-\alpha}}(y_t) = \alpha^m C_m^{\frac{\mu}{1-\alpha}}(y_{t-1}).$$

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

Pearson chi-squared test

Under standard mathematics, it is easy to prove that:

$$A = \sqrt{n} \sum_{i=1}^T \begin{pmatrix} m_1(y_i, \theta^0) \\ \vdots \\ m_{K-1}(y_i, \theta^0) \end{pmatrix} \xrightarrow[n \rightarrow \infty]{} \mathcal{N}(0, \Sigma)$$

where

$$\Sigma = \begin{bmatrix} p_1^0(1-p_1^0) & -p_1^0 p_2^0 & \dots & -p_1^0 p_{K-1}^0 \\ -p_1^0 p_2^0 & p_2^0(1-p_2^0) & \ddots & -p_2^0 p_{K-1}^0 \\ \vdots & \ddots & \ddots & \vdots \\ -p_1^0 p_{K-1}^0 & \dots & \dots & p_{K-1}^0(1-p_{K-1}^0) \end{bmatrix}$$

and:

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{p_1^0} + \frac{1}{p_K^0} & \frac{1}{p_K^0} & \cdots & \frac{1}{p_K^0} \\ \frac{1}{p_K^0} & \frac{1}{p_1^0} + \frac{1}{p_K^0} & \ddots & \\ \vdots & \ddots & \ddots & \vdots \\ \frac{1}{p_K^0} & \cdots & \cdots & \frac{1}{p_{K-1}^0} + \frac{1}{p_K^0} \end{bmatrix}$$

Following (4), the test statistic based on m_1, \dots, m_{K-1} is the well-known Pearson chi-square statistic. Let $q_j = \sum_{i=1}^T \mathbf{1}\{y_i \in C_j\}$ than:

$$\begin{aligned} \xi &= A^\top \Sigma^{-1} A = n \left(\sum_{j=1}^{K-1} \frac{(q_j - p_j^0)^2}{p_j^0 + p_K^0} + \frac{2}{p_K^0} \sum_{i < j} (q_i - p_i^0)(q_j - p_j^0) \right) \\ &= n \left(\sum_{j \neq K} \frac{(q_j - p_j^0)^2}{p_j^0} + \frac{1}{p_K^0} \left(\sum_{j=1}^{K-1} (q_j - p_j^0)^2 + 2 \sum_{i < j} (q_i - p_i^0)(q_j - p_j^0) \right) \right) \\ &= n \left(\sum_{j \neq K} \frac{(q_j - p_j^0)^2}{p_j^0} + \frac{1}{p_K^0} \left(\sum_{j=1}^{K-1} q_j - \sum_{j=1}^{K-1} p_j^0 \right)^2 \right) \\ &= n \left(\sum_{j=1}^K \frac{(q_j - p_j^0)^2}{p_j^0} \right) \end{aligned} \tag{47}$$

Proof of Proposition 4

We first compute the covariance between one component of the moment and the score function:

$$E_\theta (m_k(y, \theta) s_\theta(y)) = E_\theta (\mathbf{1}\{\mathbf{Y} \in \mathbf{C}_k\} - \mathbf{p}_k(\theta)) \left(\frac{\partial \ln p_y}{\partial \theta} \right) = \frac{\partial p_k}{\partial \theta}.$$

Moreover:

$$[E_\theta (s_\theta(y) s_\theta^\top(y))]_{ij} = \sum_{c=1}^K p_c(\theta) \frac{\partial \ln p_c}{\partial \theta_i}(\theta) \frac{\partial \ln p_c}{\partial \theta_j}(\theta) = \sum_{c=1}^K \frac{1}{p_c \theta} \frac{\partial p_c}{\partial \theta_i}(\theta) \frac{\partial p_c}{\partial \theta_j}(\theta).$$

$$\text{Let } K = \begin{bmatrix} \frac{\partial p_1}{\partial \theta_1} & \cdots & \frac{\partial p_1}{\partial \theta_s} \\ \vdots & \ddots & \vdots \\ \frac{\partial p_{K-1}}{\partial \theta_1} & \cdots & \frac{\partial p_{K-1}}{\partial \theta_s} \end{bmatrix}, P = [p_1 \dots p_{K-1}]^\top, D = \text{diag}(p_1, \dots, p_{K-1}), S = [s_{\theta_1}(y), \dots, s_{\theta_s}(y)]^\top,$$

$$M = [m_1(y, \theta), \dots, m_{K-1}(y, \theta)]^\top \text{ and } M^\top = M - K [E_\theta(SS^\top)]^{-1} S.$$

We know that $VM = D - PP^\top = \Sigma$. Using the previous results, it appears that:

$$\begin{aligned}
E_\theta(SS^\top) &= \begin{bmatrix} \frac{\partial p_K}{\partial \theta_1}(\theta) \\ \vdots \\ \frac{\partial p_K}{\partial \theta_s}(\theta) \end{bmatrix} K^\top \begin{bmatrix} D^{-1} & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} \\ \hline 0 & \dots & 0 & p_K^{-1} \end{bmatrix} \begin{bmatrix} K \\ \frac{\partial p_K}{\partial \theta_1}(\theta) \quad \dots \quad \frac{\partial p_K}{\partial \theta_s}(\theta) \end{bmatrix}^\top \\
&= K^\top D^{-1} K + \frac{1}{p_K} \begin{bmatrix} \frac{\partial p_K}{\partial \theta_1}(\theta) \\ \vdots \\ \frac{\partial p_K}{\partial \theta_s}(\theta) \end{bmatrix} \begin{bmatrix} \frac{\partial p_K}{\partial \theta_1}(\theta) \\ \vdots \\ \frac{\partial p_K}{\partial \theta_s}(\theta) \end{bmatrix}^\top
\end{aligned} \tag{48}$$

And:

$$\begin{bmatrix} \frac{\partial p_K}{\partial \theta_1}(\theta) \\ \vdots \\ \frac{\partial p_K}{\partial \theta_s}(\theta) \end{bmatrix} = -K^\top [1, \dots, 1] = -K^\top \mathbf{1} \tag{49}$$

as the sum of the derivative of the p_i 's with respect to θ is equal to zero.

Thus:

$$\begin{aligned}
E_0(SS^\top) &= K^\top D^{-1} K + \frac{1}{p_K} K^\top \mathbf{1} \mathbf{1}^\top K \\
&= K^\top (D - PP^\top)^{-1} K
\end{aligned} \tag{50}$$

We can therefore compute the asymptotic variance of the robust moments M^\top .

$$\begin{aligned}
VM^\top &= VM - E_0(MS^\top)[E_0(SS^\top)]^{-1}E_0(SM^\top) \\
&= D - PP^\top - K \left[K^\top (D - PP^\top)^{-1} K \right]^{-1} K^\top.
\end{aligned} \tag{51}$$

It follows immediately by the central limit theorem that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^T \left((D - PP^\top)^{-\frac{1}{2}} M^\top(y_i, \theta) \right) \xrightarrow{n \rightarrow +\infty} \mathcal{N}(0, Id - U(U^\top U)^{-1}U^\top)$$

where $U = \left((D - PP^\top)^{-\frac{1}{2}} K^\top \right)$.

It is easy to prove that the rank of the variance matrix is equal to $K - 1 - rk(U) = K - 1 - rk(K^\top) = K - 1 - s$ as soon as $s < K - 1$ (if not, the system of moments is degenerated). The eigenvectors are

$$e_i = (D - PP^\top)^{\frac{1}{2}} \begin{bmatrix} \frac{1}{p_i} \frac{\partial p_i}{\partial \theta_1}(\theta) - \frac{1}{p_K} \frac{\partial p_K}{\partial \theta_1}(\theta) \\ \vdots \\ \frac{1}{p_i} \frac{\partial p_i}{\partial \theta_s}(\theta) - \frac{1}{p_K} \frac{\partial p_K}{\partial \theta_s}(\theta) \end{bmatrix}.$$

Let $M_j^\top = \frac{1}{n} \sum_{i=1}^T m_j^\top(y_i, \hat{\theta})$

$$M_j^\top = q_j - p_j - \lambda \bar{s}$$

where $q_j = \frac{1}{n} \sum_{i=1}^T \mathbf{1}\{y_i = a_j\}$, $\bar{s} = \frac{1}{n} \sum_{i=1}^T s(y_i, \hat{\theta})$, $\lambda = K [K^\top (D - PP^\top)^{-1} K]^{-1}$.

We know from Rao (1973, p.188) that (see also MaCurdy and Ryu, 2003)

$$\xi = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^T \left((D - PP^\top)^{-\frac{1}{2}} M^\top(y_i, \theta) \right) \right)^2 \sim \chi^2(K - 1 - s).$$

Moreover, ξ could be simplified:

$$\begin{aligned} \xi &= n \left(\sum_{j=1}^{K-1} \frac{(M_j^\top)^2}{p_j + p_K} + \frac{2}{p_K} \sum_{i < j} (M_i^\top)(M_j^\top) \right) \\ &= n \left(\sum_{j \neq K} \frac{(M_j^\top)^2}{p_j} + \frac{1}{p_K} \left(\sum_{j=1}^{K-1} (M_j^\top)^2 + 2 \sum_{i < j} (M_i^\top)(M_j^\top) \right) \right) \\ &= n \left(\sum_{j \neq K} \frac{(M_j^\top)^2}{p_j} + \frac{1}{p_K} \left(\sum_{j=1}^{K-1} M_j^\top \right)^2 \right) \\ &= n \left(\sum_{j=1}^{K-1} \frac{(q_j - p_j - \lambda \bar{s})^2}{p_j} + \frac{(q_K - p_K + (K-1)\lambda \bar{s})^2}{p_K} \right) \end{aligned} \tag{52}$$

$\lambda = U [U^\top (D - PP^\top)^{-1} U]^{-1}$ where

$$U = \begin{bmatrix} \frac{\partial p_1}{\partial \theta_1}(\hat{\theta}) & \dots & \frac{\partial p_1}{\partial \theta_s}(\hat{\theta}) \\ \vdots & \ddots & \vdots \\ \frac{\partial p_{K-1}}{\partial \theta_1}(\hat{\theta}) & \dots & \frac{\partial p_{K-1}}{\partial \theta_s}(\hat{\theta}) \end{bmatrix}, \quad P^\top = [p_1 \dots p_{K-1}] \text{ and } D = \text{diag}(p_1, p_2, \dots, p_{K-1}).$$

Backtesting VaR models

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

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

The score function is

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

The two expectations under concern are equal to

$$V_0 s_\theta(r_t) = 2E_0 \left[\frac{\partial \ln \sigma_t(\theta)}{\partial \theta} \frac{\partial \ln \sigma_t(\theta)}{\partial \theta^\top} \right]$$

and

$$E_0(\mathbf{1}\{r_t \leq -VaR_{t|t-1}^\alpha\} s_\theta^\top(r_t)) = -q_\alpha \phi(q_\alpha) E_0 \left[\frac{\partial \ln \sigma_t(\theta)}{\partial \theta^\top} \right].$$

The unconditional robust moment is:

$$m^\perp(r_t, \theta) = \mathbf{1}\{r_t \leq -VaR_{t|t-1}^\alpha\} - \alpha - \frac{q_\alpha \phi(q_\alpha)}{2} E_0 \left[\frac{\partial \ln \sigma_t(\theta)}{\partial \theta^\top} \right] E_0 \left[\frac{\partial \ln \sigma_t(\theta)}{\partial \theta} \frac{\partial \ln \sigma_t(\theta)}{\partial \theta^\top} \right]^{-1} s_\theta(r_t).$$

The test-statistic is modified through the variance term:

$$\begin{aligned} V_0 m^\perp(r_t, \theta) &= V_0 \left(\mathbf{1}\{r_t \leq -VaR_{t|t-1}^\alpha\} \right) - V_0 \left(\frac{q_\alpha \phi(q_\alpha)}{2} E_0 \left[\frac{\partial \ln \sigma_t(\theta)}{\partial \theta^\top} \right] E_0 \left[\frac{\partial \ln \sigma_t(\theta)}{\partial \theta} \frac{\partial \ln \sigma_t(\theta)}{\partial \theta^\top} \right]^{-1} s_\theta(r_t) \right) \\ &= \alpha(1 - \alpha) - \frac{(q_\alpha \phi(q_\alpha))^2}{2} E_0 \left[\frac{\partial \ln \sigma_t(\theta)}{\partial \theta^\top} \right] E_0 \left[\frac{\partial \ln \sigma_t(\theta)}{\partial \theta} \frac{\partial \ln \sigma_t(\theta)}{\partial \theta^\top} \right]^{-1} E_0 \left[\frac{\partial \ln \sigma_t(\theta)}{\partial \theta^\top} \right]^\top. \end{aligned}$$

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

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

Figures and Tables

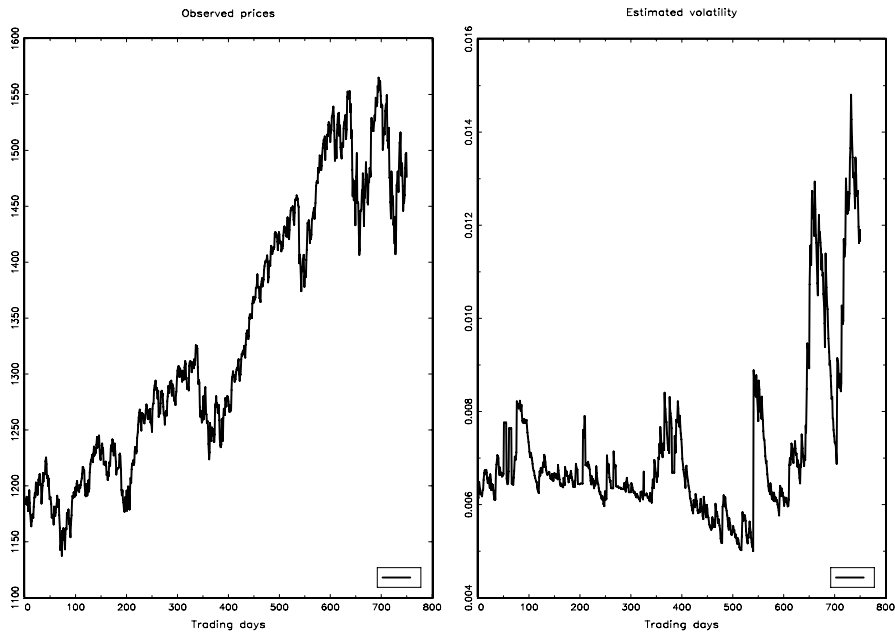


Figure 1: Prices and estimated volatility- S&P 500, 2005-2007

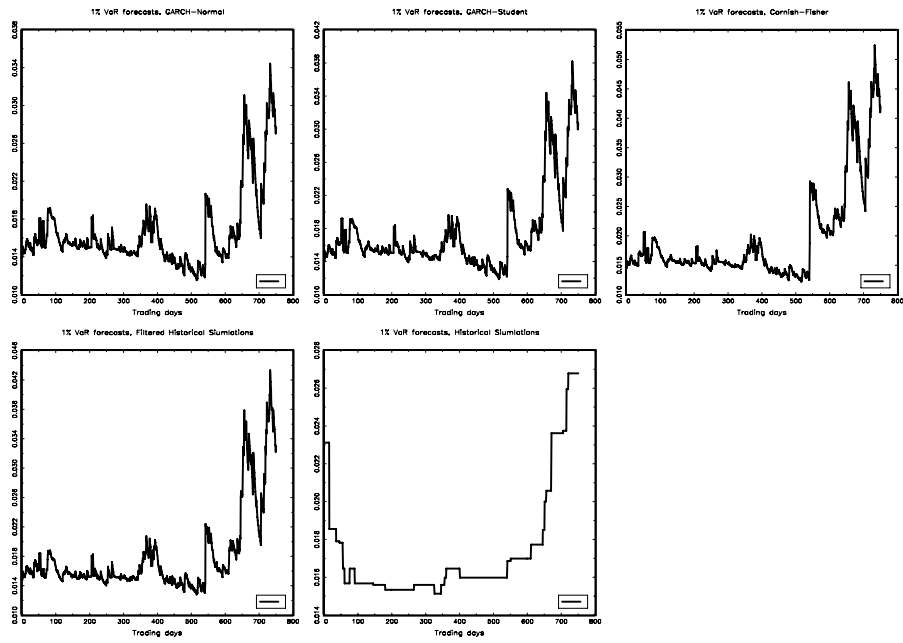


Figure 2: VaR Forecasts, S&P 500, 2005-2007

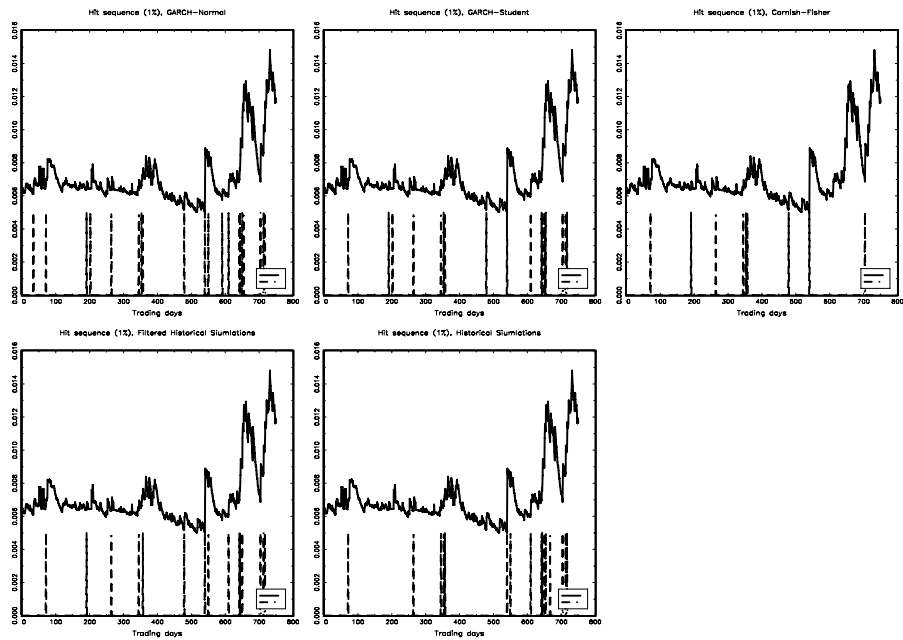


Figure 3: Hit sequences for 5 models, S&P 500, 2005-2007

Table 2

Size and Power of the tests for probit models

$h(x)$	Size	Size*	T(5)	T(10)	T(20)	Logis.	H-0.2	H-0.4	H-0.6	O-0.2	O 0.4	O 0.6
<i>n</i> = 500												
1	4.9	7.1	12.9	8.0	5.4	5.6	18.7	54.0	83.5	66.3	99.9	100.0
x	3.4	20.5	14.5	7.5	4.5	5.1	6.4	15.1	26.3	41.2	98.0	100.0
x^2	4.2	6.8	11.7	7.6	4.9	5.6	14.8	45.0	73.5	58.5	99.7	100.0
$\cos(x)$	5.0	7.1	13.1	8.2	5.6	5.3	19.9	56.8	84.9	66.8	99.9	100.0
$\cos(2x)$	5.4	6.1	7.2	5.5	5.2	5.1	17.6	47.8	73.8	52.8	81.7	99.5
Omitted	5.0	6.9	10.5	7.3	5.8	5.6	22.1	63.7	90.6	71.0	100.0	100.0
Hetero	4.0	13.8	18.5	9.0	5.6	5.2	7.8	19.8	39.1	38.5	95.9	100.0
Optilog	4.1	11.0	16.9	8.5	5.0	5.0	6.1	12.2	18.9	48.5	99.4	100.0
Opti-t5	3.8	10.5	15.6	7.8	4.8	5.0	5.9	11.7	18.2	47.5	99.3	100.0
Std	5.4	6.8	8.5	6.4	5.7	5.5	22.1	64.1	91.3	68.1	99.9	100.0
Gaussian	5.2	20.8	19.1	10.2	6.5	5.7	17.3	51.6	83.0	65.6	99.9	100.0
All	5.8	95.1	17.6	10.4	7.5	6.4	12.2	33.6	62.1	51.6	99.5	100.0
<i>n</i> = 1000												
1	4.3	5.9	16.0	10.0	6.4	5.4	33.7	84.4	99.1	95.4	100.0	100.0
x	2.7	16.0	22.2	10.1	6.4	5.2	8.3	22.3	41.6	61.9	99.9	100.0
x^2	3.7	5.7	16.1	9.0	6.1	5.4	27.1	75.8	97.1	91.8	100.0	100.0
$\cos(x)$	4.4	6.0	16.2	10.0	6.6	5.1	35.5	85.7	99.1	95.8	100.0	100.0
$\cos(2x)$	4.7	5.1	7.3	5.6	5.7	5.5	30.1	74.7	92.3	85.0	98.3	100.0
Omitted	4.5	6.1	11.6	8.0	6.1	5.3	39.6	89.6	99.7	97.4	100.0	100.0
Hetero	3.6	10.7	27.9	11.7	7.7	5.4	10.1	29.2	59.0	57.6	99.8	100.0
Optilog	3.7	9.2	26.6	11.2	7.3	5.4	7.8	16.3	28.9	71.3	100.0	100.0
Opti-t5	3.4	9.0	25.0	10.8	6.9	5.4	7.4	15.8	27.9	70.3	100.0	100.0
Std	4.7	5.9	8.1	6.6	5.7	5.2	39.3	89.6	99.8	97.6	100.0	100.0
Gaussian	4.2	15.0	26.8	12.4	8.1	5.5	30.2	81.8	98.8	95.6	100.0	100.0
All	4.8	96.4	23.7	12.1	8.5	6.4	17.7	57.3	91.5	85.1	100.0	100.0

Table 1

Ord's family and orthonormal polynomials.

Name	p_y	A	B	Q_1
Recursive relationship				
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 normalized polynomial of degree j .

Table 3

Size of the tests

	θ^0 known				θ^0 estimated by MLE					
	n	100	200	500	1000	n	100	200	500	1000
C_1	5.3	4.8	5.2	5.0						
C_2	5.1	4.7	5.2	5.1	C_2	4.7	4.8	5.0	5.1	
C_3	4.6	4.8	5.0	5.2	C_3	4.1	4.6	4.9	5.1	
C_4	2.7	3.4	3.7	4.3	C_4	2.5	3.1	3.6	4.2	
C_{2-3}	5.3	5.4	5.2	5.3	C_{2-3}	4.8	5.0	5.1	5.2	
C_{2-4}	5.1	5.2	5.5	5.4	C_{2-4}	4.5	4.9	5.3	5.4	
CR^1	5.1	5.3	5.2	5.1	CR^1	4.8	4.9	5.0	4.9	
CR^0	3.0	4.6	5.3	5.1	CR^0	3.1	4.5	5.2	5.0	

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

Table 4

Power of the tests

Binomial distribution $\mathcal{B}(k, \frac{2}{k})$														
k=10					k=15					k=20				
<i>n</i>	100	200	500	1000	<i>n</i>	100	200	500	1000	<i>n</i>	100	200	500	1000
C_2	26.2	55.2	94.2	100	C_2	12.1	24.6	59.1	89.2	C_2	8.0	14.9	35.2	63.9
C_3	0.9	1.2	2.2	4.3	C_3	1.4	1.6	1.7	2.1	C_3	2.0	2.0	2.1	2.1
C_4	0.6	0.8	1.4	2.6	C_4	0.8	0.9	1.1	1.5	C_4	1.2	1.3	1.4	1.6
C_{2-3}	13.7	35.6	85.8	99.7	C_{2-3}	6.2	13.2	41.5	78.9	C_{2-3}	4.5	7.9	21.9	48.6
C_{2-4}	8.6	26.0	77.5	99.1	C_{2-4}	3.9	8.9	31.3	69.3	C_{2-4}	3.1	5.5	15.7	37.9
CR^1	10.3	24.2	70.6	97.7	CR^1	2.6	7.3	27.8	61.7	CR^1	2.8	6.1	15.7	33.6
CR^0	3.4	18.1	73.9	98.3	CR^0	2.7	9.5	35.2	67.1	CR^0	2.5	7.0	20.1	38.4
Pascal distribution $\mathcal{Pa}(2, \delta)$														
$\delta=10$					$\delta=15$					$\delta=20$				
<i>n</i>	100	200	500	1000	<i>n</i>	100	200	500	1000	<i>n</i>	100	200	500	1000
C_2	28.7	47.2	82.2	97.9	C_2	17.1	26.7	52.5	80.0	C_2	12.3	17.9	34.7	58.1
C_3	12.8	15.5	19.2	22.8	C_3	9.4	10.9	13.1	14.4	C_3	8.0	8.9	10.4	11.3
C_4	8.9	12.5	18.1	23.4	C_4	6.2	8.3	11.5	14.2	C_4	5.1	6.6	9.1	10.8
C_{2-3}	27.1	43.4	77.0	96.4	C_{2-3}	16.8	24.9	47.1	73.6	C_{2-3}	12.7	17.4	30.8	51.3
C_{2-4}	24.6	39.6	72.9	94.9	C_{2-4}	15.3	22.6	42.9	69.0	C_{2-4}	11.8	16.0	28.2	46.8
CR^1	16.8	27.0	56.7	87.6	CR^1	9.9	14.9	29.6	54.3	CR^1	7.6	10.8	18.6	33.1
CR^0	11.5	22.0	52.0	85.6	CR^0	7.0	11.9	25.8	50.7	CR^0	5.7	8.7	16.0	30.1

Table 5

Backtesting VaR model. Size properties with a GARCH(1,1)-Normal process

θ assumed known										
	$\alpha = 0.01$					$\alpha = 0.05$				
T	250	500	750	1000	T	250	500	750	1000	
C_{uc}	10.2	7.6	4.2	6.5	C_{uc}	7.1	5.6	6.2	4.5	
C_c	4.1	4.8	8.9	8.5	C_c	3.6	6.9	10.9	12.2	
ξ_{uc}	5.7	5.3	7.3	4.8	ξ_{uc}	5.1	4.8	5.4	4.7	
$\cos(r_{t-1})$	5.8	7.9	13.1	14.5	$\cos(r_{t-1})$	8.6	11.5	14.7	19.8	
$\cos(r_{t-2})$	5.3	5.6	5.9	6.6	$\cos(r_{t-2})$	5.7	6.4	6.6	6.8	
$\cos(r_{t-3})$	4.7	5.6	5.7	5.5	$\cos(r_{t-3})$	4.2	4.1	4.1	5.5	
$\cos(r_{t-4})$	6.2	6.3	6.7	6.4	$\cos(r_{t-4})$	5.7	5.9	6.3	5.7	
$\cos(2r_{t-1})$	7.4	6.3	6.7	6.7	$\cos(2r_{t-1})$	5.1	4.5	4.7	5.7	
$\cos(2r_{t-2})$	5.8	4.5	5.6	5.2	$\cos(2r_{t-2})$	4.0	4.3	4.3	5.1	
$\cos(2r_{t-3})$	5.0	5.7	5.9	5.8	$\cos(2r_{t-3})$	6.5	5.9	6.0	4.8	
$\cos(2r_{t-4})$	5.2	5.4	4.9	6.0	$\cos(2r_{t-4})$	5.0	5.1	5.9	5.6	
Markov	1.8	4.0	6.2	6.7	Markov	5.7	7.0	8.9	9.5	

θ estimated with the last 500 obs.										
	$\alpha = 0.01$					$\alpha = 0.05$				
T	250	500	750	1000	T	250	500	750	1000	
C_{uc}	10.1	7.8	4.7	6.4	C_{uc}	7.2	4.8	4.5	4.1	
C_c	4.0	4.8	8.0	8.4	C_c	4.3	7.7	12.6	15.3	
ξ_{uc}	7.0	5.9	8.9	5.6	ξ_{uc}	5.4	5.5	3.8	4.3	
$\cos(r_{t-1})$	6.4	7.8	11.4	13.5	$\cos(r_{t-1})$	7.4	9.2	12.8	15.4	
$\cos(r_{t-2})$	6.7	6.7	6.1	6.8	$\cos(r_{t-2})$	6.5	6.7	6.2	5.9	
$\cos(r_{t-3})$	5.9	6.6	6.2	7.3	$\cos(r_{t-3})$	4.7	5.0	4.9	4.2	
$\cos(r_{t-4})$	6.6	7.2	7.5	7.7	$\cos(r_{t-4})$	6.2	6.5	6.4	5.6	
$\cos(2r_{t-1})$	8.0	7.8	8.9	8.3	$\cos(2r_{t-1})$	5.7	5.6	6.4	6.1	
$\cos(2r_{t-2})$	6.0	4.9	4.5	4.4	$\cos(2r_{t-2})$	4.3	3.9	4.2	4.5	
$\cos(2r_{t-3})$	5.3	6.4	5.9	7.1	$\cos(2r_{t-3})$	6.3	6.7	5.8	4.3	
$\cos(2r_{t-4})$	5.6	5.9	5.9	6.2	$\cos(2r_{t-4})$	3.7	4.8	5.6	5.6	
Markov	2.5	4.4	6.4	8.8	Markov	6.3	6.6	9.0	8.8	

Table 6

Backtesting VaR model. Power properties with a GARCH(1,1) process with Student (ν) innovations

$\nu = 8$										
	$\alpha = 0.01$					$\alpha = 0.05$				
T	250	500	750	1000	T	250	500	750	1000	
C_{uc}	10.2	23.5	25.0	35.6	C_{uc}	6.6	4.6	3.6	2.8	
C_c	5.6	5.7	7.9	10.0	C_c	5.0	8.7	11.2	14.4	
ξ_{uc}	18.4	23.2	35.6	35.5	ξ_{uc}	4.5	3.9	2.3	2.6	
$\cos(r_{t-1})$	8.1	8.9	10.1	9.4	$\cos(r_{t-1})$	6.4	7.4	10.9	13.6	
$\cos(r_{t-2})$	14.7	16.5	21.3	24.0	$\cos(r_{t-2})$	4.8	4.8	4.8	4.5	
$\cos(r_{t-3})$	14.7	16.3	21.8	22.5	$\cos(r_{t-3})$	5.7	3.9	4.9	4.0	
$\cos(r_{t-4})$	12.4	15.9	18.6	21.7	$\cos(r_{t-4})$	4.0	4.5	3.3	4.0	
$\cos(2r_{t-1})$	10.1	9.7	9.9	10.8	$\cos(2r_{t-1})$	4.2	5.0	5.5	4.6	
$\cos(2r_{t-2})$	9.9	10.0	8.2	9.1	$\cos(2r_{t-2})$	4.6	3.7	3.4	3.4	
$\cos(2r_{t-3})$	10.4	11.1	9.6	9.3	$\cos(2r_{t-3})$	3.6	4.6	4.1	3.6	
$\cos(2r_{t-4})$	9.0	9.8	8.0	9.4	$\cos(2r_{t-4})$	3.4	3.9	4.4	3.9	
Markov	6.5	13.6	21.6	28.2	Markov	6.1	6.7	8.2	8.9	

$\nu = 5$										
	$\alpha = 0.01$					$\alpha = 0.05$				
T	250	500	750	1000	T	250	500	750	1000	
C_{uc}	13.6	32.2	37.2	50.5	C_{uc}	8.8	8.4	8.3	9.6	
C_c	5.3	5.8	9.2	11.1	C_c	4.2	7.0	10.2	12.4	
ξ_{uc}	23.9	32.2	50.2	50.5	ξ_{uc}	5.2	5.8	5.1	7.9	
$\cos(r_{t-1})$	11.8	12.9	15.1	15.1	$\cos(r_{t-1})$	7.6	9.8	13.2	17.1	
$\cos(r_{t-2})$	16.9	24.2	31.2	36.2	$\cos(r_{t-2})$	5.4	5.6	5.5	5.0	
$\cos(r_{t-3})$	19.2	24.1	29.0	34.0	$\cos(r_{t-3})$	5.5	6.0	4.6	5.2	
$\cos(r_{t-4})$	16.5	24.1	30.0	34.8	$\cos(r_{t-4})$	4.7	4.7	5.4	5.4	
$\cos(2r_{t-1})$	11.5	11.5	10.1	10.6	$\cos(2r_{t-1})$	3.7	3.1	2.9	3.4	
$\cos(2r_{t-2})$	9.9	10.1	9.5	10.9	$\cos(2r_{t-2})$	2.4	3.7	3.5	3.2	
$\cos(2r_{t-3})$	9.8	10.8	9.8	11.7	$\cos(2r_{t-3})$	2.9	3.1	3.4	2.9	
$\cos(2r_{t-4})$	9.9	9.2	9.3	9.5	$\cos(2r_{t-4})$	3.3	3.0	2.6	3.0	
Markov	8.2	19.7	31.0	40.9	Markov	7.5	8.4	11.1	12.9	

Table 7

Backtesting VaR model. Power properties with Historical Simulation

	$\alpha = 0.01$					$\alpha = 0.05$			
T	250	500	750	1000	T	250	500	750	1000
C_{uc}	17.4	8.0	2.7	2.8	C_{uc}	15.0	9.4	4.8	1.4
C_c	5.6	8.9	12.6	14.6	C_c	5.8	13.1	17.3	23.5
ξ_{uc}	8.7	5.6	5.2	2.3	ξ_{uc}	12.1	9.7	3.9	1.6
$\cos(r_{t-1})$	9.1	11.4	18.8	22.9	$\cos(r_{t-1})$	12.7	21.1	27.6	33.4
$\cos(r_{t-2})$	7.9	12.5	15.6	18.9	$\cos(r_{t-2})$	10.9	17.6	20.8	27.7
$\cos(r_{t-3})$	6.5	10.4	15.2	15.3	$\cos(r_{t-3})$	10.6	14.2	19.4	22.3
$\cos(r_{t-4})$	6.3	7.2	10.0	12.4	$\cos(r_{t-4})$	10.3	11.8	15.0	18.3
$\cos(2r_{t-1})$	6.4	5.4	6.1	5.3	$\cos(2r_{t-1})$	5.1	5.5	5.2	5.4
$\cos(2r_{t-2})$	5.8	6.2	5.7	6.0	$\cos(2r_{t-2})$	4.4	4.9	5.2	5.2
$\cos(2r_{t-3})$	6.9	7.4	7.7	6.8	$\cos(2r_{t-3})$	6.3	6.2	5.5	4.9
$\cos(2r_{t-4})$	6.4	6.3	6.2	5.7	$\cos(2r_{t-4})$	4.8	5.5	5.1	5.4
Markov	4.8	6.6	8.1	11.2	Markov	15.0	13.1	13.4	16.3

Table 8

Index of Hit sequences for the 1% VaR Forecasts, S&P 500

G-N	34	71	191	202	264	345	353	357	479	540	550
	591	610	642	644	650	654	-	704	713	717	
G-St	-	71	191	202	264	345	353	357	479	540	-
	-	610	642	644	650	654	-	704	713	717	
CF	-	71	191	-	264	345	353	357	479	540	-
	-	-	-	-	-	-	-	704	-	-	
FHS	-	71	191	-	264	345	0	357	479	540	550
	-	610	642	644	650	-	-	704	713	717	
HS	-	71	-	-	264	345	353	357	-	540	550
	-	610	642	644	650	654	667	704	713	717	

Table 9

Backtesting 1% Var models, S&P 500

GARCH(1,1)-Normal				GARCH(1,1)-Student			
T	250	500	750	T	250	500	750
C_{uc}	0.77 (0.38)	2.61 (0.11)	14.44 (0.00)	C_{uc}	0.09 (0.76)	1.54 (0.21)	8.94 (0.00)
C_c	0.13 (0.72)	0.33 (0.57)	1.10 (0.29)	C_c	0.07 (0.79)	0.26 (0.61)	0.79 (0.37)
ξ_{uc}	0.91 (0.34)	3.23 (0.07)	21.04 (0.00)	ξ_{uc}	0.10 (0.75)	1.82 (0.18)	12.15 (0.00)
$\cos(r_{t-1})$	0.91 (0.34)	3.23 (0.07)	21.04 (0.00)	$\cos(r_{t-1})$	0.10 (0.75)	1.82 (0.18)	12.15 (0.00)
$\cos(r_{t-2})$	0.91 (0.34)	3.23 (0.07)	21.04 (0.00)	$\cos(r_{t-2})$	0.10 (0.75)	1.82 (0.18)	12.15 (0.00)
$\cos(r_{t-3})$	0.91 (0.34)	3.23 (0.07)	21.04 (0.00)	$\cos(r_{t-3})$	0.10 (0.75)	1.82 (0.18)	12.15 (0.00)
$\cos(r_{t-4})$	0.91 (0.34)	3.23 (0.07)	21.04 (0.00)	$\cos(r_{t-4})$	0.10 (0.75)	1.82 (0.18)	12.15 (0.00)
Markov	0.90 (0.64)	2.94 (0.23)	15.54 (0.00)	Markov	0.17 (0.92)	1.80 (0.41)	9.73 (0.01)
Filtered HS				Historical Simulation			
T	250	500	750	T	250	500	750
C_{uc}	0.11 (0.74)	0.19 (0.66)	5.87 (0.02)	C_{uc}	1.18 (0.28)	0.00 (1.00)	7.34 (0.01)
C_c	0.03 (0.86)	0.15 (0.70)	0.61 (0.43)	C_c	0.01 (0.93)	0.10 (0.75)	0.70 (0.40)
ξ_{uc}	0.10 (0.75)	0.20 (0.65)	7.58 (0.01)	ξ_{uc}	0.91 (0.34)	0.00 (1.00)	9.73 (0.00)
$\cos(r_{t-1})$	0.10 (0.75)	0.20 (0.65)	7.58 (0.01)	$\cos(r_{t-1})$	0.91 (0.34)	0.00 (1.00)	9.73 (0.00)
$\cos(r_{t-2})$	0.10 (0.75)	0.20 (0.65)	7.58 (0.01)	$\cos(r_{t-2})$	0.91 (0.34)	0.00 (1.00)	9.73 (0.00)
$\cos(r_{t-3})$	0.10 (0.75)	0.20 (0.65)	7.58 (0.01)	$\cos(r_{t-3})$	0.91 (0.34)	0.00 (1.00)	9.73 (0.00)
$\cos(r_{t-4})$	0.10 (0.75)	0.20 (0.65)	7.57 (0.01)	$\cos(r_{t-4})$	0.91 (0.34)	0.00 (1.00)	9.73 (0.00)
Markov	0.14 (0.93)	0.34 (0.85)	6.48 (0.04)	Markov	1.18 (0.55)	0.10 (0.95)	8.04 (0.02)
Cornish-Fisher Approximation							
T	250	500	750				
C_{uc}	0.11 (0.74)	0.72 (0.40)	0.28 (0.59)				
C_c	0.03 (0.86)	0.20 (0.66)	0.22 (0.64)				
ξ_{uc}	0.10 (0.75)	0.81 (0.37)	0.30 (0.58)				
$\cos(r_{t-1})$	0.10 (0.75)	0.81 (0.37)	0.30 (0.58)				
$\cos(r_{t-2})$	0.10 (0.75)	0.81 (0.37)	0.30 (0.58)				
$\cos(r_{t-3})$	0.10 (0.75)	0.81 (0.37)	0.30 (0.58)				
$\cos(r_{t-4})$	0.10 (0.75)	0.81 (0.37)	0.30 (0.58)				
Markov	0.14 (0.93)	0.92 (0.63)	0.50 (0.78)				

Table 10

Backtesting 5% Var models, S&P 500

GARCH(1,1)-Normal				GARCH(1,1)-Student			
T	250	500	750	T	250	500	750
C_{uc}	1.14 (0.29)	0.71 (0.40)	0.33 (0.56)	C_{uc}	1.14 (0.29)	0.71 (0.40)	0.55 (0.46)
C_c	0.60 (0.44)	0.03 (0.86)	0.02 (0.89)	C_c	0.60 (0.44)	0.03 (0.86)	0.05 (0.83)
ξ_{uc}	1.03 (0.31)	0.67 (0.41)	0.34 (0.56)	ξ_{uc}	1.03 (0.31)	0.67 (0.41)	0.57 (0.45)
$\cos(r_{t-1})$	1.03 (0.31)	0.67 (0.41)	0.34 (0.56)	$\cos(r_{t-1})$	1.03 (0.31)	0.67 (0.41)	0.57 (0.45)
$\cos(r_{t-2})$	1.03 (0.31)	0.67 (0.41)	0.34 (0.56)	$\cos(r_{t-2})$	1.03 (0.31)	0.67 (0.41)	0.57 (0.45)
$\cos(r_{t-3})$	1.03 (0.31)	0.67 (0.41)	0.34 (0.56)	$\cos(r_{t-3})$	1.03 (0.31)	0.67 (0.41)	0.57 (0.45)
$\cos(r_{t-4})$	1.03 (0.31)	0.67 (0.41)	0.34 (0.56)	$\cos(r_{t-4})$	1.03 (0.31)	0.67 (0.41)	0.57 (0.45)
Markov	1.74 (0.42)	0.74 (0.69)	0.35 (0.84)	Markov	1.74 (0.42)	0.74 (0.69)	0.59 (0.74)
Filtered HS				Historical Simulation			
T	250	500	750	T	250	500	750
C_{uc}	1.14 (0.29)	1.65 (0.20)	0.06 (0.80)	C_{uc}	4.37 (0.04)	3.02 (0.08)	2.85 (0.09)
C_c	0.60 (0.44)	0.14 (0.71)	0.00 (0.99)	C_c	0.30 (0.59)	5.79 (0.02)	1.18 (0.28)
ξ_{uc}	1.03 (0.31)	1.52 (0.22)	0.06 (0.80)	ξ_{uc}	3.56 (0.06)	2.69 (0.10)	3.09 (0.08)
$\cos(r_{t-1})$	1.03 (0.31)	1.52 (0.22)	0.06 (0.80)	$\cos(r_{t-1})$	3.56 (0.06)	2.69 (0.10)	3.09 (0.08)
$\cos(r_{t-2})$	1.03 (0.31)	1.52 (0.22)	0.06 (0.80)	$\cos(r_{t-2})$	3.56 (0.06)	2.69 (0.10)	3.09 (0.08)
$\cos(r_{t-3})$	1.03 (0.31)	1.52 (0.22)	0.06 (0.80)	$\cos(r_{t-3})$	3.56 (0.06)	2.69 (0.10)	3.09 (0.08)
$\cos(r_{t-4})$	1.03 (0.31)	1.52 (0.22)	0.06 (0.80)	$\cos(r_{t-4})$	3.56 (0.06)	2.69 (0.10)	3.09 (0.08)
Markov	1.74 (0.42)	1.78 (0.41)	0.06 (0.97)	Markov	4.66 (0.10)	8.81 (0.01)	4.04 (0.13)
Cornish-Fisher Approximation							
T	250	500	750				
C_{uc}	1.14 (0.29)	1.65 (0.20)	0.06 (0.80)				
C_c	0.60 (0.44)	0.14 (0.71)	0.00 (0.99)				
ξ_{uc}	1.03 (0.31)	1.52 (0.22)	0.06 (0.80)				
$\cos(r_{t-1})$	1.03 (0.31)	1.52 (0.22)	0.06 (0.80)				
$\cos(r_{t-2})$	1.03 (0.31)	1.52 (0.22)	0.06 (0.80)				
$\cos(r_{t-3})$	1.03 (0.31)	1.52 (0.22)	0.06 (0.80)				
$\cos(r_{t-4})$	1.03 (0.31)	1.52 (0.22)	0.06 (0.80)				
Markov	1.74 (0.42)	1.78 (0.41)	0.06 (0.97)				