

Modeling corporate defaults: Poisson autoregressions with exogenous covariates (PARX)*

Arianna Agosto[†] Giuseppe Cavaliere[‡] Dennis Kristensen[§]
Anders Rahbek[¶]

*We thank Bent J. Christensen, Richard Davis, Luca de Angelis, David Lando, Offer Lieberman, Peter C.B. Phillips, Enrique Sentana, as well as participants at the “Recent Developments in Financial Econometrics and Empirical Finance” conference held in 2014 at the University of Essex, the 2013 C.R.E.D.I.T. conference, the 2014 (EC)² conference, the 6th Italian Congress of Econometrics and Empirical Economics (ICEEE 2015), the 2015 World Congress of the Econometric Society as well as seminar/workshop participants at Columbia University, Durham University, Imperial College, Tsinghua University, Hull University, Sungkyunkwan University, University of Helsinki, University of Tasmania and University of York, for useful comments. We are also indebted to two anonymous referees for their extremely careful reading of a previous draft of the paper. The authors acknowledge support from Center for Research in Econometric Analysis of Time Series (DNR78), funded by the Danish National Research Foundation. Cavaliere and Rahbek thank the Italian Ministry of Education, University and Research (MIUR), PRIN project “Multivariate statistical models for risk assessment”, for financial support. Rahbek acknowledges research support by the Danish Council for Independent Research, Sapere Aude – DFF Advanced Grant (Grant no.: 12-124980). Kristensen acknowledges research support by the ESRC through the ESRC Centre for Microdata Methods and Practice grant RES-589-28-0001 and the European Research Council (grant no. ERC-2012-StG 312474). We also thank Moody’s Investors Service for providing us with data.

[†]Financial Risk Control Unit, Banca Carige, Via Cassa di Risparmio 15 - 16123 Genova, Italy. E-mail: agosto.arianna@gmail.com.

[‡]Corresponding author. Department of Statistical Sciences, University of Bologna, Via Belle Arti 41, I-40126 Bologna, Italy; Department of Economics, University of Copenhagen. E-mail: giuseppe.cavaliere@unibo.it.

[§]Department of Economics, University College London, WC1E 6BT, United Kingdom; Institute of Fiscal Studies; CREATES, University of Aarhus. E-mail: d.kristensen@ucl.ac.uk.

[¶]Department of Economics, University of Copenhagen, 1353 Copenhagen K, Denmark; CREATES, University of Aarhus. E-mail: anders.rahbek@econ.ku.dk.

Abstract: We develop a class of Poisson autoregressive models with exogenous covariates (PARX) that can be used to model and forecast time series of counts. We establish the time series properties of the models, including conditions for stationarity and existence of moments. These results are in turn used in the analysis of the asymptotic properties of the maximum-likelihood estimators of the models. The PARX class of models is used to analyze the time series properties of monthly corporate defaults in the US in the period 1982-2011 using financial and economic variables as exogenous covariates. Results show that our model is able to capture the time series dynamics of corporate defaults well, including the well-known default counts clustering found in data. Moreover, we find that while in general current defaults do indeed affect the probability of other firms defaulting in the future, in recent years economic and financial factors at the macro level are capable to explain a large portion of the correlation of US firms defaults over time.

Keywords: corporate defaults, count data, exogenous covariates, Poisson autoregression, estimation.

JEL codes: C13, C22, C25, G33.

1 Introduction

There is a strong ongoing interest in modelling and forecasting time series of corporate defaults. A stylized fact of defaults is that they tend to cluster over time. The default clustering phenomenon has been explored in the financial literature, giving rise to a debate about its causes, with several works trying to distinguish between “contagion effects”, by which “one firm’s default increases the likelihood of other firms defaulting” (Lando and Nielsen, 2010), and “systematic risk”, where comovements in corporate solvency are caused by common underlying macroeconomic and financial factors; see, for example, Das *et al.* (2007) and Lando and Nielsen (2010) who investigate the role of systematic risk in default correlations amongst US corporations.

We contribute to this debate by proposing a novel class of dynamic Poisson models for describing and forecasting aggregate number of corporate defaults; that is, the number of defaults within a given time period. We call this new class of models Poisson AutoRegressions with eXogeneous covariates (PARX). PARX models are an extension of the Poisson autoregression in Fokianos, Rahbek and Tjøstheim (2009) [FRT hereafter], which is here augmented by including – in addition to lagged intensity and counts – a set of exogenous covariates as predictors. This class of models provides a flexible framework within which we are able to analyze dependence of default probabilities on past number of defaults as well as on relevant financial and economic variables. These additional predictors are meant to summarize the level of uncertainty during periods of financial turmoil and/or economic downturns; that is, when corporate defaults are more likely to cluster together. We also consider the impact of auxiliary information on the estimates of persistence parameters which express the degree of dependence on the past history of the process.

Our approach to modelling defaults complements existing studies. These can broadly be divided into two categories. In the first category, firm-level data are available where default times for a cross-section of firms are recorded together with various firm-specific covariates. The default times are normally modelled by Poisson processes with macroeconomic and firm-specific covariates entering the default intensities; see, *e.g.*, Das *et al.* (2007) and Lando and Nielsen (2010). These types of models do not allow for direct modelling of contagion and only allows for indirect evidence of contagion by testing whether the Poisson model is misspecified. In the second category, to which this paper belongs, aggregate data are used where the number of defaults within a given period is observed together with various macroeconomic variables. Two recent papers in this category are Koopman, Lucas and Schwaab (2012) and Azizpour, Giesecke and Schwenkler (2015). Koopman *et al.* (2012) model default counts using a binomial specification where, similar to the PARX model, the probability of default is a time-varying functions of underlying factors. Similar to so-called frailty models, their specification involve unobserved components which have to be

integrated out in the estimation, which is generally done using computationally burdensome Monte Carlo methods. In contrast, PARX models are observation-driven in that they do not involve latent state variables. This in turn means that estimation and forecasting do not require any sophisticated numerical techniques, and is straightforward to implement in standard software packages. In particular, they can easily handle large number of exogenous covariates.

Our empirical analysis using the PARX model provides new insights into the dynamics of corporate defaults among Moody's rated US firms during the period 1982-2011. Various macroeconomic and financial variables, meant to capture the state of the US economy and financial markets, are included to investigate whether corporate defaults are driven by economic fundamentals and/or contagion effects during this period. We find that important explanators of corporate defaults are the overall volatility of the US stock market and the Leading Index of the US economy, but that contagion effects are also present in the dynamics. A structural break analysis shows that these relationships are not stable over time though and that the relative importance of the different factors have been changing over the sample period. Interestingly, we find that the contagion effects have been diminishing over time and that corporate defaults during the recent financial crisis were mostly driven by macroeconomic and financial fundamentals.

This paper also contributes to the literature on econometric and statistical analysis of Poisson autoregressions. First, we provide new results on the time series properties of PARX models, including conditions for stationarity and existence of moments. Second, we provide an asymptotic theory for the maximum likelihood estimators (MLE's) of the parameters entering the model. These results extend and complement the ones found in, among others, Rydberg and Shephard (2000), Streett (2000), Ferland *et al.* (2006) and FRT who analyze the properties of the MLE's for Poisson Autoregressive (PAR) models without covariates. Compared to these papers, we take a very different approach to establishing the asymptotic properties. Most notably, in order to establish a Law of Large Numbers [LLN] and a Central Limit Theorem [CLT] for the PARX process, we utilize the concept of τ -weak dependence (Doukhan and Winterberger, 2008). This is a relatively new stability concept which proves to be simpler to verify for discrete-valued Markov chains compared to existing stability concepts such as geometric ergodicity. This means, for example, that we avoid to deal with an augmented model where an additional error component is introduced, as done in FRT, in the asymptotic analysis. As such, our theory is completely novel.

PARX models are also related to a recent literature on GARCH models augmented by additional covariates with the aim of improving the forecast performance. These models include GARCH-X models, the so-called HEAVY model as proposed by Shephard and Shephard (2010), and the Realized GARCH model of Hansen *et al.* (2012); see also Han and

Kristensen (2014) for econometric analysis of such models. In these models, the time-varying volatility is explained by past returns and volatilities together with additional covariates, usually a realized volatility measure. PARX share the same motivation and modelling approach, but the variable of interest in our case is discrete and so the technical analysis and the applications are different.

The paper is organized as follows. In Section 2 we introduce the class of PARX models and discuss them in relation to existing models, as well as to the literature on default clustering and contagion. Time series properties of the models are investigated in Section 3. Maximum-likelihood based inference and methods for forecasting with PARX models are presented in Section 4. Specifically, large-sample properties of the maximum likelihood estimator are derived in Section 4.1, while its finite sample properties are studied in Section 4.2 through Monte Carlo simulations. Moreover, Section 4.3 illustrates how the estimated PARX specification can be used for forecasting purposes. Section 5 contains the empirical analysis of US default counts. Section 6 concludes. All auxiliary lemmas and mathematical proofs are contained in the Appendix.

2 Modelling Defaults with PARX

We here set up a general dynamic model for time series count data, motivated by the empirical application where we analyze the dynamics of US corporate defaults. Let $y_t \in \{0, 1, 2, \dots\}$, $t \geq 1$, be a time series of counts, such as the number of corporate defaults in a given period, say, a month. We then wish to model the dynamics of this process both in terms of its own past, y_{t-1}, y_{t-2}, \dots , but also in terms of d_x additional covariates $x_t := (x_{1t}, x_{2t}, \dots, x_{d_x t})' \in \mathbb{R}^{d_x}$. In the empirical analysis of Section 6 these include relevant macroeconomic and financial factors such as realized volatility measures, recession indicators, and measures of economic activity and financial stability. We do so by modelling y_t as a conditional Poisson distribution with time-varying intensity, λ_t , expressed as a function of past counts and covariates. That is,

$$y_t | \mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t), \quad t = 1, 2, \dots, T \quad (1)$$

where \mathcal{F}_{t-1} denotes the σ -field $\sigma\{y_{-p+1}, \dots, y_{t-1}, \lambda_{-q+1}, \dots, \lambda_{t-1}, x_0, \dots, x_{t-1}\}$ and $\text{Poisson}(\lambda)$ denotes a Poisson random variable with intensity parameter λ . To close the model, we propose the following specification for λ_t ,

$$\lambda_t = \omega + \sum_{i=1}^p \alpha_i y_{t-i} + \sum_{i=1}^q \beta_i \lambda_{t-i} + f(x_{t-1}, \gamma). \quad (2)$$

Note here that x_t enters the intensity through a non-negative link function $f(\cdot; \gamma) : \mathbb{R}^{d_x} \rightarrow [0, \infty)$ as chosen by the researcher; the link function is introduced to allow for possibly

negative covariates.

The parameters of interest are given by $\omega > 0$, $\alpha_i \geq 0$ ($i = 1, 2, \dots, p$) and $\beta_i \geq 0$ ($i = 1, 2, \dots, q$), together with the additional vector of parameters γ entering the function f . A possible specification of the function f , which will be extensively used in the empirical analysis of Section 5, is the additive one,

$$f(x, \gamma) := \sum_{i=1}^{d_x} \gamma_i f_i(x_i), \quad (3)$$

where $f_i : \mathbb{R} \mapsto [0, \infty)$, $i = 1, \dots, d_x$, are known functions, while $\gamma := (\gamma_1, \dots, \gamma_{d_x})' \in [0, \infty)^{d_x}$ is a vector of unknown parameters. Note that, without loss of generality, only one lag of x_t is included in the specification of λ_t since multiple, say m , lags of a given set of variables, z_t , can be included by simply stacking them into a vector of the form $x_{t-1} := (z_{t-1}, \dots, z_{t-m})'$. Observe, finally, that with $f(x_{t-1}, \gamma) \equiv 0$ the model reduces to the Poisson autoregression (PAR) considered in FRT. However, in general, the inclusion of additional covariates x_t will improve on in- and out-of-sample performance of the model and provide further insights into how exogenous covariates affect the dynamics of default counts.

The above specification allows for flexible dynamics of the number of counts in terms of past counts, captured by $\sum_{i=1}^p \alpha_i y_{t-i}$, and exogenous factors, as described by $f(x_{t-1}, \gamma)$. The term $\sum_{i=1}^q \beta_i \lambda_{t-i}$ is a parsimonious way of incorporating a large number of lags of these two components in the intensity equation, in a fashion similar to the extension of standard ARCH processes to the general GARCH process (or, similarly, to the extension of AR time series to ARMA processes). To see this, consider, for simplicity, the case of $p = q = 1$: If $\alpha_1 + \beta_1 < 1$ is satisfied together with other regularity conditions, then there exists a stationary solution to the PARX model (see Section 3), which can be represented by

$$\lambda_t = \omega(1 - \beta_1)^{-1} + \alpha_1 \sum_{i=1}^{\infty} \beta_1^{i-1} y_{t-i} + \sum_{i=1}^{\infty} \beta_1^{i-1} f(x_{t-i}, \gamma). \quad (4)$$

Thus, $\beta > 0$ allows modelling dependence of λ_t on all past lags of exogenous regressors and counts without having to introduce a large number of parameters.

Finally, we note that the PARX model share some similarities with the GARCH model with exogenous covariates, or GARCH-X; see Han and Kristensen (2014) and references therein. Specifically, in GARCH-X specifications y_t is a given return, whose conditional volatility – say h_t – follows

$$h_t = \omega + \sum_{i=1}^p \alpha_i y_{t-i}^2 + \sum_{i=1}^q \beta_i h_{t-i} + f(x_{t-1}, \gamma),$$

where x_t is a set of covariates. Special cases of the GARCH-X model are the so-called HEAVY model of Shephard and Sheppard (2010), and the realized GARCH model of Hansen *et al.*

(2012), where the exogenous variable x_{t-1} is a (realized) measure of past volatility obtained from high-frequency data.

While parts of the structure of GARCH-X type models are similar to that of the PARX model, a crucial difference is that while the former class of models is designed to capture the evolution of the (conditional) variance of a continuously distributed variable, the latter is modelling the full distribution of a count process. This also means that for the theoretical analysis of PARX models, new tools have to be developed. In section 3 below we develop one such set of tools by establishing conditions for stationarity and ergodicity which in turn can be used to derive a LLN and a Martingale CLT for PARX processes.

2.1 Related Literature

There is a large existing literature on modelling corporate defaults using firm specific data; see, *e.g.*, Das et al (2007), Duffie, Eckner, Horel and Saita (2009), Duffie, Saita and Wang (2007), and Lando and Nielsen (2010). This literature has mostly employed duration models where the default of firm i ($i = 1, \dots, n$) occurs at the first arrival time, τ_i , of a counting process $N_i(s)$ with intensity $\lambda_i(s)$, $s \geq 0$. Suppose that the counting processes N_i , $i = 1, \dots, n$, are so-called “doubly stochastic” (see, *e.g.*, Das *et al.*, 2007, Sec. I); that is, conditional on the intensities, they are mutually independent Poisson processes. Then the number of defaults within a given month $(t - 1, t]$, that is, the count variable $y_t := \#\{i : t - 1 < \tau_i \leq t\}$, follows a Poisson distribution with intensity

$$\lambda_t = \int_{t-1}^t \sum_{i=1}^n \lambda_i(s) \mathbb{I}\{\tau_i > s\} ds, \quad (5)$$

where $\mathbb{I}\{\cdot\}$ is the indicator function. In particular, this shows that the model of Das *et al.* (2007), amongst others, implies that aggregate default counts will satisfy eq. (1) and so is in agreement with our baseline PARX specification.

Suppose, moreover, that the intensity of firm i is affected by observed firm-specific, say $X_{1,i}(s)$, and economy-wide, say $X_2(t)$, covariates. Popular choices of $X_{1,i}$ include “distance to default” and stock return of firm i , while $X_2(t)$ include variables such as the US treasury bill rate, the S&P 500 return, and so forth. Note here that the doubly stochastic assumption implies that the included covariates are completely exogenous relative to the n counting processes. One specification of $\lambda_i(s)$ that allows for this analysis is

$$\lambda_i(s) = g(\omega + \beta'_1 X_{1,i}(s) + \beta'_2 X_2(s)), \quad (6)$$

for some known function g . Thus, in general this would imply that λ_t in (5) would depend on aggregated firm specific and economy-wide covariates, say x_{1t} and x_{2t} , as well as on the past default count, y_{t-1} . The PARX model is one particular specification of λ_t , and

can therefore be interpreted as an approximation of the aggregate λ_t obtained from the underlying firm-specific default model of Das *et al.* (2007), among others. Importantly, the above aggregation result shows that if the main focus of the analysis is to gain understanding of how macro-level factors and past defaults affect default probabilities, it suffices to model the aggregate number of defaults instead of individual firms' defaults. At the same time, to understand the impact of firm-specific variables on aggregate defaults, we need to obtain aggregate data on these variables.

PARX specifies λ_t as an observation-driven process. Alternatively, one can use a state space model to describe the evolution of λ_t over time. This approach is pursued by Azizpour, Giesecke and Schenkler (2015) and Giesecke and Kim (2011), amongst others, who also assume that aggregate counts satisfy eq. (1) but then proceeds to model the intensity as

$$\lambda_t = \int_{t-1}^t g(\omega + \beta'_2 X_2(s)) + \beta_3 Y(s) + \beta_4 Z(s) ds. \quad (7)$$

Here, $Y(s) = f(\{(\tau_i, d_i) : \tau_i \leq s\})$, with d_i denoting the face value of firm i 's defaulted debt, is an observed process depending on past defaults, while $Z(s)$ is a latent Markov process (the so-called "frailty") that, after controlling for observables, captures any time series dynamics in the default intensity. The frailty process $Z(s)$ captures unobserved risk – it is a common underlying factor whose clustering over time generates clustering in defaults in addition to the impact of $X_2(s)$ – while $Y(s)$ captures contagion in that past default affects its own evolution. Thus, the difference between PARX and the model of Azizpour, Giesecke and Schenkler (2015) is similar to the difference between a GARCH and a stochastic volatility model. Both approaches, however, provide an empirical device to assess the existence of default clustering channels, such as the exposure to macroeconomic and/or financial factors, and the impact of past default events.

Finally, it is worth noticing that in a recent paper, Koopman, Lucas and Schwaab (2012) model default counts in a similar fashion to Azizpour *et al.* (2015), except that they replace the baseline Poisson distribution with a negative binomial distribution (see also Koopman, Lucas and Schwaab, 2014). The underlying parameters of this discrete distribution are then modelled as a time-varying, depending on underlying economic factors and past defaults, similar to what we do here.

2.2 Contagion and systematic risk within PARX

Through the lens of the PARX model, we can differentiate between "systematic" risk, where the default probability of a given firm is affected by a set of common economic and financial risk factors, and feedback effects (or "contagion"), where current number of defaults affect the probability of other firms' future defaults, conditionally on the common factors. More

specifically, and again focusing on the PARX(1, 1) model for notational convenience, we may interpret $\sum_{i=1}^{\infty} \beta_1^{i-1} f(x_{t-i}, \gamma)$ in (4) as the risk component attributable to common macroeconomic and financial factors, while $\alpha_1 \sum_{i=1}^{\infty} \beta_1^{i-1} y_{t-i}$ captures possible feedback effects. More generally, one can interpret the value of $\sum_{i=1}^{\max\{p,q\}} \{\alpha_i + \beta_i\}$, when $\sum_{i=1}^{\max\{p,q\}} \alpha_i > 0$, as a measure of the level of dynamic contagion, since large values of $\sum_{i=1}^{\max\{p,q\}} \{\alpha_i + \beta_i\}$ imply that past defaults have a large impact on current default probabilities, after controlling for the covariates x_t . In the extreme case when $\alpha_1 + \dots + \alpha_p = 0$, the model implies conditional (on x_{t-1}) independence between current and past defaults.

It should be pointed out that our definition of contagion is specific to the PARX model and other definitions made in terms of alternative models for defaults can be found in the literature. It is worthwhile noting that Das *et al.* (2006) and Lando and Nielsen (2010) do not provide a precise definition of contagion: rather, they merely test for whether the aforementioned “doubly stochastic” assumption is supported by data or not. Conditional on all relevant covariates/risk factors having been included in their model, they attribute rejection to the presence of contagion. This is a very broad definition which basically labels any type of deviations from the “doubly stochastic” assumption as contagion. In contrast, we here, more precisely define it as the situation where past defaults affect current defaults. This measure broadly corresponds to the so-called “feedback channel” in the model of Azizpour *et al.* (2015), as discussed earlier.

Our measure of contagion may in some situations be misleading. First, it relies on the assumption that all relevant covariates, x_t , are available and so observed. If not all relevant covariates have been included, the model will be misspecified and the estimated parameters suffer from biases. In particular, in this situation, we expect the estimated α 's to be upward biased since the component $\sum_{i=1}^p \alpha_i y_{t-i}$ in eq. (2) will soak up the unexplained time series dependence generated by the missing covariate. Again, this is not specific to our approach, with the same issue being present in the framework of Das *et al.* (2006) and Lando and Nielsen (2010), among others. In fact, one of the main points of Lando and Nielsen (2010) is that by changing the specification of the firm-specific intensity employed by Das *et al.* (2006), the contagion effects reported in Das *et al.* (2006) vanish. Second, the above measure ignores feedback effects from defaults to covariates: Suppose that x_t is affected by past defaults; in this case past defaults will affect x_t which in turn will affect future defaults. That is, contagion may take place indirectly through covariates. So to get a complete picture of contagion, we would need to specify a dynamic model for x_t that incorporates potential dependence on lagged values of y_t .

3 Properties of PARX processes

In this section we provide sufficient conditions for a PARX process to be stationary and ergodic with polynomial moments of a given order. This result in turn gives us access to a LLN and CLT for (y_t, x_t) which will be used to analyze the asymptotic properties of the MLE in Section 4.

The analysis is carried out by applying results on so-called τ -weak dependence, henceforth weak dependence, recently developed in Doukhan and Wintenberger (2008). Weak dependence is a stability concept for Markov chains that implies stationarity and ergodicity and so allows us to establish, amongst other things, a (uniform) LLN for the process. It is related to alternative concepts of stability and mixing of time series such as (geometric) ergodicity (see, for example, FRT) but it is simpler to verify for discrete-valued data. Christou and Fokianos (2013) employ the same techniques in the analysis of a class of negative binomial time series models.

Weak dependence basically requires that the time series satisfies a certain Lipschitz condition (in the L^s -norm, $s \geq 1$). To establish this property for the PARX model, it is useful to rewrite the Poisson model (1) in terms of an i.i.d. sequence of Poisson processes with unit intensity, see FRT, p.1431. Specifically, for each t let $N_t(\cdot)$ be a Poisson process of unit intensity. Since for any $u > 0$, the number of events $N_t(u)$ in the interval $[0, u]$ is distributed as a Poisson random variable with intensity u , we can restate (1) in terms of $N_t(\cdot)$ as

$$y_t = N_t(\lambda_t). \quad (8)$$

where $N_t(\cdot)$ is i.i.d. over time. We complete the model by imposing a Markov-structure on the set of covariates; that is,

$$x_t = g(x_{t-1}, \varepsilon_t), \quad (9)$$

for some function $g(x, \varepsilon)$ and with ε_t being an i.i.d. error term. The above structure could be generalized to $x_t = g(x_{t-1}, \dots, x_{t-m}, \varepsilon_t)$ for some $m \geq 1$, thereby allowing for more flexible dynamics of the covariates included in the model; see the discussion in Section 2. However, we maintain eq. (9) for simplicity in the following.

We then impose the following assumptions on the complete model.

Assumption 1 (Markov) The innovations ε_t and $N_t(\cdot)$ are jointly i.i.d. over time.

Assumption 2 (Exogenous stability) $E[\|g(x; \varepsilon_t) - g(\tilde{x}; \varepsilon_t)\|^s] \leq \rho \|x - \tilde{x}\|^s$, for some $\rho < 1$, and $E[\|g(0; \varepsilon_t)\|^s] < \infty$, for some $s \geq 1$.

Assumption 3 (PARX stability) (i) $\sum_{i=1}^{\max(p,q)} (\alpha_i + \beta_i) < 1$ and (ii) $|f(x; \gamma) - f(\tilde{x}; \gamma)| \leq L \|x - \tilde{x}\|$, for some $L > 0$.

Assumption 1 implies that (y_t, x_t) can be embedded in a Markov chain and so we can employ the theory of weak dependence. Notice that Assumption 1 does not require ε_t and $N_t(\cdot)$ to be independent; on the contrary, contemporaneous dependence between current counts and innovations to the exogenous variables is allowed. Assumption 2 imposes a Lipschitz condition on $g(x, \varepsilon)$ w.r.t. x which is satisfied for many popular time series models such as (stable) linear autoregressive ones. This assumption is used to show, as a first step, that x_t is weakly dependent. Finally, Assumption 3(i) implies that the function $L(y, \lambda) = \omega + \sum_{i=1}^p \alpha_i y_i + \sum_{i=1}^q \beta_i \lambda_i$, where $y = (y_1, \dots, y_p)$ and $\lambda = (\lambda_1, \dots, \lambda_q)$, is Lipschitz with Lipschitz coefficient $\sum_{i=1}^{\max(p,q)} (\alpha_i + \beta_i)$ smaller than one. It is identical to the conditions imposed in FRT for the Poisson autoregressive model (without exogenous regressors and with $p = q = 1$) to be stationary. Assumption 3(ii) restricts how x_t can enter the Poisson intensity; it requires f to be Lipschitz and so excludes certain functions, such as the exponential one. This assumption will, however, be weakened at the end of this section.

Together the three assumptions imply that the PARX model admits a stationary and weakly dependent solution, as shown in the following theorem.

Theorem 1 *Under Assumptions 1–3, there exists a weakly dependent stationary and ergodic solution to eqs. (1)–(2) and (9), which we denote $X_t^* = (y_t^*, \lambda_t^*, x_t^{*'})'$, satisfying $E[\|X_t^*\|^s] < \infty$ with $s \geq 1$ given in Assumption 2.*

The above theorem complements the results of FRT, who derive sufficient conditions for an approximate Poisson Autoregression to be geometrically ergodic. We here allow for exogenous variables to enter the model, and provide sufficient conditions for weak dependence directly for this extended model.

One particular consequence of the above theorem is that the expected long-run number of defaults equals

$$E[y_t] = E[\lambda_t] = \mu = \frac{\omega + E[f(x_{t-1})]}{1 - \sum_{i=1}^{\max(p,q)} (\alpha_i + \beta_i)},$$

and furthermore, that $\text{Var}[y_t] > E[y_t]$. Thus, by including past values of the response as well as covariates in the evolution of intensity, PARX models generate overdispersion in the marginal distribution, a feature that is prominent in many count time series, including corporate defaults.

One further consequence of Theorem 1 is that it gives us access to the (weak) LLN for stationary and ergodic processes, $T^{-1} \sum_{t=1}^T h(X_t^*) \xrightarrow{P} E[h(X_t^*)]$ for *any* function $h(\cdot)$ of $X_t = (y_t, \lambda_t, x_t')$, provided $E[\|h(X_t^*)\|] < \infty$. In the asymptotic theory of the proposed estimators, the computation of the likelihood function is based on a set of fixed initial values for the Poisson intensity. In order to analyze the asymptotic behavior of the likelihood function in this setting, we need to generalize the LLN result to hold for any solution with

arbitrary initialization. This extension is stated in the following lemma due to Kristensen and Rahbek (2015):

Lemma 1 *Let $\{X_t\}$, $t = 0, 1, 2, \dots$, be a process with arbitrary initial value X_0 and, for $t \geq 1$, satisfying the equation $X_t = F(X_{t-1}; \xi_t)$ with ξ_t and i.i.d. sequence. Moreover, assume that $E[\|F(x; \xi_t) - F(\tilde{x}; \xi_t)\|^s] \leq \rho \|x - \tilde{x}\|$ and $E[\|F(0; \xi_t)\|^s] < \infty$ for some $s \geq 1$. Then, for any function $h(x)$ satisfying (i) $\|h(x)\|^{1+\delta} \leq C(1 + \|x\|^s)$ for some $C, \delta > 0$ and (ii) for some $c > 0$, there exists $L_c > 0$ so that $\|h(x) - h(\tilde{x})\| \leq L_c \|x - \tilde{x}\|$ for $\|x - \tilde{x}\| \leq c$, it holds that $T^{-1} \sum_{t=1}^T h(X_t) \xrightarrow{P} E[h(X_t^*)]$.*

Remark 1 The above lemma can be used to establish a Martingale CLT. Let $\{u_t\}$ be a sequence satisfying $E[u_t | \mathcal{F}_{t-1}] = 0$ and $E[u_t u_t' | \mathcal{F}_{t-1}] = h(X_{t-1})$ w.r.t. some filtration \mathcal{F}_t , where $\{X_t\}$ and h satisfy the conditions of Lemma 1. It then holds that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T u_t \xrightarrow{d} N(0, E[h(X_t^*)]). \quad (10)$$

This result follows readily from standard CLT for stationary martingale differences (see *e.g.* Brown, 1971). This CLT proves to be important for the asymptotic analysis of the maximum likelihood estimator provided in the next section. \square

We end this section by weakening the Lipschitz condition in Assumption 3, since this rules out some relevant transformations $f(x_i)$ of x_t , such as the specification in (3) with $f_i(x_i) = \exp(x_i)$ for some $1 \leq i \leq d_x$. Such transformations can be handled by introducing in the asymptotic analysis the following truncated intensity

$$\lambda_t^c = \omega + \sum_{i=1}^p \alpha_i y_{t-i}^c + \sum_{i=1}^q \beta_i \lambda_{t-i}^c + f(x_{t-1}; \gamma) \mathbb{I}\{\|x_{t-1}\| \leq c\}, \quad (11)$$

for some cut-off point $c > 0$, and with y_t^c the corresponding Poisson process. We can then relax $f(x; \gamma)$ to be locally Lipschitz in the following sense:

Assumption 3' Assumption 3 holds with condition (ii) replaced by the following: (ii') for all $c > 0$, there exists some $L_c < \infty$ such that

$$|f(x; \gamma) - f(\tilde{x}; \gamma)| \leq L_c \|x - \tilde{x}\|, \quad \|x\|, \|\tilde{x}\| \leq c.$$

By replacing Assumption 3 with Assumption 3' we now obtain, by identical arguments as in the proof of Theorem 1, that the truncated process has a weakly dependent stationary and ergodic solution. While this approach is similar to the approximation of Poisson AR process as used in FRT, the reasoning here is different. In FRT, an approximating process

was needed in order to establish geometric ergodicity of the Poisson GARCH process, while here we introduce the truncated process in order to handle the often applied practice of introducing non-bounded or exponential transformations of the regressors in the model.

In the next Lemma we formally prove that, as $c \rightarrow \infty$, the truncated process approximates the untruncated one ($c = +\infty$).

Lemma 2 *Under Assumptions 1-3' together with $E[f(x_t^*)] < \infty$,*

$$\begin{aligned} |E[\lambda_t^c - \lambda_t]| &= |E[y_t^c - y_t]| \leq \delta_1(c), \\ E[(\lambda_t^c - \lambda_t)^2] &\leq \delta_2(c), \quad E[(y_t^c - y_t)^2] \leq \delta_3(c), \end{aligned}$$

where $\delta_k(c) \rightarrow 0$ as $c \rightarrow +\infty$, $k = 1, 2, 3$.

The above result is akin to Lemma 2.1 in FRT. The additional assumption of $E[f(x_t^*)]$ being finite needs to be verified on a case-by-case basis. For example, with $f_i(x_i) = \exp(x_i)$, then this assumption holds if x_t^* has *e.g.* a Gaussian distribution, or some other distribution for which the moment generating function, or Laplace transform, is well-defined.

Remark 2 The truncation argument used here is merely a theoretical device to prove that LLNs and CLTs as the one given in Remark 1 above also hold in the presence of non-Lipschitz link functions such as the exponential function. In empirical applications it is therefore not required to truncate the conditional intensity equation as done in (11).

4 Estimation and Forecasting

In this section, we describe how the PARX model can be estimated by maximum likelihood and employed for forecasting. We also provide an asymptotic theory for the estimated parameters allowing for statistical inference, and present the results of a small simulation study investigating the finite-sample properties of the estimator.

4.1 Estimation

We consider the model for y_t as specified in eqs. (1)-(2) with $f(x; \gamma)$ specified as in eq. (3); that is, with conditional intensity given by

$$\lambda_t(\theta) = \omega + \sum_{i=1}^p \alpha_i y_{t-i} + \sum_{i=1}^q \beta_i \lambda_{t-i}(\theta) + \sum_{i=1}^{d_x} \gamma_i f_i(x_{it-1}),$$

where $\theta = (\omega, \alpha', \beta', \gamma')' \in \Theta \subset (0, \infty) \times [0, \infty)^{p+q+d_x}$, with $\alpha = (\alpha_1, \dots, \alpha_p)'$, $\beta = (\beta_1, \dots, \beta_q)'$, and $\gamma = (\gamma_1, \dots, \gamma_{d_x})'$. We let $\theta_0 = (\omega_0, \alpha'_0, \beta'_0, \gamma'_0)'$, where $\alpha_0 = (\alpha_{0,1}, \dots, \alpha_{0,p})'$, $\beta_0 =$

$(\beta_{0,1}, \dots, \beta_{0,q})'$, and $\gamma_0 = (\gamma_{0,1}, \dots, \gamma_{0,d_x})'$, denote the true, data-generating parameter value. Notice that the parameter space excludes negative parameter values; this condition (which resembles the well known non-negativity parameter constraint for GARCH models) is sufficient (albeit not necessary) for the conditional intensity to be strictly positive. It is also worth noticing that, in contrast to FRT, we do not require the parameters to be bounded away from zero¹; this condition is particularly important given that, in applications, researchers are often interested in testing whether a given parameter equals zero.

The conditional log-likelihood function of θ in terms of the observations $(y_1, x_0), \dots, (y_T, x_{T-1})$, given some initial values $\lambda_0, \lambda_{-1}, \dots, \lambda_{1-q}, y_0, \dots, y_{1-p}$ and x_0 , takes the form

$$L_T(\theta) = \sum_{t=1}^T l_t(\theta), \quad l_t(\theta) := y_t \log \lambda_t(\theta) - \lambda_t(\theta) \quad (12)$$

where we have left out any constant terms. The maximum likelihood estimator (MLE) is then computed as

$$\hat{\theta} := \arg \max_{\theta \in \Theta} L_T(\theta). \quad (13)$$

In order to analyze the large sample properties of $\hat{\theta}$, we impose the following conditions on the parameters and the exogenous regressors,

Assumption 4 Θ is compact so that for all $\theta = (\omega, \alpha, \beta, \gamma) \in \Theta$, $\beta_i \leq \beta_i^U$, $i = 1, \dots, q$, and $\omega \geq \omega_L$ for some constants $\omega_L > 0$ and $\beta_i^U > 0$ where $\sum_{i=1}^q \beta_i^U < 1$.

Assumption 5 The polynomials $A(z) := \sum_{i=1}^p \alpha_{0,i} z^i$ and $B(z) := 1 - \sum_{i=1}^q \beta_{0,i} z^i$ have no common roots; for any $a = (a_1, \dots, a_p)' \neq 0$ and $g = (g_1, \dots, g_{d_x})' \neq 0$, $\sum_{i=1}^p a_i y_{t-i}^* + \sum_{i=1}^{d_x} g_i f(x_{i,t}^*)$ has a nondegenerate distribution.

Assumption 4 imposes weak restrictions on the parameter space; these are similar to the ones imposed in the analysis of estimators of GARCH models and rule out β 's greater than one (for which $\lambda_t(\theta)$ is explosive) and ω 's equal to zero. The latter is used to ensure that $\lambda_t(\theta)$ is bounded away from zero.

Assumption 5 is an identification condition which is similar to the one found for GARCH models with exogenous regressors. The first part is the standard condition for GARCH models (see, *e.g.*, Berkes et al, 2003), while the second part rules out that the exogenous covariates are colinear with each other and the observed count process (see Han and Kristensen, 2014, for a similar condition).

Under this assumption, together with those used earlier to establish stationarity and existence of moments, we obtain the following asymptotic result for the MLE conditional on the initial values.

¹Specifically, FRT requires $\alpha_i \geq \alpha_L > 0$ and $\beta_i \geq \beta_L > 0$ for some constants α_L and β_L .

Theorem 2 *Suppose Assumptions 1–5 hold with $s \geq 2$ and $\theta = \theta_0$. Then, $\hat{\theta}$ is consistent. Furthermore, if $\theta_0 \in \text{int } \Theta$,*

$$\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, H^{-1}(\theta_0)), \quad H(\theta) := -E \left[\frac{\partial^2 l_t^*(\theta)}{\partial \theta \partial \theta'} \right],$$

where $l_t^*(\theta)$ denotes the likelihood function evaluated at the stationary solution.

Remark 3 If the model is misspecified, we expect the asymptotic properties of the MLE to remain correct except that θ_0 is now the pseudo-true value maximizing the pseudo-likelihood and the asymptotic variance takes the well-known sandwich form $H^{-1}(\theta_0) \Omega(\theta_0) H^{-1}(\theta_0)$, where

$$\Omega(\theta) = E \left[\frac{\partial l_t^*(\theta)}{\partial \theta} \frac{\partial l_t^*(\theta)}{\partial \theta'} \right];$$

see Besag (1975), White (1982) and Gouriéroux, Monfort and Trognon (1984). \square

Remark 4 The assumption $\theta_0 \in \text{int } \Theta$ rules out cases where some of the parameters are zero. We detail how this assumption can be relaxed at the end of this section. The requirement on s , as defined in Assumption 2, is used to ensure that the likelihood function has a well-defined limit and that the moments in the information matrix $H(\theta)$ exist. \square

The above theorem generalizes the result of FRT to allow for estimation of parameters associated with additional regressors in the specification of λ_t . It is established under the assumption that f is globally Lipschitz as stated in Assumption 3. By combining the arguments in FRT with Lemma 2, the asymptotic result can be extended to allow f to be locally Lipschitz, see Assumption 3'. This is proved in the following theorem.

Theorem 3 *Under Assumptions 1–3' and 4–5, and if $E[f_i(x_{it}^*)] < \infty$, $i = 1, \dots, d_x$, the conclusions of Theorem 2 remain valid.*

Remark 5 The proof of Theorem 3 is based on the following auxiliary likelihood for the approximating (or, truncated) model:

$$L_T^c(\theta) = \sum_{t=1}^T l_t^c(\theta), \quad \text{where } l_t^c(\theta) = y_t \log \lambda_t^c(\theta) - \lambda_t^c(\theta),$$

where the truncated intensity λ_t^c is defined as in (11). Then, it immediately follows that the results of Theorem 2 holds for the QMLE based on $L_T^c(\theta)$, $\hat{\theta}^c$ say. Finally, as the approximating likelihood function can be made arbitrarily close to the true likelihood as $c \rightarrow \infty$, we are able to demonstrate that Assumption 3 in Theorem 2 can be replaced by Assumption 3'.

It will often be of interest to investigate if some of the elements of θ are zero, as for example $\alpha_i = 0$, $\beta_i = 0$ or $\gamma_i = 0$. In order to allow for this, where under the null the parameter vector θ is on the boundary of the parameter space Θ , we complement the results of Theorems 2 and 3. To do so, we can apply the general theory of Andrews (1999), see also Demos and Sentana (1998) and Francq and Zakoian (2009), to obtain the following corollary where we state this explicitly for the case of testing one parameter equal to zero (more general cases of multiple parameters on the boundary can be handled as in Francq and Zakoian, 2009). Here, we denote the standard t statistic for the null hypothesis $H_0 : \theta_{i0} = 0$ against the composite alternative $H_1 : \theta_{i0} > 0$ with $t_i = \sqrt{T}\hat{\theta}_i/\hat{\sigma}_{ii}$, where $\hat{\sigma}_{ii}^2$ is a consistent estimator of the i -th diagonal element of H^{-1} as defined in Theorem 2. For instance, $\hat{\sigma}_{ii}^2$ can be taken as the i -th element of the diagonal of $H_T^{-1}(\hat{\theta})$, where

$$H_T(\theta) := -\frac{1}{T} \sum_{t=1}^T \frac{1}{\lambda_t(\theta)} \left(\frac{\partial \lambda_t(\theta)}{\partial \theta} \right) \left(\frac{\partial \lambda_t(\theta)}{\partial \theta} \right)'.$$

The likelihood ratio test for the same null hypothesis is denoted by LR_i . The following corollary of Theorem 2 holds under the null hypothesis.

Corollary 1 *Under Assumptions 1-3' and 4-5 and H_0 with $\theta_{j0} \neq 0$ for all $j \neq i$,*

$$t_i \xrightarrow{d} \max\{0, Z\}, \tag{14}$$

$$LR_i \xrightarrow{d} (\max\{0, Z\})^2, \tag{15}$$

where Z is standard normally distributed.

Remark 6 For a given significance level $\eta \in (0, 1/2)$, the $(1 - \eta)$ quantile of the asymptotic distribution in (14) equals the $(1 - 2\eta)$ quantile of the standard normal distribution. Therefore, a standard one-sided t test is (asymptotically) valid in this framework. Similarly for the LR statistic, for any $\eta \in (0, 1/2)$ the $(1 - \eta)$ quantile of the asymptotic distribution in (15) equals the $(1 - 2\eta)$ quantile of the $\chi^2(1)$ distribution. \square

4.2 Finite Sample Performance

In this section we present results from a small simulation study aimed at evaluating the finite-sample performance of the MLE presented in the previous section. We consider the PARX(1,1) model (1) with conditional intensity given by

$$\lambda_t = \omega + \alpha y_{t-1} + \beta \lambda_{t-1} + \gamma \exp(x_{t-1}) ;$$

investigation of the small sample properties of the estimator in higher order models is left out. The use of an exponential link function is motivated by the empirical application of

Section 6, where this is employed for the log-realized volatility. This particular choice of the link function is covered by our theoretical results, cf. Assumption 3' and Theorem 3 of Section 3.

We examine the performance of the MLE under two different data generating processes (DGP's) for the covariate x_t .

DGP 1 x_t is a stationary autoregressive process, $x_t = \varphi x_{t-1} + \varepsilon_t$, with $\varepsilon_t \sim \text{i.i.d.} N(0, 1)$, initialized at $x_0 \sim \text{i.i.d.} N(0, 1/(1 - \varphi^2))$; the AR parameter is set to $\varphi = 1/2$.

DGP 2 x_t is a stationary fractionally integrated process, $\Delta_+^d x_t = \varepsilon_t$, where the operator Δ_+^d is given by $\Delta_+^d z_t := \Delta^d z_t \mathbb{I}(t \geq 1) = \sum_{i=0}^{t-1} \pi_i(-d) z_{t-i}$ with $\pi_i(v) = (i!)^{-1}(v(v+1)\dots(v+i-1))$ denoting the coefficients in the usual binomial expansion of $(1-z)^{-v}$; ε_t is i.i.d. $N(0, 1)$ and $d = 1/4$.

These two DGP's represent typical time series behavior found in the factors used in the empirical application. The first DGP satisfies the theoretical conditions used in the asymptotic analysis of the MLE, while the second one does not since it is not a Markov chain. However, DGP 2 is stationary and we expect that the theory can be extended to cover non-Markov regressors as long as they are stationary.

Since the distribution of y_t is not invariant to the scale of the covariate x_t , in each case x_t has been re-scaled by its unconditional variance. We report results for $\omega = 0.10$, $\alpha = 0.30$, $\gamma = 0.5$ and three alternative scenarios for β : $\beta = 0$ (no feedback from lagged intensity to current intensity), $\beta = 0.20$ (low persistence) and $\beta = 0.70$ (high persistence). In all cases considered, the model admits a stationary solution, see section 3. Finally, we consider samples of size $T \in \{100, 250, 500, 1000\}$. For each experiment, the number of Monte Carlo replications is set to $N = 1000$.

Results for the case of DGP 1 are presented in Table 1. For each parameter, the mean and root mean square error (RMSE) of the corresponding estimator are reported. Furthermore, the p -value obtained from a Kolmogorov-Smirnov (KS) test for the hypothesis of $N(0, 1)$ distribution of each parameter estimator is reported.

The performance of the MLE for DGP 1 seems largely satisfactory for moderate and large sample sizes. For samples of $T \geq 250$ observations and for all scenarios considered, the hypothesis of normality of $\hat{\theta}_i$ is never rejected at any conventional significance level. For samples of $T = 100$ observations, the degree of persistence of the process (here captured by the β coefficient) seems to affect the distribution of the estimators. Specifically, while in the case of lowest persistence ($\beta = 0$) normality of $\hat{\theta}_i$ is never rejected, in the case of stronger persistence $\beta = 0.2$ ($\beta = 0.7$) normality is rejected for the estimated constant term ω at the 1% (10%) significance level. When $T = 100$ and $\beta = 0.7$, normality is also rejected for the estimated PAR parameters α and β . All these deviations from normality, however, do not

persist in larger sample sizes. Finally, it is worth noticing that the parameter which delivers the highest RMSE is the constant term, ω .

Next, consider the results for DGP 2 as presented in Table 2. Compared to DGP 1, x_t now has higher persistence. Despite this, for $T \geq 250$, with the only exception of the constant term ω , results do not show substantial differences relative to the ones for DGP 1; that is, the asymptotic $N(0, 1)$ approximation is largely satisfactory. In the case of high persistence ($\beta = 0.7$), normality of $\hat{\omega}$ is rejected at the 1% significance level even when $T = 1000$. This is consistent with the findings of Han and Kristensen (2014) for the GARCH-X model who also find that the intercept is less precisely estimated in the presence of persistent regressors.

[Table 1 and Table 2 about there]

4.3 Forecasting

Once the PARX model has been estimated, it can be used to forecast future number of counts, y_t . Forecasting of Poisson autoregressive processes is similar to forecasting of GARCH-X processes (see, *e.g.*, Hansen *et al.*, 2012, Section 6.2) in that it proceeds in two steps: first, a forecast of the time-varying parameter (conditional variance in the case of GARCH, conditional intensity in the case of PARX) is obtained. This is then substituted into the conditional distribution of the observed process y_t . Consider first the forecasting of λ_t . A natural one-step ahead forecast, given available information at time T and parameters θ , is

$$\lambda_{T+1|T}(\theta) = \omega + \sum_{i=1}^p \alpha_i y_{T+1-i} + \sum_{i=1}^q \beta_i \lambda_{T+1-i}(\theta) + f(x_T, \gamma). \quad (16)$$

More generally, a multi-step ahead forecast of λ_{T+h} , for some $h > 1$, can be obtained by noticing that for any $k \geq 1$ the conditional intensity equation for λ_{T+k} can be expressed as

$$\lambda_{T+k}(\theta) = \omega + \sum_{i=1}^{\max\{p,q\}} \{\alpha_i + \beta_i\} \lambda_{T+k-i}(\theta) + \sum_{i=1}^p \alpha_i \eta_{T+k-i}(\theta) + f(x_{T+k-1}, \gamma)$$

where $\eta_t(\theta) := y_t - \lambda_t(\theta)$ has (conditionally on the past) zero expectation (as is standard, we set $\alpha_i = 0$ for $i > p$ and $\beta_i = 0$ for $i > q$). By setting $\eta_t(\theta)$ to its (conditional) expectation (which is zero) and replacing x_{T+k-1} by some point forecast $x_{T+k-1|T}$, we obtain a multi-step ahead forecast of λ_{T+h} through the recursive scheme

$$\lambda_{T+k|T}(\theta) = \omega + \sum_{i=1}^{\max\{p,q\}} \{\alpha_i + \beta_i\} \lambda_{T+k-i|T}(\theta) + f(x_{T+k-1|T}, \gamma), \quad k = 2, \dots, h, \quad (17)$$

with $\lambda_{T+1|T}(\theta)$ coming from eq. (16), and $\lambda_{T+k-i|T}(\theta) = \lambda_{T+k-i}(\theta)$ for $k-i \leq 0$. Note that the above multi-step ahead forecast requires a forecasting model for x_t .

Once we have computed a (point) forecast of the underlying intensity, $\lambda_{T+h|T}(\theta)$, this can in turn be used to generate a forecast distribution of y_{T+h} ,

$$\hat{P}(y_{T+h} = y | \mathcal{F}_T) = \frac{\lambda_{T+h|T}^y(\theta) \exp(-\lambda_{T+h|T}(\theta))}{y!}, \quad y \in \{0, 1, 2, \dots\}.$$

This is related to the well-known concept of density forecasts (see Tay and Wallis, 2000, for a review) except that we are here working with a discrete-valued distribution. A simple way of representing the forecast distribution is by reporting the $100(1 - \alpha)\%$ prediction interval (as implied by the forecast distribution) for some $\alpha \in (0, 1)$. Specifically, the (symmetric) $1 - \alpha$ prediction interval takes the form

$$[Q(\alpha/2 | \lambda_{T+h|T}(\theta)), Q((1 - \alpha/2) | \lambda_{T+h|T}(\theta))],$$

where $p \mapsto Q(p | \lambda)$ denotes the quantile function of a Poisson distribution with intensity λ .

5 Empirical Analysis

The aim of this section is to provide an empirical analysis of US corporate default counts using PARX models as discussed in Section 2. Specifically, by including exogenous regressors in the intensity specification and by testing whether they cause a significant decrease in the impact of past default counts, we are able to investigate to what extent autocorrelation (as well as clustering over time) in default counts depends on common (aggregate) risk factors. That is, testing the existence of autocorrelation in default counts after correcting for common risk factors can be viewed as testing the existence of *contagion* effects over time in the sense discussed in Section 2.2.

The data set on defaults consist of monthly number of bankruptcies among Moody's rated industrial firms in the United States for the period 1982–2011 ($T = 360$ observations), collected from Moody's Credit Risk Calculator (CRC). Figure 1(a, b), which shows default counts and the corresponding autocorrelation function, reveals three important stylized facts of defaults: (i) high temporal dependence in default counts; (ii) existence of default clusters over time; (iii) overdispersion of the distribution of default counts (the empirical average is 3.51 while the empirical variance is 15.57). It will be shown later in this section that all these empirical properties are well explained using PARX specifications.

[Figure 1(a, b) about here]

The choice of covariates to be included in our PARX models is important, as they are supposed to represent the common risk factors conditional affecting firm defaults. Similar

to Lando and Nielsen (2010), we consider the following financial, credit market and macro-economic variables: Baa Moody’s rated to 10-year Treasury spread (SP), the number of Moody’s rating downgrades (DG), year-to-year change in Industrial Production Index (IP), Leading Index released by the Federal Reserve (LI), the recession indicator released by the National Bureau of Economic Research² (NB).³ Moreover, in order to shed some light on the possible impact of uncertainty in the financial markets on the number of future defaults, we also consider realized volatility (RV) on the S&P 500. RV is computed as a proxy of the S&P 500 monthly realized volatility using daily squared returns (that is, $RV_t := \sum_{i=1}^{n_t} r_{i,t}^2$ with $r_{i,t}$ denoting the i -th daily return on the S&P 500 index in month t and n_t being the number of trading days in month t).

Since Industrial Production and Leading Index take on both negative and positive values, we decompose them into their negative and positive parts and let $IP^{(+)} := \mathbb{I}_{\{IP \geq 0\}} |IP|$, $IP^{(-)} := \mathbb{I}_{\{IP < 0\}} |IP|$ and similarly for LI . This is required in order to ensure non-negativity of the additive (linear) link function adopted below, see (3) and the discussion in section 5.1. This gives us a total of eight candidate covariates.

5.1 In-sample Performance

We here provide an analysis for the full sample 1982-2011. Preliminary covariate and lag selection based on AIC and BIC and significance of the estimated coefficients – using all eight covariates – suggests the following specification of default intensity:

$$\begin{aligned} \lambda_t = & \omega + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \beta \lambda_{t-1} \\ & + \gamma_1 RV_{t-1} + \gamma_2 SP_{t-1} + \gamma_3 DG_{t-1} + \gamma_4 NB_{t-1} + \gamma_5 IP_{t-1}^{(-)} + \gamma_6 LI_{t-1}^{(-)}, \end{aligned} \tag{18}$$

which is a special case of model (1)-(2) with $p = 2$, $q = 1$ and $\beta = \beta_1$.

[Table 3 about there]

Table 3 shows the estimation results for the full PARX(2,1) model in (18), along with the PAR(2,1) model (*i.e.*, the model without covariates) and nested specifications based on

²This time series is released by the Federal Reserve Bank of St. Louis interpreting the Business Cycle Expansions and Contractions data provided by The National Bureau of Economic Research (NBER) at <http://www.nber.org/cycles/cyclesmain.html>. A value of 1 indicates a recessionary period, while a value of 0 denotes an expansionary period.

³Data are obtained from the FRED website, provided by the Federal Reserve Bank of St. Louis, <http://research.stlouisfed.org/>, except for the number of Moody’s rating downgrades, which we collect from Moody’s CRC.

subsets of the six included covariates. For each specification, we report parameter estimates together with corresponding t statistics as well as standard (AIC and BIC) information criteria⁴. Among the various models considered, the preferred PARX model, in terms of information criteria (AIC and BIC) as well as LR tests, is the one only including realized volatility and the leading index.

To our knowledge, the link between realized volatility (reflecting uncertainty in financial markets) and defaults of industrial firms has not been documented earlier in the literature. Similarly, significance of the Leading Index highlights a clear link between macroeconomic factors and corporate defaults, which is not generally found using standard econometric techniques. For instance, recent empirical results of Duffie *et al.* (2009) and Giesecke *et al.* (2011) do not show a significant role of production growth while Lando *et al.* (2013) find that, conditional on individual firm risk factors, no macroeconomic covariate seems to explain significantly individual default intensity. However, once we control for the information contained in realized volatility and the negative component of the Leading Index, none of the other four covariates (NBER recession indicator, interest rate spread, and number of downgrades) are found to be relevant in predicting future defaults.

We analyze the extent of the feedback from past defaults to current default counts (which may indicate possible contagion effects, see Section 2.2) by investigating whether by including covariates, past default counts have a smaller impact, *i.e.* there is a significant decrease in $\hat{\alpha}_1$ and $\hat{\alpha}_2$ in a given model with covariates (PARX) relative to the corresponding one without covariates (PAR). As remarked previously in Section 2.2, the (extreme) case of conditional independence over time would require that α_1 and α_2 are both zero,⁵ which would imply that conditional intensity can fully be explained by past covariates only. Indeed, the inclusion of covariates leads to a decrease in $\hat{\alpha}_1 + \hat{\alpha}_2$ for almost all the models considered. On the other hand, the null hypothesis $H_0 : \alpha_1 + \alpha_2 = 0$ is rejected for all specifications. Therefore, although part of the dependence over time in default counts can be explained by the set of covariates considered, a strong link between conditional intensity and past default counts remains. This result provides – conditional on correct choice of the exogenous regressors – strong evidence of contagion as defined in Section 2.2.

We run a number of model (mis)specification tests on the selected model. First, to check in-sample fit, we plot in Figure 2 the actual default counts (y_t) together with the fitted

⁴We do not report the LR statistic of each model relative to the maintained (general) PARX model because the null hypothesis imposes that a subset of the parameter vector lies on the boundary of the parameter space and, therefore, the null asymptotic distribution of the LR statistic is non-standard.

⁵It is worth noticing that this approach is related to empirical studies aiming at measuring the impact of covariates, such as the trading volumes, on future volatility using GARCH models (see, for instance, Lamoureux and Lastrapes, 1990 and Gallo and Pacini, 2000).

values ($\hat{y}_t := \hat{\lambda}_t = \lambda_t(\hat{\theta})$) and the corresponding confidence bands (at the 95% nominal level) based on the underlying Poisson distribution; see Section 4.3. As can be seen from this figure, the model captures the default counts dynamics well. The associated generalized, or Pearson, residuals (see Gouriéroux *et al.*, 1987; Kedem and Fokianos, 2002) – formally defined as $\hat{e}_t = \hat{\lambda}_t^{-1/2}(y_t - \hat{\lambda}_t)$ ($t = 1, \dots, T$) also appear to be uncorrelated over time; see the corresponding correlogram and correlogram of the squares in Figure 3 (using 12 lags, the corresponding Ljung-Box test has p -value 0.661 when computed using \hat{e}_t and 0.373 when computed using \hat{e}_t^2 ; similar results are obtained for other choices of the number of lags).

We also evaluate the goodness of fit of the assumed Poisson conditional distribution of y_t by comparing the observed zero counts with the corresponding model-implied probabilities, $\hat{P}(y_t = 0 | \mathcal{F}_{t-1}) = e^{-\hat{\lambda}_t}$ ($t = 1, \dots, T$), *i.e.* the (conditional on the past) probability that a Poisson($\hat{\lambda}_t$) random variable equals zero under the selected model specification. Figure 4 shows the relation between the observed zeros and such model-implied probabilities. There is a clear correspondence between periods characterized by high number of zeros and the conditional probability of observing $y_t = 0$, given the specified model.

As a final assessment of the adequacy of the model, following Davis and Liu (2014) we assess whether the randomized probability integral transform (PIT) is uniformly distributed on $[0, 1]$ using a standard Kolmogorov-Smirnov test. From the associated p -value (about 0.11) it can be seen that our preferred model passes the PIT test.

5.2 Out-of-sample Performance

We found in the previous subsection that the preferred model does a good job in terms of in-sample fit. To further examine the performance of the PARX model, we also perform a pseudo-out-of-sample forecasting exercise for the preferred model (the PARX(2, 1) with RV and $LI^{(-)}$ as included covariates) and, for comparison, the PAR(2, 1) model (where no covariates are included). The forecasting exercise is carried out along the lines of, for example, Stock and Watson (1996): we split the sample in two with the first part of the sample of size T_0 ($= 120$), $\{(y_t, x_{t-1}) : t = 1, \dots, T_0\}$ being used for initial estimation of the model, and the remaining observations $\{(y_t, x_{t-1}) : t = T_0 + 1, \dots, T\}$ being used for a forecasting exercise described below.

Let

$$\hat{\theta}_t = \arg \max_{\theta} L_t(\theta)$$

be the MLE using observations up to time $t \geq T_0$, where

$$L_t(\theta) = \sum_{s=1}^t l_s(\theta), \quad l_s(\theta) := y_s \log \lambda_s(\theta) - \lambda_s(\theta).$$

Given $\hat{\theta}_t$, we then compute the corresponding one-step-ahead forecast of λ_{t+1} using only information up to time t , $\hat{\lambda}_{t+1|t} = \lambda_{t+1}(\hat{\theta}_t)$. We then repeat the above exercise for $t = T_0 + 1, \dots, T$, thereby providing us with a time series of estimators, $\{\hat{\theta}_t : t = T_0, \dots, T\}$, and corresponding intensity forecasts, $\{\hat{\lambda}_{t+1|t} : t = T_0, \dots, T\}$. This procedure mimics what a forecaster would obtain as (s)he starts forecasting at time T_0 and updates his (her) estimates and forecasts as more data arrive. Given the forecast path $\hat{\lambda}_{t+1|t}$, we evaluate the performance of the preferred PARX specification and the corresponding PAR model through two standard forecasting loss functions: The first is the average mean-square forecasting error,

$$MSFE_t = \frac{1}{t - T_0} \sum_{s=T_0}^t (y_{s+1} - \hat{\lambda}_{s+1|s})^2, \quad t = T_0, \dots, T,$$

and the second is the average (logarithmic) forecasting score (FS), see *e.g.* Amisano and Giacomini (2008),

$$FS_t = \frac{1}{t - T_0} \sum_{s=T_0}^t (y_{s+1} \log \hat{\lambda}_{s+1|s} - \hat{\lambda}_{s+1|s}), \quad t = T_0, \dots, T.$$

The $MSFE$ loss function only measures how well a given PARX model does in terms of forecasting the level of defaults, while FS is a more comprehensive measure that evaluates how well the model does in terms of forecasting the distribution of defaults. For both measures, small (large) values indicate good (poor) forecasting performance. In Figure 5, we plot $MSFE_t$ (left panel) and FS_t (right panel) as functions of time for the PARX and PAR model. We see that in terms of $MSFE$ the two models perform very similarly with the MSFE for both models being around the same level throughout the chosen forecasting period. On the other hand, in terms of FS , the PARX model clearly dominates, providing much better probability forecasts compared to the PAR model. In conclusion, we find that if the goal is to forecast the level of defaults, covariates are not so important, while if the aim is to provide a good forecast of the default count distribution RV and $LI^{(-)}$ are important predictors.

[Figure 5 about there]

5.3 Structural Instabilities

As is also evident from Figure 5, the forecasting performances of both the PARX and PAR model vary a lot. In particular, the forecast performances of the two models in terms of FS deteriorate radically around the time when the Dot-com bubble burst in the late 1990's and again around the onset of the most recent financial crisis in 2008. If the PAR(X) model

was stable over time, we should expect $MSFE_t$ and FS_t to also remain stable over time. This is not the case, however, indicating the presence of structural instabilities in the model parameters during the sample period.

To formally test whether parameters are indeed varying over time in our sample, we compute the Nyblom (1989) test statistic (see eq. (3.1) in Nyblom, 1989) for the PARX model and clearly reject the null of parameter constancy using the critical values in Table 2 of Nyblom (1989). To further investigate the underlying causes for parameter instability in the preferred PARX model, we plot in Figure 6 the time series of rolling-window parameter estimates, $\{\hat{\theta}_t\}_{t=T_0}^T$. These graphs provide further evidence of structural breaks during the two most recent financial crises, with all parameter estimates changing radically during these periods. In particular, the impacts of lagged default counts, RV and $LI^{(-)}$ on the default intensity change dramatically over the 20 year forecasting period.

[Figure 6 about there]

Based on these findings, we split the full sample into three subsamples, 1982-1998, 1998-2007, and 2007-2011, and for each subsample re-do model selection (based on the same approach used for the full sample, see above) and estimation. In Table 4, we report the preferred model with corresponding estimated parameters for each subsample. In the early period (1982-1998), all macro factors (incl. RV and $LI^{(-)}$) are irrelevant and past default counts have a strong explanatory power ($\hat{\alpha}_1 + \hat{\alpha}_2 + \hat{\beta} = 0.65$). This may indicate that during this period, default clustering mainly depends on the clustering channel caused by past defaults, as well as on possible frailty factors (see section 2.2 and Azizpour *et al.*, 2015). During the second subsample, the feedback from past default counts to current defaults is even stronger ($\hat{\alpha}_1 + \hat{\beta} = 0.99$) and the fitted model is close to the boundary of the stationarity region established in Theorem 1. Finally, during the Great Recession (2007-2011), we find that RV and $LI^{(-)}$ are very strong explanatory variables and there are no contagion effects ($\hat{\alpha}_1 + \hat{\alpha}_2 = 0.00$). Finally, for comparison, we fit a $PAR(2, 1)$ model to the third subsample and obtain the following parameter estimates, $\hat{\omega} = 0.00$ (0.00), $\hat{\alpha}_1 = 0.43$ (5.22), $\hat{\alpha}_2 = 0.18$ (1.14), and $\hat{\beta} = 0.39$ (3.07). This shows that by, incorrectly, leaving out RV and $LI^{(-)}$ from the model, one will mistake systematic risk for contagion effects.

[Table 4 about there]

We also note that the estimated models for the three regimes match well with the parameter estimates we reported for the full sample, which are basically an average over the

estimates reported for each of the three different regimes. For instance, the full sample estimate of β is about 0.52 (see Table 3), which is close to the average of the estimated β in the three subsample, i.e. 0.0 (1982–1998 subsample), 0.79 (1998–2007 subsample) and 0.82 (2007–2001).

Ideally, we would like to find a relevant regressor that captures these structural breaks. Since the breaks occur at the onsets of the two latest financial crises, a good choice would appear to be an indicator (possibly non-stationary) for financial crises, which we leave for future research.

6 Conclusions

In this paper, we have developed a class of Poisson autoregressive models with exogenous covariates (PARX) for time series of counts. Since PARX models allow for overdispersion arising from persistence, they are suitable to model count time series of corporate defaults, which are strongly correlated over time and exhibit high peaks, known as default clustering. Our empirical analysis, where we use the PARX framework to model US default counts dynamics, reveals that our model is capable to capture the dynamic features of default counts, including the pronounced default clustering. PARX models also allow to investigate to what extent dependence over time in default counts can be attributed by the various default clustering channels, such as exposure to macroeconomic and financial factors and the impact of past default events. We find that the lagged realized volatility of financial returns, together with macroeconomic variables, significantly explains the number of defaults. A full sample analysis shows that past default counts are important explanatory variables of current default counts, even when the exogenous covariates are included; this may indicate that at the aggregate level, the so-called “conditional independence” hypothesis of firm defaults is not supported by the data. However, a further econometric investigation reveals that such dependence is not constant over time. Specifically, while in the period leading up to the Great Recession (1982–2006) all macro factors considered are not significant and, accordingly, past default counts are the main default clustering channel. However, during the Great Recession (2007–2011), we find that financial volatility and macroeconomic factors become strong explanators of defaults, while the feedback from past default counts (captured by the parameters linking current intensity to past default counts) become weaker, in fact absent. The latter result indicates that while in general current defaults do indeed affect the probability of other firms defaulting in the future, in recent years economic and financial factors at the macro level are explain most of the correlation of US firms defaults over time. It is, however, important to recall that, as for all econometric analyses of contagion and default clustering channels, also ours could be affected by misspecification error due to the

fact that our chosen set of covariates might not be exhaustive. Therefore, the results should be interpreted with caution. Similarly, our analysis also ignores possible feedback effects from past default counts to the set of covariates. As suggested in section 2.2, a more complete analysis would require us to specify a multivariate model where past default counts might affect the covariates x_t .

Further issues are left to future research. First, our analysis is limited to defaults of U.S. industrial firms. It would be interesting to assess whether similar results characterize different sectors (*e.g.*, financial) and/or countries. Second, the PARX specifications developed in this paper are univariate, in the sense that they can be used to model a single time series of default counts. The multivariate PARX case, which would permit to analyze the cross linkages between different time series of defaults, represents an obvious extension of the econometric theory proposed in this paper and is currently under investigation by the authors.

A Appendix

A.1 Proof of Theorem 1

We first verify that the process $Z_t := (y_t, x_t)'$ satisfies the conditions of Corollary 3.1 in Doukhan and Wintenberger (2008) from which the first part of the theorem will follow. To simplify the notation, assume without loss of generality $p = q$ in the following. With $\beta(z) := 1 - \sum_{i=1}^q \beta_i z^i$, $z \in \mathbb{C}$, note that Assumption 3(i) implies that $\beta(z)^{-1} = \psi(z) := \sum_{i=0}^{\infty} \psi_i z^i$ is well-defined for $|z| \leq 1 + \eta$, for some $\eta > 0$, with ψ_i exponentially decreasing and defined recursively by $\psi_0 = 1$ and $\psi_n = \sum_{i=1}^n \beta_i \psi_{n-i}$ for $n \geq 1$. Next, with $\alpha(z) := \sum_{i=1}^q \alpha_i z^{i-1}$, note that λ_t defined in (2) can be rewritten in terms of the backshift operator B as

$$\beta(B) \lambda_t = \omega + \alpha(B) y_{t-1} + f(x_{t-1}, \gamma),$$

such that with $\delta := \beta(1)$, the conditional intensity λ_t of the PARX process y_t can be represented in terms of $\{(y_{t-i}, x_{t-i})\}_{i \geq 1}$ as

$$\lambda_t = \omega/\delta + \psi(B) [\alpha(B) y_{t-1} + f(x_{t-1}, \gamma)] = \omega/\delta + \theta(B) y_{t-1} + \psi(B) f(x_{t-1}, \gamma),$$

where $\theta(z) := \psi(z) \alpha(z) = \sum_{i=1}^{\infty} \theta_i z^i$ and we have used that x_t can be extended to the infinite past since it is a weakly dependent Markov chain by Assumptions 1-2. Thus, Z_t satisfies $Z_t = F(Z_{t-1}, Z_{t-2}, \dots; \eta_t)$, where

$$F(Z_{t-1}, Z_{t-2}, \dots; \eta_t) = (N_t (\omega/\delta + \psi(B) [\alpha(B) y_{t-1} + f(x_{t-1}, \gamma)]), g(x_{t-1}, \varepsilon_t)),$$

and $\eta_t = (N_t, \varepsilon_t)'$ is an i.i.d. sequence by Assumption 1.

Define the norm of Z_t as $\|Z_t\|_w = |y_t| + w_x \|x_t\|$, for some $w_x > 0$. Then, for any two deterministic sequences $\{z_{t-i}\}_{i \geq 1}$ and $\{\tilde{z}_{t-i}\}_{i \geq 1}$, we obtain:

$$\begin{aligned}
& E [\|F(z_{t-1}, z_{t-2}, \dots; \eta_t) - F(\tilde{z}_{t-1}, \tilde{z}_{t-2}, \dots; \eta_t)\|_w] \\
&= E \left[\left| N_0(\lambda_t) - N_0(\tilde{\lambda}_t) \right| \right] + w_x E_{t-1} [\|g(x_{t-1}; \varepsilon_t) - g(\tilde{x}_{t-1}; \varepsilon_t)\|] \\
&\leq \sum_{i=0}^{\infty} \theta_i |y_{t-1-i} - \tilde{y}_{t-1-i}| + \sum_{i=0}^{\infty} \psi_i |f(x_{t-1-i}, \gamma) - f(\tilde{x}_{t-1-i}, \gamma)| + w_x \rho \|x_{t-1} - \tilde{x}_{t-1}\| \\
&\leq \sum_{i=0}^{\infty} \theta_i |y_{t-1-i} - \tilde{y}_{t-1-i}| + \rho L \sum_{i=0}^{\infty} \psi_i \|x_{t-1-i} - \tilde{x}_{t-1-i}\| + w_x \rho \|x_{t-1} - \tilde{x}_{t-1}\| \\
&= \sum_{i=0}^{\infty} \theta_i |y_{t-1-i} - \tilde{y}_{t-1-i}| + w_x \left\{ \frac{\rho L}{w_x} \sum_{i=0}^{\infty} \psi_i \|x_{t-1-i} - \tilde{x}_{t-1-i}\| + \rho \|x_{t-1} - \tilde{x}_{t-1}\| \right\} \\
&= : \sum_{i=1}^{\infty} a_i \|z_{t-i} - \tilde{z}_{t-i}\|_w,
\end{aligned}$$

where we have used that $\theta_i \geq 0$ and $\psi_i \geq 0$, $i = 1, 2, \dots$. The coefficients $\{a_i\}_{i \geq 1}$ defined above are given by $a_1 = \max\{\theta_0, \frac{\rho L}{w_x} \psi_0 + \rho\}$ and $a_i = \max\{\theta_{i-1}, \frac{\rho L}{w_x} \psi_{i-1}\}$, $i \geq 2$. Eq. (3.1) of Doukhan and Winterberger (2008) is then satisfied with $\Phi(z) = \|z\|_w$ if $\sum_{i=1}^{\infty} a_i < 1$. This inequality will in turn hold if (i) $\theta(1) = \sum_{i=0}^{\infty} \theta_i < 1$ and (ii) $\frac{\rho L}{w_x} \sum_{i=0}^{\infty} \psi_i + \rho < 1$. By the identity $\alpha(z) = \beta(z)\theta(z)$ it follows that $\theta(1) = \alpha(1)\delta^{-1} < 1$ if and only if $\sum_{i=1}^q (\alpha_i + \beta_i) < 1$ and so (i) is satisfied. Regarding (ii), we can choose w_x arbitrarily large and so this inequality will hold if $\rho < 1$ which holds by Assumption 2. We have now verified the conditions of Corollary 3.1 in Doukhan and Wintenberger (2008) which in turn implies that $\{Z_t\}$ is weakly dependent.

To show the second part of the theorem, observe that $E[|y_t^*|^s] = \sum_{j=0}^s \binom{s}{j} E[(\lambda_t^*)^j]$ where, with $\bar{y}_t = (y_t, \dots, y_{t-p+1})'$ and $\bar{\lambda}_t = (\lambda_t, \dots, \lambda_{t-p+1})'$,

$$E[\lambda_t^*] = \sum_{i=1}^p (\alpha_i + \beta_i) E[\lambda_t^*] + \gamma E[f(x_{t-1}^*)] + \omega,$$

and

$$(\lambda_t^*)^s = \sum_{j=0}^s \binom{s}{j} (\alpha \bar{y}_{t-1}^* + \beta \bar{\lambda}_{t-1}^*)^j (\omega + \gamma f(x_{t-1}^*))^{s-j}.$$

Hence,

$$\begin{aligned}
E[(\lambda_t^*)^s] &= \sum_{j=0}^s \binom{s}{j} E \left[(\alpha \bar{y}_{t-1}^* + \beta \bar{\lambda}_{t-1}^*)^j (\omega + \gamma f(x_{t-1}^*))^{s-j} \right] \\
&= E \left[(\alpha \bar{y}_{t-1}^* + \beta \bar{\lambda}_{t-1}^*)^s + E(\omega + \gamma f(x_{t-1}^*))^s \right] + E \left[r_{s-1}(\bar{y}_{t-1}^*, \bar{\lambda}_{t-1}^*, f(x_{t-1}^*)) \right],
\end{aligned}$$

with $r_{s-1}(y, \lambda, z)$ being an $(s-1)$ -order polynomial in $(\bar{y}, \bar{\lambda}, z)$ and so $E[r_{s-1}(\bar{y}_{t-1}^*, \bar{\lambda}_{t-1}^*, f(x_{t-1}^*))] < \infty$ by induction. Moreover, $E[(\omega + \gamma f(x_{t-1}^*))^s] < \infty$ by applying Doukhan and Wintenberger (2008, Theorem 3.2) on x_t together with Assumption 2. Thus, we are left with considering terms of the form

$$\begin{aligned} E[(\alpha_i y_{t-1-i}^* + \beta_i \lambda_{t-1-i}^*)^s] &= \sum_{j=0}^s \binom{s}{j} \alpha_i^j \beta_i^{s-j} E[(y_{t-1-i}^*)^j (\lambda_{t-1-i}^*)^{s-j}] \\ &= \sum_{j=0}^s \binom{s}{j} \alpha_i^j \beta_i^{s-j} \sum_{k=0}^j \binom{j}{k} E[(\lambda_t^*)^{s+(k-j)}] \\ &= \sum_{j=0}^s \binom{s}{j} \alpha_i^j \beta_i^{s-j} E[(\lambda_t^*)^s] + C \\ &= (\alpha_i + \beta_i)^s E[(\lambda_t^*)^s] + C, \end{aligned}$$

where, by induction, $E[(\lambda_t^*)^k] < \infty$, for $k < s$. Collecting terms,

$$E[(\lambda_t^*)^s] = \left[\sum_{i=1}^p (\alpha_i + \beta_i) \right]^s E[(\lambda_t^*)^s] + \tilde{C},$$

which, since $\sum_{i=1}^p (\alpha_i + \beta_i) < 1$ (Assumption 3), has a well-defined solution.

A.2 Proof of Lemma 1

See Kristensen and Rahbek (2015).

A.3 Proof of Lemma 2

The proof mimics the proof of Lemma 2.1 in FRT. We set $p = q$ without loss of generality, such that by definition

$$\lambda_t^c - \lambda_t = \sum_{i=1}^p [\alpha_i (y_{t-i}^c - y_{t-i}) + \beta_i (\lambda_{t-i}^c - \lambda_{t-i})] + \gamma e_t^c,$$

with $e_t^c := f(x_{t-1}) \mathbb{I}(\|x_{t-1}\| \geq c)$. Hence $E[\lambda_t^c - \lambda_t] = \sum_{i=0}^{t-1} \left(\sum_{j=1}^p [\alpha_j + \beta_j] \right)^i E(e_{t-i}^c)$, and as $\sum_{j=1}^p [\alpha_j + \beta_j] < 1$, $|E(e_{t-i}^c)| \leq \zeta_1(c)$ with $\zeta_1(c) \rightarrow 0$ as $c \rightarrow \infty$, the first result

holds with $\delta_1(c) := \zeta_1(c) / \left(1 - \sum_{j=1}^p [\alpha_j + \beta_j]\right)$. Next,

$$\begin{aligned}
E(\lambda_t^c - \lambda_t)^2 &= \sum_{i=1}^p \alpha_i^2 E(y_{t-i}^c - y_{t-i})^2 + \beta_i^2 E(\lambda_{t-i}^c - \lambda_{t-i})^2 + \gamma^2 E(e_t^c)^2 \\
&+ 2 \sum_{i,j=1, i < j}^p \alpha_i \beta_j E(\lambda_{t-j}^c - \lambda_{t-j}) (y_{t-i}^c - y_{t-i}) \\
&+ 2 \sum_{i=1}^p \alpha_i E[(\lambda_{t-i}^c - \lambda_{t-i}) \gamma e_t^c] + 2 \sum_{i=1}^p \beta_i \gamma E[e_t^c (y_{t-i}^c - y_{t-i})] \\
&+ 2 \sum_{i,j=1, i < j}^p \alpha_i \alpha_j E(y_{t-j}^c - y_{t-j}) (y_{t-i}^c - y_{t-i}) \\
&+ 2 \sum_{i,j=1, i < j}^p \beta_i \beta_j E(\lambda_{t-j}^c - \lambda_{t-j}) (\lambda_{t-i}^c - \lambda_{t-i})
\end{aligned}$$

With $\lambda_t^c \geq \lambda_t$, and $t \leq s$,

$$\begin{aligned}
E[(\lambda_t^c - \lambda_t)(y_s^c - y_s)] &= E[E((\lambda_t^c - \lambda_t)(y_s^c - y_s) | \mathcal{F}_{s-1})] \\
&= E[(\lambda_t^c - \lambda_t) E(N_s[\lambda_s, \lambda_s^c])] = E(\lambda_t^c - \lambda_t)(\lambda_s^c - \lambda_s),
\end{aligned}$$

where $\mathcal{F}_{s-1} = \mathcal{F}\{x_k, N_k : k \leq s-1\}$ and $N_t[\lambda_t, \lambda_t^c]$ the number of events in $[\lambda_t, \lambda_t^c]$ for the unit-intensity Poisson process N_t . Likewise for $\lambda_t \geq \lambda_t^c$. Also observe that, still for $t \leq s$,

$$\begin{aligned}
E[(y_t^c - y_t)(y_s^c - y_s)] &= E[E((y_t^c - y_t)(y_s^c - y_s) | \mathcal{F}_{s-1})] \\
&= E[(y_t^c - y_t) E((y_s^c - y_s) | \mathcal{F}_{s-1})] \\
&= E(y_t^c - y_t)(\lambda_s^c - \lambda_s),
\end{aligned}$$

For $t \geq s$, note that the recursion for $(\lambda_t^c - \lambda_t)$ above gives,

$$\begin{aligned}
\lambda_t^c - \lambda_t &= \sum_{i=1}^p [\alpha_i (y_{t-i}^c - y_{t-i}) + \beta_i (\lambda_{t-i}^c - \lambda_{t-i})] + \gamma e_t^c \\
&= \sum_{i=1}^p \beta_i \left\{ \sum_{j=1}^p [\alpha_j (y_{t-i-j}^c - y_{t-i-j}) + \beta_j (\lambda_{t-i-j}^c - \lambda_{t-i-j})] + \gamma e_{t-i}^c \right\} \\
&+ \sum_{i=1}^p [\alpha_i (y_{t-i}^c - y_{t-i}) + \gamma e_t^c] \\
&= \dots \\
&= \sum_{j=1}^{t-s} \{a_j (y_{t-j}^c - y_{t-j}) + g_j e_{t-j}\} + \sum_{j=1}^p \{c_j (\lambda_{s-j}^c - \lambda_{s-j}) + d_j e_s^c + h_j (y_{s-j}^c - y_{s-j})\}.
\end{aligned}$$

Observe that a_j, g_j, c_j, d_j and h_j are all summable. Using this, we find,

$$\begin{aligned} E [(\lambda_t^c - \lambda_t) (y_s^c - y_s)] &= E \left[\sum_{j=1}^{t-s} (a_j (y_{t-j}^c - y_{t-j}) + g_j e_{t-j}) (y_s^c - y_s) \right] \\ &\quad + E \left[\sum_{j=1}^p [c_j (\lambda_{s-j}^c - \lambda_{s-j}) + d_j e_s^c + h_j (y_{s-j}^c - y_{s-j})] (y_s^c - y_s) \right] \end{aligned}$$

Collecting terms, one finds $E [(\lambda_t^c - \lambda_t)^2]$ is bounded by, $C \sum_{j=1}^t \psi_j E (e_{t-j}^c)^2$ for some constant C , some ψ_i with $\sum_{i=1}^{\infty} \psi_i < \infty$ and which therefore tends to zero. Finally, using again the properties of the Poisson process N_t we find,

$$E [(y_t^c - y_t)^2] \leq E [(\lambda_t^c - \lambda_t)^2] + |E [\lambda_t^c - \lambda_t]| \leq E [(\lambda_t^c - \lambda_t)^2] + \delta_1(c).$$

This completes the proof of Lemma 2.

A.4 Proof of Theorem 2

We consider here the case of $p = q = d_x = 1$ and write the model as

$$\lambda_t(\theta) = \omega + \alpha y_{t-1} + \beta \lambda_{t-1}(\theta) + \gamma f(x_{t-1});$$

the following arguments are easily extended to the general case since this is solely complicated in terms of notation. We show consistency by verifying the general conditions provided in Kristensen and Rahbek (2005, Proposition 2). Given the LLN established in Lemma 1, these are easily verified apart from the condition, $E [\sup_{\theta \in \Theta} l_t^*(\theta)] < \infty$, and showing identification. Since $\lambda_t(\theta) \geq \omega_L$,

$$\begin{aligned} E \left[\sup_{\theta \in \Theta} l_t^*(\theta) \right] &\leq E \left[\lambda_t^*(\theta_0) \sup_{\theta \in \Theta} \log \lambda_t^*(\theta) \right] - \omega_L \\ &\leq \sqrt{E [\lambda_t^*(\theta_0)^2] E \left[\sup_{\theta \in \Theta} \log \lambda_t^*(\theta)^2 \right]} - \omega_L \\ &\leq \sqrt{E [\lambda_t^*(\theta_0)^2] \left(E \left[\sup_{\theta \in \Theta} \lambda_t^*(\theta)^2 \right] - 1 \right)} - \omega_L. \end{aligned}$$

Thus, the right-hand side is finite if $E [\sup_{\theta \in \Theta} \lambda_t^*(\theta)^2] < \infty$. To show this, observe that $\lambda_t^*(\theta) \leq \lambda_t^*(\theta^U)$, where $\theta^U = (\omega^U, \alpha^U, \beta^U)$ with ω^U and α^U having been chosen as the upper bounds implied by the compactness assumption on Θ while β^U is defined in Assumption 4. As $\sum_{i=1}^q \beta_i^U < 1$, it follows by the same arguments for existence of moments in the proof of Theorem 1 that $E [\lambda_t^*(\theta^U)^2] < \infty$. Regarding identification, we need to show that

$$L(\theta) := E [l_t^*(\theta)] = E [\lambda_t^*(\theta_0) \log \lambda_t^*(\theta) - \lambda_t^*(\theta)]$$

has a unique maximum at $\theta = \theta_0$. To this end, first note that

$$\begin{aligned} L(\theta) - L(\theta_0) &= E \left[\lambda_t^*(\theta_0) \log \left(\frac{\lambda_t^*(\theta)}{\lambda_t^*(\theta_0)} \right) + \lambda_t^*(\theta_0) - \lambda_t^*(\theta) \right] \\ &\leq E \left[\lambda_t^*(\theta_0) \left(\frac{\lambda_t^*(\theta)}{\lambda_t^*(\theta_0)} - 1 \right) + \lambda_t^*(\theta_0) - \lambda_t^*(\theta) \right] \\ &= 0, \end{aligned}$$

with equality if and only if

$$\lambda_t^*(\theta_0) = \lambda_t^*(\theta) \text{ almost surely.} \quad (\text{A.1})$$

The stationary solution can be represented as

$$\lambda_t^*(\theta) = \frac{\omega}{1-\beta} + \sum_{i=1}^{\infty} a_i(\theta) y_{t-i}^* + \sum_{i=1}^{\infty} b_i(\theta) f(x_{t-i}^*),$$

where $a_i(\theta) = \alpha\beta^{i-1}$ and $b_i(\theta) = \gamma\beta^{i-1}$. Suppose now that there exists $\theta \in \Theta$ so that eq. (A.1) holds. We then claim that $\omega_0 = \omega$ and $c_i(\theta_0) = c_i(\theta)$ for all $i \geq 1$, where $c_i(\theta) = (a_i(\theta), b_i(\theta))$, which in turn implies $\theta = \theta_0$. We show this by contradiction: Let $m > 0$ be the smallest integer for which $c_i(\theta_0) \neq c_i(\theta)$ (if $c_i(\theta_0) = c_i(\theta)$ for all $i \geq 1$, then obviously $\omega_0 = \omega$). Eq. (A.1) can then be rewritten as

$$a_0 y_{t-m}^* + b_0 f(x_{t-m}^*) = \omega - \omega_0 + \sum_{i=1}^{\infty} a_i y_{t-m-i}^* + \sum_{i=1}^{\infty} b_i f(x_{t-m-i}^*),$$

where $a_i := \alpha_0\beta_0^{i-1} - \alpha\beta^{i-1}$ and $b_i := \gamma_0\beta_0^{i-1} - \gamma\beta^{i-1}$, $i = 1, 2, \dots$. The right hand side belongs to \mathcal{F}_{t-m-1} and so $a_0 y_{t-m}^* + b_0 f(x_{t-m}^*) | \mathcal{F}_{t-m-1}$ is constant. This is ruled out by Assumption 5.

To establish asymptotic normality we follow Kristensen and Rahbek (2005, proof of Theorem 2) and analyze the asymptotic behavior of the score and information which is done below.

A.4.1 Score

The score $S_T(\theta) = \partial L_T(\theta) / (\partial \theta)$ is given by,

$$S_T(\theta) = \sum_{t=1}^T s_t(\theta), \quad \text{where } s_t(\theta) = \left(\frac{y_t}{\lambda_t(\theta)} - 1 \right) \frac{\partial \lambda_t(\theta)}{\partial \theta}. \quad (\text{A.2})$$

Here, with $\eta = (\omega, \alpha, \gamma)'$ and $v_t = (1, y_{t-1}, f(x_{t-1}))'$

$$\begin{aligned} \frac{\partial \lambda_t(\theta)}{\partial \eta} &= v_t + \beta \frac{\partial \lambda_{t-1}(\theta)}{\partial \eta} \\ \frac{\partial \lambda_t(\theta)}{\partial \beta} &= \lambda_{t-1}(\theta) + \beta \frac{\partial \lambda_{t-1}(\theta)}{\partial \gamma} \end{aligned}$$

In particular, with $\lambda_t = \lambda_t(\theta_0)$ and $\dot{\lambda}_t = \partial \lambda_t(\theta) / (\partial \theta)_{\theta=\theta_0}$, $s_t(\theta_0) = \dot{\lambda}_t(y_t/\lambda_t - 1)$. The score function is a Martingale difference sequence w.r.t. the filtration \mathcal{F}_t satisfying $E[s_t(\theta_0) s_t(\theta_0)' | \mathcal{F}_{t-1}] = \dot{\lambda}_t \dot{\lambda}_t' / \lambda_t$. Note that $\dot{\lambda}_t = (v_t', \lambda_{t-1})' + \beta \dot{\lambda}_{t-1}$, with $\dot{\lambda}_0 = 0$. Thus, by the same arguments as in the proof of Theorem 1, it is easily checked that the augmented process $\tilde{X}_t := (X_t', \dot{\lambda}_t)'$, with X_t defined in Theorem 1, is weakly dependent with finite second moments. Furthermore, since $\lambda_t \geq \omega$, $\|\dot{\lambda}_t \dot{\lambda}_t' / \lambda_t\| \leq \|\dot{\lambda}_t\|^2$. It now follows by the remark following Lemma 1 that $\sqrt{T} S_T(\theta_0) \xrightarrow{d} N(0, \Omega(\theta_0))$ where, with $H(\theta)$ defined in the theorem,

$$\Omega(\theta) = E[s_t^*(\theta) s_t^*(\theta)'] = E[\dot{\lambda}_t^* (\dot{\lambda}_t^*)' / \lambda_t^*] = -H(\theta).$$

A.4.2 Information

The information is defined as

$$H_T(\theta) = -\frac{1}{T} \sum_{t=1}^T \frac{\partial^2 l_t(\theta)}{\partial \theta \partial \theta'}, \quad (\text{A.3})$$

where

$$-\frac{\partial^2 l_t(\theta)}{\partial \theta \partial \theta'} = \frac{y_t}{\lambda_t^2(\theta)} \frac{\partial \lambda_t(\theta)}{\partial \theta} \frac{\partial \lambda_t(\theta)}{\partial \theta'} - \left(\frac{y_t}{\lambda_t(\theta)} - 1 \right) \frac{\partial^2 \lambda_t(\theta)}{\partial \theta \partial \theta'},$$

and

$$\begin{aligned} \frac{\partial^2 \lambda_t(\theta)}{\partial \eta \partial \beta} &= \frac{\partial \lambda_{t-1}(\theta)}{\partial \eta} + \beta \frac{\partial^2 \lambda_{t-1}(\theta)}{\partial \eta \partial \beta}, \\ \frac{\partial^2 \lambda_t(\theta)}{\partial \beta^2} &= 2 \frac{\partial \lambda_{t-1}(\theta)}{\partial \beta} + \beta \frac{\partial^2 \lambda_{t-1}(\theta)}{\partial \beta^2}, \\ \frac{\partial^2 \lambda_t(\theta)}{\partial \eta^2} &= \beta \frac{\partial^2 \lambda_t(\theta)}{\partial \eta^2} = \dots = 0. \end{aligned}$$

These recursions can be used to show that the augmented process

$$\tilde{X}_t(\theta) := (X_t'(\theta), \dot{\lambda}_t'(\theta), \text{vec}(\ddot{\lambda}_t(\theta)))'$$

is weakly dependent with second moments for $\theta \in \Theta$ in the same way that Theorem 1 was proved. In particular, for all $\theta \in \Theta$, we can apply Lemma 1 to obtain

$$H_T(\theta) = \frac{1}{T} \sum_{t=1}^T \frac{\partial^2 l_t(\theta)}{\partial \theta \partial \theta'} \xrightarrow{P} E \left[\frac{\partial^2 l_t^*(\theta)}{\partial \theta \partial \theta'} \right].$$

Moreover, $\theta \mapsto \partial^2 l_t(\theta) / (\partial \theta \partial \theta')$ is continuous and, with $\bar{\theta} = (\omega_U, \alpha_U, \beta_U, \gamma_U)$ containing the maximum values of the individual parameters, we obtain

$$\frac{\partial \lambda_t(\theta)}{\partial \beta} = \lambda_{t-1}(\theta) + \beta \frac{\partial \lambda_{t-1}(\theta)}{\partial \gamma} \leq \sum_{i=0}^{t-1} \beta_U^i \lambda_{t-1-i}(\bar{\theta}) = \frac{\partial \lambda_t(\bar{\theta})}{\partial \beta},$$

$$\frac{\partial^2 \lambda_t(\theta)}{\partial \beta^2} = 2 \frac{\partial \lambda_{t-1}(\theta)}{\partial \beta} + \beta \frac{\partial^2 \lambda_{t-1}(\theta)}{\partial \beta^2} \leq 2 \sum_{i=0}^{t-1} \beta^i \dot{\lambda}_{t-1-i}(\bar{\theta}) = \frac{\partial^2 \lambda_t(\bar{\theta})}{\partial \beta^2},$$

and similar for the other second order derivatives. while, by the same arguments as in Han and Kristensen (2014), there exists a function $B(\tilde{x})$ so that $\lambda_t(\theta_0) / \lambda_t(\theta) \leq B(\tilde{X}_t)$ for all θ in a neighborhood of θ_0 , where

$$E \left[B(\tilde{X}_t^*) \left\| \frac{\partial \lambda_t^*(\bar{\theta})}{\partial \theta} \right\|^2 \right] < \infty, \quad E \left[B(\tilde{X}_t) \left\| \frac{\partial^2 \lambda_t^*(\bar{\theta})}{\partial \theta \partial \theta'} \right\| \right] < \infty.$$

In total,

$$\left\| \frac{\partial^2 l_t(\theta)}{\partial \theta \partial \theta'} \right\| \leq \bar{D}(\tilde{X}_t(\bar{\theta})), \quad \bar{D}(\tilde{X}_t(\bar{\theta})) := B(\tilde{X}_t) \left\{ \left\| \frac{\partial \lambda_t(\bar{\theta})}{\partial \theta} \right\|^2 + \left\| \frac{\partial^2 \lambda_t(\bar{\theta})}{\partial \theta \partial \theta'} \right\| \right\},$$

where $E[\bar{D}(\tilde{X}_t^*(\theta))] < \infty$ with \tilde{X}_t^* denoting the stationary version of \tilde{X}_t . It now follows by Proposition 1 in Kristensen and Rahbek (2005) that $\sup_{\theta \in \Theta} \|H_T(\theta) - H(\theta)\| \xrightarrow{P} 0$ with $H(\theta)$ defined in the theorem.

Finally, we show that $H(\theta_0)$ is non-singular. To see this, we use the same arguments as in the proof of identification that we provided as part of showing consistency: First note that $H(\theta_0) = E[\dot{\lambda}_t^* (\dot{\lambda}_t^*)' / \lambda_t^*]$ is singular if and only if there exists $a \in \mathbb{R}^4 \setminus \{0\}$ and $t \geq 1$ such that $a' \dot{\lambda}_t^* = 0$ a.s. Since $\dot{\lambda}_t^*$ is stationary, this must hold for all t . Recall that $\dot{\lambda}_t^* \in \mathbb{R}^4$ can be written as $\dot{\lambda}_t^* = V_t^* + \beta \dot{\lambda}_{t-1}^*$, where $V_t^* = (1, y_{t-1}^*, f(x_{t-1}^*), \lambda_{t-1}^*)'$ is a vector of positive elements. So $a' \dot{\lambda}_t^* = 0$ a.s. holds if and only if $a' V_t^* = 0$ a.s. for all $t \geq 1$. However, this is ruled out by Assumption 5, cf. proof of identification.

A.5 Proof of Theorem 3

The proof follows by noting that Lemmas 3.1-3.4 in FRT carry over to our setting with only minor modifications. The only difference is that the parameter vector θ here include γ as related to the link function $f(x_{t-1})$. However, as $E[f(x_{t-1}^*)] < \infty$, all arguments remain identical as is easily seen upon inspection of the proofs of the lemmas in FRT.

A.6 Proof of Corollary 1

It suffices to verify the regularity conditions of Andrews (1999, Theorem 3). First, in the proof of Theorem 2 we establish consistency and classic convergence of the score and information. Second, the parameter set satisfies the geometric conditions needed by arguments identical to the ones in Francq and Zakoian (2009).

References

- [1] Amisano, G. and R. Giacomini (2007), “Comparing Density Forecasts via Weighted Likelihood Ratio Tests”, *Journal of Business and Economic Statistics* 25, 177–190.
- [2] Andrews, D.W.K. (1999), “Estimation when a parameter is on a boundary”, *Econometrica*, 67, 1341–1383.
- [3] Andrews, D.W.K. (2000), “Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space”, *Econometrica*, 68, 399–405.
- [4] Azizpour, S., K. Giesecke, G. Schwenkler (2015), “Exploring the Sources of Default Clustering”, working paper.
- [5] Berkes, I., L. Horváth and P. Kokoszka (2003), “GARCH processes: Structure and estimation”, *Bernoulli* 9, 201–227.
- [6] Besag, J. (1975), “Statistical Analysis of Non-Lattice Data”, *The Statistician* 24, 179–195.
- [7] Brown, B. (1971), “Martingale central limit theorems”, *Annals of Mathematical Statistics*, 42, 59–66.
- [8] Christou, V. and K. Fokianos (2013), “Quasi-likelihood inference for negative binomial time series models”, *Journal of Time Series Analysis* 35, 55–78.
- [9] Davis, R.A. and H. Liu, (2014), “Theory and Inference for a Class of Nonlinear Models with Application to Time Series of Counts”, *Statistica Sinica*, forthcoming.
- [10] Das, S.R., D. Duffie, N. Kapadia, and L. Saita (2007), “Common failings: How corporate defaults are correlated”, *Journal of Finance* 62, 93–117.
- [11] Demos, A. and E. Sentana (1998), “Testing for GARCH effects: A one-sided approach”, *Journal of Econometrics* 86, 97–127.
- [12] Doukhan, P. and O. Wintenberger (2008), “Weakly dependent chains with infinite memory”, *Stochastic Processes and their Applications* 118, 1997–2013.
- [13] Duffie, D., A. Eckner, G. Horel, and L. Saita (2009), “Frailty correlated default”, *Journal of Finance* 64, 2089–2123.
- [14] Duffie, D., Saita, L., Wang, K. (2007), “Multi-period corporate default prediction with stochastic covariates”, *Journal of Financial Economics* 83, 635–665.

- [15] Ferland, R., A. Latour and D. Oraichi (2006), “Integer-valued GARCH processes”, *Journal of Time Series Analysis*, 27, 923–942.
- [16] Fokianos, K., A. Rahbek and D. Tjøstheim (2009), “Poisson autoregression”, *Journal of the American Statistical Association*, 104, 1430–1439.
- [17] Francq, C. and J.M. Zakoïan (2009), “Testing the nullity of GARCH coefficients: Correction of the standard tests and relative efficiency comparisons”, *Journal of the American Statistical Association*, 104, 313–324.
- [18] Gallo, G.M. and B. Pacini (2000), “The effects of trading activity on market volatility”, *European Journal of Finance* 6, 163–175
- [19] Giesecke, K., F. A. Longstaff, S. Schaefer and I. Strebulaev (2011), “Corporate bond default risk: A 150-year perspective”, *Journal of Financial Economics* 102, 233–250.
- [20] Giesecke, K. and B. Kim (2011), “Systemic Risk: What Defaults Are Telling Us”, *Management Science* 57, 1387–1405.
- [21] Gouriéroux, C., A. Monfort, E. Renault and A. Trognon (1987), “Generalised residuals”, *Journal of Econometrics*, 34, 5–32.
- [22] Gouriéroux, C., A. Monfort and A. Trognon (1984), “Pseudo maximum likelihood methods: Applications to Poisson models”, *Econometrica*, 52, 701–720.
- [23] Han, H., and Kristensen, D. (2014), “Asymptotic theory for the QMLE in GARCH-X models with stationary and non-stationary covariates”, *Journal of Business and Economic Statistics* 32, 416–429.
- [24] Hansen, P.R., Z. Huang and H.W. Shek (2012) “Realized GARCH: A joint model for returns and realized measures of volatility”, *Journal of Applied Econometrics* 27, 877–906.
- [25] Kedem, B., and K. Fokianos (2002), *Regression Models for Time Series Analysis*, Hoboken, NJ: Wiley.
- [26] Koopman, S.J., A. Lucas and B. Schwaab (2012), “Dynamic factor models with macro, frailty, and industry effects for U.S. default counts: the credit crisis of 2008”, *Journal of Business and Economic Statistics* 30, 521–532.
- [27] Koopman, S.J., A. Lucas and B. Schwaab (2014), “Modeling frailty-correlated defaults using many macroeconomic covariates”, *Journal of Econometrics* 162, 312–325.

- [28] Kristensen, D. and A. Rahbek (2005), “Asymptotics of the QMLE for a class of ARCH(q) models”, *Econometric Theory*, 21, 946–961.
- [29] Kristensen, D. and A. Rahbek (2015), “Quasi-maximum likelihood estimation of multivariate GARCH models: A weak dependence approach”, working paper.
- [30] Lamoureux, C.G., and W.D. Lastrapes (1990), “Heteroskedasticity in stock return data: Volume versus GARCH effects”, *Journal of Finance*, 45, 221–229.
- [31] Lando, D., and M. Nielsen (2010), “Correlation in corporate defaults: Contagion or conditional independence?”, *Journal of Financial Intermediation*, 19, 355–372.
- [32] Lando, D., M. Medhat, M. Stenbo Nielsen and S.F. Nielsen (2013), “Additive intensity regression models in corporate default analysis”, *Journal of Financial Econometrics* 11, 443–485.
- [33] Meitz, M., and P. Saikkonen (2008), “Ergodicity, mixing and existence of moments of a class of Markov models with applications to GARCH and ACD models”, *Econometric Theory*, 24, 1291–1320.
- [34] Nyblom, J. (1989), “Testing for the constancy of parameters over time”, *Journal of The American Statistical Association* 84, 223–30.
- [35] Rydberg, T. H., and N. Shephard (2000), “A modeling framework for the prices and times of trades on the New York Stock Exchange”, in *Nonlinear and Nonstationary Signal Processing*, eds. W. J. Fitzgerlad, R. L. Smith, A. T. Walden, and P. C. Young, Cambridge: Isaac Newton Institute and Cambridge University Press, pp. 217–246.
- [36] Shephard, N. and K. Sheppard (2010), “Realising the future: Forecasting with high-frequency-based volatility (HEAVY) models”, *Journal of Applied Econometrics* 25, 197–231.
- [37] Stock, J. and M. Watson (1997) “Evidence on structural instability in macroeconomic time series relations”, *Journal of Business and Economic Statistics* 14, 11–30.
- [38] Streett, S. (2000), “Some observation driven models for time series of counts”, Ph.D. thesis, Colorado State University, Dept. of Statistics.
- [39] Tay, A.S and K.F. Wallis (2000) “Density forecasting: A survey”, *Journal of Forecasting*, 19, 235–254.
- [40] White, H. (1982) “Maximum likelihood estimation of misspecified models”, *Econometrica*, 50, 1–25.

Table 1: Results of simulations for PARX(1,1) with DGP 1.

		Scenario 1 ($\beta = 0$)				Scenario 2 ($\beta = 0.2$)			Scenario 3 ($\beta = 0.7$)		
T		True	Mean	RMSE	KS	Mean	RMSE	KS	Mean	RMSE	KS
100	ω	0.10	0.09	0.16	0.36	0.10	0.18	0.01	0.15	0.30	0.07
	α	0.30	0.28	0.13	0.32	0.27	0.11	0.97	0.18	0.15	0.00
	β	0.00	0.02	0.15	0.31	0.22	0.14	0.34	0.77	0.15	0.00
	γ	0.50	0.51	0.07	0.85	0.51	0.07	0.32	0.51	0.11	0.84
250	ω	0.10	0.09	0.07	0.85	0.10	0.08	0.19	0.13	0.21	0.13
	α	0.30	0.30	0.07	0.87	0.29	0.07	0.99	0.23	0.06	0.72
	β	0.00	0.00	0.08	0.93	0.21	0.08	0.63	0.72	0.06	0.64
	γ	0.50	0.50	0.04	0.49	0.50	0.04	0.92	0.50	0.05	0.81
500	ω	0.10	0.10	0.05	0.66	0.10	0.05	0.35	0.11	0.13	0.21
	α	0.30	0.30	0.04	0.33	0.30	0.04	0.87	0.24	0.04	0.86
	β	0.00	0.00	0.05	0.17	0.20	0.05	0.16	0.71	0.04	0.96
	γ	0.50	0.50	0.02	0.34	0.50	0.02	0.75	0.50	0.02	0.95
1000	ω	0.10	0.10	0.03	0.38	0.10	0.04	0.42	0.10	0.10	0.24
	α	0.30	0.30	0.03	0.52	0.30	0.03	0.61	0.24	0.02	0.79
	β	0.00	0.00	0.03	0.98	0.20	0.03	0.71	0.71	0.02	0.81
	γ	0.50	0.50	0.02	0.74	0.50	0.02	0.32	0.50	0.02	0.99

Table 2: Results of simulations for PARX(1,1) with DGP 2.

		Scenario 1 ($\beta = 0$)				Scenario 2 ($\beta = 0.2$)			Scenario 3 ($\beta = 0.7$)		
T		True	Mean	RMSE	KS	Mean	RMSE	KS	Mean	RMSE	KS
100	ω	0.10	0.12	0.20	0.00	0.11	0.18	0.00	0.16	0.30	0.02
	α	0.30	0.29	0.13	0.47	0.27	0.13	0.43	0.17	0.16	0.00
	β	0.00	-0.01	0.23	0.16	0.21	0.19	0.31	0.78	0.16	0.00
	γ	0.50	0.51	0.13	0.50	0.51	0.12	0.32	0.51	0.14	0.81
250	ω	0.10	0.10	0.09	0.14	0.12	0.12	0.08	0.18	0.25	0.00
	α	0.30	0.30	0.07	0.70	0.29	0.07	0.57	0.23	0.05	0.58
	β	0.00	0.00	0.10	0.33	0.20	0.10	0.81	0.71	0.06	0.84
	γ	0.50	0.50	0.06	0.39	0.50	0.07	0.85	0.51	0.14	0.30
500	ω	0.10	0.10	0.07	0.22	0.10	0.07	0.54	0.13	0.14	0.00
	α	0.30	0.30	0.05	0.95	0.30	0.05	0.96	0.24	0.04	0.47
	β	0.00	0.00	0.07	1.00	0.20	0.07	0.90	0.71	0.04	0.29
	γ	0.50	0.50	0.04	0.59	0.50	0.05	0.97	0.51	0.07	0.46
1000	ω	0.10	0.10	0.05	0.73	0.10	0.05	0.14	0.12	0.11	0.02
	α	0.30	0.30	0.03	0.81	0.30	0.03	0.56	0.24	0.02	0.95
	β	0.00	0.00	0.05	0.82	0.20	0.05	0.80	0.70	0.03	0.97
	γ	0.50	0.50	0.03	0.74	0.50	0.03	0.43	0.51	0.05	0.77

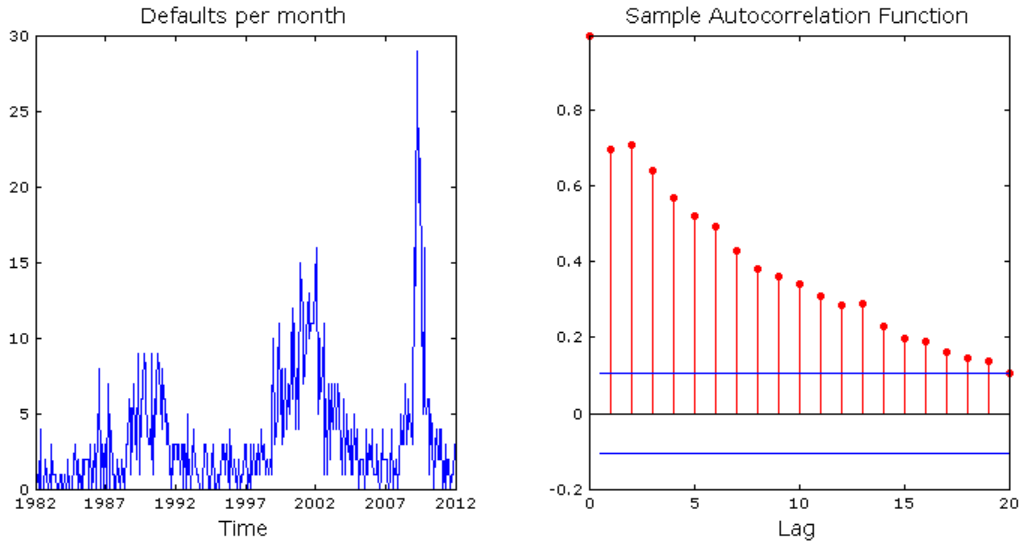


Figure 1: (a) Number of defaults per month among Moody's rated US industrial firms in the period 1982-2011. (b) Autocorrelation function of the default data.

Table 3: Estimation results of different PARX models.

	PAR	RV	SP	DG	NB	IP ⁽⁻⁾	LI ⁽⁻⁾	RV & LI ⁽⁻⁾	All
$\hat{\omega}$	0.301 (3.625)	0.169 (2.467)	0.116 (0.716)	0.206 (2.219)	0.289 (3.551)	0.202 (2.142)	0.295 (2.013)	0.232 (3.242)	0.208 (1.001)
$\hat{\alpha}_1$	0.241 (5.441)	0.197 (4.395)	0.227 (5.159)	0.221 (4.933)	0.228 (5.119)	0.213 (4.716)	0.193 (4.265)	0.185 (4.109)	0.180 (3.944)
$\hat{\alpha}_2$	0.215 (3.221)	0.179 (2.908)	0.2217 (3.348)	0.198 (3.026)	0.206 (3.138)	0.145 (2.262)	0.198 (3.117)	0.188 (3.039)	0.183 (2.898)
$\hat{\beta}$	0.459 (6.094)	0.526 (7.939)	0.4298 (5.430)	0.455 (6.063)	0.469 (6.296)	0.552 (8.173)	0.498 (6.881)	0.518 (7.547)	0.512 (7.087)
RV		63.99 (4.111)						28.09 (2.057)	24.31 (1.692)
SP			0.241 (2.802)						0.000 (0.000)
DG				0.017 (1.893)					0.006 (0.640)
NB					0.419 (2.229)				0.000 (0.000)
IP						0.695 (3.287)			0.000 (0.000)
LI							0.941 (4.194)	0.729 (3.733)	0.754 (1.561)
$\hat{\alpha}_1 + \hat{\alpha}_2$	0.465 (7.452)	0.376 (8.069)	0.449 (9.635)	0.419 (8.743)	0.434 (9.249)	0.358 (7.380)	0.391 (7.726)	0.373 (6.679)	0.363 (7.235)
AIC	-1352.04	-1368.82	-1359.86	-1352.88	-1354.94	-1360.52	-1375.06	-1377.52	-1365.84
BIC	-1336.47	-1349.36	-1340.40	-1333.42	-1335.48	-1337.17	-1351.71	-1354.17	-1319.14

Notes: t statistics in parentheses. For any significance level $\eta < 1/2$ standard critical values for one sided t tests apply, see Remark 6.

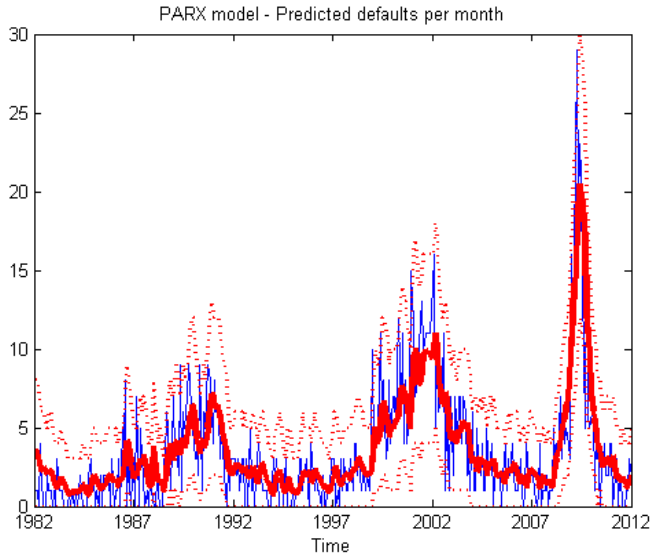


Figure 2: Actual number of defaults (blue) and estimated intensity (red).

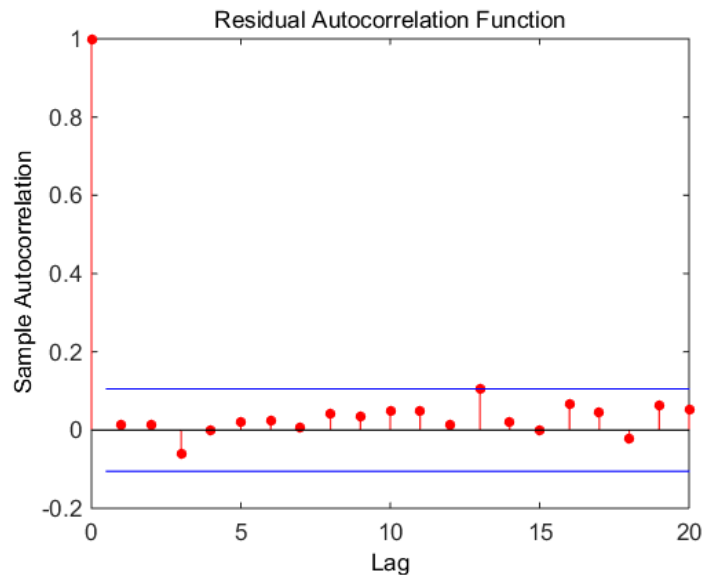


Figure 3: Sample autocorrelation function of Pearson residuals.

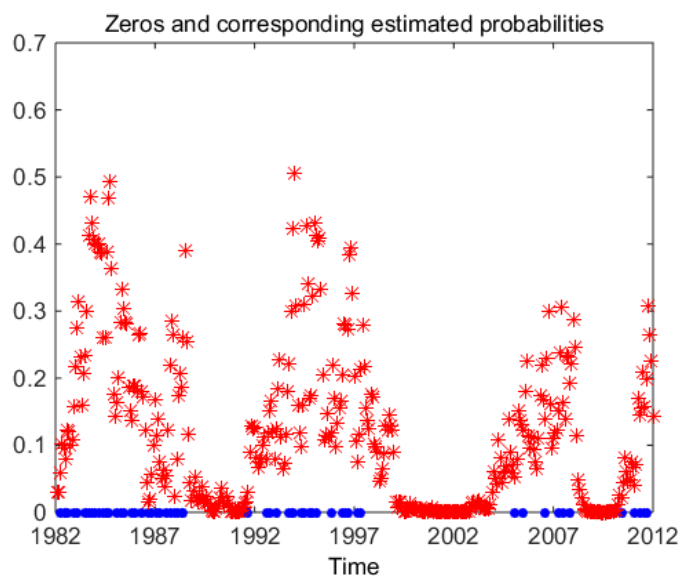


Figure 4: Empirical zero counts (asterisks) and probability of having a zero count under the estimated model (crosses).

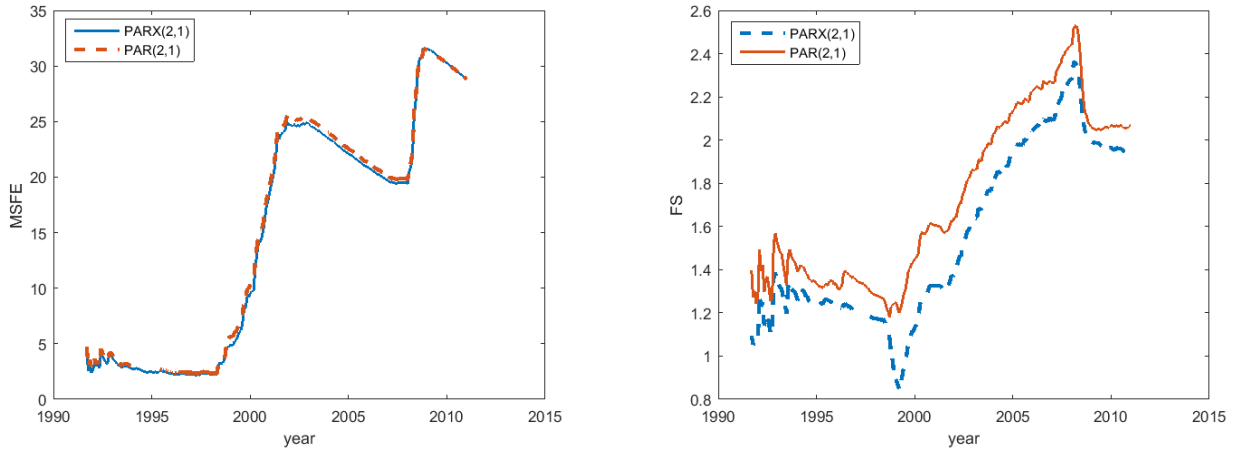


Figure 5: Rolling-window MSFE and FS of PAR and PARX model

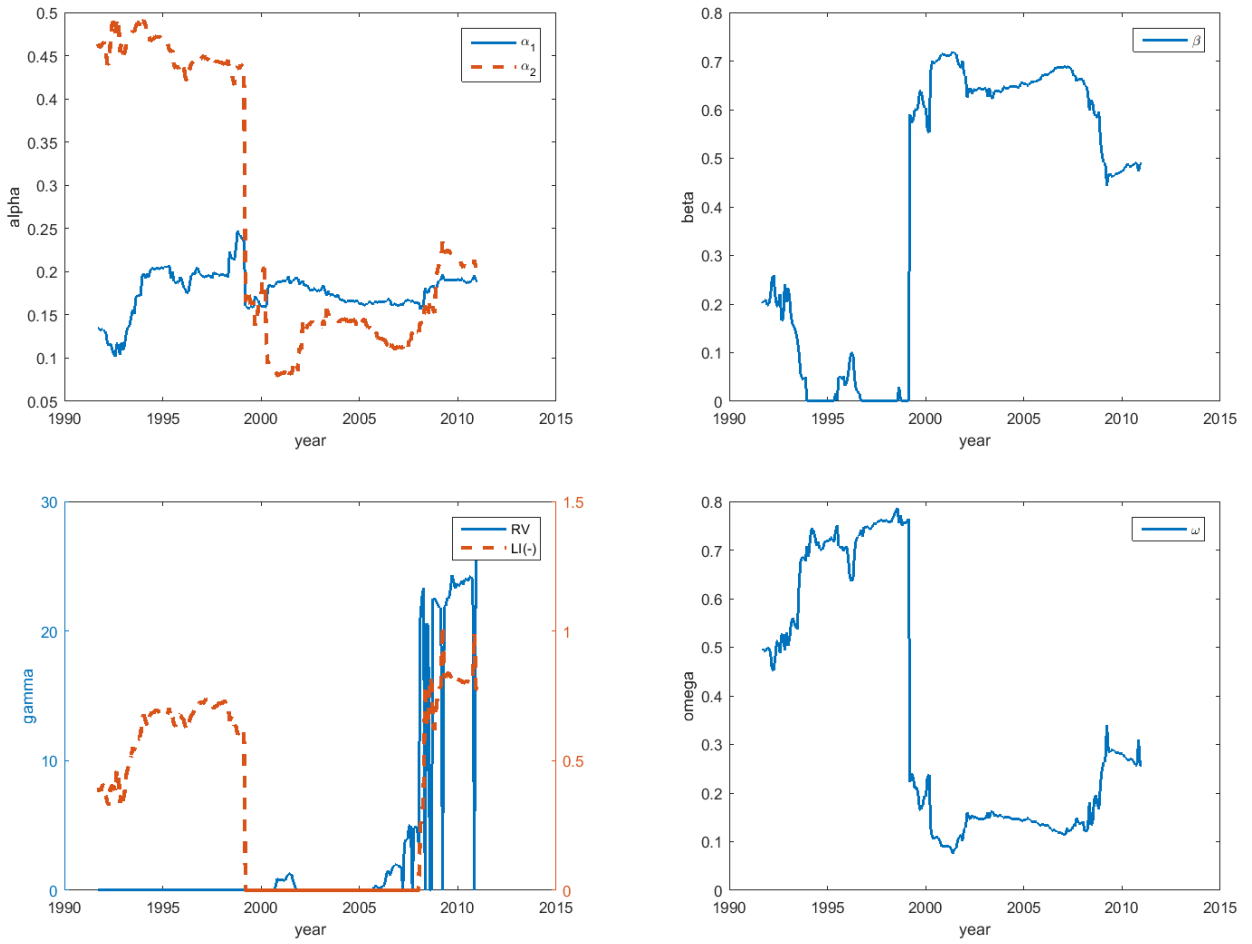


Figure 6: Rolling-window estimate of $\theta = (\omega, \alpha_1, \alpha_2, \beta, \gamma)'$ in the preferred PARX(2,1) model.

Table 4: Preferred models and their parameter estimates, 1982-1998, 1998-2007 and 2007-2011

	ω	α_1	α_2	β	RV	LI ⁽⁻⁾
1982-1998 - PAR(1,1)	0.80	0.22	0.43	-	-	-
<i>t</i> -stats	(7.04)	(5.29)	(8.32)	-	-	-
1998-2007 - PARX(1,1)	0.00	0.20	-	0.79	-	-
<i>t</i> -stats	0.10	(4.19)	-	(14.8)	-	-
2007-2011 - PARX(2,1)	0.00	-	-	0.82	99.23	0.70
<i>t</i> -stats	(0.00)	-	-	(6.81)	(2.17)	(2.38)