Quaderni di Statistica
Vol. ?, 200?

Detecting features of different financial durations through the Pareto distribution

Giovanni De Luca
Istituto di Statistica e Matematica
Università di Napoli Parthenope
E-mail: giovanni.deluca@uniparthenope.it

Paola Zuccolotto
Dipartimento Metodi Quantitativi
Università degli Studi di Brescia
E-mail: zuk@eco.unibs.it

Summary: This paper deals with Autoregressive Conditional Duration (ACD) models for ultra-high frequency financial data with attention to the analysis of durations between different market events. On a real dataset of Italian financial market these different types of durations are empirically investigated in order to highlight both their mutual relations and the informative power of a particular distributional assumption. The Pareto distribution turns out to be a good tool for detecting features of different financial durations.

Keywords: ACD models, Duration, Mixture of distributions, Hazard function.

1. Introduction

In the last decades a wide literature on financial time series analysis has produced a large variety of models. One of the most recent development is related to the availability of *tick-by-tick* data (or *ultra-high-frequency* data), collected recording every single transaction in the financial market. The main feature of ultra-high-frequency data is to be

unequally spaced in time, so that the time between consecutive records (*duration*) is itself a random variable which has to be properly modelled.

In this framework the class of Autoregressive Conditional Duration models (ACD), introduced by Engle and Russell (1998) and detailed by Engle (2000) is, at present, one of the most powerful tools, as proved by several applications to real data. Many extensions to the basic formulation have already been proposed, for example the Log-ACD (Bauwens and Giot, 2000), the Asymmetric ACD (Bauwens and Giot, 2003), the FIACD (Jasiak, 1998), a nonlinear ACD (Zhang *et al.*, 2001), the d-ACD (Zuccolotto, 2002) and the d-ACD$_\lambda$ (Zuccolotto, 2003a). The estimation of quantiles in the context of ACD models is analyzed in Zuccolotto (2003b).

Although financial durations are generally defined as the interval time between two consecutive market events, a never fully investigated topic is the suitability of the ACD framework to model durations between any two events in the financial market: a market event can be a transaction, as well as a price change, or a price variation over, for example, 0.1% or 0.2%.

Aim of this paper is to empirically investigate the behavior of financial durations between different market events, the informative contents of the estimated models, how they are related each other. Since the traditional exponential hypothesis is very robust for parameters estimation, but completely inadequate when the distributional fit is crucial (for example with quantiles estimation or hazard estimation), the analysis is carried out using a particular distributional assumption for the residuals, the Pareto II Type distribution, proposed by De Luca and Zuccolotto (2003), where the performance of a Pareto ACD model is shown to favorably compare with a traditional one with exponential errors. Relying on the results already obtained in that framework, in this paper an analogous empirical analysis is carried out in the quite different context. The better distributional assumption should hopefully be able to detect different stochastic structures characterizing different financial durations, which is the main goal of this empirical analysis.

The paper is organized as follows. In section 2 the theoretical framework of the Pareto II Type distribution is recalled, section 3 illustrates

the main diagnostics used to verify the residuals' distributional fit, an empirical analysis of real data from Italian financial market is in section 4, concluding remarks follow in section 5.

2. The Pareto ACD Model

Let t_i be the time at which the i-th event has occurred, the duration between the $(i-1)$-th and the i-th event is defined as $X_i = t_i - t_{i-1}$. Let x_i be the corresponding duration after removing the daily seasonal component[1] $\phi(t_i)$:

$$x_i = \frac{X_i}{\phi(t_i)}. \tag{1}$$

The class of ACD models assumes an autoregressive dynamic in durations, formalized in the basic ACD(q, p) model as

$$x_i/\Psi_i = \epsilon_i \qquad \epsilon_i \sim i.i.d. \text{ with } E(\epsilon_i) = 1$$

$$\Psi_i = \omega + \sum_{j=1}^{q} \alpha_j x_{i-j} + \sum_{j=1}^{p} \beta_j \psi_{i-j}$$

where Ψ_i is the expected duration at time t_i, conditionally on past information. The most popular assumptions for the probability law of the residuals ϵ_i are the exponential and the Weibull distribution, but these have often revealed inadequate and alternative proposals are present in the literature (see, for instance, Bauwens *et al.* (2004) or Bauwens and Giot (2001)).

Relying on the idea that heterogeneity of traders (different degree of information, attitudes toward risk, budget constraints, and so on) can affect hazard models, De Luca and Zuccolotto (2003) suggest the use of finite or infinite mixtures of exponential distributions, which are shown to substantially improve the distributional fit. The idea is supported by empirical evidence also in De Luca and Gallo (2004).

[1] It can be estimated using a cubic spline as in Engle (2000).

Under the finite mixture assumpion the density function of ϵ_i is

$$f_{\mathcal{E}}(\epsilon_i; \lambda_1, \cdots, \lambda_K, p_1, \cdots, p_K) = \sum_{k=1}^{K} p_k \frac{1}{\lambda_k} \exp\left\{-\frac{\epsilon_i}{\lambda_k}\right\}$$

where the mixing proportions $[p_1, \cdots, p_K]$ are constrained by $p_k > 0 \; \forall k$ and $\sum_{k=1}^{K} p_k = 1$. A further constraint, due to ensure unit mean to the residuals, regards the exponential parameters $[\lambda_1, \cdots, \lambda_K]$:

$$\sum_{k=1}^{K} p_k \lambda_k = 1.$$

For example in the simple case of two exponential distributions with proportions p_1 and p_2 the free parameters are only λ_1 and p_1, as λ_2 and p_2 are given by

$$p_2 = 1 - p_1,$$

$$\lambda_2 = \frac{1}{p_2}[1 - p_1\lambda_1].$$

With $K \to \infty$ this assumption tends to an infinite mixture of exponential distributions, which can be obtained defining a proper probability law for the parameter of the exponential distribution λ (mixing distribution), for example an Inverse Gamma, formalized as follows:

$$f(\epsilon_i|\lambda) = \frac{1}{\lambda} \exp\left\{-\frac{\epsilon_i}{\lambda}\right\}$$

with

$$f(\lambda) = \frac{\theta^\delta}{\Gamma(\delta)} \exp\left\{-\frac{\theta}{\lambda}\right\} \left(\frac{1}{\lambda}\right)^{\delta+1}$$

$\lambda > 0, \theta, \delta > 0$.

The marginal distribution of ϵ_i is easily derived integrating out λ in the joint distribution $f(\epsilon_i, \lambda) = f(\epsilon_i|\lambda)f(\lambda)$:

$$f_{\mathcal{E}}(\epsilon_i; \theta, \delta) = \int_0^\infty f(\epsilon_i|\lambda)f(\lambda)d\lambda = \int_0^\infty \frac{\theta^\delta}{\Gamma(\delta)} \exp\left\{-\frac{\theta + \epsilon_i}{\lambda}\right\} \left(\frac{1}{\lambda}\right)^{\delta+2} d\lambda$$

$$f_{\mathcal{E}}(\epsilon_i;\theta,\delta) = \frac{\theta^{\delta}}{\Gamma(\delta)}\frac{\Gamma(\delta+1)}{(\theta+\epsilon_i)^{(\delta+1)}}$$

$$f_{\mathcal{E}}(\epsilon_i;\theta,\delta) = \theta^{\delta}\delta(\theta+\epsilon_i)^{-(\delta+1)}.$$

The variable ϵ_i turns out to be a translated Pareto random variable, also known as Pareto II type, with

$$E(\epsilon_i) = \frac{\theta}{\delta-1} \quad \text{and} \quad \text{Var}(\epsilon_i) = \frac{\theta^2\delta}{(\delta-1)^2(\delta-2)}.$$

The unit mean restriction involves a constraint on one of the two parameters of the distribution, e.g.

$$\delta = \theta + 1$$

so that the density function turns out to depend only on θ and can be written as

$$f_{\mathcal{E}}(\epsilon_i;\theta) = \theta^{(\theta+1)}(\theta+1)(\theta+\epsilon_i)^{-(\theta+2)}$$

while the conditional duration $x_i|\Psi_i$ has distribution

$$f_X(x_i|\Psi_i;\omega,\alpha_1,\cdots,\alpha_q,\beta_1,\cdots,\beta_p,\theta) = \frac{\theta_1^{(\theta_1+1)}}{\Psi_i}(\theta_1+1)\left(\theta_1+\frac{x_i}{\Psi_i}\right)^{-(\theta_1+2)}.$$

Although finite mixtures performance turns out to be as good as that of infinite mixture (and sometimes even better), it seems reasonable to prefer the latter, since it is more parsimonious. The parameters of the ACD model and θ can be jointly estimated by maximizing the log-likelihood function

$$l(\omega,\alpha_1,\cdots,\alpha_q,\beta_1,\cdots,\beta_p,\theta) =$$

$$\sum_{i=1}^{n}\left[(\theta+1)\log\theta - \log\Psi_i + \log(\theta+1) - (\theta+2)\log\left(\theta+\frac{x_i}{\Psi_i}\right)\right].$$

This theoretical framework is quite general and can be used to model different types of durations, obtained by taking into account different market events. As already pointed out in the Introduction, a market event can

be a transaction, as well as a price change, or a price variation over, for example, 0.1% or 0.2%. It is reasonable to argue that different market events analyzed on the same stock can lead to durations somewhat related each other. In section 4 an empirical analysis is carried out in order to investigate these relationships through a Pareto distribution.

3. Diagnostics for residuals' distributional fit

The goodness of fit of ACD models is traditionally checked using the Ljung-Box test statistic, in order to verify the hypothesis of non-autocorrelated residuals. Since lack of autocorrelation does not necessarily imply independence, the Ljung-Box statistic is sometimes computed also over higher order moments, but experience shows that problems hardly ever occur at this step.

A second check is usually done on the residuals' mean, which should not significantly differ from unity, and also in this case experience shows that the hypothesis of unit mean is never rejected. These successes are somewhat independent from the distributional assumption on residuals and can be obtained using the simple exponential distribution, thanks to its well-known robustness in ACD parameters estimates.

But when the model is used for hazard estimation or quantiles prediction, the crucial matter, sometimes neglected, is the accuracy of the distributional hypothesis for the residuals. This can be checked in different ways.

A first rough check regards the variance. Relying on a limiting law of the distribution of the unbiased sample variance s^2, we have

$$s^2 \xrightarrow{d} N\left\{\sigma^2; \frac{\sigma^4}{n-1}\left(2+\frac{n-1}{n}\gamma_2\right)\right\}$$

where γ_2 denotes the population coefficient of kurtosis, $\frac{\mu_4}{\sigma^4} - 3$. Consequently a test on the residuals' variance can be carried out in order to verify if the sample variance is consistent with the distributional assumption according to the estimated parameters: if the exponential hypothesis is satisfied, s^2 should not significantly differ from unity, but it occurs

rather rarely. This is a first warning highlighting the possible necessity of a different assumption for the density of ϵ_i.

More accurate procedures are the Kolmogorov-Smirnov (KS) and the Cramer-Von Mises (CVM) statistics, given respectively by

$$D_n = sup_\epsilon |F_n(\epsilon) - F(\epsilon)|$$

and

$$W^2 = n \int_{-\infty}^{\infty} [F_n(\epsilon) - F(\epsilon)]^2 dF(\epsilon)$$

where $F(\epsilon)$ is the assumed cumulative distribution function and $F_n(\epsilon)$ is the empirical cumulative distribution function estimated with a sample of size n. As shown in D'Agostino and Stephens (1986), the value of W^2 can alternatively be calculated using the simpler formulation

$$W_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left[F(\epsilon_{(i)}) - \frac{2i-1}{2n} \right]^2,$$

where $F(\epsilon_{(i)})$ is the cumulative distribution function of the i-th ordered sample observation. The critical values of D_n and W_n^2 are tabulated and can be used to test the similarity between the empirical and the assumed cumulative distribution functions.

The adherence to the distributional hypothesis can be checked also with the well-known graphical tool QQ-plot, where theoretical quantiles are plotted against the corresponding empirical quantiles. If the distributional assumption is supported by empirical outcomes, the points should lie over the bisector straight line. Departures from this expected path constitute evidence of a bad hypothesis over the density function.

Among the most popular techniques used in this context, the last (but not least) is the density forecast, formalized by Diebold *et al.* (1998) and introduced in the framework of ACD models by Bauwens *et al.* (2004). It is well-known that if $F_X(x)$ is a cdf, the variable $z = F_X(x)$ has uniform distribution, regardless the function F: this property is the basis of the whole theoretical support of density forecasts. In the context of ACD models, a density forecast is the density defined for the next duration x_i given the information available up to time t_{i-1}. Let $\{\hat{f}_X(x_i|\mathcal{I}_{i-1};\hat{\theta})\}_{i=1}^{s}$

be a sequence of one-step-ahead density forecasts (in sample or out of sample) produced by an ACD model under a given distributional assumption, then the sequence of probability integral transforms $\{z_i\}_{i=1}^s$ of $\{x_i\}_{i=1}^s$ with respect to $\{\hat{f}_X(x_i|\mathcal{I}_{i-1};\hat{\theta})\}_{i=1}^s$ is *i.i.d.* uniform if the density forecasts $\hat{f}_X$ coincide with the real densities f_X, that is

$$z_i = \int_{-\infty}^{x_i} \hat{f}_X(x_i|\mathcal{I}_{i-1};\hat{\theta})dx_i \sim i.i.d. \quad U(0,1)$$

if

$$\hat{f}_X(x_i|\mathcal{I}_{i-1};\hat{\theta}) \equiv f_X(x_i|\mathcal{I}_{i-1};\theta).$$

Hence the empirical sequences $\{z_i\}_{i=1}^s$ produced by the different distributional assumptions can be tested for *i.i.d.* uniformity, using a graphical tool (where deviates from uniformity can be easily detected) or more formal goodness-of-fit test. Diebold et al. (1998), in their seminal paper on density forecasts, prefer the graphical analysis, performed using histograms with 95% bin-by-bin confidence interval, because it is informative on the nature of the violation of uniformity.

In the empirical analysis of next section all the techniques above described will be used in order to check the adequacy of the Pareto ACD model against the traditional hypotesis of exponential distribution.

4. The analysis

4.1 Data and estimation

The empirical analysis was carried out on durations between different market events for Comit, a stock largely traded in the Italian financial market, observed in February 2000 (21 trading days). We considered transaction durations, that is the durations between two consecutive transactions independently on a possible price change, and price durations defined as the times needed to observe a cumulative price change over a fixed threshold c or a percentage threshold.

We analyzed four different durations:

A. between transactions;

B. between any price changes ($c = 0$);

C. between price changes over 0.21%;

D. between price changes over 0.22%.

The sample sizes are respectively 16031 (A), 8361 (B), 8157 (C) and 7455 (D). The two percentage thresholds were chosen in order to get sufficiently large but different datasets. Actually the size difference is significant. A percentage threshold greater than 0.22% dramatically reduces the dataset and was not considered.

Durations were analyzed under two different distributional assumptions: the simple exponential and an infinite mixtures of exponentials with an Inverse Gamma as mixing distribution, that is a Pareto II type distribution, introduced in De Luca and Zuccolotto (2003).

Before estimating the ACD models, deterministic daily seasonal components were estimated and removed from the observed durations according to (1), using cubic splines with nodes set at each hour. The estimated daily seasonal components are shown in Figure 1.

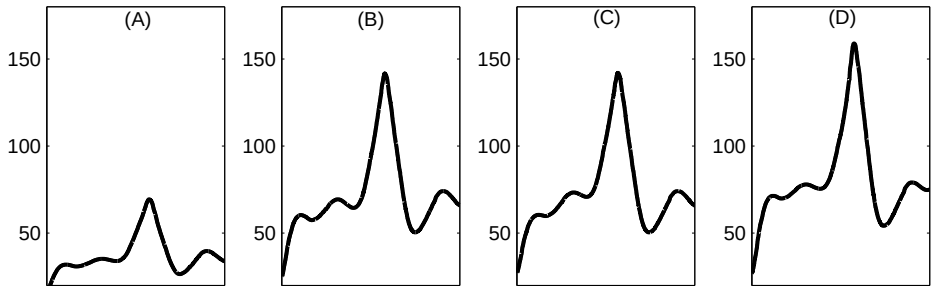

Figure 1. Daily seasonal component.

The estimates of ACD(1,1) models under the exponential and Pareto distribution assumptions are reported in Table 1. For the exponential hypothesis, in the four cases all the parameters are significant. Moreover,

Table 1. Estimates of ACD(1,1), exponential and Pareto errors.

	Estimates (*standard errors*)			
Parameter	A	B	C	D
Exponential				
ω	0.002 *(0.001)*	0.002 *(0.001)*	0.003 *(0.001)*	0.012 *(0.003)*
α	0.035 *(0.003)*	0.035 *(0.004)*	0.034 *(0.004)*	0.041 *(0.006)*
β	0.963 *(0.003)*	0.963 *(0.004)*	0.963 *(0.004)*	0.947 *(0.008)*
Pareto				
ω	0.002 *(0.001)*	0.002 *(0.001)*	0.003 *(0.001)*	0.011 *(0.004)*
α	0.036 *(0.003)*	0.035 *(0.004)*	0.034 *(0.004)*	0.040 *(0.006)*
β	0.963 *(0.003)*	0.963 *(0.004)*	0.963 *(0.004)*	0.949 *(0.008)*
θ	14.40 *(2.118)*	10.03 *(1.473)*	9.04 *(1.250)*	6.49 *(0.741)*

their values are very similar across types of durations, that is conditional expected durations seems to obey the same dynamic structure.

The bottom part of the Table contains the estimates of the models when a Pareto distribution is assumed for the residuals. The estimates of the parameters of the conditional expected duration equations are again very similar and not far from the estimates in the former case. This result corroborates the well known robustness of the simple exponential hypothesis for the estimation of the equation of Ψ_i. On the contrary, the estimates of the additional parameter θ seems to characterize the four cases. The estimates range from 6.49 (durations between price change over 0.22%) to 14.40 (durations between transactions). The probability law for the residuals is then peculiar to the definition of the market event used to compute the durations.

In the left part of Figure 2, the mixing distributions for θ are represented in the four cases. The differences involve different hazard functions (right part of the Figure), in contrast with the constant unit hazard suggested for all the datasets by the exponential distribution.

4.2 Diagnostics

In the diagnostics of ACD models, the crucial points are represented by the variance test of the residuals and goodness-of-fit tests. In the expo-

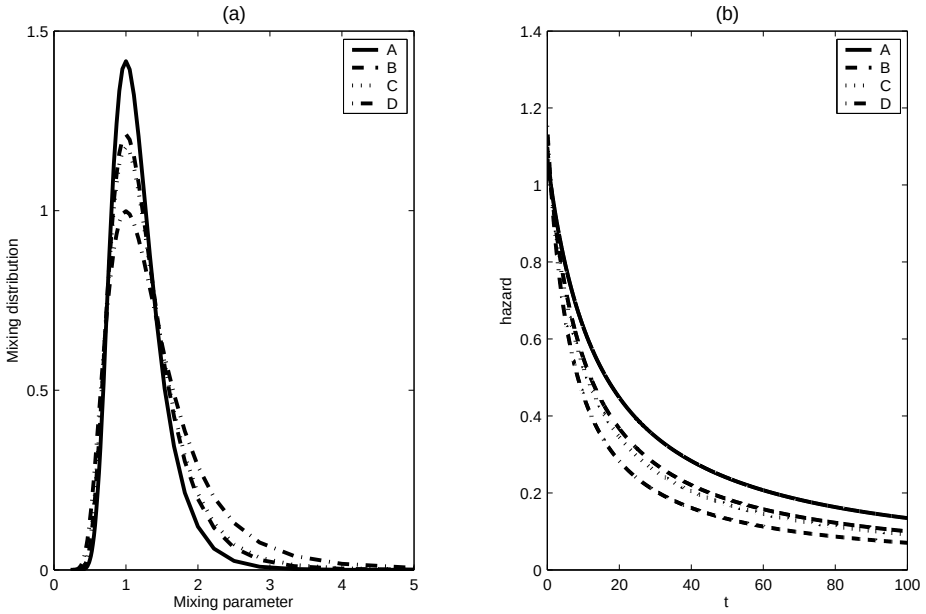

Figure 2. Mixing distribution for the infinite mixture (a) and hazard functions (b).

nential case the mean tests suggest to accept the hypothesis of unit mean (Table 2), but the hypothesis of unit variance is strongly rejected in all the cases. Moreover, the KS and CVM tests do not support the distributional assumption. The p-values are always negligible. A graphical representation of the inadequacy is contained in the QQ-plots in Figure 3.

On the whole, using the unit exponential hypothesis suggests two comments:

1. when different financial durations are analyzed, no remarkable difference appears in the parameter estimates;

2. the diagnostics on residuals are not satisfactory.

The diagnostics on the residuals of the Pareto case (Table 3) show the success of the mean and the variance tests. The goodness-of-fit statistics (KS and CVM) prove the adequacy of the Pareto hypothesis, with some

Table 2. Diagnostics of ACD(1,1) model, exponential errors.

Diagnostics on residuals				
Statistic	A	B	C	D
Mean	0.9999	1.0000	1.0004	1.0004
test	-0.0165	0.0014	0.0335	0.0272
p-value	0.9868	0.9989	0.9733	0.9783
Variance	1.1405	1.2034	1.2277	1.3779
test	6.2888	6.5749	7.2694	11.5336
p-value	0.0000	0.0000	0.0000	0.0000
$Q(20)$	27.1019	19.7551	21.0556	28.1135
p-value	0.1324	0.4733	0.3938	0.1067
likelihood	0.3958	0.3931	0.3855	0.3803
KS	0.0231	0.0294	0.0319	0.0358
p-value	$<$0.001	$<$0.001	$<$0.001	$<$0.001
CVM	2.7140	2.4852	2.8938	4.1495
p-value	$<$0.001	$<$0.001	$<$0.001	$<$0.001

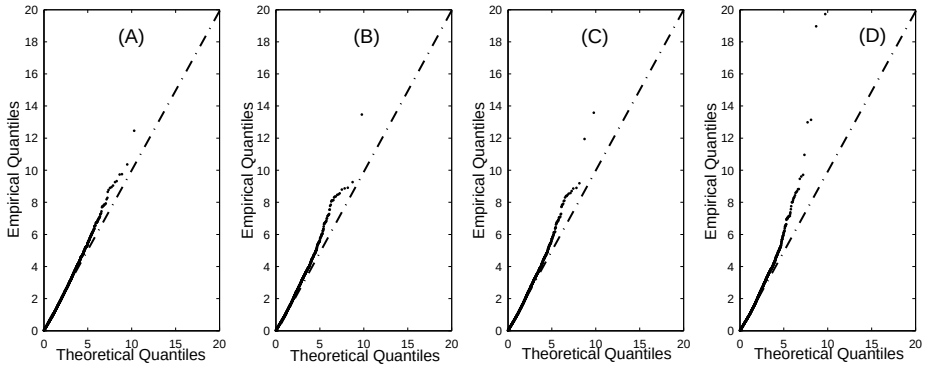

Figure 3. QQ-plot, exponential errors.

doubt for the durations between transactions. The QQ-plots in Figure 4 confirm the responses of the tests. Summarizing, when a Pareto distruibution is assumed:

1. different stochastic structures characterize different financial dura-

Table 3. Diagnostics of ACD(1,1) model, Pareto errors.

Diagnostics on residuals				
Statistic	A	B	C	D
Mean	0.9998	0.9998	1.0001	1.0006
test	-0.0288	-0.0143	0.0031	0.0416
p-value	0.9770	0.9886	0.9975	0.9668
Variance	1.1405	1.2030	1.2268	1.3790
test	-0.2714	-0.3497	-0.3889	0.1915
p-value	0.7861	0.7266	0.6974	0.8481
$Q(20)$	26.6677	19.7796	21.1124	28.9353
p-value	0.1449	0.4718	0.3905	0.0890
likelihood	0.3967	0.3947	0.3874	0.3841
KS	0.0160	0.0089	0.0089	0.0091
p-value	$<$0.001	0.532	0.549	0.582
CVM	0.3369	0.1135	0.0926	0.0569
p-value	0.106	0.529	0.623	0.822

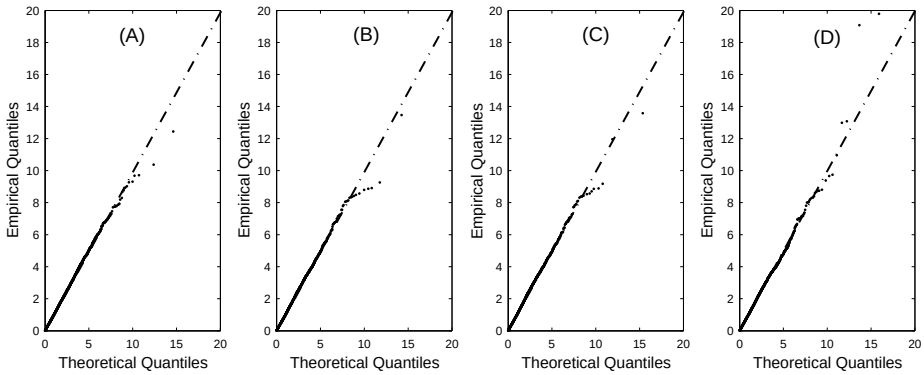

Figure 4. QQ-plot, Pareto errors.

tions;

2. the diagnostics on residuals prove the goodness of the choice.

The superiority of the Pareto assumption is confirmed by the density

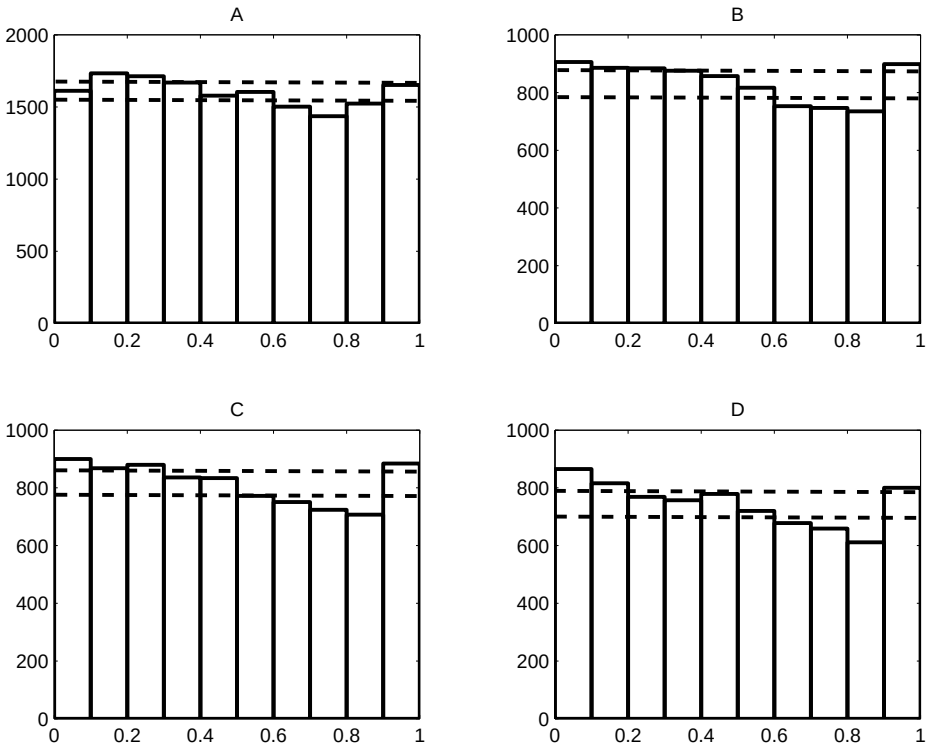

Figure 5. Histograms of empirical probability integral transforms $\{z_i\}_{i=1}^{s}$ *for the four dataset, exponential errors.*

forecast technique. According to both the histograms with 95% bin-by-bin confidence intervals for the null hypothesis of *i.i.d* $U(0,1)$ (Figures 5 and 6), the uniform distribution for the empirical sequence $\{z_i\}_{i=1}^{s}$ is acceptable for the Pareto hypothesis.

5. Concluding remarks

Datasets of durations between different market events are obtained for a stock of Italian financial market and analyzed under two particular distributional assumptions, the traditional exponential hypothesis and an

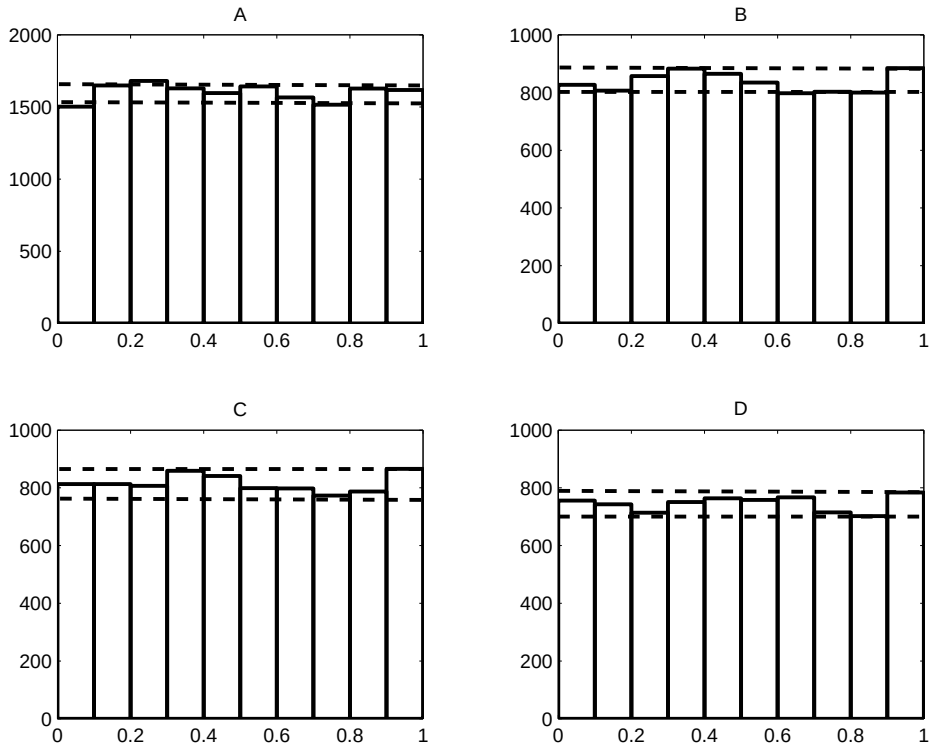

Figure 6. Histograms of empirical probability integral transforms $\{z_i\}_{i=1}^{s}$ *for the four dataset, Pareto errors.*

infinite mixtures of exponential distributions with Inverse Gamma mixing distribution, that is a Pareto II type distribution (De Luca and Zuccolotto, 2003). The results show a poor fit of the exponential distribution, as already suggested by many empirical studies. Assuming the Pareto distribution one can get a good fit regardless of the type of duration one is interested to. It is convenient to use the Pareto distribution because it is able to offer interesting interpretations on the relationships between different durations, both in terms of parameter estimation and in terms of hazard functions.

References

Bauwens L. and Giot P. (2000),The Logarithmic ACD Model: an application to the bid-ask quote process of three NYSE stocks, *Annales d'Economie et de Statistique*, 60, 117-149.

Bauwens L. and Giot P. (2001), *Econometric Modelling of Stock Market Intraday Activity*, Kluwer Academic Publisher, Boston.

Bauwens L. and Giot P. (2003), Asymmetric ACD Models: introducing price information in ACD Models, *Empirical Economics*, 28, 1-23

Bauwens L., Giot P., Grammig J. and Veredas D. (2004), A comparison of financial duration models via density forecasts, *International Journal of Forecasting*, forthcoming.

D'Agostino R.B. and Stephens M.A. (1986), *Goodness-of-fit techniques*, Marcel Dekker, Inc., New York.

De Luca G. and Gallo G. M. (2004), Mixtures of Distributions for Financial Intradaily Durations, *Studies in Nonlinearity and Dynamics in Econometrics*, Vol. 8, No. 2. http://www.bepress.com/snde/vol8/iss2/art8

De Luca G. and Zuccolotto P. (2003), Finite and infinite mixtures for financial durations, *Metron*, 61, 431-455.

Diebold F.X., Gunther T.A., Tay A.S. (1998), Evaluating density forecasts, with application to financial risk management, *International Economic Review*, 39, 863-883.

Engle R.F. (2000), The econometrics of ultra-high frequency data, *Econometrica*, 68, 1-22.

Engle R.F. and Russell J.E. (1998), Autoregressive Conditional Duration: a new model for irregularly spaced transaction data, *Econometrica*, 66, 1127-1162.

Jasiak J. (1998), Persistence in intertrade durations, *Finance*, 19, 166-195.

Zhang M.Y., Russell J.R. and Tsay R.S. (2001), A nonlinear autoregressive conditional duration model with application to financial transaction data, *Journal of Econometrics*, 104, 179-207.

Zuccolotto P. (2002), Modelling the impact of open volume on intertrade autoregressive durations, *Metron*, 60, 51-65.

Zuccolotto P. (2003a), La relazione tra volumi di apertura ed intensità

delle contrattazioni nel mercato finanziario: introduzione di un fattore di smorzamento nel modello ACD giornaliero, *Statistica & Applicazioni*, 1, 69-86.

Zuccolotto P. (2003b), Quantiles estimation in ultra-high frequency financial data: a comparison between parametric and semiparametric approach, *Statistical Methods & Applications*, 12, 243-257.

REMARK. The paper is the outcome of the assiduous collaboration of the authors. However, sections 1, 2 and 4.2 can be attributed to Paola Zuccolotto, while sections 3, 4.1 and 5 can be attributed to Giovanni De Luca.