No. 2010/19

Capturing the Zero: A New Class of Zero-Augmented Distributions and Multiplicative Error Processes

Nikolaus Hautsch, Peter Malec,
and Melanie Schienle

Center for Financial Studies
Goethe-Universität Frankfurt ■ House of Finance
Grüneburgplatz 1 ■ 60323 Frankfurt ■ Deutschland

Telefon: +49 (0)69 798-30050
Fax: +49 (0)69 798-30077
http://www.ifk-cfs.de ■ E-Mail: info@ifk-cfs.de

Center for Financial Studies

The *Center for Financial Studies* is a nonprofit research organization, supported by an association of more than 120 banks, insurance companies, industrial corporations and public institutions. Established in 1968 and closely affiliated with the University of Frankfurt, it provides a strong link between the financial community and academia.

The CFS Working Paper Series presents the result of scientific research on selected topics in the field of money, banking and finance. The authors were either participants in the Center´s Research Fellow Program or members of one of the Center´s Research Projects.

If you would like to know more about the *Center for Financial Studies*, please let us know of your interest.

Prof. Michalis Haliassos, Ph.D. Prof. Dr. Jan Pieter Krahnen Prof. Dr. Uwe Walz

Center for Financial Studies
Goethe-Universität ■ House of Finance
Grüneburgplatz 1 ■ 60323 Frankfurt am Main ■ Deutschland

Telefon: +49 (0)69 798-30050
Fax: +49 (0)69 798-30077
http://www.ifk-cfs.de ■ E-Mail: info@ifk-cfs.de

CFS Working Paper No. 2010/19

Capturing the Zero: A New Class of Zero-Augmented Distributions and Multiplicative Error Processes*

Nikolaus Hautsch[1], Peter Malec[2], and Melanie Schienle[3]

November 2010

Abstract:
We propose a novel approach to model serially dependent positive-valued variables which realize a non-trivial proportion of zero outcomes. This is a typical phenomenon in financial time series observed on high frequencies, such as cumulated trading volumes or the time between potentially simultaneously occurring market events. We introduce a flexible point-mass mixture distribution and develop a semiparametric specification test explicitly tailored for such distributions. Moreover, we propose a new type of multiplicative error model (MEM) based on a zero-augmented distribution, which incorporates an autoregressive binary choice component and thus captures the (potentially different) dynamics of both zero occurrences and of strictly positive realizations. Applying the proposed model to high-frequency cumulated trading volumes of liquid NYSE stocks, we show that the model captures both the dynamic and distribution properties of the data very well and is able to correctly predict future distributions.

JEL Classification: C22, C25, C14, C16, C51

Keywords: High-frequency Data, Point-mass Mixture, Multiplicative Error Model, Excess Zeros, Semiparametric Specification Test, Market Microstructure

* For helpful comments and discussions we thank the participants of workshops at Humboldt-Universit at zu Berlin. This research is supported by the Deutsche Forschungsgemeinschaft (DFG) via the Collaborative Research Center 649 "Economic Risk".

1 Institute for Statistics and Econometrics and Center for Applied Statistics and Economics (CASE), Humboldt-Universität zu Berlin, Quantitative Products Laboratory (QPL), Berlin, and Center for Financial Studies (CFS), Frankfurt.
Address: Humboldt Universität zu Berlin, CASE, Spandauer Str. 1, 10178 Berlin, Germany. Email: nikolaus.hautsch@wiwi.hu-berlin.de..

2 Corresponding author. Institute for Statistics and Econometrics, Humboldt-Universität zu Berlin, Spandauer Str. 1, 10178 Berlin, Germany. Email: malecpet@hu-berlin.de.

3 Institute for Statistics and Econometrics and Center for Applied Statistics and Economics (CASE), Humboldt-Universität zu Berlin, Spandauer Str. 1, 10178 Berlin, Germany. Email: melanie.schienle@wiwi.hu-berlin.de.

1 Introduction

The availability and increasing importance of high-frequency data in empirical finance and financial practice has triggered the development of new types of econometric models capturing the specific properties of these observations. Typical features of financial data observed on high frequencies are strong serial dependencies, irregular spacing in time, price discreteness and the non-negativity of various (trading) variables. To account for these properties, models have been developed which contain features of both time series approaches and microeconometric specifications, see, e.g., Engle and Russell (1998), Russell and Engle (2005) or Rydberg and Shephard (2003), among others.

This paper proposes a novel type of model capturing a further important property of high-frequency data which is present in many situations but not taken into account in extant approaches: the occurrence of a non-trivial part of zeros in the data – henceforth referred to as "excess zeros" – which is a typical phenomenon in both irregularly-spaced as well as aggregated (regularly spaced) financial time series on high frequencies. For instance, when modeling trade durations, many transaction data sets reveal a high proportion of simultaneous occurrences of trades (and thus zero durations). They are induced either by (large) trades which are simultaneously executed against several (small) orders, by imprecise recording which assigns identical time stamps to fast sequences of trades[1] or by actual simultaneous occurrences of trading events. As it is mostly quite difficult or even impossible to exactly identify and to disentangle such effects, the often employed pragmatic solution of just aggregating all simultaneous events together – and thus ultimately discarding zero occurrences – is not appropriate. A further example, which serves as the major motivation for our study, arises in the context of high-frequency time aggregates (e.g., 15 sec or 30 sec data), as often used in high-frequency trading. Here, measures of trading activity, such as cumulated trading volumes, naturally reveal a high proportion of zero observations since even for liquid stocks there is not necessarily trading in each time interval. As a representative example, Figure 1 depicts the empirical distribution of cumulated trading volumes per 15 seconds

[1] For example, the transaction time stamp provided by the widely used Trade and Quote (TAQ) database is accurate to one second only.

of the Disney stock traded at the New York Stock Exchange (NYSE). Though Disney is a highly liquid security, no-trade intervals amount to a proportion of about 20%, leading to a significant spike at the leftmost bin.

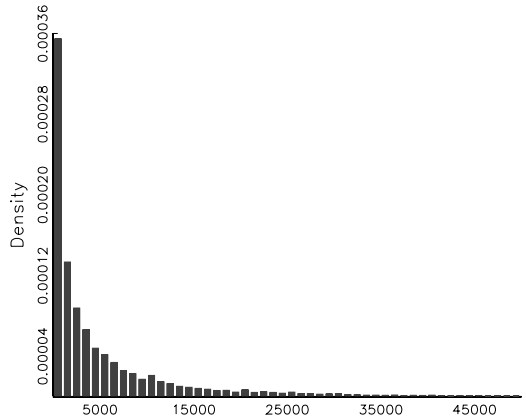


Figure 1: Histogram of 15 Sec Cumulated Volumes of the Disney Stock (NYSE), January 16 to February 10, 2006

The occurrence of such high proportions of zero observations is obviously not appropriately captured by any standard distribution for non-negative random variables, such as the exponential distribution, generalizations thereof as well as various types of truncated models (c.f. Johnson et al., 1994). This has serious consequences in a dynamic framework, as, e.g., in the multiplicative error model (MEM) introduced by Engle (2002) which is commonly used to model positive-valued autocorrelated data. In such a framework, employing distributions which do not explicitly account for excess zeros induces severe distributional misspecifications causing inefficiency and in many cases even inconsistency of parameter estimates. These misspecification effects become even more evident when zero occurrences – and thus (no) trading probabilities – follow their own dynamics. Moreover, standard distributions are clearly inappropriate whenever density forecasts are in the core of interest since they are not able to explicitly predict zero outcomes.

To our best knowledge, existent literature does not provide any systematic and self-contained framework to model, test and predict serially-dependent positive-valued data realizing a non-trivial part of excess zeros. Therefore, our main contributions can be summarized as follows. First, we introduce a new type of discrete-continuous mixture distribution capturing a clustering of observations at zero. The idea is to decompose the distribution into a point-mass at zero and a flexible continuous distribution for strictly positive values. Second, we propose a novel semiparametric density test, which is

tailored to distributions based on point-mass mixtures. Both the proposed distribution as well as the specification test might be valuable not only in the context of financial data but also in engineering or natural sciences, see, e.g., Weglarczyk et al. (2005). Third, we employ the above mixture distribution to specify a so-called zero-augmented MEM (ZA-MEM) that allows for maximum likelihood estimation in the presence of zero observations. Finally, we explicitly account for serial dependencies in zero occurrences by introducing an augmented MEM structure which captures the probability of zeros based on a dynamic binary choice component. The resulting so-called Dynamic ZA-MEM (DZA-MEM) yields a specification which allows to explicitly predict zero outcomes and thus is able to produce appropriate density forecasts.

A zero augmented model is an important complement to current approaches which reveal clear deficiencies and weaknesses in the presence of zeros. For instance, the quasi log-likelihood function of a MEM based on a (standard) gamma distribution (see, e.g., Drost and Werker, 2004) cannot be evaluated in the case of zero observations. An analogous argument holds for the log-normal distribution yielding QML estimates of a logarithmic MEM (Allen et al., 2008). Consequently, the only feasible distribution yielding QML estimates is the exponential distribution. The latter, however, is heavily misspecified in cases as shown in Figure 1 and thus yields clearly inefficient parameter estimates. This inefficiency can be harmful if a model is applied to time-aggregated data and is (re-)estimated over comparably short time intervals as, e.g., a day (for instance, to be not affected by possible structural breaks). Moreover, using exponential QML or the generalized methods of moments (GMM) as put forward by Brownlees et al. (2010) does not allow to explicitly estimate (and thus to predict) point masses at zero. This limits the applicability of such approaches, for instance, in VWAP applications where intraday trading volumes have to be predicted. Finally, from an economic viewpoint, no-trade intervals contain own-standing information. E.g., in the asymmetric information-based market microstructure model by Easley and O'Hara (1992), the absence of a trade indicates lacking information in the market. Indeed, the question whether to trade and (if yes) how much to trade are separate decisions which do not necessarily imply that no-trade intervals can be considered as the extreme case of low trading volumes. Consequently, the binary process of no-trading might follow its own dynamics others than that of (non-zero) volumes.

This paper contributes to several strings of literature. Firstly, it adds to the literature on point-mass mixture distributions. An important distinguishing feature of the existing specifications is whether the point-mass at zero is held constant (e.g., Weglarczyk et al., 2005) or explained by a standard (static) binary-choice model (e.g., Duan et al., 1983).

We extend these approaches by allowing for a dynamic model for zero occurrences. In an MEM context, De Luca and Gallo (2004) or Lanne (2006) employ mixtures of continuous distributions which are typically motivated by economic arguments, such as trader heterogeneity. The idea of employing a point-mass mixture distribution to model zero values is only mentioned, but not applied, by Cipollini et al. (2006).

Secondly, our semiparametric specification test contributes to the class of kernel-based specification tests, as e.g., proposed by Fan (1994), Fernandes and Grammig (2005) or Hagmann and Scaillet (2007). None of the existing methods, however, is suitable for distributions including a point-mass component. If applied to MEM residuals, our approach also complements the literature on diagnostic tests for MEM specifications.

Third, since the proposed dynamic zero-augmented MEM comprises a MEM and a dynamic binary-choice part, we also extend the literature on component models for high-frequency data, as, e.g., Rydberg and Shephard (2003) or Liesenfeld et al. (2006), among others. While the latter focus on transaction price changes, our model is applicable to various transaction characteristics, as it decomposes a (nonnegative) persistent process into the dynamics of zero values and strictly positive realizations. For instance, the approach can explain the trading probability in a first stage and, given that a trade has occurred, models the corresponding cumulated volume.

We apply the proposed model to 15 second cumulative volumes of two liquid stocks traded at the NYSE. Using the developed semiparametric specification test, we show that the ZA-MEM captures the distributional properties of the data revealing distinct excess zeros very well. Moreover, a density forecast analysis shows that the novel type of MEM structure is successful in explaining the dynamics of zero values and appropriately predicting the entire distribution. The best performance is shown for a DZA-MEM specification where the zero outcomes are modeled using an autoregressive conditional multinomial (ACM) model as proposed by Russell and Engle (2005). In fact, we observe that trading probabilities are quite persistent following their own dynamics. Our results show that the proposed model can serve as a workhorse for the modeling and prediction of various high-frequency variables and can be extended in different directions. Moreover, the introduced class of zero-augmented distributions and semiparametric specification test might be useful also in other areas where (continuous) data reveal a clustering at single realizations. Examples might come from genomics and proteomics (e.g., Taylor and Pollard, 2009) as well as hydrology (e.g., Weglarczyk et al., 2005).

The remainder of this paper is structured as follows. In Section 2, we introduce a novel point-mass mixture distribution and develop a corresponding semiparametric specification test which is applied to evaluate the goodness-of-fit based on MEM residuals.

Section 3 presents the dynamic zero-augmented MEM capturing serial dependencies in zero occurrences. We evaluate the extended model and benchmark its performance against the basic zero-augmented MEM by density forecast methods. Finally, Section 4 concludes.

2 A Discrete-Continuous Mixture Distribution

2.1 Data and Motivation

We analyze high-frequency trading volume data for the two stocks Disney (DIS) and Johnson & Johnson (JNJ) traded at the New York Stock Exchange. The transaction data is extracted from the Trade and Quote (TAQ) database released by the NYSE and covers one trading week from February 6 to February 10, 2006. We filter the raw data by deleting transactions that occurred outside regular trading hours from 9:30 am to 4:00 pm, as well as observations recorded during the first and last 30 minutes of each trading day to reduce the impact of opening and closure effects. The tick-by-tick data is aggregated by computing cumulated trading volumes over 15 second intervals, resulting in 6595 observations for both stocks. Modeling and forecasting cumulated volumes on high frequencies is, for instance, crucial for trading strategies replicating the (daily) volume weighted average price (VWAP), see, e.g., Brownlees et al. (2010). To account for the well-known intraday seasonalities (see, e.g., Hautsch (2004) for an overview), we divide the cumulated volumes by a seasonality component which is pre-estimated employing a cubic spline function.

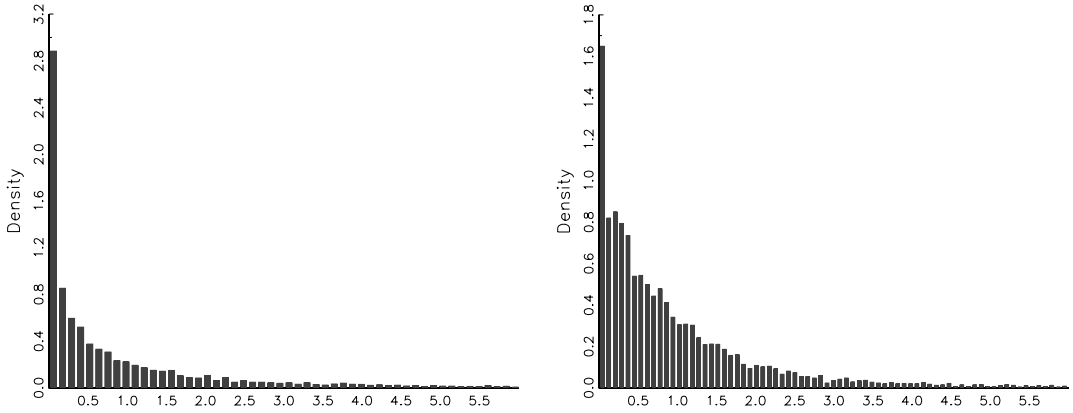
An important feature of the data is the high number of zeros induced by non-trading intervals. The summary statistics in Table 1 and the histograms depicted in Figure 2 report a non-trivial share of zero observations of about 21% for DIS and roughly 7% for JNJ.

A further major feature of cumulated volumes is their strong autocorrelation and high persistence as documented by the Q-statistics in Table 1 and the autocorrelation functions (ACFs) displayed in Figure 3.

To account for these strong empirical features, we first propose a distribution capturing the phenomenon of excess zeros and, secondly, implement it in a MEM setting.

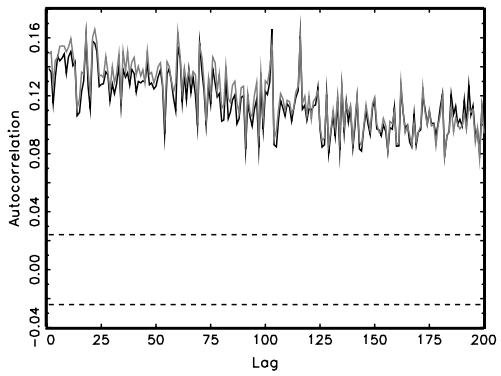
Table 1: Summary Statistics of Cumulated Trading Volumes
All statistics are reported for the raw and seasonally adjusted time series. SD: standard deviation, SK: skewness, q_5 and q_{95}: 5%- and 95%-quantile, respectively. n_z/n: share of zero observations. $Q(l)$: Ljung-Box statistic associated with l lags. The 5% critical values associated with lag lengths 20, 50 and 100 are 31.41, 67.51 and 124.34, respectively.

	Disney (DIS)		**Johnson & Johnson (JNJ)**	
	Raw	Adj.	Raw	Adj.
Obs	6595	6595	6595	6595
Mean	7071.3	1.02	4250.4	1.01
SD	13588.2	1.99	6293.3	1.48
SK	7.31	8.54	8.92	8.95
q_5	0	0.00	0	0.00
q_{95}	30000	4.24	14000	3.29
n_z/n	20.7%	20.7%	7.1%	7.1%
$Q(20)$	2710.90	2421.30	554.03	1006.02
$Q(50)$	6601.78	5916.07	867.30	1801.22
$Q(100)$	12025.76	10745.00	1173.65	2485.09

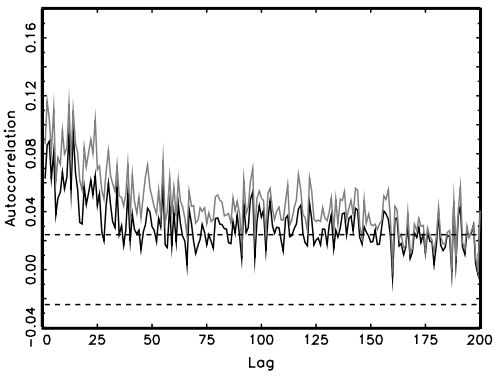


2.1: DIS 2.2: JNJ

Figure 2: Sample Histograms of Deseasonalized Cumulated Volumes



3.1: DIS



3.2: JNJ

Figure 3: Sample Autocorrelograms
Sample autocorrelation functions of raw (grey line) and diurnally adjusted (black line) cumulated trading volumes. Horizontal lines indicate the limits of 95% confidence intervals ($\pm 1.96/\sqrt{n}$).

2.2 A Zero-Augmented Distribution for Non-Negative Variables

We consider a non-negative random variable X with independent observations $\{X_t\}_{t=1}^n$, corresponding, e.g., to the residuals of an estimated time series model. In the presence of zero observations, a natural choice is the exponential distribution as it is also defined for zero outcomes and, as a member of the standard gamma family, provides consistent QML estimates of the underlying conditional mean function (e.g., specified as a MEM). However, in case of high proportions of zero realizations (as documented in Section 2.1), this distribution is severely misspecified making QML estimation quite inefficient. To account for excess zeros we assign a discrete probability mass to the exact zero value. Hence, we define the probabilities

$$\pi := P(X > 0), \qquad 1 - \pi := P(X = 0). \tag{1}$$

Conditional on $X > 0$, X follows a continuous distribution with density $g_X(x) := f_X(x|X > 0)$, which is continuous for $x \in (0, \infty)$. Consequently, the unconditional distribution of X is semicontinuous with a discontinuity at zero, implying the density

$$f_X(x) = (1 - \pi)\,\delta(x) + \pi\; g_X(x)\,\mathbb{1}_{(x>0)}, \tag{2}$$

where $0 \leq \pi \leq 1$, $\delta(x)$ is a point probability mass at $x = 0$, while $\mathbb{1}_{(x>0)}$ denotes an indicator function taking the value 1 for $x > 0$ and 0 else. The probability π is treated as a parameter of the distribution determining how much probability mass is assigned

to the strictly positive part of the support. Note that the above point-mass mixture assumes zero values to be "true" zeros, i.e., they originate from another source than the continuous component and do not result from truncation. This assumption is valid, e.g., in case of cumulative trading volumes, where zero values correspond to non-trade intervals and originate from the decision whether or whether not to trade.

The log-likelihood function implied by the mixture density (2) is

$$\mathcal{L}(\vartheta) = n_z \ln(1-\pi) + n_{nz} \ln \pi + \sum_{t,nz} \ln g_X(x_t; \vartheta^g), \tag{3}$$

where $\vartheta = (\pi, \vartheta^g)'$, while n_z and n_{nz} denote the number of zero and nonzero observations, respectively. If no dependencies between π and ϑ^g are introduced, componentwise estimation is possible and the estimate of π is given by the empirical frequency of zero observations.

The conditional density $g_X(x)$ can be specified according to any distribution defined on positive support. We consider the generalized F (GF) distribution, since it nests most of the distributions frequently used in high-frequency applications (see, e.g., Hautsch, 2003). The corresponding conditional density is given by

$$g_X(x) = \frac{a\, x^{a\,m-1} \left[\eta + (x/\lambda)^a\right]^{(-\eta-m)} \eta^\eta}{\lambda^{a\,m}\, \mathcal{B}(m,\eta)}, \tag{4}$$

where $a > 0, m > 0, \eta > 0$ and $\lambda > 0$. $\mathcal{B}(\cdot)$ describes the full Beta function with $\mathcal{B}(m,\eta) := \frac{\Gamma(m)\Gamma(\eta)}{\Gamma(m+\eta)}$. The conditional noncentral moments implied by the GF distribution are

$$E[X^s | X > 0] = \lambda^s\, \eta^{s/a} \frac{\Gamma(m + s/a)\, \Gamma(\eta - s/a)}{\Gamma(m)\, \Gamma(\eta)}; \quad a\,\eta > s. \tag{5}$$

Accordingly, the distribution is based on three shape parameters a, m and η, as well as a scale parameter λ. The support of the GF distribution includes the exact zero only if the parameters satisfy the condition $a\,m \geq 1$ with the limiting case of an exponential distribution. A detailed discussion of special cases and density shapes implied by different parameter values can be found in Lancaster (1997).

The unconditional density of the zero-augmented generalized F (ZAF) distribution follows from (2) and (4) as

$$f_X(x) = (1-\pi)\, \delta(x) + \pi\, \frac{a\, x^{a\,m-1} \left[\eta + (x/\lambda)^a\right]^{(-\eta-m)} \eta^\eta}{\lambda^{a\,m}\, \mathcal{B}(m,\eta)}\, \mathbb{1}_{(x>0)}, \tag{6}$$

which reduces to the GF density for $\pi = 1$. The unconditional moments can be obtained by exploiting eq. (5), i.e.

$$\begin{aligned} E[X^s] &= \pi\ E[X^s|X>0] + (1-\pi)\ E[X^s|X=0]\,, \\ &= \pi\ \lambda^s\,\eta^{s/a}\frac{\Gamma(m+s/a)\,\Gamma(\eta-s/a)}{\Gamma(m)\,\Gamma(\eta)}; \quad a\,\eta > s. \end{aligned} \tag{7}$$

The log-likelihood function of the ZAF distribution is given by

$$\begin{aligned} \mathcal{L}\,(\vartheta) = n_z \ln(1-\pi) + n_{nz} \ln \pi + \sum_{t,nz} \Big\{ &\ln a + (am-1)\ \ln x_t + \eta\ \ln \eta \\ &- (\eta+m)\ \ln\left\{\eta + \left[x_t\,\lambda^{-1}\right]^a\right\} - \ln \mathcal{B}(m,\eta) - am \ln \lambda \Big\}, \end{aligned} \tag{8}$$

where $\vartheta = (\pi, a, m, \eta, \lambda)'$.

2.3 A New Semiparametric Specification Test

We introduce a specification test that is tailored to point-mass mixture distributions on nonnegative support like (2) and hence, complements the methods for dealing with excess zero effects as described in Section 2.1. Instead of, e.g., checking a number of moment conditions, we consider a kernel-based semiparametric approach, which allows to formally examine whether the entire distribution is correctly specified. Compared to similar smoothing specification tests for densities with left-bounded support, as, e.g., proposed by Fernandes and Grammig (2005) and Hagmann and Scaillet (2007), the assumption of a point-mass mixture under the null and alternative hypothesis is a novelty. Furthermore, estimation in our procedure is optimized for densities which are locally concave for small positive values as in Section 2.1.

In this setting, an appropriate semiparametric benchmark estimator for the unconditional density $f_X(x)$ must have the point mass mixture structure as in (2). Since the support of the discrete and continuous component is disjoint, we can estimate both parts separately without further functional form assumptions. In particular, we use the empirical frequency $\hat{\pi} = n^{-1}\sum_t \mathbb{1}_{(x_t>0)}$ as an estimate for the probability $X > 0$. The conditional density g_X is estimated using a nonparametric kernel smoother

$$\hat{g}_X(x) = \frac{1}{nb}\sum_{t=1}^{n} K_{x,b}\,(X_t)\,, \tag{9}$$

where K is a kernel function integrating to unity. The estimator is generally consistent on unbounded support for bandwidth choices $b = O(n^{-\nu})$ with $\nu < 1$. Though, if

the support of the density is bounded, in our case from below at zero, standard fixed kernel estimators assign weight outside the support at points close to zero and therefore yield inconsistent results at points near the boundary. Thus instead, we consider a gamma kernel estimator as proposed in Chen (2000) whose flexible form ensures that it is boundary bias free, while density estimates are always nonnegative in contrast to some boundary correction methods for fixed kernels such as boundary kernels (Jones, 1993) or local-linear estimation (Cheng et al., 1997). The asymmetric gamma kernel is defined on the positive real line and is based on the density of the gamma distribution with shape parameter $x/b+1$ and scale parameter b

$$K^{\gamma}_{x/b+1,b}(u) = \frac{u^{x/b}\exp(-u/b)}{b^{x/b}\,\Gamma(x/b+1)}. \tag{10}$$

For the final standard gamma kernel estimator set $K_{x,b}(X_t) = K^{\gamma}_{x/b+1,b}(X_t)$ in (9). Note that if the density is locally concave near zero, it is statistically favorable to employ the standard gamma kernel (10) and not the modified version as also proposed in Chen (2000) or other boundary correction techniques such as reflection methods (e.g. Schuster, 1958) or cut-and-normalized kernels (Gasser and Müller, 1979). In this case, signs of first and second derivative of the density in this region are opposed causing the leading term of the vanishing bias of the standard gamma kernel estimator to be of smaller absolute value than the pure second derivative in the corresponding term for the modified estimator and the other estimators (see Zhang (2010) for details). With the same logic, however, the opposite is true for locally convex densities near zero, as for, e.g., income distributions (Hagmann and Scaillet, 2007), which we do not consider here.

While for estimation at points further away from the boundary the variance of gamma kernel estimators is smaller compared to symmetric fixed kernels, their finite sample bias is generally larger. We therefore apply a semiparametric correction factor technique as in Hjort and Glad (1995) or Hagmann and Scaillet (2007) to enhance the precision of the gamma kernel estimator in the interior of the support. This approach is semiparametric in that the unknown density $g_X(x)$ is decomposed as the product of the initial parametric model $g_X\big(x,\vartheta\big)$ and a factor $r(x)$ which corrects for the potentially misspecified parametric start. The estimate of the parametric start is given by $g_X\big(x,\widehat{\vartheta}\big)$, where $\widehat{\vartheta}$ is the maximum likelihood estimator. The correction factor is estimated by kernel smoothing, such that $\hat{r}(x) = \frac{1}{n}\sum_{t=1}^{n} K_{x/b+1,b}(x_t)\,/g_X\big(X_t,\widehat{\vartheta}\big)$. Therefore, the

bias-corrected gamma kernel estimator is

$$\tilde{g}_X(x) = \frac{1}{nb} \sum_{t=1}^{n} K^{\gamma}_{x/b+1,b}\left(X_t\right) \; \frac{g_X\big(x, \widehat{\vartheta}\big)}{g_X\big(X_t, \widehat{\vartheta}\big)}, \tag{11}$$

which reduces to the uncorrected estimator if the uniform density is chosen as the initial model. Hjort and Glad (1995) show that a corrected kernel estimator yields a smaller bias than its uncorrected counterpart, whenever the correction function is less "rough" then the original density. Their proof is for fixed symmetric kernels, but the argument also holds true for gamma-type kernels with slightly modified calculations.

The formal test of the parametric model $f_X\big(x, \vartheta\big)$ against the semiparametric alternative $f_X(x)$ measures discrepancies in squared distances integrated over the support. As the discrete parts coincide in both cases, it is based on

$$I = \pi \int_0^{\infty} \left\{ g_X(x) - g_X\big(x, \vartheta\big) \right\}^2 dx, \tag{12}$$

where $g_x(x)$ and $g_X\big(x, \vartheta\big)$ denote the general and parametric conditional densities respectively. The null and alternative hypothesis are

$$H_0\colon \; P\Big\{ \hat{f}_X(x) = f_X\big(x, \widehat{\vartheta}\big) \Big\} = 1 \quad H_1\colon \; P\Big\{ \hat{f}_X(x) = f_X\big(x, \widehat{\vartheta}\big) \Big\} < 1, \tag{13}$$

where $\hat{f}_x(x)$ and $f_X\big(x, \widehat{\vartheta}\big)$ are the semiparametric and parametric density estimates with respective continuous conditional parts $g_X\big(x, \widehat{\vartheta}\big)$ and $\tilde{g}_X(x)$ as in (11). The feasible test statistic is given by

$$T_n = n \sqrt{b} \; \hat{\pi} \int_0^{\infty} \Big\{ \tilde{g}_X(x) - g_X\big(x, \widehat{\vartheta}\big) \Big\}^2 dx. \tag{14}$$

Asymptotic normality of statistic (14) could be shown using the results of Fernandes and Monteiro (2005). But it is well-documented that non- and semiparametric tests suffer from size distortions in finite samples (e.g. Fan, 1998). Thus, we employ a bootstrap procedure as in Fan (1998) to compute size-corrected p-values. This is outlined in detail in the following subsection for a specific model of serial dependence in the data.

We choose the bandwidth b according to least squares cross-validation, which is fully data-driven and automatic. Thus, for the bias-corrected gamma kernel estimator (11)

the bandwidth b must minimize

$$CV(b) = \frac{1}{n^2} \sum_i \sum_j \frac{1}{g_X\big(x_i, \widehat{\vartheta}\big) g_X\big(x_j, \widehat{\vartheta}\big)} \int_0^\infty g_X\big(x, \widehat{\vartheta}\big)^2 \, K^{\gamma}_{x/b+1,b}\left(x_i\right) \, K^{\gamma}_{x/b+1,b}\left(x_j\right) \, dx$$
$$- \frac{2}{n\,(n-1)} \sum_i \sum_{j \neq i} K^{\gamma}_{x_i/b+1,b}\left(x_j\right) \frac{g_X\big(x_i, \widehat{\vartheta}_{(i)}\big)}{g_X\big(x_j, \widehat{\vartheta}_{(i)}\big)}, \tag{15}$$

where $\widehat{\vartheta}_{(i)}$ denotes the maximum likelihood estimate computed without observation X_i. The cross-validation objective function is directly derived from requiring the bandwidth to minimize the integrated squared distance between semiparametric and parametric estimate. For the uncorrected gamma kernel estimator, the corresponding objective function is analogous to (15), but does not involve density terms.

Our test differs from related methods not only by being designed for point-mass mixtures. Fan (1994) uses fixed kernels with the respective boundary consistency problems. Fully nonparametric (uncorrected) gamma kernel-based tests as Fernandes and Grammig (2005) have a larger finite sample bias near the boundary for locally concave densities and generally also in the interior of the support. The semiparametric test of Hagmann and Scaillet (2007) suffers from the same problem near zero. Furthermore, weighting with the inverse of the parametric density in their test statistic yields a particularly poor fit in regions with sparse probability, which is an issue in our application, as the distributions are heavily right-skewed.

2.4 Empirical Evidence: Testing a Zero-Augmented MEM

To apply the proposed specification test to our data, we have to appropriately capture the serial dependence in cumulated volumes. This task is performed by specifying a multiplicative error model (MEM) based on a zero-augmented distribution. Accordingly, cumulated volumes, y_t, are given by

$$y_t = \mu_t \, \varepsilon_t, \quad \varepsilon_t \sim \text{i.i.d. } \mathcal{D}(1)\,, \tag{16}$$

where μ_t denotes the conditional mean given the past information set $\mathcal{F}_{t-1}$, while ε_t is a disturbance following a distribution $\mathcal{D}(1)$ with positive support and $E[\varepsilon_t] = 1$. A deeper discussion of the properties of MEMs is given by Engle (2002) or Engle and Gallo (2006). We specify μ_t in terms of a Log-MEM specification proposed by Bauwens and Giot (2000) which does not require parameter constraints to ensure the positivity

of μ_t. Accordingly, μ_t is given by

$$\ln \mu_t = \omega + \sum_{i=1}^{p} \alpha_i \, \ln \varepsilon_{t-i} \, \mathbb{1}_{(x_{t-i}>0)} + \sum_{i=1}^{p} \alpha_i^0 \, \mathbb{1}_{(x_{t-i}=0)} + \sum_{i=1}^{q} \beta_i \, \ln \mu_{t-i}, \tag{17}$$

where the additional dummy variables prevent the computation of $\ln \varepsilon_{t-i}$ whenever $\varepsilon_{t-i} = 0$. The lag structure is chosen according to the Schwartz information criterion (SIC). For more details on the properties of the logarithmic MEM, we refer to Bauwens and Giot (2000) and Bauwens et al. (2003). A comprehensive survey of additional MEM specifications can be found in Bauwens and Hautsch (2008).

We evaluate whether the ZAF distribution provides an appropriate parametric specification for the distribution of MEM innovations for cumulated volumes. Define the zero-augmented MEM (ZA-MEM) as a MEM where ε_t is distributed according to the ZAF density (6) with scale parameter $\lambda = (\pi \, \xi)^{-1}$ and

$$\xi := \eta^{1/a} \left[\Gamma(m + 1/a) \, \Gamma(\eta - 1/a)\right] \left[\Gamma(m) \, \Gamma(\eta)\right]^{-1}. \tag{18}$$

Recalling (7), this constraint on λ ensures that the unit mean assumption for ε_t is fulfilled. The MEM structure (16) implies that, conditionally on the past information set $\mathcal{F}_{t-1}$, y_t follows a ZAF distribution with $\lambda_t = \mu_t \, (\pi \, \xi)^{-1}$. Note that the latter constraint prevents componentwise optimization of the corresponding log-likelihood and thus requires joint estimation of all parameters.

To generate residuals $\hat{\varepsilon}_t := y_t / \hat{\mu}_t$ which are consistent estimates of the errors ε_t, we estimate the model by exponential QML. Alternatively, we could obtain consistent error estimates using the semiparametric methods by Drost and Werker (2004) or employing GMM as in Brownlees et al. (2010). The corresponding estimates are given by Table 2. The consistency and parametric rate of convergence of the conditional mean estimates enable us to use the residuals as inputs for the semiparametric specification test without affecting the asymptotics of the kernel estimators discussed in Section 2.3. A similar procedure is applied by Fernandes and Grammig (2005) for their nonparametric specification test.

Figure 4 depicts the error densities implied by the quasi maximum likelihood estimates and the semiparametric approach based on the uncorrected gamma kernel. For both stocks, the parametric and semiparametric density are quite close to each other. However, there is a noticeable discrepancy near the mode, which can be explained by the increased bias of the gamma kernel as compared to standard fixed kernels in the interior of the support. To refine the semiparametric density estimate, we employ the bias-

Table 2: Estimation Results – ZA-MEM
Quasi maximum likelihood estimates and t-statistics of the zero-augmented Log-MEM. Lag structure is determined using SIC.

	DIS			JNJ		
	Coef.	T-Stat.	P-Val.	Coef.	T-Stat.	P-Val.
ω	0.0100	4.866	0.000	0.026	8.192	0.000
α_1	0.026	7.432	0.000	0.057	9.218	0.000
β_1	0.978	299.709	0.000	0.925	98.489	0.000
α_1^0	0.000	0.076	0.939	0.002	0.114	0.909
m	2.715	9.685	0.000	1.111	5.989	0.000
η	55.410	5.427	0.000	2.948	4.049	0.000
a	0.487	18.438	0.000	1.151	8.392	0.000
π	0.793	159.235	0.000	0.930	294.951	0.000
$\mathcal{L}$		-8796.096			-7870.592	
SIC		17662.544			15811.536	

corrected gamma kernel estimator (11), choosing the ZAF distribution as parametric start. The plots in Figure 5 show that, in both cases, the discrepancy near the mode vanishes, as the parametric density now generally lies within the 95%-confidence region of the semiparametric estimate.

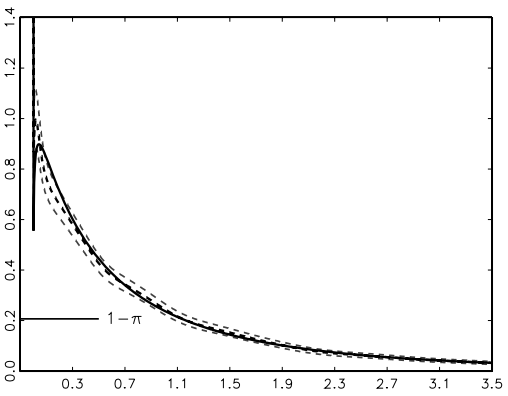
The estimation results suggest that the ZAF distribution provides a superior way to model MEM disturbances for cumulated volumes. This graphical intuition can be formally assessed by the proposed semiparametric specification test (14). In the MEM setting (16), we obtain applicable finite sample p-values via the following bootstrap procedure:

Step 1: Draw a random sample $\{\varepsilon_t^*\}_{t=1}^n$ from the parametric ZAF distribution with density $f_\varepsilon\big(\varepsilon, \widehat{\vartheta}\big)$, where $\widehat{\vartheta}$ is the maximum likelihood estimate of the ZAF parameters ϑ as in (8) from the original data. From this, generate the bootstrap sample $\{y_t^*\}_{t=1}^n$ as $y_t^* = \hat{\mu}_t\, \varepsilon_t^*$, where $\hat{\mu}_t$ is the fitted conditional mean as in (17) based on the maximum likelihood estimates from the original data.

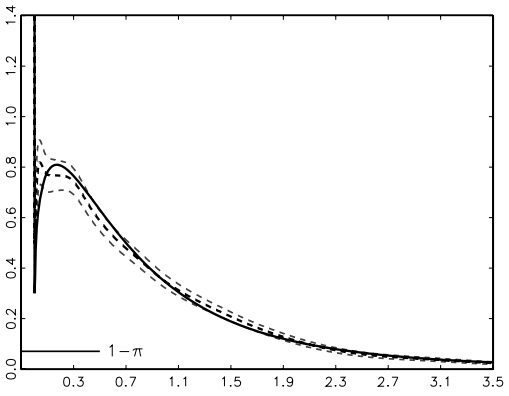
Step 2: Use $\{y_t^*\}_{t=1}^n$ to compute the statistic T_n, which we denote as T_n^*. This requires the re-evaluation of both the parametric and semiparametric estimates of $f_\varepsilon(\varepsilon)$.

Step 3: Steps 1 and 2 are repeated B times and critical values are obtained from the empirical distribution of $\big\{T_{n,r}^*\big\}_{r=1}^B$.

Table 3 displays the test results based on $B = 500$ bootstrap replications for the critical values. In both cases, the statistic is insignificant at the 5%-level, which implies that we cannot reject the null hypothesis as given in (13). These results confirm that the

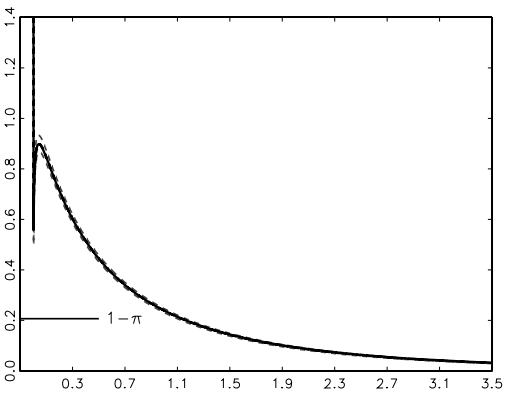


4.1: DIS

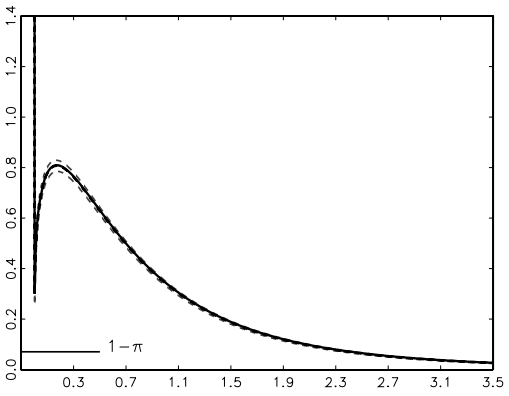


4.2: JNJ

Figure 4: Estimates of Error Density with Gamma KDE
The black solid line represents the error density implied by the QML estimates of the ZA-MEM. The black dashed line is the semiparametric estimate based on the gamma kernel estimator. The grey dashed lines are 95% confidence bounds of the kernel density estimator. LSCV bandwidths: 0.0212 (DIS), 0.0210 (JNJ). Estimates of $1-\pi$ based on sample percentage of zeros values: 0.2068 (DIS), 0.0705 (JNJ).



5.1: DIS



5.2: JNJ

Figure 5: Estimates of Error Density with Corrected Gamma KDE
The black solid line represents the error density implied by the QML-estimates of the ZA-MEM. The black dashed line is the semiparametric estimate based on the bias-corrected gamma kernel estimator. The grey dashed lines are 95% confidence bounds of the kernel density estimator. LSCV bandwidths: 1.3055 (DIS), 1.3067 (JNJ). Estimates of $1-\pi$ based on sample percentage of zeros values: 0.2068 (DIS), 0.0705 (JNJ).

ZA-MEM is able to capture the distributional properties of high-frequency cumulated volumes.

3 Dynamic Zero-Augmented Multiplicative Error Models

3.1 Motivation

Assumption (1) implies that, conditional on past information, the trading probability is constant or, more formally,

$$\pi := P(\varepsilon_t > 0|\mathcal{F}_{t-1}) = P(y_t > 0|\mathcal{F}_{t-1}) = P(\mathcal{I}_t = 1|\mathcal{F}_{t-1})\,, \tag{19}$$

where $\mathcal{I}_t$ is a "trade indicator" taking the value 1 for $y_t > 0$ and 0 else. Given that (non-zero) cumulative volume is clearly time-varying and reveals persistent serial dependencies, the assumption of constant no-trade probabilities appears to be rather restrictive. Moreover, it is at odds with the well-known empirical evidence of autocorrelated trading intensities, see, e.g., Engle and Russell (1998). Table 4 shows the results of a simple runs test based on the trade indicator $\mathcal{I}_t$ suggesting that the null hypothesis of no serial correlation in no-trade probabilities is clearly rejected. Consequently, we propose an augmented version of ZA-MEM accounting also for dynamics in zero occurrences.

3.2 A ZA-MEM with Dynamic Zero Probabilities

Assume that, given the past information set $\mathcal{F}_{t-1}$, the conditional probability of the disturbance ε_t being zero depends on a restricted information set $\mathcal{H}_{t-1} \subset \mathcal{F}_{t-1}$. Moreover, π_t is assumed to depend on $\mathcal{H}_{t-1}$ by a function $\pi(\cdot;\psi)$ with parameter vector ψ,

$$\pi_t := P(\varepsilon_t > 0|\mathcal{F}_{t-1}) = P(\varepsilon_t > 0|\mathcal{H}_{t-1}) = \pi(\mathcal{H}_{t-1};\psi)\,. \tag{20}$$

Table 3: Semiparametric Test for Specification of Error Density
Results of the semiparametric specification test from Section 2.3 applied to the MEM errors ε_t. The reported p-values are based on the empirical distribution of the test statistic resulting from 500 simulated bootstrap samples.

DIS		**JNJ**	
T_n	P-Val.	T_n	P-Val
0.049	0.337	0.160	0.066

Table 4: Runs Test for Trade Indicator
Results of the two-sided runs test for serial dependence of the indicator for nonzero aggregated volumes. Under the null of no serial dependence, the statistic $Z = \frac{R-E(R)}{V(R)}$ (R: number of runs) is asymptotically standard normal.

DIS		**JNJ**	
Z	P-Val.	Z	P-Val
-6.523	0.000	-6.432	0.000

As a consequence of this assumption, the disturbances lose the i.i.d. property and, conditionally on $\mathcal{H}_{t-1}$, are independently but not identically distributed. Thus, the dynamics of the endogenous variable, y_t, are not fully captured by the conditional mean μ_t, as past information contained in $\mathcal{H}_{t-1}$ affects the innovation distribution. Similar generalizations of the MEM error structure have been considered, e.g., by Zhang et al. (2001) or Drost and Werker (2004). The resulting dynamic zero-augmented MEM (DZA-MEM) can be formally written as

$$y_t = \mu_t \, \varepsilon_t; \quad \varepsilon_t | \mathcal{H}_{t-1} \sim \text{i.n.i.d. } \mathcal{PMD}(1) \,, \tag{21}$$

where $\mathcal{PMD}(1)$ denotes a point-mass mixture as in (2) with assumption (1) replaced by (20) and $E[\varepsilon_t|\mathcal{H}_{t-1}] = E[\varepsilon_t] = 1$. Hence, the conditional density of ε_t given $\mathcal{H}_{t-1}$ is

$$f_\varepsilon(\varepsilon_t|\mathcal{H}_{t-1}) = (1-\pi_t)\,\delta(\varepsilon_t) + \pi_t \; g_\varepsilon(\varepsilon_t|\mathcal{H}_{t-1}) \; \mathbb{1}_{(\varepsilon_t>0)}, \tag{22}$$

where the conditional density for $\varepsilon_t > 0$, $g_\varepsilon(\varepsilon_t|\mathcal{H}_{t-1})$, depends on $\mathcal{H}_{t-1}$ through the probability π_t, as the unit mean assumption in (21) requires

$$\kappa_t := E[\varepsilon_t | \varepsilon_t > 0; \mathcal{H}_{t-1}] = \pi_t^{-1}, \tag{23}$$

such that

$$E[\varepsilon_t] = E\{E[\varepsilon_t|\mathcal{H}_{t-1}]\} = E[\pi_t \, \kappa_t] = 1. \tag{24}$$

Since the function $\pi(\cdot;\psi)$ is equivalent to a binary-choice specification for the trade indicator $\mathcal{I}_t$ defined in (19), the log-likelihood of the DZA-MEM consists of a MEM

and a binary-choice part,

$$\mathcal{L}(\vartheta) = \sum_{t=1}^{n} \{\mathcal{I}_t \ln \pi_t + (1 - \mathcal{I}_t) \ln (1 - \pi_t)\} + \sum_{t,nz} \{\ln f_\varepsilon(y_t/\mu_t | \mathcal{H}_{t-1}; \vartheta^g) - \ln \mu_t\}, \quad (25)$$

where $\vartheta = (\psi, \vartheta^g, \vartheta^\mu)'$ with ϑ^μ denoting the parameter vector of the conditional mean μ_t. As in the previous section, a separate optimization of the two parts is infeasible, since the constraint (23) implies that both components depend on the parameters of the binary-choice specification, ψ.

If we use the ZAF distribution as point-mass mixture $\mathcal{PMD}(1)$, we obtain the conditional density of ε_t given $\mathcal{H}_{t-1}$ as

$$f_\varepsilon(\varepsilon_t | \mathcal{H}_{t-1}) = (1 - \pi_t)\, \delta(\varepsilon_t) + \pi_t \, \frac{a\, \varepsilon_t^{a\,m-1} \left[\eta + (\varepsilon_t \, \pi_t \, \xi)^a\right]^{(-\eta-m)} \eta^\eta}{(\pi_t \, \xi)^{-a\,m} \, \mathcal{B}(m, \eta)} \, \mathbb{1}_{(\varepsilon_t > 0)}, \quad (26)$$

where we set $\lambda_t = (\pi_t \, \xi)^{-1}$, with ξ defined as in (18), to meet the constraint (23). The corresponding log-likelihood function is

$$\begin{aligned} \mathcal{L}(\vartheta) = \sum_{t=1}^{n} \{\mathcal{I}_t \ln \pi_t + (1 - \mathcal{I}_t) \ln (1 - \pi_t)\} + \sum_{t,nz} \Bigg\{ & \log a + (am - 1) \ln x_t \\ & - (\eta + m) \ln \left[\eta + \left(\frac{x_t}{\mu_t} \, \pi_t \, \xi\right)^a\right] + \eta \ln \eta - am \ln \left(\mu_t \, \pi_t^{-1} \, \xi^{-1}\right) - \ln \mathcal{B}(m, \eta) \Bigg\}, \end{aligned} \quad (27)$$

where $\vartheta = (\psi, a, m, \eta, \vartheta^\mu)'$.

3.3 Dynamic Models for the Trade Indicator

To allow the trade indicator $\mathcal{I}_t$ to follow a dynamic process, we propose two alternative specifications: a parsimonious autologistic specification and a more flexible parameterization using autoregressive conditional multinomial (ACM) dynamics as proposed by Russell and Engle (2005). By considering the general logistic link function

$$\pi_t = \pi(\mathcal{H}_{t-1}; \psi) = \frac{\exp(h_t)}{1 + \exp(h_t)}, \quad (28)$$

the autologistic specification for $h_t = \ln[\pi_t / (1 - \pi_t)]$ is given by

$$h_t = \theta_0 + \sum_{i=1}^{l} \theta_i \, \Delta_{t-i} + g_t, \quad \text{and } g_t = \sum_{i=1}^{d} \gamma_i \, \mathcal{I}_{t-i}, \quad (29)$$

where Δ_t denotes an indicator for large values of the endogenous variable of y_t and is defined as

$$\Delta_t := \max(y_t - \mathcal{I}_t, 0)\,. \tag{30}$$

This type of transformation was suggested in a similar setting by Rydberg and Shephard (2003) and accounts for the multicollinearity between the lags of y_t and $\mathcal{I}_t$. The autologistic model has advantages in terms of tractability, such as the concavity of the log-likelihood function, making numerical maximization straightforward. However, since this process does not include a moving average component, it is not able to capture persistent dynamics in the binary sequence. Therefore, as an alternative specification, we propose an ACM specification given by

$$h_t = \varpi + \sum_{j=1}^{v} \rho_j \, s_{t-j} + \sum_{j=1}^{w} \zeta_j \, h_{t-j}, \tag{31}$$

where

$$s_{t-j} = \frac{\mathcal{I}_{t-j} - \pi_{t-j}}{\sqrt{\pi_{t-j}\,(1-\pi_{t-j})}} \tag{32}$$

denotes the standardized trade indicator. The process $\{s_t\}$ is a martingale difference sequence with zero mean and unit conditional variance, which implies that $\{h_t\}$ follows an ARMA process driven by a weak white noise term. Consequently, $\{h_t\}$ is stationary if all values of z satisfying $1 - \zeta_1 z - \ldots - \zeta_w z^w = 0$ lie outside the unit circle. For more details, see Russell and Engle (2005).

An appealing feature of the ACM specification in the given framework is its similarity to a MEM. Actually, analogously to a MEM specification, it imposes a linear autoregressive structure for the logistic transformation of the probability π_t, which, in turn, equals the conditional mean of the trade indicator $\mathcal{I}_t$ given the restricted information set $\mathcal{H}_{t-1}$, i.e., $E\left[\mathcal{I}_t|\mathcal{H}_{t-1}\right]$.

The DZA-MEM dynamics can be straightforwardly extended by covariates which allow to test specific market microstructure hypotheses. Moreover, a further natural extension of the DZA-MEM is to allow for dynamic interaction effects between the conditional mean of y_t, μ_t, and the probability of zero values, π_t. For instance, by allowing for spillovers between both dynamic equations, the DZA-MEM can be modified

as

$$h_t = \varpi + \sum_{j=1}^{v} \rho_j \, s_{t-j} + \sum_{j=1}^{w} \zeta_j \, h_{t-j} + \sum_{j=1}^{m} \tau_j \, \mu_{t-j}, \tag{33}$$

$$\ln \mu_t = \omega + \sum_{i=1}^{p} \alpha_i \, \ln \varepsilon_{t-i} \, \mathbb{1}_{(x_{t-i}>0)} + \sum_{i=1}^{p} \alpha_i^0 \, \mathbb{1}_{(x_{t-i}=0)} + \sum_{i=1}^{q} \beta_i \, \ln \mu_{t-i} + \sum_{i=1}^{n} \varrho_i \, \pi_{t-i}.$$

In the resulting model, the intercepts ϖ and ω are not identified without additional restrictions. Hence, identification requires, for instance, $\varpi = 0$. Alternatively, or additionally, dynamic spillover effects might be also modeled by the inclusion of the lagged endogenous variables of the two equations, see, e.g., Russell and Engle (2005) in an ACD-ACM context.

3.4 Diagnostics

The evaluation of the DZA-MEM is complicated by the fact that the disturbances are not i.i.d. In particular, the non-identical distribution makes an application of the semiparametric specification test from Section 2.3 impossible. Moreover, since the disturbances are not i.i.d. even given the restricted information set $\mathcal{H}_{t-1}$, we cannot employ a transformation that provides standardized i.i.d. innovations as in De Luca and Zuccolotto (2006). As an alternative, we suggest evaluating the model based on density forecasts as developed by Diebold et al. (1998) and firstly applied to MEM-type models by Bauwens et al. (2004). One difficulty is that this method is designed for continuous random variables, while we have to deal with a discrete probability mass at zero. Therefore, following Liesenfeld et al. (2006) and Brockwell (2007), we employ a modified version of the test. The idea is to add random noise to the discrete component, making sure that the c.d.f. is invertible. Thus, we compute randomized probability integral transforms

$$z_t = \begin{cases} U_t \, F_Y(y_t|\mathcal{F}_{t-1}) & \text{if } y_t = 0, \\ F_Y(y_t|\mathcal{F}_{t-1}) & \text{if } y_t > 0, \end{cases} \tag{34}$$

where $F_Y(y_t|\mathcal{F}_{t-1})$ denotes the conditional c.d.f. of y_t implied by the DZA-MEM, while U_t are random variables with $\{U_t\}_{t=1}^n$ being i.i.d. $U(0,1)$. Using equation (22), we obtain

$$z_t = \begin{cases} U_t \, (1-\pi_t) & \text{if } y_t = 0, \\ (1-\pi_t) + \pi_t \, G_\varepsilon(y_t/\mu_t|\mathcal{H}_{t-1}) & \text{if } y_t > 0, \end{cases} \tag{35}$$

where $G_\varepsilon(y_t/\mu_t|\mathcal{H}_{t-1})$ is the conditional c.d.f. of the disturbances ε_t for $\varepsilon_t > 0$ evaluated at y_t/μ_t. For a DZA-MEM based on the ZAF distribution, it follows that

$$z_t = \begin{cases} U_t\ (1-\pi_t) & \text{if } y_t = 0, \\ (1-\pi_t) + \pi_t\ [\mathcal{B}(c\,;m,\eta)\,/\mathcal{B}(m,\eta)] & \text{if } y_t > 0, \end{cases} \tag{36}$$

where $\mathcal{B}(c\,;m,\eta) := \int_0^c t^{m-1}\,(1-t)^{\eta-1}\,dt$ is the incomplete beta function evaluated at

$$c := \left(y_t\,\mu_t^{-1}\,\pi_t\,\xi\right)^a \left[\eta + \left(y_t\,\mu_t^{-1}\,\pi_t\,\xi\right)^a\right]^{-1}. \tag{37}$$

If the conditional distribution of y_t is correct, the z_t sequence is i.i.d. $U(0,1)$, see Brockwell (2007) for a proof. While Diebold et al. (1998) recommend a visual inspection of the properties of the z_t's, we also check for uniformity using Pearson's χ^2-test.

3.5 Empirical Evidence on DZA-MEM Processes

We apply a DZA-MEM by parameterizing the conditional mean function μ_t based on the Log-MEM specification (17). The lag orders in both dynamic components are chosen according to the Schwarz information criterion. Table 5 shows the estimation results for the DZA-MEM with autologistic binary-choice component. No consistent effects for the large volume indicator m_t across the two stocks are found. Large volumes increase the trading probability in the case of DIS, while reducing it for JNJ. However, the lagged trade indicators are significantly positive in almost every case. Thus, trade occurrences are positively autocorrelated, which is in line with empirical studies on market microstructure (see, e.g., Engle, 2000).

For both stocks, all Q-statistics of the autologistic residuals

$$u_t = \frac{\mathcal{I}_t - \hat{\pi}_t}{\sqrt{\hat{\pi}_t\ (1-\hat{\pi}_t)}} \tag{38}$$

are significant at the 5% level, showing that an autologistic specification does not completely capture the dynamics and is too parsimonious.

As shown by Table 6, dynamically modeling trade occurrences by an ACM specification yields significantly lower Q-statistics. Hence, the ACM specification seems to fully capture the persistence in the trade indicator series, with the parameter estimates underlining the strong persistence in the process. Actually, the coefficient ζ_1 is close to unity for both stocks suggesting that the underlying process is very persistent.

We evaluate the model using the density forecast approach discussed in Section 3.4. Table 7 shows the results of the χ^2-test for uniformity of the randomized probability

Table 5: Estimation Results – DZA-MEM with Autologistic Component
Maximum likelihood estimates of the DZA-MEM based on the ZAF distribution with autologistic specification for the binary choice component. The Q-statistics refer to the residuals of the autologistic component. 5% critical values of the Q-statistics with 20, 50 and 100 lags are 31.410, 67.505 and 124.342, respectively. The autologistic residuals are defined as:

$$u_t = \frac{\mathcal{I}_t - \hat{\pi}_t}{\sqrt{\hat{\pi}_t\,(1-\hat{\pi}_t)}}.$$

	DIS			**JNJ**		
	Coef.	T-Stat.	P-Val.	Coef.	T-Stat.	P-Val.
ω	0.012	5.132	0.000	0.026	8.184	0.000
α_1	0.026	7.652	0.000	0.054	9.08	0.000
β_1	0.978	305.292	0.000	0.929	104.145	0.000
α_1^0	-0.009	-1.253	0.210	-0.020	-1.187	0.235
m	2.612	9.839	0.000	1.100	6.010	0.000
η	55.464	5.301	0.000	2.917	4.053	0.000
a	0.500	8.446	0.000	1.161	8.397	0.000
θ_0	0.285	2.412	0.016	0.953	4.352	0.000
θ_1	0.059	2.529	0.011	-0.083	-3.331	0.001
γ_1	0.511	7.584	0.000	1.089	7.981	0.000
γ_2	0.355	5.245	0.000	0.151	0.947	0.344
γ_3	0.107	1.580	0.114	0.591	3.957	0.000
γ_4	-0.093	-1.365	0.172	-	-	-
γ_5	0.203	2.980	0.003	-	-	-
γ_6	0.241	3.541	0.000	-	-	-
$\mathcal{L}$		-8729.730			-7833.532	
SIC		17591.371			15772.592	
$Q(20)$		42.387			35.483	
$Q(50)$		86.260			85.031	
$Q(100)$		173.606			174.753	

integral transforms (PITs) implied by the ZA-MEM and the DZA-ACM-MEM. For the former specification, the χ^2-statistics are significant at the 5% level. In the case of the DZA-ACM-MEM, the null hypothesis of a uniform distribution cannot be rejected at any conventional significance level for both stocks. For DIS, the χ^2-statistic nearly halves indicating an impressive performance improvement. These findings are underlined by the histograms of the PITs depicted in Figure 6. For both stocks, the PIT histogram implied by the DZA-ACM-MEM is close to a uniform distribution, while almost none of the bars is outside the 95% confidence region.

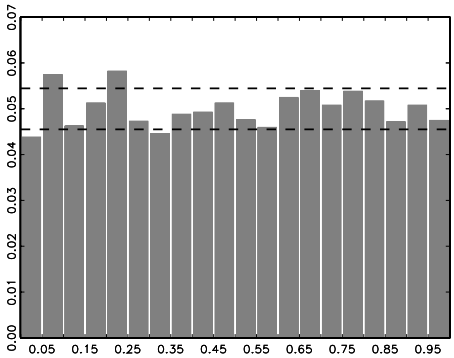
Table 6: Estimation Results – DZA-MEM with ACM Component
Maximum likelihood estimates of the DZA-MEM based on the ZAF distribution with ACM specification for the binary choice component. The Q-statistics refer to the residuals of the ACM component. 5% critical values of the Q-statistics with 20, 50 and 100 lags are 31.410, 67.505 and 124.342, respectively. The ACM residuals are defined as:

$$u_t = \frac{\mathcal{I}_t - \hat{\pi}_t}{\sqrt{\hat{\pi}_t\,(1-\hat{\pi}_t)}}.$$

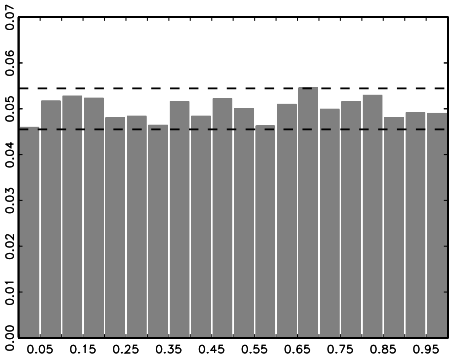
	DIS			**JNJ**		
	Coef.	T-Stat.	P-Val.	Coef.	T-Stat.	P-Val.
ω	0.013	5.721	0.000	0.026	8.374	0.000
α_1	0.025	7.780	0.000	0.054	8.987	0.000
β_1	0.979	327.030	0.000	0.930	104.084	0.000
α_1^0	-0.018	-2.588	0.010	-0.031	-1.825	0.068
m	2.652	9.493	0.000	1.109	6.555	0.000
η	55.332	4.121	0.000	2.949	4.398	0.000
a	0.495	17.256	0.000	1.152	9.176	0.000
ϖ	0.023	1.970	0.049	0.112	2.056	0.040
ρ_1	0.202	7.820	0.000	0.237	6.940	0.000
ρ_2	-0.138	-4.870	0.000	-0.146	-3.815	0.000
ζ_1	0.983	116.111	0.000	0.957	45.964	0.000
$\mathcal{L}$		-8726.831			-7838.810	
SIC		17550.398			15774.355	
$Q(20)$		30.672			20.425	
$Q(50)$		52.548			51.552	
$Q(100)$		109.633			124.276	

Table 7: χ^2-Test for Uniformity of the PITs
Results of the χ^2-test for uniformity of the randomized probability integral transforms of the estimated ZA-MEM and DZA-ACM-MEM specifications.

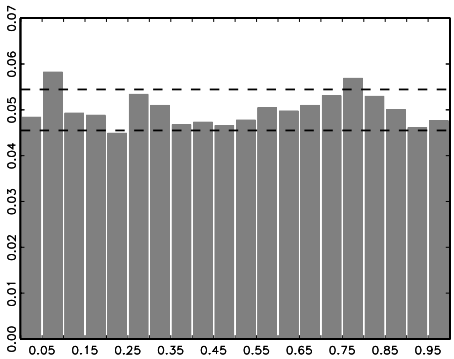
	DIS		**JNJ**	
	χ^2	P-Val.	χ^2	P-Val
ZA-MEM	38.883	0.005	30.719	0.043
DZA-ACM-MEM	15.035	0.720	22.386	0.265



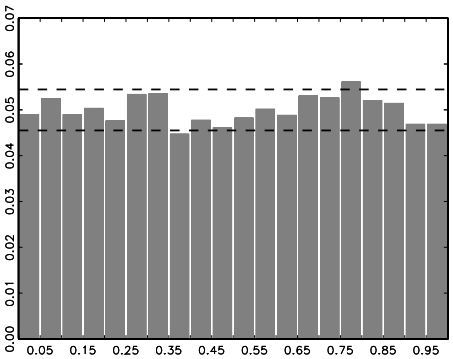
6.1: DIS, ZA-MEM



6.2: DIS , DZA-ACM-MEM



6.3: JNJ, ZA-MEM



6.4: JNJ, DZA-ACM-MEM

Figure 6: Histograms of Computed PIT-Sequences
Histograms of the randomized probability integral transforms of the estimated ZA-MEM and DZA-ACM-MEM specifications. The dashed lines represent 95% confidence bounds.

4 Conclusions

We introduce a new approach for modeling autoregressive positive-valued variables with excess zero outcomes. These properties are typical for both irregularly-spaced and time-aggregated financial high-frequency data and cannot be appropriately handled in extant approaches.

To capture the clustering of observations at zero, we propose a new point-mass mixture distribution, which consists of a discrete component at zero and a flexible continuous distribution for the strictly positive part of the support. To evaluate such a distribution, a novel semiparametric specification test tailored for point-mass mixture distributions is introduced. Finally, to accommodate serial dependencies in the data we incorporate the proposed point-mass mixture into a new type of multiplicative error model (MEM) capturing the dynamics of both zero occurrences and strictly positive values.

The empirical evidence based on cumulated trading volumes of two NYSE stocks shows that a zero-augmented MEM relying on the proposed point-mass mixture is able to capture the occurrence of excess zeros very well. The best fit is shown for a specification incorporating a two-state ACM component for the trade indicator. Besides MEM dynamics in the volumes, the model also explains the (own-standing) dynamics in trade occurrences and produces good density forecasts.

The model is sufficiently flexible to be extended in various ways, e.g., to allow for dynamic spillovers between the two types of dynamics or to incorporate regressors. Moreover, the model is straightforwardly extended to a multivariate framework, in the spirit of, e.g. Manganelli (2005), Cipollini et al. (2006) or Hautsch (2008).

Finally, the proposed ZAF distribution, its dynamic inclusion and the new semi-parametric test for point-mass mixtures could be relevant also in other areas, such as hydrology or genetics (see, e.g., Weglarczyk et al., 2005; Taylor and Pollard, 2009).

References

Allen, D., F. Chan, M. McAleer, and S. Peiris (2008): "Finite sample properties of the QMLE for the Log-ACD model: application to Australian stocks," *Journal of Econometrics*, 147, 163–185.

Bauwens, L., F. Galli, and P. Giot (2003): "The moments of Log-ACD models," CORE Discussion Paper.

Bauwens, L. and P. Giot (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., P. Giot, J. Grammig, and D. Veredas (2004): "A comparison of financial duration models via density forecasts," *International Journal of Forecasting*, 20, 589–609.

Bauwens, L. and N. Hautsch (2008): "Modeling financial high-frequency data with point processes," in *Handbook of Financial Time Series*, ed. by T. G. Andersen, R. A. Davis, J.-P. Kreiss, and T. Mikosch.

Brockwell, A. (2007): "Universal residuals: a multivariate transformation," *Stat Probab Lett.*, 77, 1473–1478.

Brownlees, C. T., F. Cipollini, and G. M. Gallo (2010): "Intra-daily volume modeling and prediction for algorithmic trading," *Journal of Financial Econometrics*, 8, 1–30.

Chen, S. (2000): "Probability density function estimation using gamma kernels," *Annals of the Institute of Statistical Mathematics*, 52, 471–480.

Cheng, M., J. Fan, and J. Marron (1997): "On automatic boundary corrections," *The Annals of Statistics*, 25, 1691–1708.

Cipollini, F., R. F. Engle, and G. M. Gallo (2006): "Vector multiplicative error models: representation and inference," NYU Working Paper No. Fin-07-048.

De Luca, G. and G. M. Gallo (2004): "Mixture Processes for Financial Intradaily Durations," *Studies in Nonlinear Dynamics & Econometrics*, 8.

De Luca, G. and P. Zuccolotto (2006): "Regime-switching pareto distributions for ACD models," *Computational Statistics & Data Analysis*, 51, 2179–2191.

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

Drost, F. C. and B. J. M. Werker (2004): "Semiparametric duration models," *Journal of Business and Economic Statistics*, 22, 40–50.

Duan, N. J., W. G. Manning, C. N. Morris, and J. P. Newhouse (1983): "A comparison of alternative models for the demand for medical care," *Journal of Business and Economic Statistics*, 1, 115–126.

Easley, D. and M. O'Hara (1992): "Time and the process of security price adjustment," *Journal of Finance*, 47, 577–605.

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

——— (2002): "New frontiers for ARCH models," *Journal of Applied Econometrics*, 17, 425–446.

Engle, R. F. and G. M. Gallo (2006): "A multiple indicators model for volatility using intra-daily data," *Journal of Econometrics*, 131, 3–27.

Engle, R. F. and J. Russell (1998): "Autoregressive conditional duration: a new model for irregularly spaced transaction data," *Econometrica*, 66, 1127–1162.

Fan, Y. (1994): "Testting the goodness of of fit of a parametric density function by kernel method," *Econometric Theory*, 10, 316–356.

——— (1998): "Goodness-of-fit tests based on kernel density estimators with fixed smoothing parameters," *Econometric Theory*, 14, 604–621.

Fernandes, M. and J. Grammig (2005): "Nonparametric specification tests for conditional duration models," *Journal of Econometrics*, 127, 35–68.

Fernandes, M. and P. Monteiro (2005): "Central limit theorem for asymmmetric kernel functionals," *Annals of the Institute of Statistical Mathematics*, 57, 425–442.

Gasser, T. and H. Müller (1979): "Kernel estimation of regression functions," in *Lecture Notes in Mathematics 757*, ed. by T. Gasser and M. Rosenblatt, Heidelberg: Springer, 23–68.

Hagmann, M. and O. Scaillet (2007): "Local multiplicative bias correction for asymmetric kernel density estimators," *Journal of Econometrics*, 141, 213–249.

Hautsch, N. (2003): "Assessing the Risk of Liquidity Suppliers on the Basis of Excess Demand Intensities," *Journal of Financial Econometrics*, 1, 189–215.

——— (2004): *Modelling Irregularly Spaced Financial Data: Theory and Practice of Dynamic Duration Models*, Berlin: Springer.

——— (2008): "Capturing common components in high-frequency financial time series: A multivariate stochastic multiplicative error model," *Journal of Economic Dynamics & Control*, 32, 3978–4009.

Hjort, N. L. and I. K. Glad (1995): "Nonparametric density estimation with a parametric start," *The Annals of Statistics*, 23, 882–904.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1994): *Continuous Univariate Distributions, Volumes I and II*, New York: John Wiley and Sons, second ed.

Jones, M. (1993): "Simple boundary correction for kernel density estimation," *Statistics and Computing*, 3, 135–146.

Lancaster, T. (1997): *The Econometric Analysis of Transition Data*, Cambridge: Cambridge University Press.

Lanne, M. (2006): "A mixture multiplicative error model for realized volatility," *Journal of Financial Econometrics*, 4, 594–616.

Liesenfeld, R., I. Nolte, and W. Pohlmeier (2006): "Modeling financial transaction price movements: a dynamic integer count model," *Empirical Economics*, 30, 795–825.

Manganelli, S. (2005): "Duration, Volume and volatility impact of trades," *Journal of Financial Markets*, 8, 377–399.

Russell, J. R. and R. F. Engle (2005): "A discrete-state continuous-time model of financial transactions prices and times: the autoregressive conditional multinomial-autoregressive conditional duration model," *Journal of Business & Economic Statistics*, 23, 166–180.

Rydberg, T. and N. Shephard (2003): "Dynamics of trade-by-trade price movements: decomposition and models," *Journal of Financial Econometrics*, 1, 2–25.

Schuster, E. (1958): "Incorporating support constraints into nonparametric estimators of densities," *Communications in Statistics, Part A - Theory and Methods*, 14, 1123–1136.

Taylor, S. and K. Pollard (2009): "Hypothesis tests for point-mass mixture data with application to omics data with many zero values," *Statistical Applications in Genetics and Molecular Biology*, 8, 1–43.

Weglarczyk, S., W. G. Stupczewski, and V. P. Singh (2005): "Three parameter discontinuous distributions for hydrological samples with zero values," *Hydrological Processes*, 19, 2899–2914.

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

Zhang, S. (2010): "A note on the performance of gamma kernel estimators at the boundary," *Statistics and Probability Letters*, 80, 548–557.

CFS Working Paper Series:

No.	Author(s)	Title
2010/18	**Horst Entorf** **Christian Knoll** **Liliya Sattarova**	Measuring Confidence and Uncertainty during the Financial Crisis: Evidence from the CFS Survey
2010/17	**Nikolaus Hautsch** **Mark Podolskij**	Pre-Averaging Based Estimation of Quadratic Variation in the Presence of Noise and Jumps: Theory, Implementation, and Empirical Evidence
2010/16	**Tullio Jappelli**	Economic Literacy: An International Comparison
2010/15	**Bartholomäus Ende** **Marco Lutat**	Trade-throughs in European Cross-traded Equities After Transaction Costs – Empirical Evidence for the EURO STOXX 50 –
2010/14	**Terrence Hendershott** **Albert J. Menkveld**	Price Pressures
2010/13	**Sarah Draus**	Does Inter-Market Competition Lead to Less Regulation?
2010/12	**Kai-Oliver Maurer** **Carsten Schäfer**	Analysis of Binary Trading Patterns in Xetra
2010/11	**Annamaria Lusardi** **Olivia S. Mitchell**	How Ordinary Consumers Make Complex Economic Decisions: Financial Literacy and Retirement Readiness
2010/10	**Annamaria Lusardi** **Daniel Schneider** **Peter Tufano**	The Economic Crisis and Medical Care Usage
2010/09	**Annamaria Lusardi** **Olivia S. Mitchell** **Vilsa Curto**	Financial Literacy among the Young: Evidence and Implications for Consumer Polic

Copies of working papers can be downloaded at http://www.ifk-cfs.de