Habilitation Thesis
October 2023

Dynamic Score-Driven Models

An R Package with Applications

Vladimír Holý

Prague University of Economics and Business
Faculty of Informatics and Statistics
Department of Econometrics

Title: Dynamic Score-Driven Models: An R Package with Applications

Author: Mgr. Vladimír Holý, Ph.D.

Abstract: Score-driven models are a class of time series models that leverage the score, i.e. the gradient of the log-likelihood function, to iteratively update time-varying parameters of the underlying probability distribution. This thesis comprises seven papers devoted to score-driven models. The first paper introduces an R package called gasmodel, which is designed to facilitate the estimation, forecasting, and simulation of a wide range of score-driven models. The remaining papers focus on the development of specific score-driven models within the fields of sport statistics, operations research, and finance.

Keywords: Generalized Autoregressive Score Models, Dynamic Conditional Score Models, Score-Driven Models.

JEL Classification: C22, C87.

The research presented in this habilitation thesis was supported by the Czech Science Foundation under project 23-06139S: "Dynamic Score-Driven Models in Operations Research".

I would like to express my gratitude to Francisco Blasques, Michal Černý, Josef Jablonský, Ondřej Sokol, Petra Tomanová, and Jan Zouhar for their collaboration and invaluable comments. Additionally, I extend my thanks to Kateřina Holá for her unwavering support and to Metoděj Holý for being a constant source of inspiration.

Vladimír Holý

Foreword

Score-driven models, also known as generalized autoregressive score (GAS) models and dynamic conditional score (DCS) models, constitute a contemporary framework for time series modeling. Within this framework, dynamic models can be built upon any given probability distribution, allowing for any parameter to be time-varying. The pivotal element enabling this versatility is the incorporation of the score, i.e. the gradient of the log-likelihood function, in model dynamics. The score-driven model based on the normal distribution with time-varying mean corresponds to the autoregressive moving average (ARMA) model, while the score-driven model based on the normal distribution with time-varying volatility corresponds to the generalized autoregressive conditional heteroskedasticity (GARCH) model. Utilizing other distributions, however, leads to the development of entirely novel models suitable for a wide array of univariate and multivariate data types, including non-negative, count, integer, and ranking data. Score-driven models are classified as observation-driven, but several empirical studies have found that, in general, they have comparable performance to parameter-driven models while offering straightforward estimation through the maximum likelihood method.

This habilitation thesis consists of seven papers dedicated to score-driven models, either authored entirely by me or in which I have made significant contributions. Additionally to writing these papers, I have developed an R package named gasmodel, which is designed to facilitate the estimation, forecasting, and simulation of a wide range of score-driven models. The first paper in this collection presents the framework of score-driven models, reviews the score-driven literature, and explains the gasmodel package, while the subsequent papers delve into specific score-driven models and their respective applications.

The contents of the papers in this collection are outlined below:

1. Holý V (2023). "gasmodel: An R Package for Generalized Autoregressive Score Models." In review in *Journal of Statistical Software*.

This paper serves as a companion to the gasmodel package. It provides a comprehensive description of the package's functionality and offers practical illustrations of its usage. The aim of the package is to provide flexible customization, enabling users to incorporate various parametrizations, exogenous variables, joint and separate modeling of exogenous variables and dynamics, higher score and autoregressive orders, custom and unconditional initial values of time-varying parameters, fixed and bounded values of coefficients, and handling missing values. It offers a selection of 26 distributions, catering to various univariate and multivariate data types such as binary, categorical, ranking, count, integer, circular, interval, compositional, duration, and real data. Model estimation is performed using the maximum likelihood method and the Hessian matrix. Furthermore, the package offers a range of functionalities, including forecasting, simulation, bootstrapping, and assessment of parameter uncertainty. Two case studies are presented to showcase the package's utility: the analysis of the timing of bookshop orders and the analysis of ice hockey rankings. The package is also compared to an alternative package called GAS, which allows only for basic model formulation without exogenous variables and offers a limited range of distributions.

2. Holý V, Zouhar J (2022). "Modelling Time-Varying Rankings with Autoregressive and Score-Driven Dynamics." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **71**(5), 1427–1450. ISSN 0035-9254. https://doi.org/10.1111/rssc.12584.

This paper introduces an innovative score-driven model designed for dynamic rankings, utilizing the Plackett-Luce distribution, which is based on the Luce's choice axiom, with time-varying worth parameters. The model's effectiveness is demonstrated through its application to the outcomes of the Ice Hockey World Championships, and potential applications in other domains are explored. This contribution holds substantial significance, particularly in light of the limited existing literature addressing the dynamics of ranking data. The model can be used with a large number of ranked items, accommodates exogenous time-varying covariates and partial rankings, and is estimated via the maximum likelihood in a straightforward manner. Simulation experiments show that the small-sample properties of the maximum-likelihood estimator improve rapidly with the length of the time series and suggest that statistical inference relying on conventional Hessian-based standard errors is usable even for medium-sized samples. The empirical application of the model to the Ice Hockey World Championships from 1998 to 2019 underscores its practicality. It is found that the mean-reverting model offers a superior fit to the data compared to both the static and random walk models. This approach presents several key advantages, including the compilation of the ultimate (long-term) ranking of teams, the straightforward estimation of the probabilities of specific rankings (e.g., podium positions), and the prediction of future rankings. Furthermore, this paper discusses potential applications to rankings based on underlying indices, repeated surveys, and non-parametric efficiency analysis.

3. Holý V (2023). "Ranking-Based Second Stage in Data Envelopment Analysis: An Application to Research Efficiency in Higher Education." https://arxiv.org/abs/2307.01869. In review in *Annals of Operations Research*.

This paper is a follow-up study that explores the use of the score-driven ranking model in the context of two-stage data envelopment analysis (DEA). In DEA research, it is common to follow efficiency measurements with a second-stage regression analysis using efficiency scores as dependent variables and contextual (or environmental) variables as independent variables. Often, efficiency is assessed annually, requiring a panel regression as the second-stage model to account for time-varying contextual factors. The most commonly used panel methods for the second stage include panel linear regression and panel Tobit regression. This paper proposes an alternative approach to the second stage of DEA, suggesting the use of rankings instead of efficiency scores. The score-driven ranking model proves valuable as it avoids problems specific to efficiency scores. It is also somewhat robust to the chosen DEA technique. The empirical part of the paper focuses on assessing research efficiency in higher education among European Union (EU) countries by analyzing scientific publications from 2005 to 2020. In the first stage, DEA analysis is conducted for each year independently, using gross domestic expenditure on research and development (R&D) and the number of researchers as inputs to reflect financial and human resources, respectively. For outputs, the number of publications and the number of citations are used to reflect the quantity and quality of scientific research, respectively. In the second stage, rather than relying solely on efficiency scores, the paper advocates for incorporating rankings using the score-driven ranking model, emphasizing its potential as a robustness check. When employed to assess research efficiency in the higher education sector and its connection with good governance, the approach confirms a positive relation with the Voice and Accountability indicator found in standard panel linear regression, while suggesting caution regarding the Government Effectiveness indicator.

4. Holý V, Tomanová P (2022). "Modeling Price Clustering in High-Frequency Prices." *Quantitative Finance*, **22**(9), 1649–1663. ISSN 1469-7688. https://doi.org/10.1080/14697688.2022.2050285.

In finance, the price clustering refers to an increased occurrence of specific prices. Notably, it arises from the activities of distinct agent types, trading exclusively in particular multiples of the tick size, resulting in the frequent appearance of these multiples in price levels. As a case in point, stocks on prominent exchanges like NYSE and NASDAQ exhibit precision trading to the nearest cent, yet multiples of five and ten cents manifest more frequently in price levels. This phenomenon, observed across various financial instruments and markets, however, is rarely integrated into existing price models. To address this behavior, this paper introduces a novel discrete score-driven model for prices, leveraging a dynamic mixture of double Poisson distributions that accommodates both dynamic volatility and the evolving proportions of agent types. An empirical study of 30 Dow Jones Industrial Average (DJIA) stocks from the first half of 2020 reveals intriguing findings. Analyzing price clustering daily, in alignment with prevailing literature approaches, shows that daily volatility positively influences price clustering. However, upon employing the high-frequency price model, a contrasting observation emerges: instantaneous volatility inversely affects price clustering. Hence, data aggregation levels critically influence the relationship between price clustering and volatility. At a granular level, while heightened daily volatility corresponds with intensified price clustering, increased instantaneous volatility yields the reverse effect. Concurrently, volume amplifies the impact on price clustering, whereas factors like price and last trade duration remain statistically inconsequential.

5. Holý V (2023). "An Intraday GARCH Model for Discrete Price Changes and Irregularly Spaced Observations." https://arxiv.org/abs/2211.12376. In review in *Annals of Operations Research*.

This study presents an innovative approach to modeling high-frequency time series of prices, addressing their unique characteristics such as irregularly spaced observations, simultaneous transactions, discrete price levels, and the presence of market microstructure noise. The proposed model leverages smoothing splines to capture the relation between trade durations and price volatility, as well as intraday patterns of trade durations and price volatility. Grounded in the zero-inflated Skellam distribution with a newly proposed overdispersion parametrization, this dynamic model incorporates time-varying volatility within the score-driven framework, and effectively filters market microstructure noise using a moving average component. While other models in the literature also address these issues, this is the first model to integrate all four components. Empirical analysis, conducted on data of the IBM stock traded on the New York Stock Exchange (NYSE), demonstrates the model's ability to provide a robust fit to the observed high-frequency price data. It is found that volatility per second decreases with increasing trade duration, which is consistent with the literature. Beyond its utility in modeling intraday volatility, this model also proves valuable for measuring daily realized volatility as a parametric alternative to realized kernels and similar measures, filtering both diurnal patterns and market microstructure noise. Additionaly, the results for the CA, CSCO, EA, INTC, MA, and MCD stocks traded on the NYSE and NASDAQ exchanges are also reported providing further empirical evidence.

6. Blasques F, Holý V, Tomanová P (2022). "Zero-Inflated Autoregressive Conditional Duration Model for Discrete Trade Durations with Excessive Zeros." https://arxiv.org/abs/1812.07318. In review in *Studies in Nonlinear Dynamics & Econometrics*.

Simultaneous transactions must also be considered in models aiming to capture the time intervals between transactions within the autoregressive conditional duration (ACD) literature. Typically, in the ACD literature, models are constructed based on continuous distributions that do not accommodate zero values within their support. All zero durations are usually assumed to correspond to split transactions and are subsequently discarded from the dataset. The present study, however, advocates against the removal of zero durations from the data. It posits that zero durations can be associated not only with split transactions but also with independent transactions. Furthermore, it highlights that split transactions can produce both zero and positive values of durations. In response to these considerations, the paper proposes a discrete model capable of effectively handling an abundance of zero values. This novel model employs the zero-inflated negative binomial distribution with score dynamics, incorporating mean, overdispersion, and zero-inflation parameters, all of which are treated as time-varying. This model enables the distinction between the processes generating split and independent transactions. The study leverages the asymptotic theory on score models to establish the invertibility of the score filter and verify that sufficient conditions hold for the consistency and asymptotic normality of the maximum likelihood of the model parameters, in the case of time-varying mean. An empirical investigation, encompassing data from six stocks traded on the Euronext, NYSE, and NASDAQ exchanges, uncovers that split transactions account for approximately 92 to 98 percent of durations smaller than 0.01 seconds. Intriguingly, the loss of decimal places in the proposed approach is less severe than the incorrect treatment of zero values in continuous models.

7. Tomanová P, Holý V (2021). "Clustering of Arrivals in Queueing Systems: Autoregressive Conditional Duration Approach." *Central European Journal of Operations Research*, **29**(3), 859–874. ISSN 1435-246X. https://doi.org/10.1007/s10100-021-00744-7.

This paper explores the application of score-driven ACD models in the context of queueing systems. Contrary to the typical assumption that arrivals are independent and exponentially distributed, the empirical analysis of an online bookshop demonstrates an underlying autocorrelation structure in inter-arrival times. To account for diurnal and seasonal variations, a cubic spline approach is employed, with parameter estimation executed via the weighted ordinary least square method. Following this adjustment, it becomes apparent that the score-driven model, based on the generalized gamma distribution and its special cases fit, provides a more faithful depiction compared to their static counterparts. A simulation study underscores that ignoring the autocorrelation structure leads to biased performance measures within queueing systems with single and multiple servers. The number of customers in the system, the busy periods of the servers, and the response times, exhibit higher means and variances as well as heavier tails for the proposed dynamic arrivals model than for the standard static model. By relying on a conventional static model, businesses risk making suboptimal decisions, which could culminate in lost profits. Ultimately, this research underscores the economic imperative of correctly modeling arrival dependencies, offering invaluable insights for process simulations, optimization, and quality assessment.

gasmodel: An R Package for Generalized Autoregressive Score Models

Vladimír Holý
Prague University of Economics and Business
Winston Churchill Square 1938/4, 130 67 Prague 3, Czechia
vladimir.holy@vse.cz

Abstract: Generalized Autoregressive Score (GAS) models are a class of observation-driven time series models that employ the score to dynamically update time-varying parameters of the underlying probability distribution. GAS models have been extensively studied and numerous variants have been proposed in the literature to accommodate diverse data types and probability distributions. This paper introduces the **gasmodel** package, which has been designed to facilitate the estimation, forecasting, and simulation of a wide range of GAS models. The package provides a rich selection of distributions, offers flexible options for specifying dynamics, and allows to incorporate exogenous variables. Model estimation utilizes the maximum likelihood method and the Hessian matrix.

Keywords: Generalized Autoregressive Score Models, Dynamic Conditional Score Models, Score-Driven Models, R.

JEL Classification: C22, C87.

1 Introduction

The generalized autoregressive score (GAS) models, introduced by Creal *et al.* (2013) and Harvey (2013), have emerged as a valuable and contemporary framework for time series modeling. These models, also referred to as dynamic conditional score (DCS) models or score-driven models, offer flexibility by accommodating various underlying probability distributions and time-varying parameters. GAS models are observation-driven, effectively capturing the dynamic behavior of time-varying parameters through the autoregressive term and the score, i.e., the gradient of the log-likelihood function. Within the GAS framework, it is possible to formulate a wide range of dynamic models for any type of data.

There are several packages and code available in R that handle GAS models. One notable package is **GAS** developed by Ardia *et al.* (2019), which provides functionality for both univariate and multivariate GAS models. The current version, 0.3.4, supports 16 distributions. However, the model specification in the **GAS** package is somewhat limited, only allowing for basic dynamics without the inclusion of exogenous variables. Additionally, this package lacks distributions for certain more specialized data types, such as circular, compositional, and ranking data. The package thus supports only a limited selection of GAS models found in the literature[1]. Another relevant R package is **betategarch** by Sucarrat (2013), which deals specifically with the Beta-Skew-t-EGARCH model, a GAS model for time-varying volatility based on the Student's t-distribution. In Python, the **PyFlux** library by Taylor (2018) deals with time series analysis and features various GAS models including the Beta-Skew-t-EGARCH model, standard GAS models, GAS random walk models, GAS pairwise comparison models, and GAS regression models. In Julia, the **ScoreDrivenModels.jl** package by Bodin *et al.* (2020) provides a framework for standard GAS models. The **Time Series Lab** program by Lit *et al.* (2021) is a stand-alone GUI application designed to model and forecast time series, including standard GAS models, GAS pairwise comparison models, and GAS regression models. Additional R, MATLAB, and Ox code for some specific GAS models, often associated with individual research papers, can be found on the `www.gasmodel.com` website.

In this paper, we present the **gasmodel** package, which is designed to provide comprehensive functionality that encompasses a wide range of GAS models documented in the existing literature. It

[1]For a more detailed comparison of the **gasmodel** and **GAS** packages, see Appendix A.

Table 1: List of available distributions and their parametrizations. First parametrization is the default.

Label	Distribution	Dimension	Data Type	Parametrizations
`alaplace`	Asymmetric Laplace	Univariate	Real	`meanscale`
`bernoulli`	Bernoulli	Univariate	Binary	`prob`
`beta`	Beta	Univariate	Interval	`conc`, `meansize`, `meanvar`
`bisa`	Birnbaum–Saunders	Univariate	Duration	`scale`
`cat`	Categorical	Multivariate	Categorical	`worth`
`dirichlet`	Dirichlet	Multivariate	Compositional	`conc`
`dpois`	Double Poisson	Univariate	Count	`mean`
`exp`	Exponential	Univariate	Duration	`scale`, `rate`
`gamma`	Gamma	Univariate	Duration	`scale`, `rate`
`gengamma`	Generalized Gamma	Univariate	Duration	`scale`, `rate`
`geom`	Geometric	Univariate	Count	`mean`, `prob`
`laplace`	Laplace	Univariate	Real	`meanscale`
`mvnorm`	Multivariate Normal	Multivariate	Real	`meanvar`
`mvt`	Multivariate Student's t	Multivariate	Real	`meanvar`
`negbin`	Negative Binomial	Univariate	Count	`nb2`, `prob`
`norm`	Normal	Univariate	Real	`meanvar`
`pluce`	Plackett–Luce	Multivariate	Ranking	`worth`
`pois`	Poisson	Univariate	Count	`mean`
`skellam`	Skellam	Univariate	Integer	`meanvar`, `diff`, `meandisp`
`t`	Student's t	Univariate	Real	`meanvar`
`vonmises`	von Mises	Univariate	Circular	`meanconc`
`weibull`	Weibull	Univariate	Duration	`scale`, `rate`
`zigeom`	Zero-Inflated Geometric	Univariate	Count	`mean`
`zinegbin`	Zero-Inflated Negative Binomial	Univariate	Count	`nb2`
`zipois`	Zero-Inflated Poisson	Univariate	Count	`mean`
`ziskellam`	Zero-Inflated Skellam	Univariate	Integer	`meanvar`, `diff`, `meandisp`

offers versatile model specification and core features available for the entire spectrum of implemented distributions. The current version of the package, 0.5.1, offers a selection of 26 distributions, catering to various univariate and multivariate data types such as binary, categorical, ranking, count, integer, circular, interval, compositional, duration, and real data. A comprehensive list of these distributions is provided in Table 1. Model specification within the package allows for flexible customization, enabling users to incorporate different parametrizations, exogenous variables, joint and separate modeling of exogenous variables and dynamics, higher score and autoregressive orders, custom and unconditional initial values of time-varying parameters, fixed and bounded values of coefficients, and missing values. Model estimation is performed by the maximum likelihood method and the Hessian matrix. Furthermore, the package offers a range of functionalities including forecasting, simulation, bootstrapping, and assessment of parameter uncertainty. Comprehensive documentation is provided with the package, offering details on each distribution and its corresponding parametrizations.

The **gasmodel** package is accessible on CRAN at `cran.r-project.org/package=gasmodel`. Additionally, users can find the development version of the package on GitHub at `github.com/vladimirholy/gasmodel`, providing them with the opportunity to report any bugs or issues they encounter.

The rest of the paper is as follows. In Section 2, we outline the key characteristics of GAS models. In Section 3, we present an overview of the **gasmodel** package. In Section 4, we present two case studies demonstrating the practical application of the package. In Section 5, we discuss limitations and customization. We conclude the paper in Section 6.

2 Generalized Autoregressive Score Models

2.1 Background

The concept of utilizing the score as a driving mechanism for dynamics in time series was independently developed at both Vrije Universiteit Amsterdam and the University of Cambridge. At

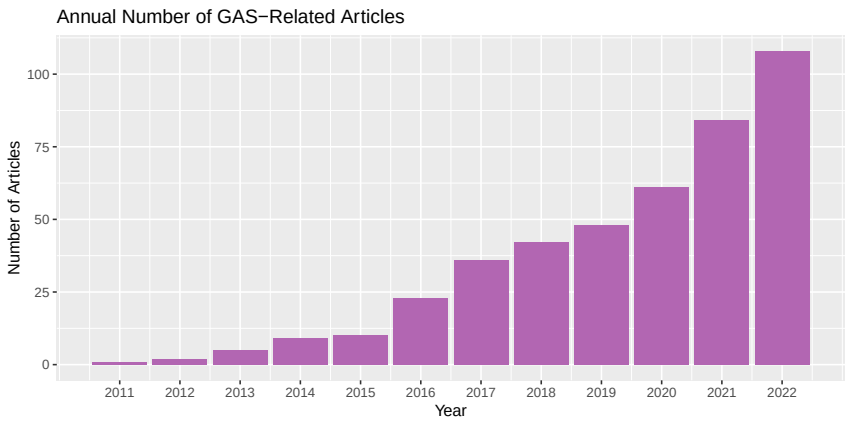

Figure 1: The annual number of articles containing phrase "generalized autoregressive score" or "dynamic conditional score" from 2011 to 2022 according to Scopus.

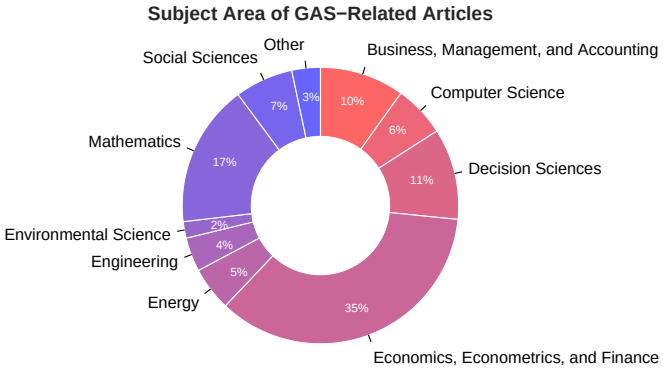

Figure 2: The subject area of articles containing phrase "generalized autoregressive score" or "dynamic conditional score"" from 2011 to 2022 according to Scopus.

Vrije Universiteit Amsterdam, researchers established a comprehensive general methodology that encompasses various models driven by the score, known as the generalized autoregressive score (GAS) models. The initial findings were presented in a working paper Creal *et al.* (2008), which was subsequently published as Creal *et al.* (2013). At the University of Cambridge, the initial focus was on a specific model that employed the Student's t-distribution with dynamic volatility, named Beta-t-(E)GARCH. This approach was introduced in a working paper Harvey and Chakravarty (2008). The book by Harvey (2013) explores a variety of dynamic location and scale models driven by the score, referring to them as dynamic conditional score (DCS) models. Both Creal *et al.* (2013) and Harvey (2013) are widely recognized as seminal contributions to the literature on GAS models. More recently, in order to reconcile different terminologies used in the literature, the term "score-driven models" has also emerged as a synonymous label.

Figures 1 and 2 illustrate the continuous growth of the GAS literature, encompassing a wide range of subject areas. The Scopus database reports 429 articles containing phrase "generalized autoregressive score" or "dynamic conditional score", as of December 31, 2022. The website `www.gasmodel.com` lists 288 articles, working papers, and books on GAS models, as of September 28, 2022.

2.2 Basic Notation

The goal is to model time series y_t, $t = 1, \ldots, T$, which can be univariate or multivariate, continuous or discrete. Let f_t denote the vector of time-varying parameters and g the vector of static parameters. Let $p(y_t|f_t, g)$ denote the density function in the case of a continuous variable, or the probability mass function in the case of a discrete variable.

Constructing a model involves two main components: selecting an appropriate distribution and specifying the dynamics of its time-varying parameters.

2.3 Score as the Key Ingredient

In GAS models, the key ingredient driving the dynamics of f_t is the score, i.e., the gradient of the log-likelihood function,

$$\nabla(y_t, f_t) = \frac{\partial \ln p(y_t|f_t, g)}{\partial f_t}. \tag{1}$$

The score has zero expected value and its variance is known as the Fisher information,

$$\mathcal{I}(f_t) = \mathrm{E}\left[\left(\frac{\partial \ln p(y_t|f_t, g)}{\partial f_t}\right)^2 \middle| f_t, g\right]. \tag{2}$$

The score quantifies the discrepancy between the fitted distribution, determined by the parameter f_t, and a particular observation y_t. As such, it can be employed as a correction term following the realization of an observation. When the score is positive, it suggests that the parameter of interest should be increased to better accommodate the observed data. Conversely, when the score is negative, decreasing the parameter would help in aligning the distribution with the observation. When the score is zero, it indicates that the current parameter value represents the optimal fit for the specific observation at hand.

An advantage of the score is that it takes into account the shape of the distribution. To illustrate this point, Creal *et al.* (2013) consider two GARCH models: one based on the normal distribution and another based on the Student's t-distribution. Now, imagine an extreme observation occurs. Due to its heavier tails, the Student's t-distribution assigns a higher probability to such extreme observations compared to the normal distribution. Crucially, this distinction is also mirrored in the score. Specifically, when assuming the normal distribution, the score for the extreme observation will have a significantly higher absolute value compared to when assuming the Student's t-distribution. The dynamics can thus reflect the shape of the distribution.

The simple difference between expectation and realization, commonly used as a correction term in various time series models, may not always be effective for distributions with specific support. Harvey *et al.* (2019) highlight this limitation in the context of circular time series. To illustrate this, let us suppose the expected value of an observation is 0.01, but the actual observation turns out to be 6.27. Although the numerical difference between these values is substantial, their corresponding angles are very similar as values 0 and 2π represent the exact same angle. This discrepancy highlights the inadequacy of using a simple difference metric. On the other hand, the score respects the circular nature of the data. For instance, when working with the von Mises distribution characterized by a time-varying location parameter μ_t and a static concentration parameter ν, the score for μ_t is equal to $\nu \sin(y_t - \mu_t)$. By employing the sine function, the score accounts for the circularity of the data and ensures that the angular differences are appropriately considered during the analysis.

2.4 Dynamics of Time Varying Parameters

In GAS models, time-varying parameters f_t follow the recursion

$$f_t = \omega + \sum_{j=1}^{P} \alpha_j S(f_{t-j}) \nabla(y_{t-j}, f_{t-j}) + \sum_{k=1}^{Q} \varphi_k f_{t-k}, \tag{3}$$

where ω is a vector of constants, α_j are score parameters, φ_k are autoregressive parameters, and $S(f_t)$ is a scaling function for the score. In the majority of empirical studies, it is common practice to set the score order P and the autoregressive order Q to 1. Furthermore, one of three scaling functions is typically chosen: the unit function, the inverse of the Fisher information, or the square root of the inverse of the Fisher information. When the latter is used, the scaled score has unit variance. However, the choice of the scaling function is not always a straightforward task and is closely tied to the underlying distribution. For a detailed discussion on this matter, see Holý (2020).

The dynamics of the model can be expanded to incorporate exogenous variables as

$$f_t = \omega + \sum_{i=1}^{M} \beta_i x_{ti} + \sum_{j=1}^{P} \alpha_j S(f_{t-j}) \nabla(y_{t-j}, f_{t-j}) + \sum_{k=1}^{Q} \varphi_k f_{t-k}, \tag{4}$$

where β_i are regression parameters associated with the exogenous variables x_{ti}. Alternatively, a different model can be obtained by defining the recursion in the fashion of regression models with dynamic errors as

$$f_t = \omega + \sum_{i=1}^{M} \beta_i x_{ti} + e_t, \quad e_t = \sum_{j=1}^{P} \alpha_j S(f_{t-j}) \nabla(y_{t-j}, f_{t-j}) + \sum_{k=1}^{Q} \varphi_k e_{t-k}. \tag{5}$$

The key distinction between the two models lies in the impact of exogenous variables on f_t. Specifically, in the latter model formulation, the exogenous variables influence solely the concurrent parameter f_t, while in the former model, they additionally affect all future parameters through the autoregressive term. In a stationary model without exogenous variables, the two specifications are equivalent, although with differently parametrized intercept. When numerically finding the values of the parameters, the latter model can converge faster as ω is "disconnected" from φ_k.

Other model specifications can be obtained by imposing various restrictions on ω, β_i, α_j, or φ_k. In addition, it is possible to have different orders P and Q for individual parameters when multiple parameters are time-varying. Furthermore, the set of exogenous variables can also vary for different parameters.

The recursive nature of f_t necessitates the initialization of the first few elements $f_1, \ldots, f_{\max\{P,Q\}}$. A sensible approach is to set them to the long-term value (in the case of a stationary model omitting exogenous variables),

$$\bar{f} = \begin{cases} \frac{1}{1-\sum_{k=1}^{Q} \varphi_k} & \text{in model (4)}, \\ \omega & \text{in model (5)}. \end{cases} \tag{6}$$

Alternatively, if additional information is available, the initial elements can be set to a specified value.

2.5 Maximum Likelihood Estimation

GAS models can be straightforwardly estimated by the maximum likelihood method. Let $\theta = (\omega, \beta_1, \ldots, \beta_M, \alpha_1, \ldots, \alpha_P, \varphi_1, \ldots, \varphi_Q, g)'$ denote the vector of all parameters to be estimated. The estimate $\hat{\theta}$ is then obtained by maximizing the full log-likelihood as

$$\hat{\theta} \in \arg\max_{\theta} \sum_{t=1}^{T} \ln p(y_t | f_t, g). \tag{7}$$

Alternatively, the conditional log-likelihood can be maximized, which excludes the initial $\max\{P, Q\}$ terms. The maximization of the log-likelihood function can be accomplished using various general-purpose algorithms designed for solving nonlinear optimization problems.

The standard errors of the estimated parameters can be obtained using the standard maximum likelihood asymptotics. Under appropriate regularity conditions, the maximum likelihood estimator $\hat{\theta}$ is consistent and asymptotically normal. Specifically, it satisfies:

$$\sqrt{T}\big(\hat{\theta} - \theta_0\big) \xrightarrow{\mathrm{d}} \mathrm{N}\big(0, -H^{-1}\big), \tag{8}$$

where θ_0 represents the true parameter values and H denotes the asymptotic Hessian of the log-likelihood, defined as

$$H = \operatorname*{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \frac{\partial^2 \ln p(y_t|f_t, g)}{\partial\theta_0 \partial\theta_0'}. \tag{9}$$

In finite samples, the asymptotic Hessian H can be approximated by the empirical Hessian of the log-likelihood evaluated at the estimated parameter values $\hat{\theta}$. This empirical Hessian provides an estimate of the curvature of the log-likelihood function and serves as a practical substitute for the true asymptotic Hessian when finite-sample inference is required.

The conditions for the consistency and asymptotic normality of the estimator depend on the specific distributional assumptions and dynamics of the model and need to be verified on a case-by-case basis. Each distribution may have its own specific characteristics and requirements for maximum likelihood estimation. For the general asymptotic theory regarding GAS models and maximum likelihood estimation, see Blasques *et al.* (2014), Blasques *et al.* (2018), and Blasques *et al.* (2022b).

2.6 Theoretical and Empirical Properties

The use of the score for updating time-varying parameters is optimal in an information theoretic sense. For an investigation of the optimality properties of GAS models, see Blasques *et al.* (2015) and Blasques *et al.* (2021).

Generally, the GAS models perform quite well when compared to alternatives, including parameter-driven models. For a comparison of the GAS models to alternative models, see Koopman *et al.* (2016) and Blazsek and Licht (2020).

2.7 Notable Models

The GAS class includes many well-known econometric models, such as the generalized autoregressive conditional heteroskedasticity (GARCH) model of Bollerslev (1986) based on the normal distribution, the autoregressive conditional duration (ACD) model of Engle and Russell (1998) based on the exponential distribution, and the count model of Davis *et al.* (2003) based on the Poisson distribution.

More recently, a variety of novel score-driven models has been proposed, such as the Beta-t-(E)GARCH model of Harvey and Chakravarty (2008), a multivariate Stu- dent's t volatility model of Creal *et al.* (2011), a Dirichlet model of Calvori *et al.* (2013), the GRAS copula model of De Lira Salvatierra and Patton (2015), the realized Wishart-GARCH model of Hansen *et al.* (2016), a bimodal Birnbaum–Saunders model of Fonseca and Cribari-Neto (2018), a Skellam model of Koopman *et al.* (2018), a circular model of Harvey *et al.* (2019), a Bradley–Terry model of Gorgi *et al.* (2019), a bivariate Poisson model of Koopman and Lit (2019), a censoring model of Harvey and Ito (2020), a zero-inflated negative binomial model of Blasques *et al.* (2022a), a double Poisson mixture model of Holý and Tomanová (2022), a ranking model of Holý and Zouhar (2022), and a Tobit model of Harvey and Liao (2023).

For an overview of various GAS models, see Artemova *et al.* (2022) and Harvey (2022).

3 Features of the Package

3.1 Model Specification and Estimation

The heart of the **gasmodel** package is the `gas()` function, which serves as a powerful tool for estimating both univariate and multivariate GAS models. This function offers extensive flexibility with its wide range of arguments:

```
R> gas(y, x = NULL, distr, param = NULL, scaling = "unit", regress = "joint",
+    p = 1L, q = 1L, par_static = NULL, par_link = NULL, par_init = NULL,
+    lik_skip = 0L, coef_fix_value = NULL, coef_fix_other = NULL,
+    coef_fix_special = NULL,  = NULL, coef_bound_upper = NULL, coef_start = NULL,
+    optim_function = wrapper_optim_nloptr, optim_arguments = list(opts =
```

```
+    list(algorithm = "NLOPT_LN_NELDERMEAD", xtol_rel = 0, maxeval = 1e+06)),
+    hessian_function = wrapper_hessian_stats, hessian_arguments = list(),
+    print_progress = FALSE)
```

However, at its core, it only requires two essential inputs: a time series `y` and a distribution `distr`. All other arguments come with default values, ensuring that the function can be readily used even with minimal specifications.

A time series `y` can be represented as either a vector of length T or a $T \times 1$ matrix in the case of univariate series. In the multivariate case, it should be a $T \times N$ matrix, where N denotes the dimension of the series.

Additionally, there is an option to include exogenous variables `x`. When incorporating a single variable that is common for all time-varying parameters, a numeric vector of length T can be provided. For multiple variables that are common for all time-varying parameters, a $T \times M$ numeric matrix can be used. In cases where there are individual variables for each time-varying parameter, a list of numeric vectors or matrices following the aforementioned formats can be utilized. To control whether the variables are included in the dynamics equation together, as in (4), the arguemnt `regress` can be set to `"joint"`. Alternatively, if separate equations for dynamics and regression are preferred, as in (5), `regress` can be set to `"sep"`.

The selection of the distribution in the `gas()` function is determined by the `distr` argument. Some distributions have multiple parametrizations available, which can be specified using the `param` argument. It is important to note that certain parameters may have restrictions imposed on them, and these restrictions should be considered in the model dynamics. However, it may not always be possible to satisfy these restrictions, or it may require additional constraints on the coefficients controlling the dynamics. To handle parameter restrictions, it is generally recommended to use a link function that transforms the parameters into unrestricted real numbers. By default, the logarithmic function is applied to time-varying parameters in the interval $(0, \infty)$, while the logistic function is used for time-varying parameters in the interval $(0, 1)$. The static parameters are unaffected. This behavior can be modified by the `par_link` argument, which takes the form of a logical vector. The `TRUE` values indicate that the logarithmic/logistic link is applied to the corresponding parameters. The list of available distributions and their parametrizations can be obtained using the `distr()` function. Alternatively, Table 1 provides the relevant information.

The determination of time-varying and static parameters is guided by the `par_static` argument, which takes the form of a logical vector. The `TRUE` values indicate static parameters. By default, the first parameter of the distribution is considered time-varying, while the remaining parameters are treated as static. The score order P and the autoregressive order Q are selected by the `p` and `q` arguments respectively. These arguments can take either a single non-negative integer or a vector of non-negative integers when different orders are required for different parameters.

The choice of scaling function for the score is determined by the `scaling` argument. The supported scaling options include the unit scaling (`scaling = "unit"`), the scaling based on the inverse of the Fisher information matrix (`scaling = "fisher_inv"`), and the scaling based on the inverse square root of the Fisher information matrix (`scaling = "fisher_inv_sqrt"`). The latter two scalings utilize the Fisher information for the time-varying parameters exclusively. If the preference is to use the full Fisher information matrix, which includes both time-varying and static parameters, the `"full_fisher_inv"` or `"full_fisher_inv_sqrt"` scaling options can be selected. For the individual Fisher information associated with each parameter, the `"diag_fisher_inv"` and `"diag_fisher_inv_sqrt"` scaling options are available. It should be noted that when the parametrization is orthogonal (see `distr()`), there are no differences among these scaling variants.

The first $\max\{P, Q\}$ initial values of the time-varying parameters are by default set to their long-term values (6). It is also possible to assign specific values to the initial parameters using the `par_init` argument. During the maximization of the log-likelihood, the initial values can be included, resulting in the computation of the full likelihood, which is the default option. Alternatively, the initial values can be omitted, leading to the computation of the conditional likelihood by specifying `lik_skip = NULL`. To exclude a specified number of first few values from the likelihood calculation, a non-negative integer can be provided to `lik_skip`.

Restrictions on estimated coefficients can be enforced using several arguments. The `coef_fix_value` argument allows coefficients to be fixed at specific values, using a numeric vector where `NA` values indicate coefficients that are not fixed. To set coefficients as linear combinations of other coefficients, the `coef_fix_other` argument can be used. It requires a square matrix with multiples of the estimated coefficients, which are added to the fixed coefficients. A coefficient given by row is fixed on coefficient given by column. By this logic, all rows corresponding to the estimated coefficients should contain only `NA` values. All columns corresponding to the fixed coefficients should also contain only `NA` values. For convenience, common coefficient structures can be specified by name using the `coef_fix_special` argument. Examples include `panel_structure`, `zero_sum_intercept`, and `random_walk`. Section 4.2 provides demonstrations of their usage. To impose lower and upper bounds on coefficients, the `coef_bound_lower` and `coef_bound_upper` arguments can be utilized, respectively.

The `coef_start` argument allows for the specification of the starting values of coefficients used in the optimization process. If no values are provided, the starting values are automatically selected from a small grid of values. To define the optimization function, the `optim_function` argument is used. The function should be formatted according to the required specifications. Two wrapper functions are available for convenience: `wrapper_optim_stats()`, which utilizes the `optim()` function from the **stats** package, and `wrapper_optim_nloptr()`, which utilizes the `nloptr()` function from the **nloptr** package. Additional arguments can be passed to the optimization function as a list using the `optim_arguments` argument. Similarly, the Hessian matrix can be computed using the function specified in the `hessian_function` argument. Three wrappers are available: `wrapper_hessian_stats` for the `optimHess()` function from the **stats** package, `wrapper_hessian_pracma` for the `hessian()` function from the **pracma** package, and `wrapper_hessian_numderiv` for the `hessian()` function from the **numDeriv** package. Additional arguments for the Hessian function can be passed as a list using the `hessian_arguments` argument. If desired, a detailed computation report can be continuously printed by setting the `print_progress` argument to `TRUE`.

The function returns a list of S3 class `gas`. This list consists of five components: `data`, `model`, `control`, `solution`, and `fit`, each of which is also a list. The `data` component contains the supplied time series and exogenous variables. The `model` component contains the specification of the model structure and size. The `control` component contains the settings that control the optimization and Hessian computation. The `solution` component contains the raw results of the optimization and Hessian computation. Lastly, and most importantly, the `fit` component contains comprehensive estimation results. When an object of the `gas` class is printed, it provides a concise summary similar to the `summary.lm()` function from the **stats** package (refer to Sections 4.1 and 4.2 for examples). Various generic functions can be applied to `gas` objects, including `summary()`, `plot()`, `coef()`, `vcov()`, `fitted()`, `residuals()`, `logLik()`, `AIC()`, `BIC()`, and `confint()`.

3.2 Forecasting

Forecasting of GAS models is performed using the `gas_forecast()` function. This function offers two forecasting methods. The `mean_path` method filters the time-varying parameters based on zero score and then generates the mean of the time series. The `simulated_paths` method repeatedly simulates time series, simultaneously filters time-varying parameters, and then estimates mean, standard deviation, and quantiles. See Blasques *et al.* (2016b) for more details on this method.

To use the `gas_forecast()` function, an estimated GAS model is required. Typically, the output of the `gas()` function (a `gas` object) can be supplied via the `gas_object` argument:

```
R> gas_forecast(gas_object, method = "mean_path", t_ahead = 1L, x_ahead = NULL,
+    rep_ahead = 1000L, quant = c(0.025, 0.975))
```

Alternatively, multiple arguments including the data, model specification, and estimated coefficients can be manually specified:

```
R> gas_forecast(method = "mean_path", t_ahead = 1L, x_ahead = NULL,
+    rep_ahead = 1000L, quant = c(0.025, 0.975), y, x = NULL, distr, param = NULL,
```

```
+    scaling = "unit", regress = "joint", p = 1L, q = 1L, par_static = NULL,
+    par_link = NULL, par_init = NULL, coef_est = NULL)
```

The forecasting method is determined by the `method` argument. The number of observations to forecast can be specified using the `t_ahead` argument. If exogenous variables are utilized, their values must be provided for the forecasted period using the `x_ahead` argument. For the `simulated_paths` method, the number of simulations can be controlled using the `rep_ahead` argument, and the desired quantiles can be specified using the `quant` argument.

The function returns a list of S3 class `gas_forecast` with three components: `data`, `model`, `forecast`. The `data` component contains the supplied time series and exogenous variables. The `model` component contains the specification of the model structure and size. The `forecast` component contains the mean of the forecasted observations, along with standard deviations and quantiles if the `simulated_paths` method is used. Available generic functions are `summary()` and `plot()`.

3.3 Simulation

Basic simulation of GAS models is handled by the `gas_simulate()` function.

The `gas_simulate()` function requires suppling the coefficients using the `coef_est` argument and specifying the model using arguments `distr`, `param`, `scaling`, `regress`, `p`, `q`, `par_static`, `par_link`, `par_init`, and, in the case of multivariate models, the dimension `n`:

```
R> gas_simulate(t_sim = 1L, x_sim = NULL, distr, param = NULL, scaling = "unit",
+    regress = "joint", n = NULL, p = 1L, q = 1L, par_static = NULL,
+    par_link = NULL, par_init = NULL, coef_est = NULL)
```

Alternatively, only a `gas` object containing a model estimated by the `gas()` function can be provided using the `gas_object` argument:

```
R> gas_simulate(gas_object, t_sim = 1L, x_sim = NULL)
```

The number of observations to simulate can be specified using the `t_sim` argument. If exogenous variables are utilized, their values must be provided for the simulation sample using the `x_sim` argument.

The function returns a list of S3 class `gas_simulate` with three components: `data`, `model`, `simulation`. The `data` component contains the exogenous variables, if supplied. The `model` component contains the specification of the model structure and size. The `simulation` component contains the simulated time series, time-varying parameters, and scores. Available generic functions are `summary()` and `plot()`.

3.4 Bootstrapping

To compute standard deviations and confidence intervals of the estimated coefficients in GAS models, the package provides the `gas_bootstrap()` function. This function employs the bootstrapping technique to estimate the uncertainty associated with the coefficients. The `parametric` method involves repeatedly simulating time series using the parametric model and re-estimating the coefficients based on the simulated data. The `simple_block`, `moving_block`, and `stationary_block` methods execute the circular block bootstrap with fixed non-overlapping blocks, fixed overlapping blocks, and randomly sized overlapping blocks, respectively.

The `gas_bootstrap()` function requires an estimated GAS model with optimization settings as inputs. The best way is to simply supply a `gas` object to the `gas_object` argument:

```
R> gas_bootstrap(gas_object, method = "parametric", rep_boot = 1000L,
+    block_length = NULL, quant = c(0.025, 0.975), parallel_function = NULL,
+    parallel_arguments = list())
```

Alternatively, the individual arguments including the data, model specification, estimated coefficients, and optimization setting can be provided:

```
R> gas_bootstrap(method = "parametric", rep_boot = 1000L, block_length = NULL,
+    quant = c(0.025, 0.975), y, x = NULL, distr, param = NULL, scaling = "unit",
+    regress = "joint", p = 1L, q = 1L, par_static = NULL, par_link = NULL,
+    par_init = NULL, lik_skip = 0L, coef_fix_value = NULL, coef_fix_other = NULL,
+    coef_fix_special = NULL, coef_bound_lower = NULL, coef_bound_upper = NULL,
+    coef_est = NULL, optim_function = wrapper_optim_nloptr, optim_arguments = list(opts =
+    list(algorithm = "NLOPT_LN_NELDERMEAD", xtol_rel = 0, maxeval = 1e+06)),
+    parallel_function = NULL, parallel_arguments = list())
```

The bootstrapping method is determined by the `method` argument. The number of bootstrap samples is specified by the `rep_boot` argument. For the `simple_block` and `moving_block` methods, the fixed size of blocks must be specified by the `block_length` argument. For the `stationary_block` method, the mean size of blocks must be specified by the `block_length` argument. The desired quantiles can be specified using the `quant` argument. As boostrapping can be computationally very demanding, parallelization is achievable by employing the `parallel_function` argument, which expects a function similar to `lapply()`, allowing the application of a function over a list. Two wrapper functions are available for convenience: `wrapper_parallel_multicore()`, which utilizes the multicore parallelization functionality from the **parallel** package, and `wrapper_parallel_snow()`, which utilizes the snow parallelization functionality from the **parallel** package. Additional arguments can be passed to the parallelization function as a list using the `parallel_arguments` argument. If `parallel_function` is set to `NULL`, no parallelization is employed and `lapply()` is used.

The function returns a list of S3 class `gas_bootstrap` with three components: `data`, `model`, `bootstrap`. The `data` component contains the supplied time series and exogenous variables. The `model` component contains the specification of the model structure and size. The `bootstrap` component contains the full set of bootstrapped coefficients as well as the basic statistics derived from them. Available generic functions are `summary()`, `plot()`, `coef()`, and `vcov()`.

3.5 Filtered Parameters

The filtered time-varying parameters of an estimated model can be directly obtained from the output of the `gas()` function. However, to investigate the uncertainty associated with these parameters, the `gas_filter()` function can be used. This function also supports forecasting and provides two methods. The `simulated_coefs` method calculates a path of time-varying parameters for each simulated coefficient set, assuming asymptotic normality with a given variance-covariance matrix. See Blasques *et al.* (2016b) for more details on this method. The `given_coefs` methods computes a path of time-varying parameters for each supplied coefficient set. Suitable sets of coefficients can be obtained, for example, through the use of the `gas_bootstrap()` function.

An estimated GAS model can be supplied as a `gas` object to the `gas_object` argument:

```
R> gas_filter(gas_object, method = "simulated_coefs", coef_set = NULL,
+    rep_gen = 1000L, t_ahead = 0L, x_ahead = NULL, rep_ahead = 1000L,
+    quant = c(0.025, 0.975))
```

Alternatively, the individual arguments including the data, model specification, and estimated coefficients with variance-covariance matrix can be provided:

```
R> gas_filter(method = "simulated_coefs", coef_set = NULL, rep_gen = 1000L,
+    t_ahead = 0L, x_ahead = NULL, rep_ahead = 1000L, quant = c(0.025, 0.975), y,
+    x = NULL, distr, param = NULL, scaling = "unit", regress = "joint", p = 1L,
+    q = 1L, par_static = NULL, par_link = NULL, par_init = NULL,
+    coef_fix_value = NULL, coef_fix_other = NULL, coef_fix_special = NULL,
+    coef_bound_lower = NULL, coef_bound_upper = NULL, coef_est = NULL,
+    coef_vcov = NULL)
```

The `method` argument determines the approach for capturing uncertainty. For the `given_coefs` method, the `coef_set` argument in the form a numeric matrix of coefficient sets in rows must be provided. For the `simulated_coefs` method, the `rep_gen` argument representing the number of generated coefficient sets must be provided. If forecasting is desired, the number of observations to

forecast can be specified using the `t_ahead` argument, values of exogenous variable for the forecasted period can be provided using the `x_ahead` argument, and the number of simulation repetitions in the forecasted sample can be controlled using the `rep_ahead` argument. The desired quantiles can be specified using the `quant` argument.

The function returns a list of S3 class `gas_filter` with three components: `data`, `model`, `filter`. The `data` component contains the supplied time series and exogenous variables. The `model` component contains the specification of the model structure and size. The `filter` component contains in-sample and possibly out-of-sample means, standard deviations, and quantiles of the time-varying parameters and scores. Available generic functions are `summary()` and `plot()`.

3.6 Supplementary Functions for Distributions

The `distr()` function can be utilized to retrieve a list of distributions and their parametrizations supported by the `gas()` function. To narrow down the output and focus on specific distributions, arguments such as `filter_distr`, `filter_param`, `filter_type`, `filter_dim`, `filter_orthog`, and `filter_default` can be specified. The output is in the form of a data.frame with columns providing information on the distributions such as the data type, dimension, orthogonality, and default parameterization.

To work with individual distributions, the **gasmodel** package offeres several functions. The `distr_density()` function computes the density of a given distribution, the `distr_mean()` function computes the mean of a given distribution, the `distr_var()` function computes the variance of a given distribution, the `distr_score()` function computes the score of a given distribution, the `distr_fisher()` function computes the Fisher information of a given distribution, and the `distr_random()` function generates random observations from a given distribution. Each of these function can be supplied with arguments specifying the distribution and the parametrization, namely `distr`, `param`, `par_link`. It is important to note that while the `gas()` function may automatically set the logarithmic/logistic link for time-varying parameters, it must be set manually for the distribution functions. Additionaly, a vector of parameter values must be provided to the `f` argument. Some functions may also require an observation to be provided to the `y` argument. For detailed usage instructions, please refer to the documentation for each individual function.

4 Case Studies

4.1 Bookshop Orders

In the first case study, we demonstrate the estimation of a univariate GAS model, complemented by bootstrapping and simulation techniques.

We loosely follow Tomanová and Holý (2021) and analyze the timing of orders from a Czech antiquarian bookshop. Besides seasonality and diurnal patterns, one would expect the times of orders to be independent of each other. However, this is not the case and we use a GAS model to capture dependence between the times of orders.

A strand of financial econometrics is devoted to analyzing the timing of transactions by the so-called autoregressive conditional duration (ACD) model introduced by Engle and Russell (1998). For a textbook treatment of such financial point processes, see e.g., Hautsch and Huang (2012).

Let us prepare the analyzed data. We use the `bookshop_sales` dataset containing times of orders from June 8, 2018 to December 20, 2018. We calculate differences of subsequent times, i.e., durations. To avoid zero durations, we set them to 0.5 second.

```
R> library("dplyr")
R> library("tidyr")
R> library("ggplot2")
R> library("hms")
R> library("gasmodel")

R> data("bookshop_sales")
```

```
R> data_dur <- bookshop_sales %>%
+    as_tibble() %>%
+    rename(datetime = time) %>%
+    mutate(date = as.Date(datetime)) %>%
+    mutate(time = as_hms(datetime)) %>%
+    mutate(duration = as.numeric(datetime - lag(datetime)) / 60) %>%
+    mutate(duration = recode(duration, "0" = 0.5)) %>%
+    drop_na()
```

We adjust the observed durations for diurnal pattern and extract the time series to be analyzed.

```
R> model_spl <- smooth.spline(as.vector(data_dur$time), data_dur$duration, df = 10)

R> data_dur <- data_dur %>%
+    mutate(duration_spl = predict(model_spl, x = as.vector(time))$y) %>%
+    mutate(duration_adj = duration / duration_spl)

R> y <- data_dur$duration_adj
```

The following distributions are available for our data type. We utilize the generalized gamma family.

```
R> distr(filter_type = "duration", filter_dim = "uni")

         distr_title param_title    distr param     type dim orthog default
6  Birnbaum-Saunders       Scale     bisa scale duration uni   TRUE    TRUE
10       Exponential        Rate      exp  rate duration uni   TRUE   FALSE
11       Exponential       Scale      exp scale duration uni   TRUE    TRUE
12             Gamma        Rate    gamma  rate duration uni  FALSE   FALSE
13             Gamma       Scale    gamma scale duration uni  FALSE    TRUE
14 Generalized Gamma        Rate gengamma  rate duration uni  FALSE   FALSE
15 Generalized Gamma       Scale gengamma scale duration uni  FALSE    TRUE
31           Weibull        Rate  weibull  rate duration uni  FALSE   FALSE
32           Weibull       Scale  weibull scale duration uni  FALSE    TRUE
```

First, we estimate the model based on the exponential distribution. By default, the logarithmic link for the time-varying scale parameter is adopted. In this particular case, the Fisher information is constant and the three scalings are therefore equivalent.

```
R> est_exp <- gas(y = y, distr = "exp")
R> est_exp

GAS Model: Exponential Distribution / Scale Parametrization / Unit Scaling

Coefficients:
                     Estimate  Std. Error   Z-Test  Pr(>|Z|)
log(scale)_omega  -0.00085202  0.00114896  -0.7416    0.4584
log(scale)_alpha1  0.04888439  0.00650562   7.5142 5.727e-14 ***
log(scale)_phi1    0.96343265  0.00910508 105.8126 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -5608.518, AIC: 11223.04, BIC: 11243.01
```

Second, we estimate the model based on the Weibull distribution. Compared to the exponential distribution, it has an additional shape parameter. By default, the first parameter is assumed time-varying while the remaining are assumed static. In our case, the model features the time-varying scale parameter with the constant shape parameter. However, it is possible to modify this behavior using the `par_static` argument.

```
R> est_weibull <- gas(y = y, distr = "weibull")
R> est_weibull
```

```
GAS Model: Weibull Distribution / Scale Parametrization / Unit Scaling

Coefficients:
                  Estimate Std. Error   Z-Test  Pr(>|Z|)
log(scale)_omega  -0.0019175  0.0013552  -1.4149    0.1571
log(scale)_alpha1  0.0562619  0.0082010   6.8604 6.867e-12 ***
log(scale)_phi1    0.9622643  0.0102230  94.1278 < 2.2e-16 ***
shape              0.9442209  0.0094299 100.1300 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -5591.442, AIC: 11190.88, BIC: 11217.51
```

Third, we estimate the model based on the gamma distribution. This is another generalization of the exponential distribution with an additional shape parameter.

```
R> est_gamma <- gas(y = y, distr = "gamma")
R> est_gamma

GAS Model: Gamma Distribution / Scale Parametrization / Unit Scaling

Coefficients:
                   Estimate Std. Error   Z-Test  Pr(>|Z|)
log(scale)_omega  0.0013296  0.0013395   0.9926    0.3209
log(scale)_alpha1 0.0518896  0.0071672   7.2399 4.491e-13 ***
log(scale)_phi1   0.9634327  0.0093853 102.6532 < 2.2e-16 ***
shape             0.9420850  0.0153854  61.2325 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -5601.735, AIC: 11211.47, BIC: 11238.1
```

Fourth, we estimate the model based on the generalized gamma distribution. The generalized gamma distribution encompasses all three aforementioned distributions as special cases.

```
R> est_gengamma <- gas(y = y, distr = "gengamma")
R> est_gengamma

GAS Model: Generalized Gamma Distribution / Scale Parametrization / Unit Scaling

Coefficients:
                   Estimate Std. Error  Z-Test  Pr(>|Z|)
log(scale)_omega  -0.049164   0.018903 -2.6009  0.009299 **
log(scale)_alpha1  0.069834   0.011670  5.9841 2.176e-09 ***
log(scale)_phi1    0.951761   0.015024 63.3493 < 2.2e-16 ***
shape1             1.764362   0.150759 11.7032 < 2.2e-16 ***
shape2             0.682971   0.033690 20.2723 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -5562.092, AIC: 11134.18, BIC: 11167.47
```

By comparing the Akaike information criterion (AIC), we find that the most general model, i.e., the one based on the generalized gamma distribution, is the most suitable. For this purpose, we use generic function `AIC()`. Alternatively, the AIC of an estimated model is stored in `est_gengamma$fit$aic`.

```
R> AIC(est_exp, est_weibull, est_gamma, est_gengamma)

             df      AIC
est_exp       3 11223.04
est_weibull   4 11190.88
est_gamma     4 11211.47
est_gengamma  5 11134.18
```

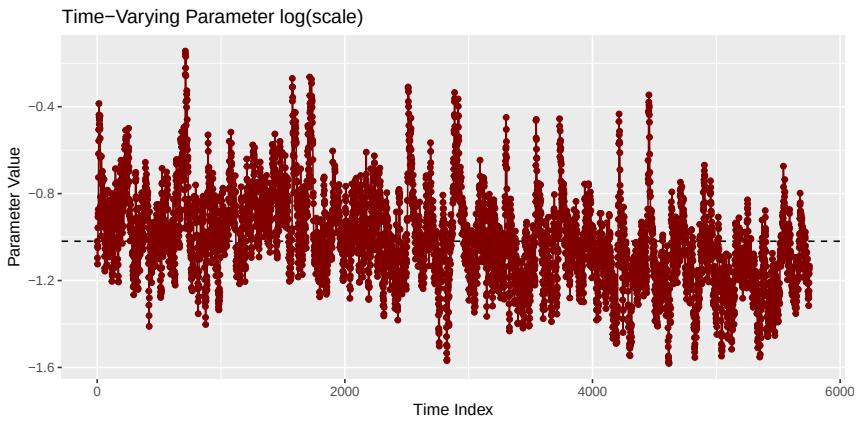

Figure 3: Time-varying parameters based on the generalized gamma model.

Let us take a look on the time-varying parameters of the generalized gamma model (Figure 3).

```
R> plot(est_gengamma)
```

We can see a slight negative trend in time-varying parameters. We can try including a trend as an exogenous variable for all four considered distributions.

```
R> x <- as.integer(data_dur$date) - as.integer(data_dur$date[1])

R> est_exp_tr <- gas(y = y, x = x, distr = "exp", reg = "sep")
R> est_exp_tr

GAS Model: Exponential Distribution / Scale Parametrization / Unit Scaling

Coefficients:
                    Estimate  Std. Error  Z-Test  Pr(>|Z|)
log(scale)_omega   0.29683416  0.04509203  6.5829 4.615e-11 ***
log(scale)_beta1  -0.00304957  0.00037137 -8.2118 < 2.2e-16 ***
log(scale)_alpha1  0.05401728  0.00802442  6.7316 1.678e-11 ***
log(scale)_phi1    0.91358230  0.02146703 42.5575 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -5583.723, AIC: 11175.45, BIC: 11202.08

R> est_weibull_tr <- gas(y = y, x = x, distr = "weibull", reg = "sep")
R> est_weibull_tr

GAS Model: Weibull Distribution / Scale Parametrization / Unit Scaling

Coefficients:
                    Estimate  Std. Error  Z-Test  Pr(>|Z|)
log(scale)_omega   0.26955739  0.04763575  5.6587 1.525e-08 ***
log(scale)_beta1  -0.00302892  0.00039014 -7.7638 8.244e-15 ***
log(scale)_alpha1  0.06215424  0.00992563  6.2620 3.801e-10 ***
log(scale)_phi1    0.90950196  0.02399584 37.9025 < 2.2e-16 ***
shape              0.94858384  0.00949927 99.8586 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -5569.405, AIC: 11148.81, BIC: 11182.1

R> est_gamma_tr <- gas(y = y, x = x, distr = "gamma", reg = "sep")
R> est_gamma_tr
```

```
GAS Model: Gamma Distribution / Scale Parametrization / Unit Scaling

Coefficients:
                   Estimate  Std. Error  Z-Test  Pr(>|Z|)
log(scale)_omega   0.35024097  0.04910603  7.1323 9.868e-13 ***
log(scale)_beta1  -0.00304957  0.00038142 -7.9954 1.292e-15 ***
log(scale)_alpha1  0.05698059  0.00874363  6.5168 7.182e-11 ***
log(scale)_phi1    0.91358230  0.02204841 41.4353 < 2.2e-16 ***
shape              0.94799429  0.01549052 61.1983 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -5578.303, AIC: 11166.61, BIC: 11199.89

R> est_gengamma_tr <- gas(y = y, x = x, distr = "gengamma", reg = "sep")
R> est_gengamma_tr

GAS Model: Generalized Gamma Distribution / Scale Parametrization / Unit Scaling

Coefficients:
                    Estimate  Std. Error  Z-Test  Pr(>|Z|)
log(scale)_omega  -0.70489163  0.19283438 -3.6554 0.0002568 ***
log(scale)_beta1  -0.00292746  0.00039123 -7.4827 7.280e-14 ***
log(scale)_alpha1  0.08164957  0.01387329  5.8854 3.971e-09 ***
log(scale)_phi1    0.87684184  0.03506612 25.0054 < 2.2e-16 ***
shape1             1.76342697  0.15253550 11.5608 < 2.2e-16 ***
shape2             0.68568220  0.03426457 20.0114 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -5541.097, AIC: 11094.19, BIC: 11134.14
```

The trend variable is significant in all cases. The AIC also confirms improvement of the fit.

```
R> AIC(est_exp_tr, est_weibull_tr, est_gamma_tr, est_gengamma_tr)

                df      AIC
est_exp_tr       4 11175.45
est_weibull_tr   5 11148.81
est_gamma_tr     5 11166.61
est_gengamma_tr  6 11094.19
```

Note that the time-varying parameters returned by the `gas()` function include the effect of exogenous variables. By using the `plot()` function, the now modeled trend can be clearly seen (Figure 4).

```
R> plot(est_gengamma_tr)
```

To assess the suitability of standard deviations based on asymptotics for our finite sample, we employ the `gas_bootstrap()` function. This function conducts a parametric bootstrap, allowing us to calculate standard errors and quantiles. It's important to note that this could be computationally very intensive, depending on the number of repetitions, the quantity of observations, the complexity of the model structure, and the optimizer used. The function supports parallelization through arguments `parallel_function` and `parallel_arguments`. For example, for the `snow` parallelization functionality with 4 cores, you can call `gas_bootstrap(est_gengamma_tr, parallel_function = wrapper_parallel_snow, parallel_arguments = list(spec = 4))`.

```
R> set.seed(42)
R> boot_gengamma_tr <- gas_bootstrap(est_gengamma_tr, method = "parametric")
R> boot_gengamma_tr
```

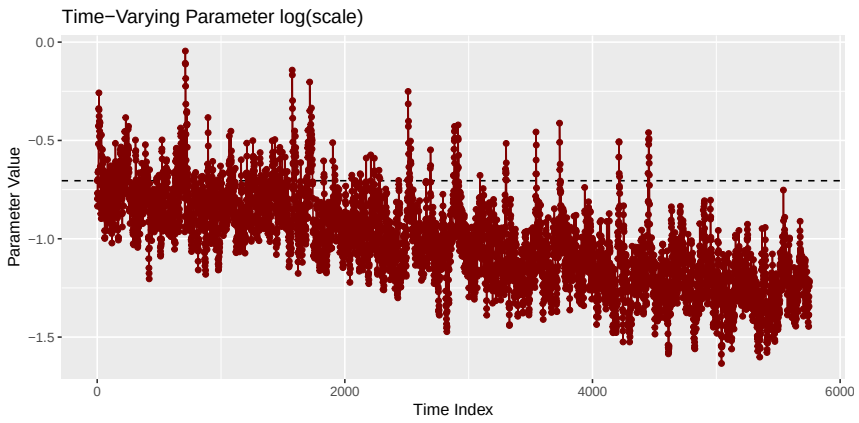

Figure 4: Time-varying parameters based on the generalized gamma model with trend.

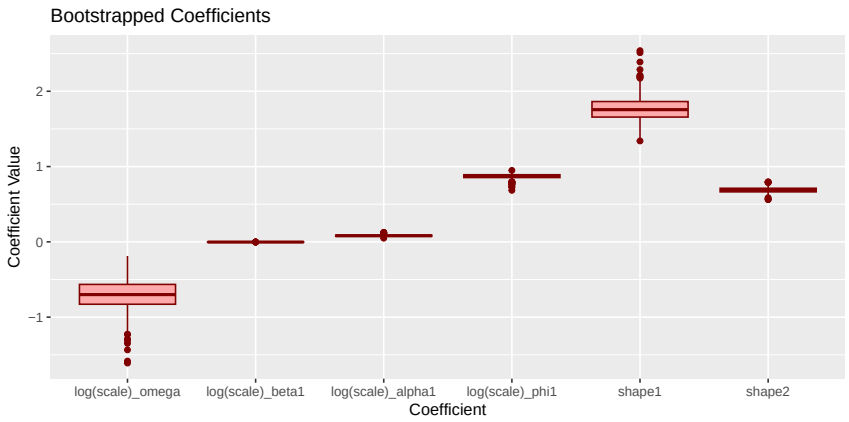

Figure 5: Boxplot of bootstrapped coefficients based on the generalized gamma model with trend.

```
GAS Model: Generalized Gamma Distribution / Scale Parametrization / Unit Scaling

Method: Parametric Bootstrap

Number of Bootstrap Samples: 1000

Bootstrapped Coefficients:
                    Original         Mean   Std. Error P-Value        2.5%       97.5%
log(scale)_omega  -0.704891626 -0.705226357 0.1980063054       0 -1.110703445 -0.34772409
log(scale)_beta1  -0.002927462 -0.002932047 0.0003883289       0 -0.003698968 -0.00217132
log(scale)_alpha1  0.081649573  0.081648774 0.0116918758       0  0.059179321  0.10575927
log(scale)_phi1    0.876841843  0.871223702 0.0297172317       0  0.806282230  0.91973863
shape1             1.763426971  1.765458333 0.1596470324       0  1.476831790  2.09983786
shape2             0.685682201  0.688266197 0.0354034049       0  0.623633255  0.76097862
```

The results can also be viewed in a boxplot (Figure 5).

```
R> plot(boot_gengamma_tr)
```

Given that the number of observations in our model is 5752 (accessible through `est_gengamma_tr$model$t`), it is reasonable to anticipate that standard deviations based on asymptotics would yield precise results. Fortunately, this holds true in our scenario. Note that standard deviations can also be obtained using the `vcov()` generic function for both `est_gengamma_tr` and `boot_gengamma_tr`.

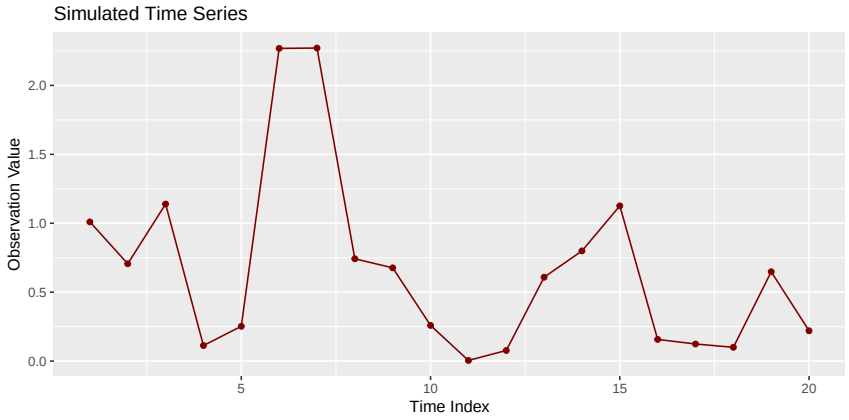

Figure 6: Simulated time series based on the generalized gamma model with trend.

```
R> est_gengamma_tr$fit$coef_sd - boot_gengamma_tr$bootstrap$coef_sd

 log(scale)_omega  log(scale)_beta1 log(scale)_alpha1   log(scale)_phi1            shape1
    -5.171928e-03      2.900388e-06      2.181417e-03      5.348887e-03     -7.111530e-03
           shape2
    -1.138839e-03
```

Lastly, we highlight the utilization of simulation techniques. Simulation is executed using the `gas_simulate()` function, which can be supplied with either an estimated model or a custom model structure.

```
R> t_sim <- 20
R> x_sim <- rep(max(x) + 1, t_sim)

R> set.seed(42)
R> sim_gengamma_tr <- gas_simulate(est_gengamma_tr, t_sim = t_sim, x_sim = x_sim)
R> sim_gengamma_tr

GAS Model: Generalized Gamma Distribution / Scale Parametrization / Unit Scaling

Simulations:
         t1          t2          t3          t4          t5          t6          t7
1.009836881 0.706070572 1.139254609 0.112834862 0.252712188 2.268641670 2.271065825
         t8          t9         t10         t11         t12         t13         t14
0.742231695 0.676595922 0.259042333 0.004836128 0.077080566 0.608510890 0.799449725
        t15         t16         t17         t18         t19         t20
1.126124047 0.157351783 0.124067217 0.100168697 0.648121920 0.219983546
```

The simulated time series can be plotted using the generic `plot()` function (Figure 6).

```
R> plot(sim_gengamma_tr)
```

The simulated time series can be employed, for example, to assess the impact of order arrivals on queuing systems, as demonstrated by Tomanová and Holý (2021).

4.2 Ice Hockey Rankings

In the second case study, we showcase the estimation of a multivariate GAS model, followed by forecasting and assessing uncertainty in the filtered time-varying parameters.

We present the empirical study of Holý and Zouhar (2022) which analyzes the results of the Ice Hockey World Championships. Our main object of interest is the annual ranking of 16 teams participating in the championships. While there exists a comprehensive statistical toolkit for ranking

data, as described e.g., by Alvo and Yu (2014), it is worth noting that the time perspective is often overlooked in the ranking literature, as highlighted by Yu *et al.* (2019). This is precisely where the GAS model emerges as a valuable tool in our analysis.

Our analyzed data are supplied in the `ice_hockey_championships` dataset. We restrict ourselves to years 1998–2019 just as Holý and Zouhar (2022). In 1998, the number of teams in the tournament increased from 12 to 16. In 2020, the championship was cancelled due to Covid-19 pandemic. We start by creating two variables – the final ranking of 16 participating teams in each year `y` and the dummy variable indicating which country (or countries) hosted the championship in each year `x`.

```
R> library("dplyr")
R> library("ggplot2")
R> library("gasmodel")

R> data("ice_hockey_championships")

R> t <- 22
R> n <- ncol(ice_hockey_championships$host)
R> y <- ice_hockey_championships$rankings[1:t, ]
R> x <- setNames(lapply(1:n, function(i) { ice_hockey_championships$host[1:t, i] }),
+    colnames(y))
```

We look at some basic statistics. In our sample, nine countries have participated each year.

```
R> participate <- colSums(is.finite(y))
R> names(participate)[participate == t]

[1] "CAN" "CHE" "CZE" "FIN" "LVA" "RUS" "SVK" "SWE" "USA"
```

The following countries hosted the championships, some of them multiple times.

```
R> host <- sapply(x, FUN = sum)
R> host[host > 0L]

AUT BLR CAN CHE CZE DEU FIN FRA LVA NOR RUS SVK SWE
  1   1   1   2   2   3   3   1   1   1   3   2   3
```

In the years under analysis, the gold medals were awarded to the following countries.

```
R> gold <- colSums(y == 1L)
R> gold[gold > 0L]

CAN CZE FIN RUS SVK SWE
  5   5   2   4   1   5
```

The **gasmodel** package provides a single distribution on rankings – the Plackett–Luce distribution.

```
R> distr(filter_type = "ranking")

     distr_title param_title distr param    type   dim orthog default
24 Plackett-Luce       Worth pluce worth ranking multi  FALSE    TRUE
```

It is a convenient and simple probability distribution on rankings utilizing a worth parameter for each item to be ranked. It originates from Luce's choice axiom and is also related to the Thurstone's theory of comparative judgment, see Luce (1977) and Yellott (1977). For more details on this distribution, see Plackett (1975), Stern (1990), and Critchlow *et al.* (1991).

We consider a total of three different models. We incorporate `x` as an exogenous variable in our model to capture possible home advantage. For each model, we assume a panel-like structure where each worth parameter has its own intercept, while the regression and dynamics parameters remain the same for all worth parameters. In the **gasmodel** package, this structure can be achieved

using the `coef_fix_value` and `coef_fix_other` arguments. Alternatively, for convenience, the value `panel_structure` can be included in the `coef_fix_special` argument. It is important to note that the worth parameters in the Plackett–Luce distribution are not identifiable, and it is common practice to impose a standardizing condition. In our model, we enforce the condition that the sum of all ω_i is 0. This can be accomplished by including the value `zero_sum_intercept` in the `coef_fix_special` argument.

First, we estimate the static model where there are no dynamics involved. In this case, we set both the autoregressive and score orders to zero. Either a single integer can be provided to determine the order for all parameters, or a vector of integers can be supplied to specify the order for individual parameters.

```
R> est_static <- gas(y = y, x = x, distr = "pluce", p = 0, q = 0,
+    coef_fix_special = c("zero_sum_intercept", "panel_structure"))
```

Second, we estimate the standard mean-reverting GAS model of order one. In order to expedite the numerical optimization process, we incorporate starting values based on the static model.

```
R> est_stnry <- gas(y = y, x = x, distr = "pluce",
+    coef_fix_special = c("zero_sum_intercept", "panel_structure"),
+    coef_start = as.vector(rbind(est_static$fit$par_unc / 2, 0, 0.5, 0.5)))
```

Third, we estimate the random walk model. In other words, we set the autoregressive coefficient to 1. The easiest way to specify this is by including the value `random_walk` in the `coef_fix_special` argument. In our random walk model, we consider the initial values of the worth parameters to be parameters to be estimated. While the `par_init` argument does not directly support this, we can set `regress = "sep"` and use cumulative sums of exogenous variables to achieve this initialization for this particular model. However, it is generally not recommended to estimate initial parameter values as it introduces additional variables, lacks reasonable asymptotics, and can lead to overfitting in finite samples. It is important to approach the random walk model with caution, as it is not stationary and the standard maximum likelihood asymptotics are not valid.

```
R> est_walk <- gas(y = y, x = lapply(x, cumsum), distr = "pluce", regress = "sep",
+    coef_fix_special = c("zero_sum_intercept", "panel_structure", "random_walk"),
+    coef_start = as.vector(rbind(est_static$fit$par_unc, 0, 0.5, 1)))
```

To avoid redundancy, we will omit the output of the `gas()` function, which contains rows for each coefficient of each worth parameter. Since most coefficients are the same due to the assumed panel structure, it is unnecessary to display them all. Instead, we print only one set of the home advantage and dynamics coefficients.

```
R> cbind(est_static = c("beta1" = unname(coef(est_static)[2]), "alpha1" = 0, "phi1" = 0),
+    est_stnry = coef(est_stnry)[2:4], est_walk = coef(est_walk)[2:4])

       est_static est_stnry   est_walk
beta1   0.1707329 0.2274378 0.09873335
alpha1  0.0000000 0.3919432 0.34300137
phi1    0.0000000 0.5062479 1.00000000
```

In all three models, coefficient β_1 representing the home advantage is positive but not significant.

```
R> cbind(est_static = c("beta1" = unname(est_static$fit$coef_pval)[2], "alpha1" = 0, "phi1" = 0),
+    est_stnry = est_stnry$fit$coef_pval[2:4], est_walk = est_walk$fit$coef_pval[2:4])

       est_static    est_stnry     est_walk
beta1    0.514887 3.772815e-01 5.995825e-01
alpha1   0.000000 2.141361e-06 2.634611e-09
phi1     0.000000 6.463353e-04 0.000000e+00
```

We compare the models using the Akaike information criterion (AIC). The `gas` class allows for generic function `AIC()`. In terms of AIC, the mean-reverting model outperformed the remaining two by a wide margin.

```
R> AIC(est_static, est_stnry, est_walk)

           df      AIC
est_static 24 1299.600
est_stnry  26 1274.391
est_walk   25 1300.851
```

Our models enable us to construct the "ultimate" or long-run ranking. The rankings produced by both models are in agreement for all but the first three positions. However, the long-term strength estimates for these three teams are very close to each other, making the final ranking less clear-cut.

```
R> tibble(team = colnames(y)) %>%
+    mutate(stnry_strength = est_stnry$fit$par_unc) %>%
+    mutate(stnry_rank = rank(-stnry_strength)) %>%
+    mutate(static_strength = est_static$fit$par_unc) %>%
+    mutate(static_rank = rank(-static_strength)) %>%
+    arrange(stnry_rank)

# A tibble: 24 × 5
   team  stnry_strength stnry_rank static_strength static_rank
   <chr>          <dbl>      <dbl>           <dbl>       <dbl>
 1 CAN           3.72            1          3.72            2
 2 FIN           3.70            2          3.66            3
 3 SWE           3.65            3          3.84            1
 4 CZE           3.47            4          3.41            4
 5 RUS           3.25            5          3.17            5
 6 USA           1.83            6          2.18            6
 7 CHE           1.67            7          1.76            7
 8 SVK           1.65            8          1.55            8
 9 LVA           0.862           9          0.822           9
10 DEU           0.280          10          0.311          10
11 BLR           0.254          11          0.109          11
12 NOR           0.0335         12         -0.0743         12
13 DNK          -0.0732         13         -0.175          13
14 FRA          -0.405          14         -0.509          14
15 AUT          -0.833          15         -0.886          15
16 ITA          -1.02           16         -1.10           16
17 UKR          -1.34           17         -1.52           17
18 SVN          -1.75           18         -1.64           18
19 KAZ          -1.83           19         -1.78           19
20 JPN          -1.99           20         -1.94           20
21 HUN          -3.28           21         -3.20           21
22 GBR          -3.92           22         -3.89           22
23 POL          -3.95           23         -3.90           23
24 KOR          -3.96           24         -3.91           24
```

Additionally, we can examine the evolution of the worth parameters for individual teams over the years. The point estimates of time-varying parameter values can be directly obtained from the `gas()` function. Using the generic `plot()` function allows us to visualize the time-varying parameters of individual models. When multiple parameters are time-varying, as in our scenario, the function plots them in sequence. For the purpose of this document, we will only display figures specific to the Canada team (Figures 7, 8, and 9).

```
R> plot(est_static, which = 3)
R> plot(est_stnry, which = 3)
R> plot(est_walk, which = 3)
```

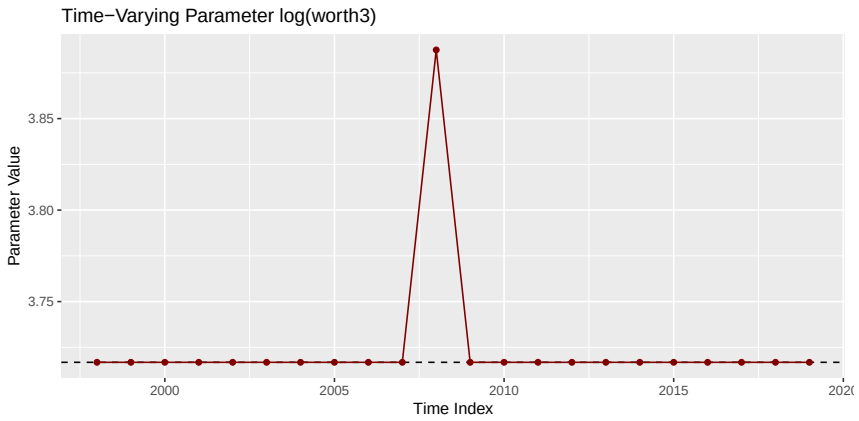

Figure 7: Time-varying parameters of the Canada team based on the static model.

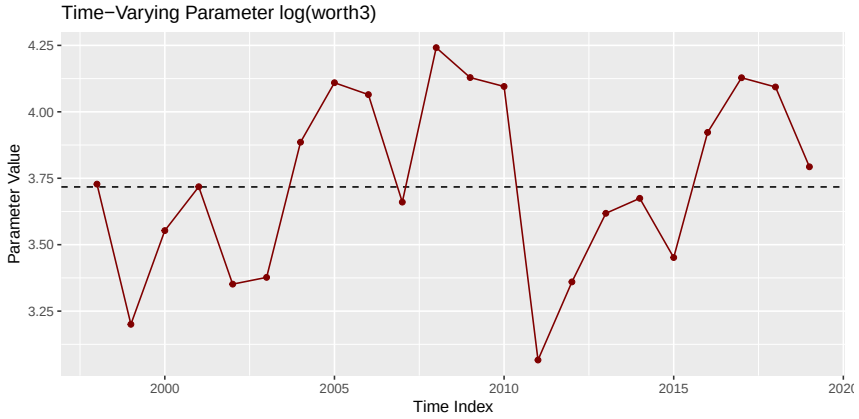

Figure 8: Time-varying parameters of the Canada team based on the stationary model.

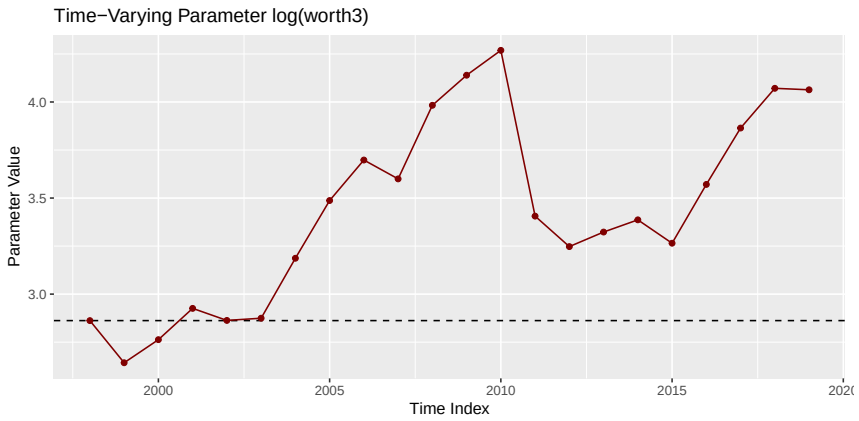

Figure 9: Time-varying parameters of the Canada team based on the random walk model.

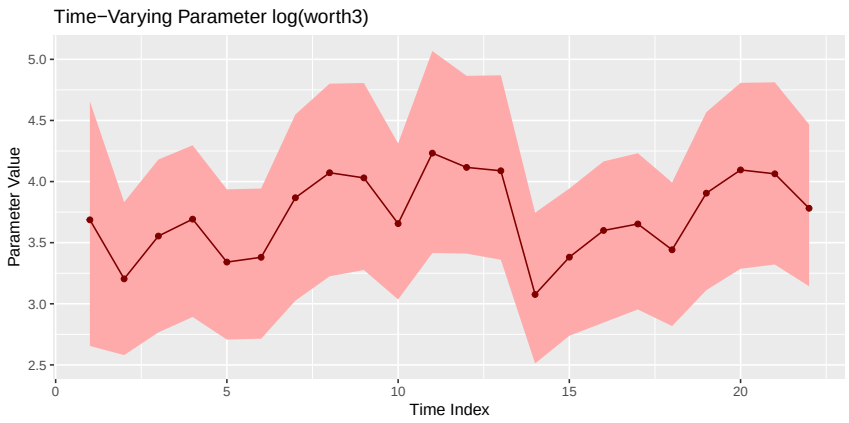

Figure 10: Confidence bands of time-varying parameters of the Canada team based on the stationary model.

However, it is important to note that these estimates are subject to uncertainty. To capture the uncertainty, we can utilize simulations by leveraging the `gas_filter()` function, which accepts the output of the `gas()` function as an argument. This allows us to obtain the standard deviations and quantiles for the worth parameter estimates, providing a more comprehensive understanding of the parameter dynamics over time.

```
R> set.seed(42)
R> flt_stnry <- gas_filter(est_stnry)
```

To visualize time-varying parameters with confidence band, we can use the `plot()` on the `gas_filter` object (Figure 10).

```
R> plot(flt_stnry, which = 3)
```

Finally, we perform one-year-ahead forecasts. We use the `gas_forecast()` function, which can again take the estimated model as an argument.

```
R> fcst_stnry <- gas_forecast(est_stnry, t_ahead = 1, x_ahead = 0)

R> tibble(team = colnames(y)) %>%
+    mutate(fcst_strength = fcst_stnry$forecast$par_tv_ahead_mean[1, ]) %>%
+    mutate(fcst_gold = exp(fcst_strength) / sum(exp(fcst_strength))) %>%
+    mutate(fcst_rank = rank(-fcst_strength)) %>%
+    mutate(real_rank = ice_hockey_championships$rankings[24, ]) %>%
+    arrange(real_rank)

# A tibble: 24 × 5
   team  fcst_strength fcst_gold fcst_rank real_rank
   <chr>         <dbl>     <dbl>     <dbl>     <dbl>
 1 CAN           3.97  0.234             2         1
 2 FIN           3.97  0.235             1         2
 3 USA           2.09  0.0356            6         3
 4 DEU           0.742 0.00929          10         4
 5 RUS           3.43  0.137             3         5
 6 CHE           1.82  0.0272            7         6
 7 CZE           3.41  0.134             4         7
 8 SVK           1.58  0.0214            8         8
 9 SWE           3.40  0.133             5         9
10 KAZ          -2.05  0.000569         19        10
11 LVA           0.978 0.0117            9        11
12 DNK           0.229 0.00556          11        12
13 NOR           0.125 0.00501          12        13
```

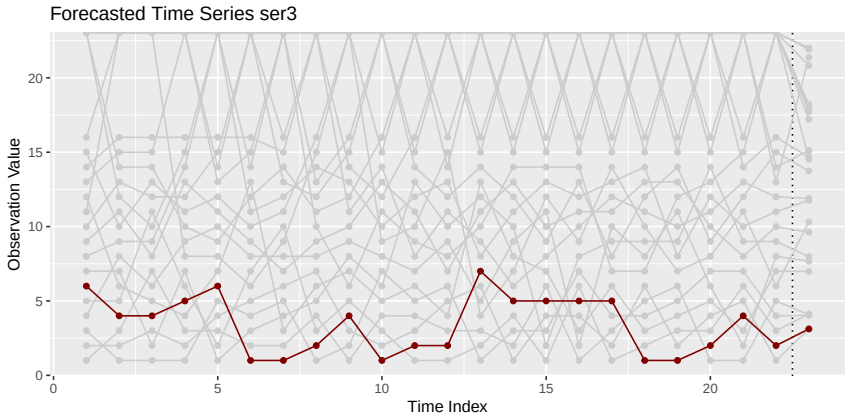

Figure 11: One-step ahead forecasts of the Canada team based on the stationary model.

```
14 GBR          -3.55  0.000127        22        14
15 BLR          -0.704 0.00219         14        15
16 ITA          -0.957 0.00170         16        16
17 AUT          -0.832 0.00192         15       Inf
18 FRA          -0.490 0.00271         13       Inf
19 HUN          -3.31  0.000162        21       Inf
20 JPN          -2.21  0.000486        20       Inf
21 KOR          -3.81  0.0000982       23       Inf
22 POL          -3.99  0.0000821       24       Inf
23 SVN          -1.95  0.000631        18       Inf
24 UKR          -1.69  0.000813        17       Inf
```

The forecasted values can be displayed using the generic `plot()` function (Figure 11).

```
R> plot(fcst_stnry, which = 3)
```

5 Limitations and Customization

5.1 Adding a New Distribution

Despite providing a reasonable range of distributions (refer to Table 1), the current version of the **gasmodel** package does not include certain distributions found in the GAS literature. Notable examples are copula models (see, e.g., De Lira Salvatierra and Patton, 2015; Koopman *et al.*, 2018), matrix models (see, e.g., Hansen *et al.*, 2016; Opschoor *et al.*, 2018), and censoring models (see, e.g., Harvey and Ito, 2020; Harvey and Liao, 2023).

Users are encouraged to customize the package by adding new distributions. To incorporate a new distribution into the package, please follow these steps:

1. Choose a name for the distribution and parametrization, such as `newdistr` and `newparam`, respectively.

2. Create an R file in the `R` directory, such as `R/distr_newdist_newparam.R`, which will contain all the necessary functions for the new distribution.

3. Implement the following functions in the R file, adhering to the structure used for other distributions in the package:

   - `distr_newdistr_newparam_parameters()` listing the parameters,
   - `distr_newdistr_newparam_density()` computing the density,

- `distr_newdistr_newparam_loglik()` computing the log-likelihood,
- `distr_newdistr_newparam_mean()` computing the mean,
- `distr_newdistr_newparam_var()` computing the variance,
- `distr_newdistr_newparam_score()` computing the score,
- `distr_newdistr_newparam_fisher()` computing the Fisher information,
- `distr_newdistr_newparam_random()` generating random variables,
- `distr_newdistr_newparam_start()` estimating starting values of the parameters.

4. Update the `distr_table.xlsx` file located in the `data-raw` directory by adding a new row to the table that includes the names of the distribution and parametrization.

5. Run the `distr_table.R` script located in the `data-raw` directory. This script saves the content of the `distr_table.xlsx` table to the `distr_table` dataset in the package.

By following these steps, users will be able to add a novel distribution to the package, integrating it with the existing framework.

5.2 Interaction Between Parameters and Non-Linear Dependence

The dynamics are implemented in the standard form (3) of Creal *et al.* (2013), which has been further extended to include exogenous variables in the form (4) or (5). However, it is worth noting that existing literature includes models with interactions between different time-varying parameters or nonlinear forms of dependence on past values (see, e.g., Harvey and Sucarrat, 2014; Holý and Tomanová, 2022). Incorporating such complex dynamics would significantly complicate the interface of the functions, so we have opted to keep the dynamics simple for ease of use.

Nevertheless, the source code can be modified to accommodate specific cases. This can be achieved by using a placeholder exogeneous variable and making a manual adjustment within the `likelihood_evaluate()` function in the `helper_likelihood.R` file. Specifically, the value of the placeholder variable can be can hard-coded to a desirable transformation of any concurrent or lagged parameter.

5.3 Non-Standard Structure of Time Series

The package focuses on the standard form of time series. However, certain applications, such as those in the field of sports statistics, may require a specialized structure for modeling time series data. In these cases, the individual matches between teams or players in a specific league are often modeled using distributions like Bernoulli, Skellam, or bivariate Poisson (see, e.g., Gorgi *et al.*, 2019; Koopman and Lit, 2019). Time series should therefore represent the outcomes of matches. However, at each observation, different teams may be participating. This unique characteristic cannot be adequately captured by the standard form of univariate (or bivariate) time series, and a more sophisticated data structure is required to account for the varying teams involved.

To address this limitation, an R package that specifically caters to the use of score-driven models in sports statistics is currently being developed. This specialized package will provide the necessary tools and data structures to effectively model and analyze the unique dynamics present in these applications. However, there are other options beyond R that already exist. Notably, GAS pairwise comparison models can be estimated using the **PyFlux** package in Python (Taylor, 2018), as well as through the stand-alone GUI application **Time Series Lab** (Lit *et al.*, 2021).

5.4 Other Dynamic Models Using Score

In the literature on GAS models, the score has been employed in a wide range of dynamic models. Some of these models fall outside the scope of this package. Examples of such models include semi-parametric models (see, e.g., Blasques *et al.*, 2016a; Patton *et al.*, 2019), Markov regime switching models (see, e.g., Bazzi *et al.*, 2017; Blazsek and Haddad, 2022), and spatio-temporal models (see, e.g., Catania and Billé, 2017; Gasperoni *et al.*, 2023).

6 Conclusion

The purpose of the **gasmodel** package is to provide researchers, analysts, and data scientists with a versatile toolkit for a broad spectrum of GAS models in R. While it is important to note that not all GAS models found in the literature are supported by the package due to their diverse nature, the package still provides a solid foundation. For some specific GAS models, modifications of the package may be required, or an alternative specialized package/code may prove to be a better option. Nevertheless, the **gasmodel** package offers considerable flexibility for specifying dynamics, and it boasts an extensive array of probability distribution options. This ensures that users have a diverse set of tools at their disposal when working with GAS models, enabling them to tailor their analyses to their specific needs.

Funding

The work on this paper was supported by the Czech Science Foundation under project 23-06139S and the personal and professional development support program of the Faculty of Informatics and Statistics, Prague University of Economics and Business.

References

Alvo M, Yu PLH (2014). *Statistical Methods for Ranking Data*. Springer, New York. ISBN 978-1-4939-1470-8. https://doi.org/10.1007/978-1-4939-1471-5.

Ardia D, Boudt K, Catania L (2019). "Generalized Autoregressive Score Models in R: The GAS Package." *Journal of Statistical Software*, **88**(6), 1–28. ISSN 1548-7660. https://doi.org/10.18637/jss.v088.i06.

Artemova M, Blasques F, van Brummelen J, Koopman SJ (2022). "Score-Driven Models: Methods and Applications." *Oxford Research Encyclopedia of Economics and Finance*. https://doi.org/10.1093/acrefore/9780190625979.013.671.

Bazzi M, Blasques F, Koopman SJ, Lucas A (2017). "Time-Varying Transition Probabilities for Markov Regime Switching Models." *Journal of Time Series Analysis*, **38**(3), 458–478. ISSN 0143-9782. https://doi.org/10.1111/jtsa.12211.

Blasques F, Gorgi P, Koopman SJ, Wintenberger O (2018). "Feasible Invertibility Conditions and Maximum Likelihood Estimation for Observation-Driven Models." *Electronic Journal of Statistics*, **12**(1), 1019–1052. ISSN 1935-7524. https://doi.org/10.1214/18-ejs1416.

Blasques F, Holý V, Tomanová P (2022a). "Zero-Inflated Autoregressive Conditional Duration Model for Discrete Trade Durations with Excessive Zeros." https://arxiv.org/abs/1812.07318.

Blasques F, Ji J, Lucas A (2016a). "Semiparametric Score Driven Volatility Models." *Computational Statistics & Data Analysis*, **100**(2013), 58–69. ISSN 0167-9473. https://doi.org/10.1016/j.csda.2015.04.003.

Blasques F, Koopman SJ, Łasak K, Lucas A (2016b). "In-Sample Confidence Bands and Out-of-Sample Forecast Bands for Time-Varying Parameters in Observation-Driven Models." *International Journal of Forecasting*, **32**(3), 875–887. ISSN 0169-2070. https://doi.org/10.1016/j.ijforecast.2015.11.018.

Blasques F, Koopman SJ, Lucas A (2014). "Stationarity and Ergodicity of Univariate Generalized Autoregressive Score Processes." *Electronic Journal of Statistics*, **8**(1), 1088–1112. ISSN 1935-7524. https://doi.org/10.1214/14-ejs924.

Blasques F, Koopman SJ, Lucas A (2015). "Information-Theoretic Optimality of Observation-Driven Time Series Models for Continuous Responses." *Biometrika*, **102**(2), 325–343. ISSN 0006-3444. https://doi.org/10.1093/biomet/asu076.

Blasques F, Lucas A, van Vlodrop AC (2021). "Finite Sample Optimality of Score-Driven Volatility Models: Some Monte Carlo Evidence." *Econometrics and Statistics*, **19**, 47–57. ISSN 2452-3062. https://doi.org/10.1016/j.ecosta.2020.03.010.

Blasques F, van Brummelen J, Koopman SJ, Lucas A (2022b). "Maximum Likelihood Estimation for Score-Driven Models." *Journal of Econometrics*, **227**(2), 325–346. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2021.06.003.

Blazsek S, Haddad MFC (2022). "Score-Driven Multi-Regime Markov-Switching EGARCH: Empirical Evidence Using the Meixner Distribution." *Studies in Nonlinear Dynamics and Econometrics*. ISSN 1558-3708. https://doi.org/10.1515/snde-2021-0101.

Blazsek S, Licht A (2020). "Dynamic Conditional Score Models: A Review of Their Applications." *Applied Economics*, **52**(11), 1181–1199. ISSN 0003-6846. https://doi.org/10.1080/00036846.2019.1659498.

Bodin G, Saavedra R, Fernandes C, Street A (2020). "ScoreDrivenModels.jl: a Julia Package for Generalized Autoregressive Score Models." https://arxiv.org/abs/2008.05506.

Bollerslev T (1986). "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics*, **31**(3), 307–327. ISSN 0304-4076. https://doi.org/10.1016/0304-4076(86)90063-1.

Calvori F, Cipollini F, Gallo GM (2013). "Go with the Flow: A GAS Model for Predicting Intra-Daily Volume Shares." https://ssrn.com/abstract=2363483.

Catania L, Billé AG (2017). "Dynamic Spatial Autoregressive Models with Autoregressive and Heteroskedastic Disturbances." *Journal of Applied Econometrics*, **32**(6), 1178–1196. ISSN 0883-7252. https://doi.org/10.1002/jae.2565.

Creal D, Koopman SJ, Lucas A (2008). "A General Framework for Observation Driven Time-Varying Parameter Models." https://www.tinbergen.nl/discussion-paper/2649.

Creal D, Koopman SJ, Lucas A (2011). "A Dynamic Multivariate Heavy-Tailed Model for Time-Varying Volatilities and Correlations." *Journal of Business & Economic Statistics*, **29**(4), 552–563. ISSN 0735-0015. https://doi.org/10.1198/jbes.2011.10070.

Creal D, Koopman SJ, Lucas A (2013). "Generalized Autoregressive Score Models with Applications." *Journal of Applied Econometrics*, **28**(5), 777–795. ISSN 0883-7252. https://doi.org/10.1002/jae.1279.

Critchlow DE, Fligner MA, Verducci JS (1991). "Probability Models on Rankings." *Journal of Mathematical Psychology*, **35**(3), 294–318. ISSN 0022-2496. https://doi.org/10.1016/0022-2496(91)90050-4.

Davis RA, Dunsmuir WTM, Street SB (2003). "Observation-Driven Models for Poisson Counts." *Biometrika*, **90**(4), 777–790. ISSN 0006-3444. https://doi.org/10.1093/biomet/90.4.777.

De Lira Salvatierra I, Patton AJ (2015). "Dynamic Copula Models and High Frequency Data." *Journal of Empirical Finance*, **30**, 120–135. ISSN 0927-5398. https://doi.org/10.1016/j.jempfin.2014.11.008.

Engle RF, Russell JR (1998). "Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data." *Econometrica*, **66**(5), 1127–1162. ISSN 0012-9682. https://doi.org/10.2307/2999632.

Fonseca RV, Cribari-Neto F (2018). "Bimodal Birnbaum-Saunders Generalized Autoregressive Score Model." *Journal of Applied Statistics*, **45**(14), 2585–2606. ISSN 0266-4763. https://doi.org/10.1080/02664763.2018.1428734.

Gasperoni F, Luati A, Paci L, D'Innocenzo E (2023). "Score-Driven Modeling of Spatio-Temporal Data." *Journal of the American Statistical Association*, **118**(542), 1066–1077. ISSN 0162-1459. https://doi.org/10.1080/01621459.2021.1970571.

Gorgi P, Koopman SJ, Lit R (2019). "The Analysis and Forecasting of Tennis Matches by Using a High Dimensional Dynamic Model." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **182**(4), 1393–1409. ISSN 0964-1998. https://doi.org/10.1111/rssa.12464.

Hansen PR, Janus P, Koopman SJ (2016). "Realized Wishart-Garch: A Score-Driven Multi-Asset Volatility Model." https://doi.org/10.2139/ssrn.2821497. https://www.ssrn.com/abstract=2821497.

Harvey A, Hurn S, Thiele S (2019). "Modeling Directional (Circular) Time Series." https://doi.org/10.17863/cam.43915.

Harvey A, Liao Y (2023). "Dynamic Tobit Models." *Econometrics and Statistics*, **26**, 72–83. ISSN 2452-3062. https://doi.org/10.1016/j.ecosta.2021.08.012.

Harvey A, Sucarrat G (2014). "EGARCH Models with Fat Tails, Skewness and Leverage." *Computational Statistics & Data Analysis*, **76**, 320–338. ISSN 0167-9473. https://doi.org/10.1016/j.csda.2013.09.022.

Harvey AC (2013). *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series*. First Edition. Cambridge University Press, New York. ISBN 978-1-107-63002-4. https://doi.org/10.1017/cbo9781139540933.

Harvey AC (2022). "Score-Driven Time Series Models." *Annual Review of Statistics and Its Application*, **9**(1), 321–342. ISSN 2326-8298. https://doi.org/10.1146/annurev-statistics-040120-021023.

Harvey AC, Chakravarty T (2008). "Beta-t-(E)GARCH." https://doi.org/10.17863/cam.5286.

Harvey AC, Ito R (2020). "Modeling Time Series When Some Observations Are Zero." *Journal of Econometrics*, **214**(1), 33–45. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2019.05.003.

Hautsch N, Huang R (2012). "The Market Impact of a Limit Order." *Journal of Economic Dynamics and Control*, **36**(4), 501–522. ISSN 0165-1889. https://doi.org/10.1016/j.jedc.2011.09.012.

Holý V (2020). "Impact of the Parametrization and the Scaling Function in Dynamic Score-Driven Models: The Case of the Negative Binomial Distribution." In *Proceedings of the 38th International Conference Mathematical Methods in Economics*, 173–179. Mendel University in Brno, Brno. ISBN 978-80-7509-734-7. https://mme2020.mendelu.cz/wcd/w-rek-mme/mme2020{_}conference{_}proceedings{_}final.pdf.

Holý V, Tomanová P (2022). "Modeling Price Clustering in High-Frequency Prices." *Quantitative Finance*, **22**(9), 1649–1663. ISSN 1469-7688. https://doi.org/10.1080/14697688.2022.2050285.

Holý V, Zouhar J (2022). "Modelling Time-Varying Rankings with Autoregressive and Score-Driven Dynamics." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **71**(5), 1427–1450. ISSN 0035-9254. https://doi.org/10.1111/rssc.12584.

Koopman SJ, Lit R (2019). "Forecasting Football Match Results in National League Competitions Using Score-Driven Time Series Models." *International Journal of Forecasting*, **35**(2), 797–809. ISSN 0169-2070. https://doi.org/10.1016/j.ijforecast.2018.10.011.

Koopman SJ, Lit R, Lucas A, Opschoor A (2018). "Dynamic Discrete Copula Models for High-Frequency Stock Price Changes." *Journal of Applied Econometrics*, **33**(7), 966–985. ISSN 0883-7252. https://doi.org/10.1002/jae.2645.

Koopman SJ, Lucas A, Scharth M (2016). "Predicting Time-Varying Parameters with Parameter-Driven and Observation-Driven Models." *Review of Economics and Statistics*, **98**(1), 97–110. ISSN 0034-6535. https://doi.org/10.1162/rest_a_00533.

Lit R, Koopman SJ, Harvey AC (2021). "Time Series Lab - Score Edition." https://timeserieslab.com.

Luce RD (1977). "The Choice Axiom after Twenty Years." *Journal of Mathematical Psychology*, **15**(3), 215–233. ISSN 0022-2496. https://doi.org/10.1016/0022-2496(77)90032-3.

Opschoor A, Janus P, Lucas A, van Dijk D (2018). "New HEAVY Models for Fat-Tailed Realized Covariances and Returns." *Journal of Business & Economic Statistics*, **36**(4), 643–657. ISSN 0735-0015. https://doi.org/10.1080/07350015.2016.1245622.

Patton AJ, Ziegel JF, Chen R (2019). "Dynamic Semiparametric Models for Expected Shortfall (and Value-at-Risk)." *Journal of Econometrics*, **211**(2), 388–413. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2018.10.008.

Plackett RL (1975). "The Analysis of Permutations." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **24**(2), 193–202. ISSN 0035-9254. https://doi.org/10.2307/2346567.

Stern H (1990). "Models for Distributions on Permutations." *Journal of the American Statistical Association*, **85**(410), 558–564. ISSN 0162-1459. https://doi.org/10.1080/01621459.1990.10476235.

Sucarrat G (2013). "betategarch: Simulation, Estimation and Forecasting of Beta-Skew-t-EGARCH Models." *R Journal*, **5**(2), 137–147. ISSN 2073-4859. https://doi.org/10.32614/rj-2013-034.

Taylor R (2018). "PyFlux: An Open Source Time Series Library for Python." https://github.com/rjt1990/pyflux.

Tomanová P, Holý V (2021). "Clustering of Arrivals in Queueing Systems: Autoregressive Conditional Duration Approach." *Central European Journal of Operations Research*, **29**(3), 859–874. ISSN 1435-246X. https://doi.org/10.1007/s10100-021-00744-7.

Yellott JI (1977). "The Relationship Between Luce's Choice Axiom, Thurstone's Theory of Comparative Judgment, and the Double Exponential Distribution." *Journal of Mathematical Psychology*, **15**(2), 109–144. ISSN 0022-2496. https://doi.org/10.1016/0022-2496(77)90026-8.

Yu PLH, Gu J, Xu H (2019). "Analysis of Ranking Data." *Wiley Interdisciplinary Reviews: Computational Statistics*, **11**(6), e1483:1–e1483:26. ISSN 1939-5108. https://doi.org/10.1002/wics.1483.

A Comparison with the GAS Package

In this appendix, we compare the **gasmodel** package with the **GAS** package. First, let us try replicate the results from the case studies in Section 4. The **GAS** package offers the exponential and gamma distributions but does not support the Weibull and generalized gamma distributions. The exponential distribution is parametrized in terms of the rate parameter with the logistic link function in the **GAS** package. The **gasmodel** package allows for both the scale and rate parametrizations as well as the identical and logarithmic link functions. When the logarithmic link function is used, however, the only difference between the scale and rate parametrizations is in the sign of the constant.

We can therefore compare the exponential model using the scale parametrization estimated by the **gasmodel** package in Section 4.1 with the following exponential model using the rate parametrization estimated by the **GAS** package.

```
R> spec_exp <- UniGASSpec(Dist = "exp", GASPar = list(location = TRUE))
R> fit_exp <- UniGASFit(spec_exp, data = y)
R> fit_exp

----------------------------------------
-          Univariate GAS Fit           -
----------------------------------------

Model Specification:
T =  5752
Conditional distribution:  exp
Score scaling type:  Identity
Time varying parameters:  location
----------------------------------------
Estimates:
           Estimate  Std. Error     t value     Pr(>|t|)
kappa1 0.0008535178 0.001149039   0.7428099 2.287984e-01
a1     0.0488830441 0.006513237   7.5051838 3.064216e-14
b1     0.9634339711 0.009119708 105.6430684 0.000000e+00

----------------------------------------
Unconditional Parameters:
location
1.023616

----------------------------------------
Information Criteria:
      AIC       BIC        np       llk
11223.036 11243.008     3.000 -5608.518

----------------------------------------
Convergence: 0
----------------------------------------

Elapsed time: 0.02 mins
```

The results are nearly identical, within a reasonable level of precision. Other than the inverted sign of the constant, the only difference lies in the reported p-values: the **GAS** package seems to employ one-tailed hypotheses, whereas the **gasmodel** package uses two-tailed hypotheses. The visual representation of time-varying parameters is also comparable, albeit with inverted signs (Figure 12).

```
plot(fit_exp, which = 1)
```

Next, we estimate the model with the gamma distribution and the rate parametrization.

```
R> spec_gamma <- UniGASSpec(Dist = "gamma", GASPar = list(scale = TRUE, shape = FALSE))
R> fit_gamma <- UniGASFit(spec_gamma, data = y)
R> fit_gamma

----------------------------------------
-          Univariate GAS Fit           -
----------------------------------------

Model Specification:
T =  5752
Conditional distribution:  gamma
Score scaling type:  Identity
Time varying parameters:  scale
----------------------------------------
Estimates:
```

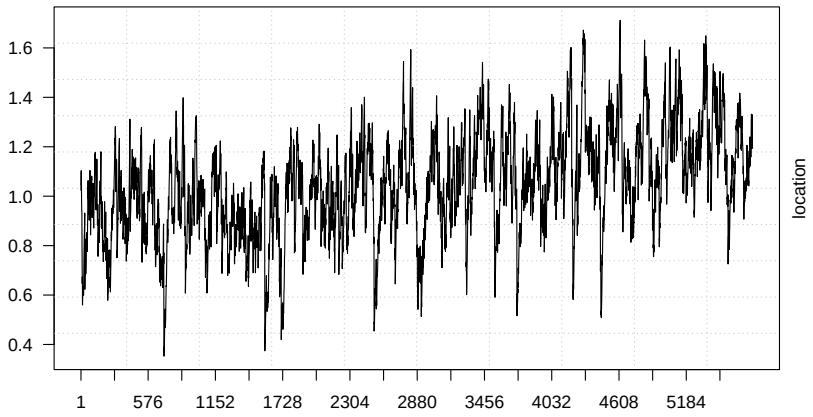

Figure 12: Time-varying parameters based on the exponential model with the rate parametrization from the **GAS** package.

```
          Estimate  Std. Error   t value     Pr(>|t|)
kappa1 -0.01301586 0.005587936 -2.329279 9.922151e-03
kappa2 -0.06533321 0.021302585 -3.066915 1.081403e-03
a1      0.05420745 0.009282091  5.840005 2.609969e-09
b1      0.82496872 0.058251327 14.162231 0.000000e+00

-----------------------------------------
Unconditional Parameters:
    scale     shape
0.9283346 0.9367553

-----------------------------------------
Information Criteria:
      AIC       BIC        np       llk
11330.024 11356.653     4.000 -5661.012

-----------------------------------------
Convergence: 0
-----------------------------------------

Elapsed time: 0.03 mins
```

This result significantly contrasts with the outcome obtained from the gamma model using the scale parametrization, as estimated by the **gasmodel** package in Section 4.1. The default optimizer within the **GAS** package identifies a suboptimal solution, yielding a significantly lower log-likelihood compared to the exponential model. Note that the gamma distribution is a generalization of the exponential distribution and should therefore result in the same or better fit. A visual examination of the time-varying parameters further underscores the substantial disparity between the estimated gamma and exponential models (Figure 13).

```
R> plot(fit_gamma, which = 1)
```

The default optimizer within the **gasmodel** package finds a considerably superior solution, likely the optimal one, albeit demanding more computational resources. In both packages, it is possible to alter the optimizers. However, in the **GAS** package, the optimizer's parameters cannot be directly provided through the `UniGASFit()` function. Instead, a complete replacement of the optimizer is necessary, rendering it a more intricate process to manage.

After performing parameter estimation for the Weibull and generalized gamma distributions, the case study presented in Section 4.1 proceeds by introducing a trend into the model. Regrettably, the **GAS** package lacks the capacity for accommodating exogenous variables, thus preventing this

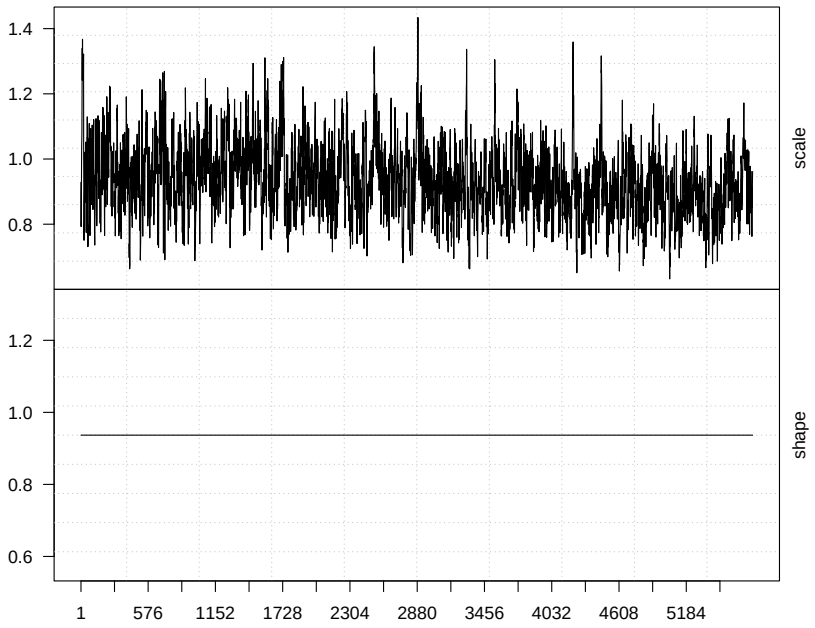

Figure 13: Time-varying parameters based on the gamma model with the rate parametrization from the **GAS** package.

extension. This shortcoming stands as a substantial limitation that considerably restricts the package's potential applications. The case study in Section 4.1 also utilizes the bootstrapping function provided by the **gasmodel** package. Such a feature is absent in the **GAS** package. This primarily affects convenience, as bootstrapping can still be executed using custom code from the user along with specialized packages. The functionality for simulation of GAS processes is very similar in both packages.

The second case study presented in Section 4.2 is not replicable at all using the **GAS** package due to its lack of support for the Plackett–Luce distribution or any distribution based on rankings. Furthermore, the **GAS** package does not facilitate the imposition of constraints on coefficients, which is useful, for instance, in creating random walk models or multivariate models with a panel structure. In the same case study, the process of forecasting and deriving confidence bands on time-varying parameters is illustrated. Similar functionality is also offered by the **GAS** package.

Table 2 compares the supported distributions, while Table 3 contrasts the available functionalities in both packages. In general, the **gasmodel** package offers much broader range of GAS models, encompassing various probability distributions and model specifications. The **gasmodel** package (version 0.5.1) supports 26 distributions, whereas the **GAS** package (version 0.3.4) includes only 16 distributions. The **GAS** package features asymmetric and skewed versions of the normal and Student's t distributions, which are currently absent in the **gasmodel** package. Conversely, the **gasmodel** package incorporates 14 distributions tailored for count, duration, categorical, circular, compositional, and ranking data, which are not present in the **GAS** package. While the **GAS** package caters primarily to standard GAS models without the ability to handle missing values, the **gasmodel** package offers enhanced flexibility, allowing for various model structures, incorporation of exogenous variables, and the handling of missing values in time series. Apart from differences in probability distributions and model specification, both packages provide analogous functionalities for inference, forecasting, and simulation. The **GAS** package also computes the probability integral transform and offers certain capabilities for backtesting one-step ahead density and Value-at-Risk. However, these functionalities are limited to continuous distributions, which constitute only a subset of GAS models. Furthermore, such functionalities can be derived from the output generated by the **gasmodel** package. For these reasons, we have decided not to implement them in **gasmodel**.

Table 2: Comparison of the supported distributions in the **gasmodel** and **GAS** packages.

Distribution	**gasmodel**	**GAS**
Asymmetric Laplace	✓	✓
Asymmetric Student's t with One Tail Decay	✗	✓
Asymmetric Student's t with Two Tail Decay	✗	✓
Bernoulli	✓	✓
Beta	✓	✓
Birnbaum–Saunders	✓	✗
Categorical	✓	✗
Dirichlet	✓	✗
Double Poisson	✓	✗
Exponential	✓	✓
Gamma	✓	✓
Generalized Gamma	✓	✗
Geometric	✓	✗
Laplace	✓	✗
Multivariate Normal	✓	✓
Multivariate Student's t	✓	✓
Negative Binomial	✓	✓
Normal	✓	✓
Plackett–Luce	✓	✗
Poisson	✓	✓
Skellam	✓	✓
Skewed Normal	✗	✓
Skewed Student's t	✗	✓
Student's t	✓	✓
von Mises	✓	✗
Weibull	✓	✗
Zero-Inflated Geometric	✓	✗
Zero-Inflated Negative Binomial	✓	✗
Zero-Inflated Poisson	✓	✗
Zero-Inflated Skellam	✓	✗

Table 3: Comparison of the available functionality in the **gasmodel** and **GAS** packages.

Functionality	**gasmodel**	**GAS**
Various parametrizations and link functions	✓	✗
Exogenous variables	✓	✗
Higher score and autoregressive orders	✓	✗
Custom initial values of time-varying parameters	✓	✗
Fixed and bounded values of coefficients	✓	✗
Missing values	✓	✗
Custom optimization function	✓	✓
Hessian-based inference	✓	✓
Probability integral transform	✗	✓
Confidence bands	✓	✓
Forecasting	✓	✓
Backtesting and rolling re-estimation	✗	✓
Basic simulation	✓	✓
Bootstrapping	✓	✗
Easy visualization	✓	✓

Modelling Time-Varying Rankings with Autoregressive and Score-Driven Dynamics

Vladimír Holý
Prague University of Economics and Business
Winston Churchill Square 1938/4, 130 67 Prague 3, Czechia
vladimir.holy@vse.cz

Jan Zouhar
Prague University of Economics and Business
Winston Churchill Square 1938/4, 130 67 Prague 3, Czechia
zouharj@vse.cz

Abstract: We develop a new statistical model to analyse time-varying ranking data. The model can be used with a large number of ranked items, accommodates exogenous time-varying covariates and partial rankings, and is estimated via the maximum likelihood in a straightforward manner. Rankings are modelled using the Plackett-Luce distribution with time-varying worth parameters that follow a mean-reverting time series process. To capture the dependence of the worth parameters on past rankings, we utilise the conditional score in the fashion of the generalised autoregressive score (GAS) models. Simulation experiments show that the small-sample properties of the maximum-likelihood estimator improve rapidly with the length of the time series and suggest that statistical inference relying on conventional Hessian-based standard errors is usable even for medium-sized samples. In an empirical study, we apply the model to the results of the Ice Hockey World Championships. We also discuss applications to rankings based on underlying indices, repeated surveys, and non-parametric efficiency analysis.

Keywords: Ranking Data, Random Permutation, Plackett-Luce Distribution, Generalised Autoregressive Score Model, Ice Hockey Rankings.

JEL Classification: C32, C46, L83.

1 Introduction

The rankings of universities, scientific journals, sports teams, election candidates, top-visited websites, or products preferred by customers are all examples of ranking data. Statistical models of ranking data have a long history, dating back at least to Thurstone (1927). Since then, the breadth of the statistical toolkit for ranking data has increased rapidly; see, e.g., Marden (1995) and Alvo and Yu (2014) for an in-depth textbook overview. However, a recent survey of the ranking literature by Yu *et al.* (2019) draws attention to the lack of the time perspective in rankings and calls for research in this particular direction. This paper aims to heed this call and contribute to the thin strand of literature on time-varying ranking data. Unlike the existing models for time variation in rankings, our approach aims to provide a flexible tool for the modelling of time-varying ranking data that is similar to the autoregressive moving average (ARMA) model in the case of continuous variables.

Our model builds upon the (static) *Plackett-Luce distribution* of Luce (1959) and Plackett (1975), a convenient and simple probability distribution on rankings utilising a *worth parameter* for each item to be ranked. It originates from Luce's *choice axiom* and is also related to the Thurstone's *theory of comparative judgment* (see Luce, 1977 and Yellott, 1977 for details). Although it is not without limitations, the Plackett-Luce distribution is widely used as a base for statistical models that are used to analyse ranking data.

This also holds true for the scarce literature devoted to models with time-varying ranks. Baker and Mchale (2015) utilise the Plackett-Luce model and consider the individual worth parameters behind the rankings to be time-varying – but deterministically so – in an application to golf tournament

results. Glickman and Hennessy (2015) also base their model on the Plackett-Luce distribution but consider worth parameters following the Gaussian random walk in an application to women's alpine downhill skiing results. Asfaw *et al.* (2017) take a different path and include the lagged ranking as the current modal ranking in the Mallows model in an application to the academic performance of high school students. Finally, Henderson and Kirrane (2018) employ the Plackett-Luce model with observations weighted in time in an application to Formula One results. The latter three papers adopt a Bayesian approach.

The generalised autoregressive score (GAS) models of Creal *et al.* (2013), which are also called *dynamic conditional score* (DCS) models by Harvey (2013), have established themselves as a useful modern framework for time series modelling. The GAS models are observation-driven models allowing for any underlying probability distribution with any time-varying parameters. They capture the dynamics of time-varying parameters using the autoregressive term and the lagged score, i.e., the gradient of the log-likelihood function. The GAS class includes many well-known econometric models, such as the *generalised autoregressive conditional heteroskedasticity* (GARCH) model of Bollerslev (1986), which is based on the normal distribution with time-varying variance; the *autoregressive conditional duration* (ACD) model of Engle and Russell (1998), which is based on the exponential distribution with a time-varying scale; and the count model of Davis *et al.* (2003), which is based on the Poisson distribution with a time-varying mean. The GAS models can be straightforwardly estimated by the maximum likelihood method (see, e.g., Blasques *et al.*, 2018 for details on the asymptotic theory). Generally, the GAS models perform very well when compared to alternatives (see, e.g., Koopman *et al.*, 2016 for an extensive empirical and simulation study). Currently, the website `www.gasmodel.com` lists over 200 scientific papers devoted to the GAS models.

In the paper, we propose a dynamic model for rankings based on the Plackett-Luce distribution with time-varying worth parameters following the GAS score-driven dynamics. Our formulation allows for exogenous covariates and corresponds to the setting of a panel linear regression with fixed effects. We also consider the case of partial rankings. The proposed model is described in Section 2.

Using simulations, we investigate the finite-sample performance of the maximum likelihood estimator of our model. First, we demonstrate the convergence of the estimated coefficients for exogenous variables and of the GAS dynamics to their true values along the time dimension. Second, we show that confidence intervals based on the standard maximum likelihood asymptotics appear to be usable even if the dimensions of data are moderate (such as 20 items ranked in 20 time periods). The simulation study is conducted in Section 3.

To demonstrate the proposed methodology, we analyse the results of the Ice Hockey World Championships from 1998 to 2019. We find that the proposed mean-reverting model fits the data better than the static and random walk models. The benefits of our approach include a compilation of the ultimate (long-term) ranking of teams, the straightforward estimation of the probabilities of specific rankings (e.g., podium positions), and the prediction of future rankings. The empirical study is presented in Section 4.

Besides sports statistics, we discuss several other possible applications of the proposed model. Notably, we argue that our approach can be used to model rankings based on underlying indices (such as various country rankings) and captures the interaction between items, which the univariate models used directly for indices do not account for. Furthermore, we note that our model is suitable for repeated surveys that are designed as rankings. Finally, we show how our approach can be utilised for the rankings of decision-making units obtained by non-parametric efficiency analysis. These applications are discussed in Section 5.

2 Dynamic Score-Driven Ranking Model

2.1 Plackett–Luce Distribution

Let us consider a set of N items $\mathcal{Y} = \{1, \dots, N\}$. Our main object of interest is a complete permutation of this set $y = (y(1), \dots, y(N))$, known as a *ranking*, and its inverse $y^{-1} = \big(y^{-1}(1), \dots, y^{-1}(N)\big)$, known as an *ordering*. Element $y(i)$ represents the rank given to item i while $y^{-1}(r)$ represents the

item with rank r; to enhance readability, in subscripts we will simply write r^{th} instead of $y^{-1}(r)$ to denote the item ranked r^{th}.

We assume that a random permutation Y follows the *Plackett-Luce distribution* of Luce (1959) and Plackett (1975). According to this distribution, a ranking is constructed by successively selecting the best item, the second best item, the third best item, and so on. The probability of selecting a specific item in any stage is equal to the ratio of its worth parameter and the sum of the worth parameters of all items that have not yet been selected. Therefore, the probability of a complete ranking y is

$$\mathrm{P}\left[Y = y|f\right] = \prod_{r=1}^{N} \frac{\exp f_{r^{\text{th}}}}{\sum_{s=r}^{N} \exp f_{s^{\text{th}}}}, \tag{1}$$

where $f = (f_1, \ldots, f_N)'$ are the items' worth parameters. We use a parametrization allowing for arbitrary values of f_i, which facilitates subsequent modelling. Note that the probability mass function (1) is invariant to the addition of a constant to all parameters f_i. Therefore, we employ the standardisation

$$\sum_{i=1}^{N} f_i = 0. \tag{2}$$

The log-likelihood function is

$$\ell\left(f|y\right) = \sum_{i=1}^{N} f_i - \sum_{r=1}^{N} \ln\left(\sum_{s=r}^{N} \exp f_{s^{\text{th}}}\right). \tag{3}$$

For a random sample of rankings, a necessary and sufficient condition for the log-likelihood to have a unique maximum is that in every possible partition of $\mathcal{Y}$ into two non-empty subsets, some item in the second set ranks higher than some item in the first set at least once (Hunter, 2004). This condition, for example, rules out that there is an item always ranked first (in maximum likelihood estimation, this would result in an infinite worth parameter). To overcome the limitations of this condition in practical applications, Luo and Qin (2019) propose a penalised maximum likelihood estimator that adds a small perturbation to the log-likelihood.

For further details regarding the Plackett-Luce distribution, see Luce (1977), Yellott (1977), Stern (1990), and Critchlow *et al.* (1991).

2.2 Conditional Score

The key ingredient in our dynamic model is the score, i.e., the gradient of the log-likelihood function, which is defined as

$$\nabla\left(f|y\right) = \frac{\partial \ell\left(f|y\right)}{\partial f}. \tag{4}$$

The score represents the direction for improving the fit of the distribution with a given f to a specific observation y and indicates the sensitivity of the log-likelihood to the parameter f. For a complete ranking y following the Plackett-Luce distribution, the score is given by

$$\nabla_i\left(f|y\right) = 1 - \sum_{r=1}^{y(i)} \frac{\exp f_i}{\sum_{s=r}^{N} \exp f_{s^{\text{th}}}}, \qquad i = 1, \ldots, N. \tag{5}$$

An example with three items, in which the score is easily obtained, is given in Appendix A. Appendix B rewrites the score formula using the softmax function and shows additional steps in its derivation. In general, the score function has zero expected value and its variance is equal to the Fisher information:

$$\mathcal{I}(f) = \mathrm{E}\left[\nabla\left(f|y\right)\nabla\left(f|y\right)'\middle|f\right]. \tag{6}$$

Although the Fisher information is available in a closed form for the Plackett-Luce distribution, it is computationally very intensive for larger N as it includes a sum over all possible permutations of $\mathcal{Y}$.

The score has an appealing interpretation. In essence, it reflects the discrepancy between the items' worth parameters and the eventual ranking. This information can be exploited in a time-series context where the worth parameters are updated with each new observation. For example, consider a tournament with teams A, B, and C with worth parameters $f = (2, 0, -2)'$; here, f_i can be interpreted as a measure of team i's strength. If the tournament goes as expected and the order of the players is (A, B, C) – which happens with a probability of 76.3% – the score is close to zero for all players: $\nabla[f|(A, B, C)] = (0.13, 0.0019, -0.14)'$. However, if the order is reversed (an outcome occurring with a probability of a mere 0.2%), the score is $\nabla[f|(C, B, A)] = (-1.75, 0.76, 0.98)'$. For team A, which failed despite the high expectations, the score is negative; for those who beat A, the score is positive, with the largest score obtained for the unlikely winner C. The score can therefore serve as a basis for the correction of worth parameters after an observation is realised.

Figure 1 extends the previous example: it shows the score for $f = (c, 0, -c)'$ at different levels of $c > 0$ under all six orderings. For large values of c, the score appears to converge to integer values. This is no coincidence: the score is bounded by integer values. As we show in Appendix B, in the general case of N items, the score always lies in $(1 - r, 1)$ for the item with rank $r = 1, \dots, N - 1$ and in $(1 - N, 0)$ for the item with rank N.

2.3 Score-Driven Dynamics

Let us observe the rankings y_t in times $t = 1, \dots, T$. Furthermore, let us assume that individual worth parameters evolve over time and denote them $f_t = (f_{1,t}, \dots, f_{N,t})'$ for $t = 1, \dots, T$. Specifically, let the time-varying parameter $f_{i,t}$ follow the generalised autoregressive score (GAS) dynamics of Creal *et al.* (2013) and Harvey (2013) with a score order P and an autoregressive order Q. Let it also linearly depend on exogenous covariates $x_1, \dots, x_M$. The parameter $f_{i,t}$ is then given by the recursion

$$f_{i,t} = \omega_i + \sum_{j=1}^{M} \beta_j x_{i,t,j} + \sum_{k=1}^{P} \alpha_k \nabla_i \left(f_{t-k} | y_{t-k}\right) + \sum_{l=1}^{Q} \varphi_l f_{i,t-l}, \qquad i = 1, \dots, N, \quad t = 1, \dots, T, \tag{7}$$

where ω_i is item i's individual fixed effect, β_j is the regression parameter on x_j, α_k is the score parameter for lag k, φ_l is the autoregressive parameter for lag l, and $x_{i,t,j}$ is the value of x_j for item i at time t. Note that this formulation corresponds to the setting of a panel linear regression with fixed effects. In most of the GAS literature (and the GARCH and ACD literature, as a matter of fact), only the first lags are utilised, i.e., $P = Q = 1$.

In the GAS framework, the score function can be scaled by the inverse of the Fisher information or the inverse of the square root of the Fisher information, although the unit scaling is often utilised as well (see Creal *et al.*, 2013). The right choice of scaling can make estimators robust by mitigating the effect of outlying realisations of y_t on time-varying parameters. A well-known case is the Beta-t-GARCH model of Harvey and Chakravarty (2008), the GAS counterpart to Bollerslev's (1987) GARCH-t model: by applying the inverse-information scaling in Beta-t-GARCH, one obtains a model with a milder response of the variance to a large $|y_t|$ than that in GARCH-t (Harvey, 2013, Ch. 4). In the case of the Plackett-Luce distribution, analogous robustness properties are already in place with unit scaling (i.e., no scaling) thanks to the boundedness of the Plackett-Luce score. In fact, it turns out that inverse-information scaling will typically make the effect of outlying observations on the worth parameters more pronounced. Moreover, obtaining the Fisher information is computationally very intensive even for moderate N, as it involves a sum over $N!$ permutations. For these reasons, we only consider unit scaling.

Standardisation (2) cannot be enforced at each time t without deforming the dynamics given by the recursion (7). Instead, we use the standardisation

$$\sum_{i=1}^{N} \omega_i = 0. \tag{8}$$

In the case of mean-reverting dynamics with no exogenous covariates, this corresponds to a zero sum

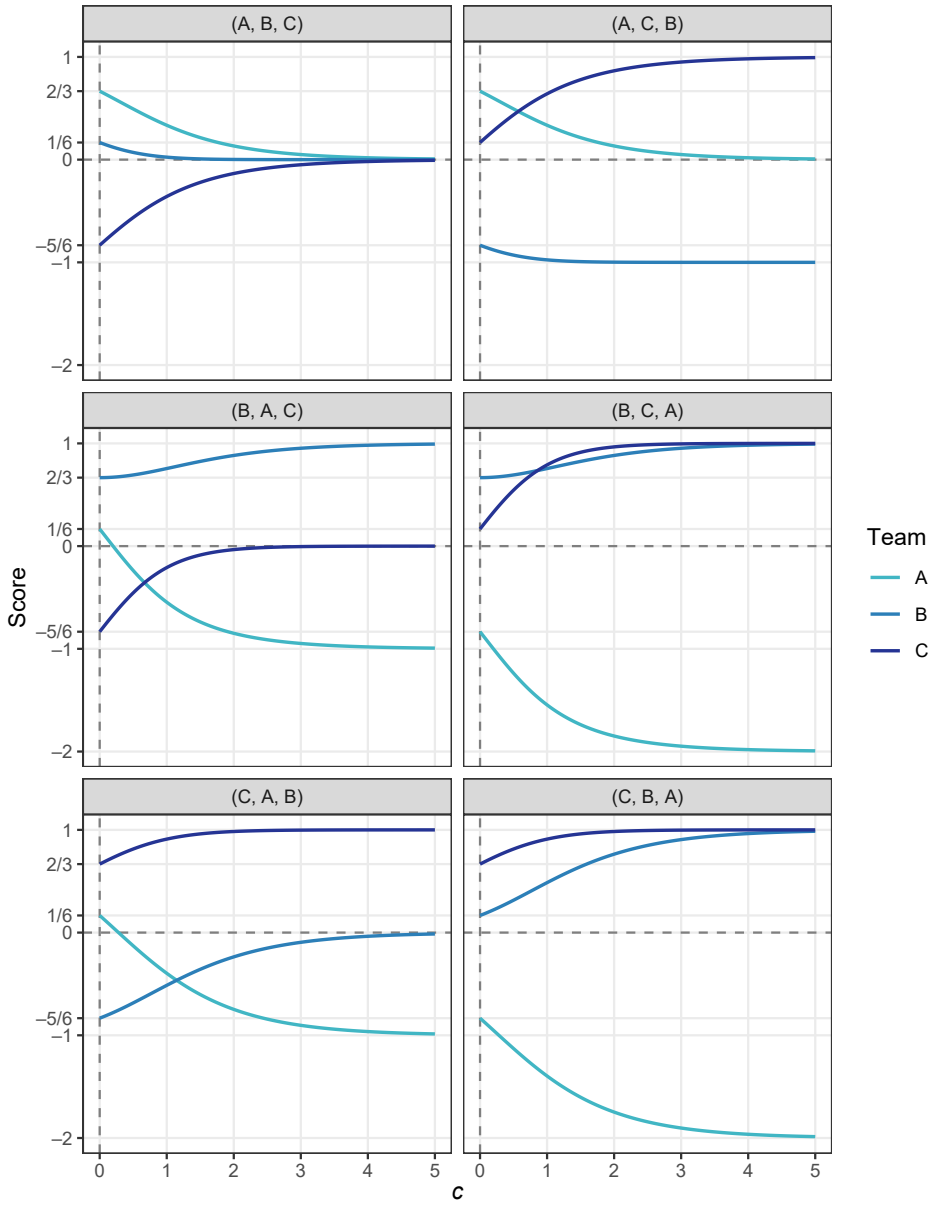

Figure 1: Scores in the Plackett-Luce distribution for three items (A, B, and C), worth parameters $f = (c, 0, -c)'$ for $c \in [0, 5]$ (horizontal axis), and all possible orderings (panel titles).

of the unconditional values:

$$\bar{f}_i = \frac{\omega_i}{1 - \sum_{l=1}^{Q} \varphi_l}, \qquad i = 1, \ldots, N. \tag{9}$$

2.4 Maximum Likelihood Estimation and Inference

For the estimation of the proposed dynamic model, we utilise the maximum likelihood estimator. Let $\theta = (\omega_1, \ldots, \omega_{N-1}, \beta_1, \ldots, \beta_M, \alpha_1, \ldots, \alpha_P, \varphi_1, \ldots, \varphi_Q)'$ be the vector of the $N + M + P + Q - 1$ parameters to be estimated, with ω_N being obtained from (8) as $\omega_N = -\sum_{i=1}^{N-1} \omega_i$. The estimate $\hat{\theta}$ is obtained from the conditional log-likelihood as

$$\hat{\theta} \in \arg\max_{\theta} \sum_{t=1}^{T} \ell\left(f_t | y_t\right). \tag{10}$$

The recursive nature of f_t requires the initialisation of the first few elements of the conditional score and worth parameter time series. A reasonable approach is to set the initial conditional scores $\nabla(f_0|y_0)$, …, $\nabla(f_{-P+1}|y_{-P+1})$ to zero, i.e., their expected value, and the initial parameters f_0, …, f_{-Q+1} to the unconditional value $\bar{f}$ given by (9). Alternatively, if additional information about the initial worth parameters is available, it can be used instead. For instance, in a related GAS-type model for the binary outcomes of tennis matches, Gorgi *et al.* (2019) use current ranking points to initialise the worth parameters. On the other hand, they also note that other initialisation methods yielded very similar parameter estimates. The initial worth parameters can also be considered as additional parameters to be estimated. This would, however, significantly increase the number of variables in the maximisation problem.

From a computational perspective, it is possible to utilise any general-purpose algorithm to solve nonlinear optimisation problems. In our simulation study and empirical application, we employ the *Broyden–Fletcher–Goldfarb–Shanno* (BFGS) algorithm. The optimisation performance can be improved, however, by exploiting the special structure of the problem. Concerning the GAS models, Creal *et al.* (2013) recall the work of Fiorentini *et al.* (1996) and suggest an algorithm that computes the gradient of the likelihood recursively and simultaneously with the time-varying parameters. Concerning the Plackett-Luce distribution, Hunter (2004) presents an iterative *minorization–maximization* (MM) algorithm, which is further developed by Caron and Doucet (2012). These ideas might prove to be a useful starting point for a specialised likelihood-maximisation algorithm tailored to our model; however, the development of such an algorithm is beyond the scope of this paper.

Our implementation of statistical inference tasks is based on standard maximum likelihood asympotics. Recall that under suitable regularity conditions, the maximum likelihood estimator $\hat{\theta}$ is consistent and asymptotically normal, i.e., it satisfies

$$\sqrt{T}\big(\hat{\theta} - \theta_0\big) \xrightarrow{\mathrm{d}} \mathrm{N}\big(0, -H^{-1}\big), \tag{11}$$

where H denotes the asymptotic Hessian of the log-likelihood, defined as

$$H = \operatorname*{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \frac{\partial^2 \ln \mathrm{P}\left[Y_t = y_t | f_t\right]}{\partial\theta_0 \partial\theta_0'}. \tag{12}$$

In finite samples, standard errors are often computed using the empirical Hessian of the log-likelihood evaluated at $\hat{\theta}$, and the normal c.d.f. is used for statistical inference.

The truth is that establishing the asymptotic theory for GAS-type models is difficult in general. At a minimum, it is necessary that the filter f_t is invertible (see, e.g., Blasques *et al.*, 2018 for more details). The invertibility property ensures, among other things, that the initialisation does not matter in the long run. The theoretical derivation of the conditions restricting the parameter space in order to obtain consistency and asymptotic normality is, however, beyond the scope of this paper. Indeed, it is very challenging in general to obtain any asymptotic results for the case of multivariate variables with multiple time-varying parameters. In the following, we base our inference on the asymptotics outlined above and rely on simulations to verify their validity.

2.5 Extension to Partial Rankings

The distribution can be extended to the case in which the ranking of only the top $\tilde{N} < N$ items is observed. The set of ranked items is then $\tilde{\mathcal{Y}} = \{y^{-1}(1), \ldots, y^{-1}(\tilde{N})\}$. We denote the partial ranking of items $i \in \tilde{\mathcal{Y}}$ as $\tilde{y}$ and the partial ordering as $\tilde{y}^{-1}$. The probability mass function of the Plackett-Luce distribution for the partial ranking $\tilde{y}$ is then

$$\mathrm{P}\left[\tilde{Y} = \tilde{y} \middle| f\right] = \prod_{r=1}^{\tilde{N}} \frac{\exp f_{r^{\text{th}}}}{\sum_{s=r}^{\tilde{N}} \exp f_{s^{\text{th}}} + \sum_{j \notin \tilde{\mathcal{Y}}} \exp f_j}. \tag{13}$$

The score function for the partial ranking $\tilde{y}$ is

$$\nabla_i \left(f|\tilde{y}\right) = \begin{cases} 1 - \sum_{r=1}^{\tilde{y}(r)} \frac{\exp f_i}{\sum_{s=r}^{\tilde{N}} \exp f_{s^{\text{th}}} + \sum_{j \notin \tilde{\mathcal{Y}}} \exp f_j} & \text{for } i \in \tilde{\mathcal{Y}}, \\ - \sum_{r=1}^{\tilde{N}} \frac{\exp f_i}{\sum_{s=r}^{\tilde{N}} \exp f_{s^{\text{th}}} + \sum_{j \notin \tilde{\mathcal{Y}}} \exp f_j} & \text{for } i \notin \tilde{\mathcal{Y}}. \end{cases} \tag{14}$$

The probability mass function and the score function for partial rankings can be straightforwardly plugged into the dynamics from previous sections.

3 Finite-Sample Performance

3.1 Simulation Design

We conduct a simulation study in order to investigate the behaviour of the maximum likelihood estimator over two dimensions – the number of items N and the time horizon T. In particular, N varies between 10, 20, and 30, and T ranges from 10 to 100. For each combination of N and T, we conduct 100,000 replications.

The simulations employ the following toy model with the parameters selected to resemble those estimated in the empirical study in Section 4. As we consider different values of N, the number of ω_i parameters differs. To somewhat standardise the item-specific fixed effects, we set $\omega_i = 4(i-1)/(N-1) - 2$, $i = 1, \ldots, N$, i.e., the parameters ω_i range from -2 to 2 for any N. We include a single exogenous covariate independently generated from the standard normal distribution. The regression parameter is then set to $\beta_1 = 1$. Finally, the order of the GAS model is chosen as $P = Q = 1$, with the dynamics parameters set to $\alpha_1 = 0.4$ and $\varphi_1 = 0.5$. Such parameter values result in the unconditional values $\bar{f}_i$ given by (9), which range from -4 to 4. In the following, we drop unnecessary subscripts for brevity, and simply refer to β_1 as β, α_1 as α, and φ_1 as φ.

3.2 Simulation Results

The results of the simulation study are reported in Figure 2. First, we investigate the accuracy of the estimators $\hat{\omega}_i$, $\hat{\beta}$, $\hat{\alpha}$, and $\hat{\varphi}$; to enhance readability, the results for $\hat{\omega}_i$ are averaged across all items (i). In the first column of Figure 2, we report the mean absolute errors (MAE) between the estimated coefficients and their true values. All estimates converge to their true values along the time dimension. The score parameter α proves to be hard to estimate for small T as it has much higher MAE than the autoregressive parameter φ with a comparable nominal value. Nevertheless, even in a medium sample with $N = 20$ and $T = 20$, the errors are not that substantial, with values of 0.22 for $\hat{\omega}_i$, 0.08 for $\hat{\beta}$, 0.12 for $\hat{\alpha}$, and 0.05 for $\hat{\varphi}$. In a large sample with $N = 30$ and $T = 100$, the errors further decrease to 0.08 for $\hat{\omega}_i$, 0.02 for $\hat{\beta}$, 0.02 for $\hat{\alpha}$, and 0.01 for $\hat{\varphi}$.

Second, we assess the usability of the maximum likelihood asymptotics for finite-sample inference. Specifically, in the second column of Figure 2, we report the fraction of the samples in which the 95 percent confidence intervals contained the true parameter values, i.e., we present the estimated coverage probabilities. For all parameters, the coverage probability converges to the target value of 0.95 from below along the time dimension. As in the case of MAE, the convergence of the coverage probability is the slowest for the score parameter α. In a medium sample with $N = 20$ and $T = 20$,

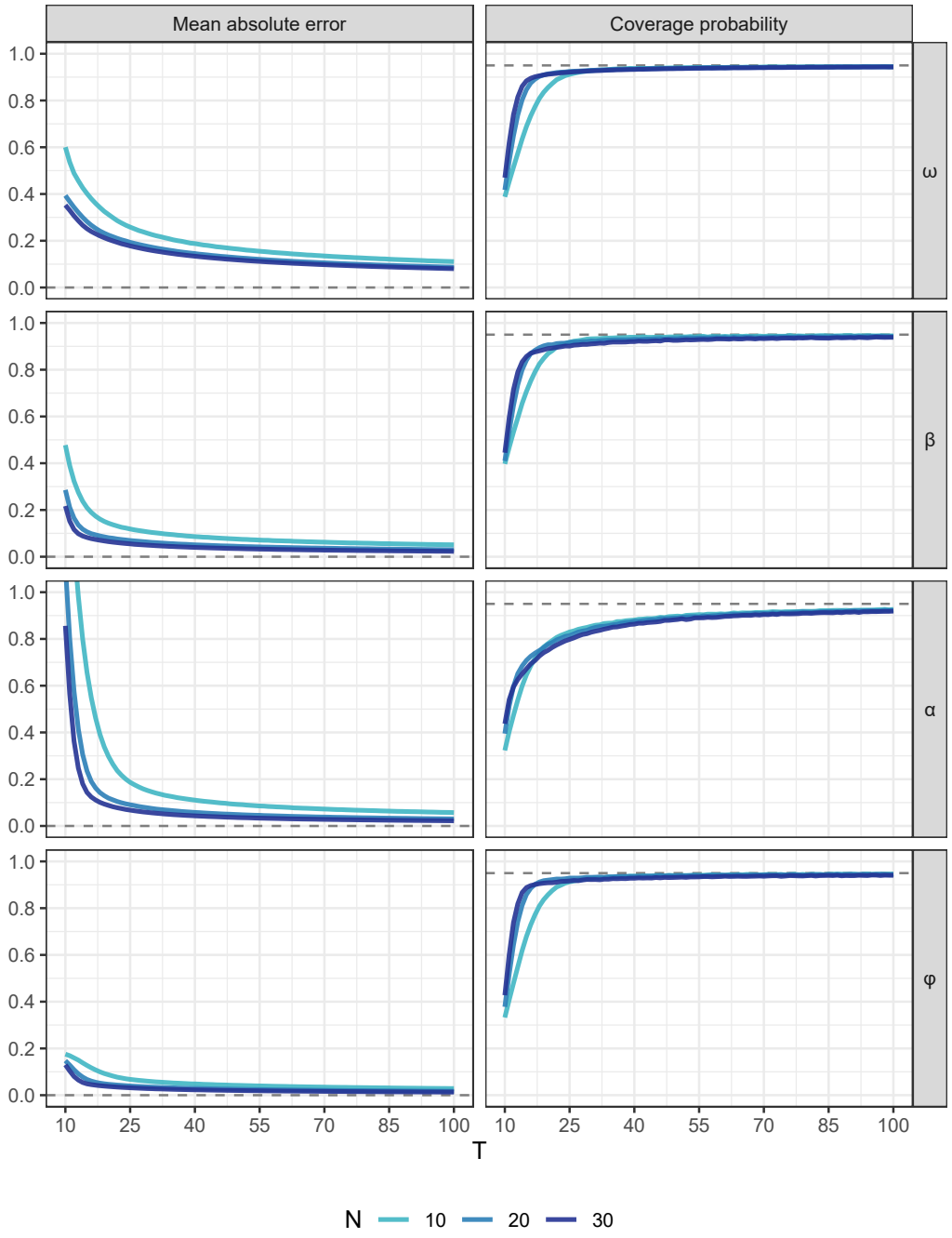

Figure 2: Mean absolute errors of the estimated coefficients and coverage probabilities of the 95% confidence intervals. For the item-specific fixed effects, $\hat{\omega}_i$, the results are averaged across all items. Dashed horizontal lines are drawn at the 0 and 0.95 vertical coordinates to show the limit values of mean absolute errors and the coverage probabilities under standard maximum-likelihood asymptotics.

the coverage probabilities are 0.91 for $\hat{\omega}_i$, 0.91 for $\hat{\beta}$, 0.78 for $\hat{\alpha}$, and 0.92 for $\hat{\varphi}$, while in a large sample with $N = 30$ and $T = 100$, they amount to 0.94 for $\hat{\omega}_i$, 0.94 for $\hat{\beta}$, 0.92 for $\hat{\alpha}$, and 0.94 for $\hat{\varphi}$.

4 Application to Ice Hockey Rankings

4.1 Data Set

We demonstrate the use of our model using data on the results of the Ice Hockey World Championships between the years 1998 and 2019. In 1998, the sanctioning body of the championships, the International Ice Hockey Federation (IIHF), increased the number of teams in the tournament from 12 to 16, and has kept the number of teams at that level since then; hence, 1998 was chosen as the starting year. For each year, the IIHF provides a complete ranking of all 16 participants. Over the years, 24 different teams made it through the qualification process, and they comprise the set of ranked items in our model.

For each year, we obtained information about the host country of the championships. In order to account for the home-ice advantage, we included a *home ice* covariate, which is a time-varying indicator variable (it is equal to 1 for home teams in the respective years, and it is equal to 0 otherwise).

4.2 Model Specification

The general structure of the team strength dynamics given by (7) includes an array of different model specifications that can be obtained by (i) choosing the order of the GAS model (P, Q) and (ii) imposing specific restrictions on the parameter space. As for the former, with the limited size of our data set, it seems impractical to consider anything beyond the canonical $P = Q = 1$ model.

Setting $P = Q = 0$, on the other hand, yields a static strength model, which is equivalent to the standard *ranked-order logit* (ROL) – a common go-to model for sports rankings. Recent applications to sport rankings include, e.g., Caron and Doucet (2012) and Henderson and Kirrane (2018). The latter use time-weighted observations to improve forecasts, but their model is intrinsically static. Both referenced studies use a Bayesian approach to the estimation of the ROL model. We estimate the **static model** to provide a benchmark for the models with score-driven dynamics.

In the model with $P = Q = 1$, we generally expect the autoregressive parameter to lie in the $(0, 1)$ interval, implying a certain degree of persistence in team strengths with a mean-reverting tendency. In our data set, this is indeed the result we obtain if we leave the parameters unrestricted in the likelihood-maximisation procedure. We refer to this variant as the **mean-reverting model**. The need for this type of a sports ranking model has recently been recognised by Baker and Mchale (2015). In their analysis of golf tournament rankings, they note that while their model's deterministic dynamics do sufficiently capture the time variation in an individual player's performance, a mean-reverting random process would be more appropriate for teams. The performance of individual players tends to follow long-term trends, potentially with breakpoints (due to injuries, the long-term evolution of self-confidence, ageing, etc.). Teams, on the other hand, do not have a fixed membership structure; players come and go on a relatively flexible basis, depending on their current performance. Massive exogenous shocks with a persistent effect are less common.

A GAS setting similar to ours has recently been used in the context of sports statistics by Koopman and Lit (2019). Rather than dealing with ranking data, Koopman and Lit focus on individual football matches, modelling either the qualitative (win-draw-loss) outcomes or match scores. Their estimates indicate high persistence levels in strength dynamics, with a typical value of (the equivalent to our) $\hat{\varphi}$ of around 0.998 for back-to-back matches in most of the estimated models. The authors note that this corresponds to a yearly persistence of about 0.90.

If strong persistence is expected, it might be reasonable to restrict the autoregressive parameter to unity, making the team strengths follow a random walk pattern. A ranking model of this type was presented by Glickman and Hennessy (2015) in the context of women's alpine downhill skiing competitions. Their model, however, does not make use of the score-driven component and is estimated in a Bayesian framework. In a GAS setting, a model with a random walk behaviour of the team

Table 1: Selected estimates for the Ice Hockey World Championships data (1998–2019).

	Mean-Reverting	Static	Random Walk
Home ice ($\hat{\beta}$)	0.227	0.171	0.099
	(0.258)	(0.262)	(0.188)
Score parameter ($\hat{\alpha}$)	0.392***		0.343***
	(0.083)		(0.058)
Autoregressive parameter ($\hat{\varphi}$)	0.506***		
	(0.149)		
log-likelihood	−611.195	−625.800	−625.425
AIC	1274.391	1299.600	1300.851

Notes: (i) Estimates of ω_i are omitted from the table to enhance readability. (ii) Standard errors in parentheses. (iii) $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

strength (henceforth, **random walk model**) should be approached with caution. The f_t filter in this case is not invertible, making the consistency of the maximum likelihood estimator dubious. That said, in our simulation experiments, the mean absolute error of the coefficient estimates was roughly comparable to the mean-reverting model of equal sample size, provided that the model specifications agree with the underlying data-generating processes. We also note that a GAS model with random walk strength dynamics has recently been presented by Gorgi *et al.* (2019) to predict the outcomes of individual tennis matches. Gorgi et al. view mean-reverting processes as the dynamics of choice for team sports and non-stationary dynamic processes as more appropriate for individual sports.

Using the general framework developed in Section 2, we can easily estimate all three model variants (the static model, mean-reverting model, and random walk model) and compare their fits using information-theoretic criteria. As only a subset of all teams participated in each championship, we employ the form of the likelihood function for partial rankings developed in Section 2.5.

The computation is performed using R package `gasmodel` for estimation, forecasting, and simulation of GAS models based on various distributions including the Plackett-Luce distribution. The package includes the analyzed Ice Hockey World Championships data and a vignette describing our modelling approach. It is available at `https://github.com/vladimirholy/gasmodel`.

4.3 Empirical Results

In the observed period, 1998–2019, only 9 teams participated in all 22 Ice Hockey World Championships. These included all the teams from the so-called Big Six (Canada, Czechia, Finland, Russia, Sweden, and the United States) along with Latvia, Slovakia, and Switzerland. Three teams – Great Britain, Poland, and South Korea – only appeared once. The dominance of the Big Six is evident when looking at the podium positions: out of the 66 medals, only six were handed out to teams outside the Big Six (four were awarded to Slovakia and two to Switzerland). Hosting was also unevenly distributed among the countries: only 14 of the teams experienced the home-ice advantage, with Czechia, Slovakia, and Switzerland hosting the championships twice and Germany, Finland, Russia, and Sweden hosting them three times each.

Table 1 presents the results for all three estimated models. In terms of the Akaike information criterion (AIC), the mean-reverting model outperformed the remaining two by a wide margin, with ΔAIC exceeding 25 in both cases. This (i) implies that the introduction of strength dynamics can improve the model fit dramatically and (ii) provides empirical support for the conjecture of Gorgi *et al.* (2019) about the suitability of mean-reverting dynamics for team sports. The 95% confidence interval for the autoregressive parameter in the mean-reverting model, $[0.21, 0.80]$, indicates the presence of moderate persistence and leads us to reject the null hypotheses of both a random-walk behaviour and no serial dependence.

Despite the differences in AIC values, for parameters that are shared across the models, the

Table 2: Unconditional team strength estimates and ultimate ranking in the mean-reverting and static models. Teams are sorted by the ultimate ranking obtained from the mean-reverting model.

	Mean-Reverting		Static	
Country	Strength	Rank	Strength	Rank
Canada	3.72	1	3.72	2
Finland	3.70	2	3.66	3
Sweden	3.65	3	3.84	1
Czechia	3.47	4	3.41	4
Russia	3.25	5	3.17	5
United States	1.83	6	2.18	6
Switzerland	1.67	7	1.76	7
Slovakia	1.65	8	1.55	8
Latvia	0.86	9	0.82	9
Germany	0.28	10	0.31	10
Belarus	0.25	11	0.11	11
Norway	0.03	12	−0.07	12
Denmark	−0.07	13	−0.17	13
France	−0.41	14	−0.51	14
Austria	−0.83	15	−0.89	15
Italy	−1.02	16	−1.10	16
Ukraine	−1.34	17	−1.52	17
Slovenia	−1.75	18	−1.64	18
Kazakhstan	−1.83	19	−1.78	19
Japan	−2.00	20	−1.94	20
Hungary	−3.28	21	−3.20	21
Great Britain	−3.92	22	−3.89	22
Poland	−3.95	23	−3.90	23
South Korea	−3.96	24	−3.91	24

estimates are qualitatively similar. In both the mean-reverting and random walk model, the values of the score coefficient, $\hat{\alpha}$, are positive and significant. This implies that the score component of our model does help in explaining the ranking dynamics. A positive sign of $\hat{\alpha}$ is in line with the interpretation of the conditional score outlined in Section 2.2: a surprising success will positively affect the team's strength estimate for the next season and vice versa.

In accordance with expectations, point estimates in all three models suggest the existence of a home-ice advantage ($\hat{\beta} > 0$), but the *home ice* is not statistically significant in either model. To assess the effect size implied by $\hat{\beta}$, we need to know the team strengths. Table 2 shows the estimates of the unconditional team strength from the mean-reverting model, which are obtained based on (9), and the $\hat{\omega}_i$ for the static model. The differences in successive strength values suggest that an increase of 0.23 (the home-ice advantage estimate in the mean-reverting model) moves a team 0–2 places ahead in the ranking.

For the mean-reverting and static models, the estimates of ω_i can be used to provide the 'ultimate' (or long-run) ranking. Both models confirm the dominance of the Big Six. Indeed, the rankings in both models agree in all but the first three places; the long-term strength estimates for these three teams are very close to one another, though, making the eventual ranking less clear cut.

Figure 3 presents the estimated values of the worth parameters $f_{i,t}$ (referred to here as the strength) in the mean-reverting model. Even though the serial dependence, given by the autoregressive parameter α, is mild, it is clearly discernible in the plots. For teams that only appeared in a handful of championships (i.e., the weaker teams, located at the bottom of the figure), we can

Table 3: One-step-ahead rank prediction and medal probabilities for the Big Six in the mean-reverting model. No home-ice advantage assumed.

Country	Strength	Predicted rank	P[gold medal]	P[podium position]
Finland	3.974	1	0.235	0.630
Canada	3.970	2	0.234	0.629
Russia	3.431	3	0.137	0.431
Czechia	3.415	4	0.134	0.426
Sweden	3.400	5	0.133	0.421
United States	2.086	6	0.036	0.128

see prolonged periods with unchanging values of the strength value, which correspond to the absent observations. Similar figures for the static and random walk models are given in Figures 4 and 5.

If the values of the explanatory variables at time $T+1$ are known, they can be plugged into (7) to obtain the values of the worth parameters at $T+1$. These can in turn serve to make one-step-ahead predictions or estimate the probabilities of specific rankings or ranking-based events. Applications in betting are straightforward. For instance, one can easily obtain the probability that a particular team will win a medal or that the podium will be occupied by a given list of teams.

An example is presented in Table 3. Assuming that none of the Big Six countries host the upcoming championships, we calculated the future value of the team strength, $f_{i,T+1}$, and the associated rank prediction for the Big Six based on our estimates of the mean-reverting model. Even though the team strengths have the mean-reverting tendency, short-run predictions can differ from the unconditional mean substantially: even though the Big Six occupy the first six places according to both the predicted rankings for $T+1$ and the ultimate rankings in Table 2, the rankings themselves are notably different.

Plugging the values of $f_{i,T+1}$ into (13) yields the estimated probability of a partial ordering of interest at $T+1$. For instance, the estimated probability of the partial ordering (Finland, Canada, Russia) – the predicted podium outcome – is 1.85 percent. This probability is low mainly because the predicted strengths happen to be quite similar across the first five teams. Analogously, we can obtain the probability of winning a gold medal, which is presented in the fourth column of Table 3. Note that the winning probabilities are markedly different despite the similar team strengths. In practical applications, one might be interested in general ranking-based events, such as the probability that a team finishes on the podium; these probabilities can easily be obtained by combining suitable elementary events. An example is given in the last column of Table 3.

5 Discussion of Other Applications

5.1 Underlying Index

There is often an underlying index or score behind a ranking. For example, the *Times Higher Education World University Rankings* are based on the score weighted over 13 individual indicators grouped into five categories – industry income, international diversity, teaching, research, and citations. International rankings based on various indices such as the *Global Competitiveness Index*, *Bloomberg Innovation Index*, *Human Development Index*, *Climate Change Performance Index*, and *Good Country Index* are compiled in a similar fashion. Naturally, an analysis of these rankings and indices is a popular subject of scientific research; e.g., Saisana *et al.* (2005) assess the robustness of country rankings, Paruolo *et al.* (2013) measure the importance of individual variables in composite indicators, and Varin *et al.* (2016) investigate the role of citation data in the ratings of scholarly journals.

The time aspect is inherent in these rankings, as they are typically compiled annually. Leckie and Goldstein (2009) highlight the need for the prediction of ratings in the context of school choice based on league tables. They model the test scores of individual students nested in schools using

Figure 3: Mean-reverting model – estimated team strength for all teams over the entire observed period. Teams are ordered by the estimated unconditional ranking.

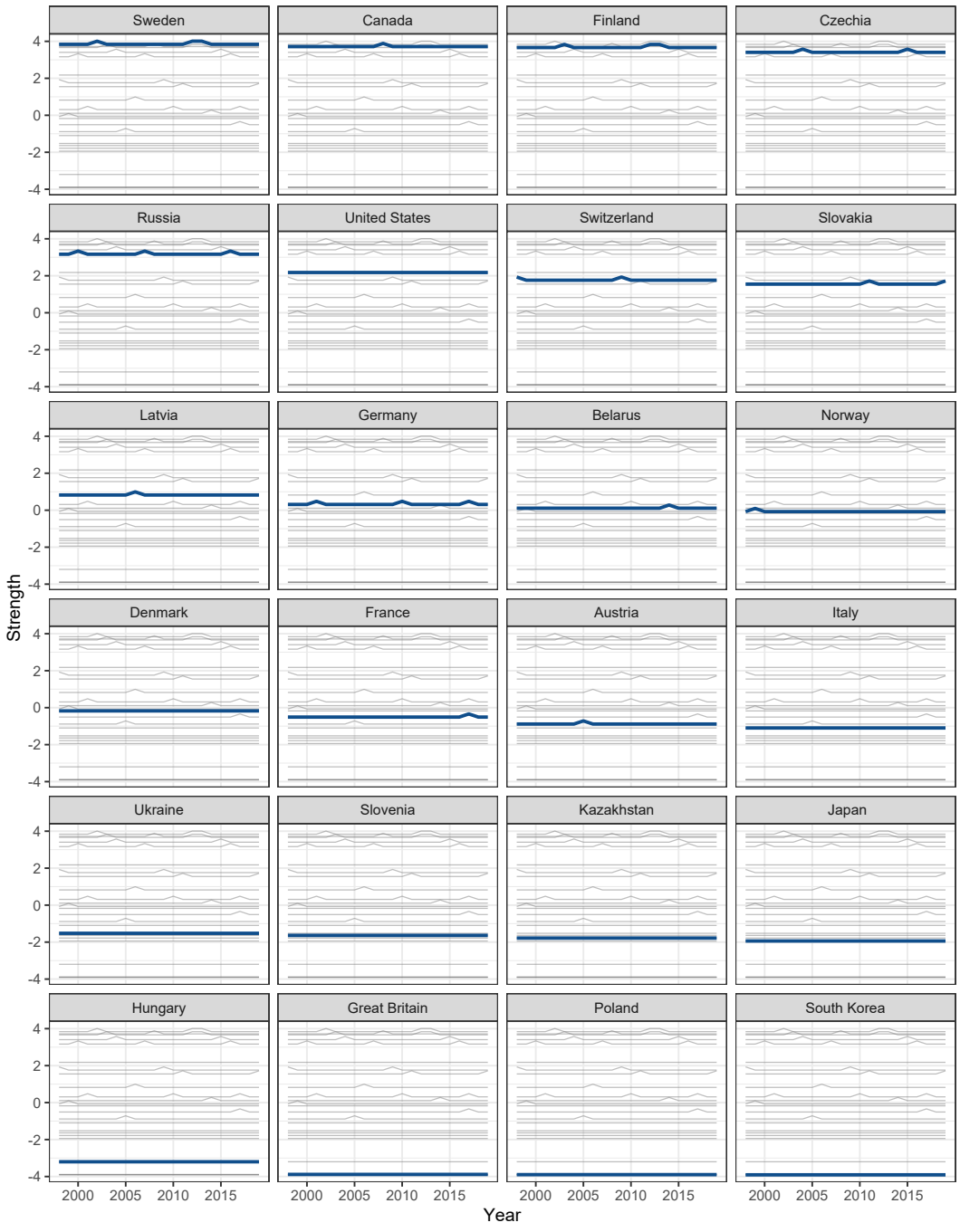

Figure 4: Static model – estimated team strength for all teams over the entire observed period. Teams are ordered by the estimated unconditional ranking. In the static model, team strengths only vary with the home-ice advantage, producing little bumps in the plots.

Figure 5: Random walk model – estimated team strength for all teams over the entire observed period. Teams are ordered by the mean strength estimate.

the multilevel random-intercepts model. As they note, the main goal is to obtain relative ratings of schools rather than changes in the mean and variance over time. The conclusion is that there is a substantial uncertainty in test scores and their ability to forecast school performance is therefore very limited. Nevertheless, it may prove to be interesting to model the rankings of schools using our proposed model.

In general, rankings can be modelled directly – by a model for permutations – or indirectly – by a model for the underlying index. If it is reasonable to assume that the indices of individual items are independent, the best option might be to model just the underlying index using a univariate model. In many real-life applications, the independence of the items' index values is questionable, as the items may interact in various ways or share a common pool of resources. When there is a potential relationship, our dynamic model for rankings might be more suitable, as it naturally captures dependence between the items. Furthermore, in the end, the reader is often only interested in the eventual rankings anyway, as they are more illustrative and attractive than the underlying indices.

5.2 Repeated Surveys

A common way of obtaining ranking data is through a survey in which respondents are asked to rank items. Many surveys are repeated on several occasions, forming time-series data. For the statistical methodology dealing with repeated surveys, see Scott and Smith (1974) and Steel and McLaren (2009). By asking ranking questions in repeated surveys, we arrive at a time series of rankings. For example, customers of a retail shop may be periodically asked to rank products according to their preferences. In this case, the proposed dynamic ranking model could be a useful tool.

5.3 Non-Parametric Efficiency Analysis

Another interesting application is modelling the rankings of decision making units (DMUs) obtained by non-parametric efficiency analysis, such as the *data envelopment analysis (DEA)* pioneered by Charnes *et al.* (1978) and Banker *et al.* (1984). Typically, DEA is applied in fields such as banking, health care, agriculture, transportation, and education to analyse the performance of banks, hospitals, farms, airlines, and schools, respectively (Liu *et al.*, 2013). The goal of such analyses is to separate efficient and inefficient DMUs, assign efficiency scores to them, and determine their ranking. Many empirical papers also study the determinants of efficiency. The analysis is usually carried out by first obtaining the efficiency scores and then analysing them using regression in the second phase. Simar and Wilson (2007) (i) point out that a vast majority of these analyses ignore the inherent dependence between efficiency scores in their second phase, and (ii) develop bootstrap procedures to fix invalid inference.

Simar and Wilson (2007) focus on the cross-sectional case in which dependence only occurs between the DMUs, not over time, which also greatly facilitates bootstrapping. The extension to panel data is not straightforward to say the least. Nevertheless, in many empirical studies, DMUs are observed annually, with the intention of both assessing the way efficiency evolved over time and providing a list of units that proved to be capable of sustaining efficiency over a long period. For this type of analysis, it may be beneficial to model the dynamics of DEA rankings using our model. If the long-term efficiency is of interest, it can be measured via the unconditional ranking. A major limitation of this approach is, however, the use of the Plackett-Luce distribution, as DEA rankings do not obey Luce's choice axiom. Other, more complex distributions on rankings could prove more appropriate here. For example, a richer dependence structure can be provided by Thurstone order statistics models based on the multivariate normal distribution (see Thurstone, 1927 and Yu, 2000) or multivariate extreme value distributions (see McFadden, 1978 and Joe, 2001). Note that the latter class contains the Plackett-Luce model as a special case.

Modelling rankings instead of efficiency scores may also enhance the robustness with regard to method selection. For example, a novel DEA approach utilising the Chebyshev distance proposed by Hladík (2019) offers alternative efficiency scores to the classical DEA models of Charnes *et al.*

(1978) and Banker *et al.* (1984), but it has been shown to produce the exact same ranking. Modelling rankings instead of efficiency scores thus eliminates differences between the two methods.

6 Conclusion

Our new modelling approach brings two main features that have not been utilised in the analysis of time-varying rankings so far: (i) it allows a general autoregressive scheme for the process that governs the items' worth parameters, and (ii) new observations can update the worth parameters through a score-driven mechanism. Both of these features proved useful in our case study dealing with ice hockey team rankings. We believe that empiricists in diverse application areas can benefit from these features as well. These empiricists will hopefully also appreciate other practical merits of the model, such as the ability to include time-varying covariates or the straightforward maximum likelihood estimation.

This paper has presented the first results of ongoing research. Future efforts should mainly cover the following areas. First, we hope to see more complex results regarding both finite-sample performance and limit behavior. We doubt that comprehensive analytical treatment of the maximum likelihood asymptotics is tractable, but we aim to extend the current simulation results substantially. Second, for applications with a very large number of items, empiricists would surely benefit from a specialised algorithm for likelihood maximisation that exploits the specific structure of the likelihood function. As we mentioned above, Creal *et al.* (2013) and Caron and Doucet (2012) might provide useful inspiration in this respect. Finally, for applications to rankings where ties are possible, the model can be extended using the approach of Firth *et al.* (2019) and Turner *et al.* (2020).

Acknowledgements

We would like to thank Michal Černý for his comments. Computational resources were supplied by the project 'e-Infrastruktura CZ' (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

Funding

This research was supported by the Internal Grant Agency of Prague University of Economics and Business under project F4/27/2020 and the Czech Science Foundation under project 19-08985S.

References

Alvo M, Yu PLH (2014). *Statistical Methods for Ranking Data*. Springer, New York. ISBN 978-1-4939-1470-8. https://doi.org/10.1007/978-1-4939-1471-5.

Asfaw D, Vitelli V, Sørensen Ø, Arjas E, Frigessi A (2017). "Time-Varying Rankings with the Bayesian Mallows Model." *Stat*, **6**(1), 14–30. ISSN 2049-1573. https://doi.org/10.1002/sta4.132.

Baker RD, Mchale IG (2015). "Deterministic Evolution of Strength in Multiple Comparisons Models: Who Is the Greatest Golfer?" *Scandinavian Journal of Statistics*, **42**(1), 180–196. ISSN 0303-6898. https://doi.org/10.1111/sjos.12101.

Banker RD, Charnes A, Cooper WW (1984). "Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis." *Management Science*, **30**(9), 1078–1092. ISSN 0025-1909. https://doi.org/10.1287/mnsc.30.9.1078.

Blasques F, Gorgi P, Koopman SJ, Wintenberger O (2018). "Feasible Invertibility Conditions and Maximum Likelihood Estimation for Observation-Driven Models." *Electronic Journal of Statistics*, **12**(1), 1019–1052. ISSN 1935-7524. https://doi.org/10.1214/18-ejs1416.

Bollerslev T (1986). "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics*, **31**(3), 307–327. ISSN 0304-4076. https://doi.org/10.1016/0304-4076(86)90063-1.

Bollerslev T (1987). "A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return." *Review of Economics and Statistics*, **69**(3), 542–547. ISSN 0034-6535. https://doi.org/10.2307/1925546.

Caron F, Doucet A (2012). "Efficient Bayesian Inference for Generalized Bradley-Terry Models." *Journal of Computational and Graphical Statistics*, **21**(1), 174–196. ISSN 1061-8600. https://doi.org/10.1080/10618600.2012.638220.

Charnes A, Cooper WW, Rhodes E (1978). "Measuring the Efficiency of Decision Making Units." *European Journal of Operational Research*, **2**(6), 429–444. ISSN 0377-2217. https://doi.org/10.1016/0377-2217(78)90138-8.

Creal D, Koopman SJ, Lucas A (2013). "Generalized Autoregressive Score Models with Applications." *Journal of Applied Econometrics*, **28**(5), 777–795. ISSN 0883-7252. https://doi.org/10.1002/jae.1279.

Critchlow DE, Fligner MA, Verducci JS (1991). "Probability Models on Rankings." *Journal of Mathematical Psychology*, **35**(3), 294–318. ISSN 0022-2496. https://doi.org/10.1016/0022-2496(91)90050-4.

Davis RA, Dunsmuir WTM, Street SB (2003). "Observation-Driven Models for Poisson Counts." *Biometrika*, **90**(4), 777–790. ISSN 0006-3444. https://doi.org/10.1093/biomet/90.4.777.

Engle RF, Russell JR (1998). "Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data." *Econometrica*, **66**(5), 1127–1162. ISSN 0012-9682. https://doi.org/10.2307/2999632.

Fiorentini G, Calzolari G, Panattoni L (1996). "Analytic Derivatives and the Computation of GARCH Estimates." *Journal of Applied Econometrics*, **11**(4), 399–417. ISSN 0883-7252. https://doi.org/10.2307/2284932.

Firth D, Kosmidis I, Turner HL (2019). "Davidson-Luce Model for Multi-Item Choice with Ties." https://arxiv.org/abs/1909.07123.

Glickman ME, Hennessy J (2015). "A Stochastic Rank Ordered Logit Model for Rating Multi-Competitor Games and Sports." *Journal of Quantitative Analysis in Sports*, **11**(3), 131–144. ISSN 2194-6388. https://doi.org/10.1515/jqas-2015-0012.

Gorgi P, Koopman SJ, Lit R (2019). "The Analysis and Forecasting of Tennis Matches by Using a High Dimensional Dynamic Model." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **182**(4), 1393–1409. ISSN 0964-1998. https://doi.org/10.1111/rssa.12464.

Harvey AC (2013). *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series*. First Edition. Cambridge University Press, New York. ISBN 978-1-107-63002-4. https://doi.org/10.1017/cbo9781139540933.

Harvey AC, Chakravarty T (2008). "Beta-t-(E)GARCH." https://doi.org/10.17863/cam.5286.

Henderson DA, Kirrane LJ (2018). "A Comparison of Truncated and Time-Weighted Plackett-Luce Models for Probabilistic Forecasting of Formula One Results." *Bayesian Analysis*, **13**(2), 335–358. ISSN 1936-0975. https://doi.org/10.1214/17-ba1048.

Hladík M (2019). "Universal Efficiency Scores in Data Envelopment Analysis Based on a Robust Approach." *Expert Systems with Applications*, **122**, 242–252. ISSN 0957-4174. https://doi.org/10.1016/j.eswa.2019.01.019.

Hunter DR (2004). "MM Algorithms for Generalized Bradley-Terry Models." *The Annals of Statistics*, **32**(1), 384–406. ISSN 0090-5364. https://doi.org/10.1214/aos/1079120141.

Joe H (2001). "Multivariate Extreme Value Distributions and Coverage of Ranking Probabilities." *Journal of Mathematical Psychology*, **45**(1), 180–188. ISSN 0022-2496. https://doi.org/10.1006/jmps.1999.1294.

Koopman SJ, Lit R (2019). "Forecasting Football Match Results in National League Competitions Using Score-Driven Time Series Models." *International Journal of Forecasting*, **35**(2), 797–809. ISSN 0169-2070. https://doi.org/10.1016/j.ijforecast.2018.10.011.

Koopman SJ, Lucas A, Scharth M (2016). "Predicting Time-Varying Parameters with Parameter-Driven and Observation-Driven Models." *Review of Economics and Statistics*, **98**(1), 97–110. ISSN 0034-6535. https://doi.org/10.1162/rest_a_00533.

Leckie G, Goldstein H (2009). "The Limitations of Using School League Tables to Inform School Choice." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **172**(4), 835–851. ISSN 0964-1998. https://doi.org/10.1111/j.1467-985x.2009.00597.x.

Liu JS, Lu LYY, Lu WM, Lin BJY (2013). "A Survey of DEA Applications." *Omega*, **41**(5), 893–902. ISSN 0305-0483. https://doi.org/10.1016/j.omega.2012.11.004.

Luce RD (1959). *Individual Choice Behavior: A Theoretical Analysis*. First Edition. Wiley, New York. ISBN 978-0-486-44136-8. https://books.google.com/books/about/Individual{_}choice{_}behavior.html?id=a80DAQAAIAAJ.

Luce RD (1977). "The Choice Axiom after Twenty Years." *Journal of Mathematical Psychology*, **15**(3), 215–233. ISSN 0022-2496. https://doi.org/10.1016/0022-2496(77)90032-3.

Luo J, Qin H (2019). "A Note on Ranking in the Plackett-Luce Model for Multiple Comparisons." *Acta Mathematicae Applicatae Sinica*, **35**(4), 885–892. ISSN 0168-9673. https://doi.org/10.1007/s10255-019-0857-z.

Marden JI (1995). *Analyzing and Modeling Rank Data*. Chapman and Hall / CRC Press, New York. ISBN 978-0-429-19249-4. https://doi.org/10.1201/b16552.

McFadden D (1978). "Modeling the Choice of Residential Location." In A Karlqvist, F Snickars, J Weibull (Eds.), *Spatial Interaction Theory and Planning Models*, 75–96. North-Holland, Amsterdam. https://econpapers.repec.org/paper/cwlcwldpp/477.htm.

Paruolo P, Saisana M, Saltelli A (2013). "Ratings and Rankings: Voodoo or Science?" *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **176**(3), 609–634. ISSN 0964-1998. https://doi.org/10.1111/j.1467-985X.2012.01059.x.

Plackett RL (1975). "The Analysis of Permutations." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **24**(2), 193–202. ISSN 0035-9254. https://doi.org/10.2307/2346567.

Saisana M, Saltelli A, Tarantola S (2005). "Uncertainty and Sensitivity Analysis Techniques as Tools for the Quality Assessment of Composite Indicators." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **168**(2), 307–323. ISSN 0964-1998. https://doi.org/10.1111/j.1467-985x.2005.00350.x.

Scott AJ, Smith TM (1974). "Analysis of Repeated Surveys Using Time Series Methods." *Journal of the American Statistical Association*, **69**(347), 674–678. ISSN 0162-1459. https://doi.org/10.1080/01621459.1974.10480187.

Simar L, Wilson PW (2007). "Estimation and Inference in Two-Stage, Semi-Parametric Models of Production Processes." *Journal of Econometrics*, **136**(1), 31–64. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2005.07.009.

Steel D, McLaren C (2009). "Design and Analysis of Surveys Repeated over Time." In *Handbook of Statistics*, Volume 29, Chapter 33, 289–313. Elsevier, Amsterdam. ISBN 978-0-444-53438-5. https://doi.org/10.1016/S0169-7161(09)00233-8.

Stern H (1990). "Models for Distributions on Permutations." *Journal of the American Statistical Association*, **85**(410), 558–564. ISSN 0162-1459. https://doi.org/10.1080/01621459.1990.10476235.

Thurstone LL (1927). "A Law of Comparative Judgment." *Psychological Review*, **34**(4), 273–286. ISSN 0033-295X. https://doi.org/10.1037/h0070288.

Turner HL, van Etten J, Firth D, Kosmidis I (2020). "Modelling Rankings in R: The PlackettLuce Package." *Computational Statistics*, **35**, 1027–1057. ISSN 0943-4062. https://doi.org/10.1007/s00180-020-00959-3.

Varin C, Cattelan M, Firth D (2016). "Statistical Modelling of Citation Exchange Between Statistics Journals." *Journal of the Royal Statistical Society: Series A (General)*, **179**(1), 1–63. ISSN 0035-9238. https://doi.org/10.1111/rssa.12124.

Yellott JI (1977). "The Relationship Between Luce's Choice Axiom, Thurstone's Theory of Comparative Judgment, and the Double Exponential Distribution." *Journal of Mathematical Psychology*, **15**(2), 109–144. ISSN 0022-2496. https://doi.org/10.1016/0022-2496(77)90026-8.

Yu PLH (2000). "Bayesian Analysis of Order-Statistics Models for Ranking Data." *Psychometrika*, **65**(3), 281–299. ISSN 0033-3123. https://doi.org/10.1007/bf02296147.

Yu PLH, Gu J, Xu H (2019). "Analysis of Ranking Data." *Wiley Interdisciplinary Reviews: Computational Statistics*, **11**(6), e1483:1–e1483:26. ISSN 1939-5108. https://doi.org/10.1002/wics.1483.

A Placket-Luce Probabilities, Log-Likelihood, and Score with Three Items

We present an example of equations (1), (3), and (5) for the case of three items. The probability mass function is given by

$$\mathrm{P}\left[Y = y|f\right] = \frac{\exp f_{1\mathrm{st}}}{\exp f_{1\mathrm{st}} + \exp f_{2\mathrm{nd}} + \exp f_{3\mathrm{rd}}} \cdot \frac{\exp f_{2\mathrm{nd}}}{\exp f_{2\mathrm{nd}} + \exp f_{3\mathrm{rd}}}. \tag{15}$$

The log-likelihood function is given by

$$\ell\left(f|y\right) = f_{1\mathrm{st}} + f_{2\mathrm{nd}} - \ln\left(\exp f_{1\mathrm{st}} + \exp f_{2\mathrm{nd}} + \exp f_{3\mathrm{rd}}\right) - \ln\left(\exp f_{2\mathrm{nd}} + \exp f_{3\mathrm{rd}}\right). \tag{16}$$

The score function is given by

$$\begin{aligned}
\nabla_{1\mathrm{st}}\left(f|y\right) &= 1 - \frac{\exp f_{1\mathrm{st}}}{\exp f_{1\mathrm{st}} + \exp f_{2\mathrm{nd}} + \exp f_{3\mathrm{rd}}}, \\
\nabla_{2\mathrm{nd}}\left(f|y\right) &= 1 - \frac{\exp f_{2\mathrm{nd}}}{\exp f_{1\mathrm{st}} + \exp f_{2\mathrm{nd}} + \exp f_{3\mathrm{rd}}} - \frac{\exp f_{2\mathrm{nd}}}{\exp f_{2\mathrm{nd}} + \exp f_{3\mathrm{rd}}}, \\
\nabla_{3\mathrm{rd}}\left(f|y\right) &= -\frac{\exp f_{3\mathrm{rd}}}{\exp f_{1\mathrm{st}} + \exp f_{2\mathrm{nd}} + \exp f_{3\mathrm{rd}}} - \frac{\exp f_{3\mathrm{rd}}}{\exp f_{2\mathrm{nd}} + \exp f_{3\mathrm{rd}}}.
\end{aligned} \tag{17}$$

It is obvious that $\nabla_{1\mathrm{st}}(f|y) + \nabla_{2\mathrm{nd}}(f|y) + \nabla_{3\mathrm{rd}}(f|y) = 0$, as fractions in (17) with the same denominator sum to one. Moreover, it is easily seen that $\nabla_{1\mathrm{st}} \in (0, 1)$, $\nabla_{2\mathrm{nd}} \in (-1, 1)$, and $\nabla_{3\mathrm{rd}} \in (-2, 0)$, as each fraction has a value between 0 and 1.

B Deriving Properties of the Plackett-Luce Model via the Softmax and Logsumexp Functions

The analysis of the Plackett-Luce model is facilitated by the use of the softmax and logsumexp functions, which are denoted here as $\sigma(\cdot)$ and $\mathrm{lse}(\cdot)$, respectively. In this appendix, we first employ these functions to easily derive the shape of the likelihood and score, and then use them to study the properties of the score.

Recall that for an n-vector z, $\sigma(z)_i = \exp(z_i)/\sum_{j=1}^n \exp z_j$ and $\mathrm{lse}(z) = \log \sum_{j=1}^n \exp z_j$. It is easily verified that

$$\log \sigma(z)_i = z_i - \mathrm{lse}(z), \tag{18}$$

$$\frac{\partial \,\mathrm{lse}(z)}{\partial z_i} = \sigma(z)_i \,. \tag{19}$$

To simplify formulas, we introduce the shorthand notation $f_{\geq r}$ for a vector containing worth parameters of items ranked r^{th} or worse, i.e., $f_{\geq r} = (f_i)_{\{i \in \mathcal{Y} : y^{-1}(i) \geq r\}}$. With this notation, we can rewrite (1) as

$$\mathrm{P}\left[Y = y | f\right] = \prod_{r=1}^{N} \sigma(f_{\geq r})_r. \tag{20}$$

Combining (18) and (20) immediately yields

$$\ell\left(f | y\right) = \sum_{i=1}^{N} f_i - \sum_{r=1}^{N} \mathrm{lse}(f_{\geq r}), \tag{21}$$

and using (19), we obtain the score for player i in the form

$$\nabla_i \left(f | y\right) = 1 - \sum_{r=1}^{y(i)} \sigma(f_{\geq r})_i \,. \tag{22}$$

Since (i) values of the softmax function lie in the $(0, 1)$ interval and (ii) $\sigma(f_{\geq N}) = 1$, we can easily establish the following bounds for the score. The score lies in $(1 - r, 1)$ for an item with rank $r = 1, \ldots, N-1$, and in $(1 - N, 0)$ for the item ranked last (N^{th}). The bounds are tight, as the following example with a dominant item demonstrates. Consider the identity ranking $y(i) = i$ and worth parameters

$$f_i = \begin{cases} c & \text{if } i = d, \\ \frac{-c}{N-1} & \text{otherwise,} \end{cases}$$

where $c > 0$ and $d \in \mathcal{Y}$ is a dominant item. (It is easily verified that $\sum_{i=1}^N f_i = 0$.) For $r \leq \min(i, d)$ we obtain

$$\lim_{c \to \infty} \sigma\big(f_{\geq r}\big)_i = \begin{cases} 1 & \text{if } i = d, \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

Combining (22) and (23) yields $\lim_{c\to\infty} \nabla_d \left(f | y\right) = 1 - d$, which demonstrates that the lower bound for the score is tight. Setting $d = N$ (the dominant item unexpectedly ranks last) yields $\lim_{c\to\infty} \nabla_i \left(f | y\right) = 1$ for $i = 1, \ldots, N-1$, which demonstrates the tightness of the upper bound of the score for all but the last item. (The tightness of the upper bound for the last item can be shown using a similar example with an inferior item that ranks last, as expected.)

Ranking-Based Second Stage in Data Envelopment Analysis: An Application to Research Efficiency in Higher Education

Vladimír Holý
Prague University of Economics and Business
Winston Churchill Square 1938/4, 130 67 Prague 3, Czechia
vladimir.holy@vse.cz

Abstract: An alternative approach for the panel second stage of data envelopment analysis (DEA) is presented in this paper. Instead of efficiency scores, we propose to model rankings in the second stage using a dynamic ranking model in the score-driven framework. We argue that this approach is suitable to complement traditional panel regression as a robustness check. To demonstrate the proposed approach, we determine research efficiency in the higher education sector by examining scientific publications and analyze its relation to good governance. The proposed approach confirms positive relation to the Voice and Accountability indicator, as found by the standard panel linear regression, while suggesting caution regarding the Government Effectiveness indicator.

Keywords: Two-Stage DEA, Ranking Model, Score-Driven Model, Research & Development.

JEL Classification: C33, C44, I23, O32.

1 Introduction

In operations research, data envelopment analysis (DEA) is a non-parametric method used to measure the relative efficiency of decision-making units (DMUs) that convert inputs into outputs. It compares DMUs by calculating their efficiency scores based on a set of inputs and outputs. The method has been widely applied in the fields of agriculture, education, energy, finance, government, healthcare, manufacturing, retail, sport, and transportation.

In DEA research, it is common to follow the efficiency measurement with a second-stage regression analysis that uses efficiency scores as the dependent variable and includes contextual (or environmental) variables as independent variables. This approach is known as two-stage DEA. In many cases, efficiency is assessed annually, which may require a panel regression as the second-stage model to account for time-varying contextual variables. The most frequently employed panel methods for the second stage are panel linear regression (see, e.g., Chen *et al.*, 2019; Mamatzakis *et al.*, 2013) and panel Tobit regression (see, e.g., Borozan, 2018; Fonchamnyo and Sama, 2016). In linear regression, log transformations of efficiency scores are often used (see, e.g., Poveda, 2011; Zhang *et al.*, 2018). Other panel methods include panel quantile regression (see, e.g., Frýd and Sokol, 2021; Zhang *et al.*, 2018), panel fractional regression (see, e.g., Da Silva e Souza and Gomes, 2015; Fonchamnyo and Sama, 2016), and panel beta regression (see, e.g., Pirani *et al.*, 2018; Song *et al.*, 2016).

The standard two-stage DEA has been subject to criticism by Simar and Wilson (2007), Simar and Wilson (2011), and Kneip *et al.* (2015). The criticisms mainly stem from three issues: (1) correlation among the estimated efficiency scores due to the complex structure of the data generating process, (2) the use of estimated efficiency scores as dependent variable instead of the true unobserved efficiency scores, and (3) the potential inseparability between the frontier production and the impact of contextual variables. These issues can significantly affect the validity of inference. When dealing with repeated assessement of efficiency, there is also the issue of temporal dependence. Nevertheless, some authors such as Banker and Natarajan (2008), McDonald (2009), and Banker *et al.* (2019) argue for the use of linear regression. For a survey on statistical approaches in nonparametric frontier models, see Moradi-Motlagh and Emrouznejad (2022).

In this paper, we present an alternative approach for the panel second stage of DEA. Instead of modeling efficiency scores, we propose to model the rankings. In the recent literature, Holý and Zouhar (2022) developed a time series model for rankings that utilize the Plackett–Luce distribution

and incorporates autoregressive and score dynamics. This model is based on the modern framework of score-driven models introduced by Creal *et al.* (2013) and Harvey (2013). While Holý and Zouhar (2022) applied the model to the results of the Ice Hockey World Championships, they also suggested its potential use in the second stage of DEA. Following this call, we devote this paper to exploring the use of this dynamic ranking model in DEA.

The motivation for using the score-driven dynamic ranking model in the second stage of DEA arises from the following properties:

- *Relevance of Rankings.* Rankings preserve the important information of mutual comparison among DMUs. In certain scenarios, the primary objective of DEA may even be to obtain rankings of DMUs, in which case modeling rankings directly is more appropriate. The long-term behavior of DMUs may also be of interest, in which case the long-term ranking may have a clearer interpretation than an aggregate of efficiency scores.

- *Robustness to DEA Model.* Consider two DEA models: the super-efficiency DEA model of Andersen and Petersen (1993) and the universal DEA model of Hladík (2019), both with either constant returns to scale (CRS) of Charnes *et al.* (1978) or variable returns to scale (VRS) of Banker *et al.* (1984). Despite producing different efficiency scores, these models generate the exact same ranking. By modeling rankings instead of efficiency scores in the second stage, any differences between these models are eliminated. An additional consideration when modeling efficiency scores is whether to use the logarithmic transformation. However, since the log transformation preserves rankings, this is not a concern when using a ranking model.

- *Robustness to Outliers.* Outliers, in the form of extreme values of efficiency scores, can significantly influence the coefficients in a second-stage regression model. However, using rankings can mitigate this issue, as a DMU with an extremely low or high efficiency score would simply be ranked last or first, respectively. Thus, a ranking model can effectively handle such outliers.

- *Simple yet Powerful.* The model of Holý and Zouhar (2022) is straightforward to work with. The Plackett–Luce distribution, unlike its alternatives, is available in a closed form (see Alvo and Yu, 2014) and the dynamics are observation-driven (see Cox, 1981). As a result, the model can be estimated using the maximum likelihood method, and conventional Hessian-based standard errors can be used. Moreover, the model only requires a modest number of parameters, consisting of individual effects of DMUs, regression coefficients common for all DMUs, and two additional parameters controling dynamics common for all DMUs.

Our approach also faces the following limitations:

- *Loss of Information.* While using rankings instead of efficiency scores can provide robustness to DEA model and outliers (as discussed above), it also leads to loss of information. This loss can be beneficial in some scenarios, but it is still important to recognize that it occurs. One drawback of using rankings alone is that it is not possible to determine the boundary between inefficient and efficient DMUs. Efficiency scores, on the hand, provide a clear distinction between the two groups.

- *Different Data Generating Process.* Our approach does not address the criticism of Simar and Wilson (2007), Simar and Wilson (2011), and Kneip *et al.* (2015). Indeed, the dependence between the DMUs is not captured by the Plackett–Luce distribution, which assumes the property known as the independence of irrelevant alternatives. The data generating process assumed by the model of Holý and Zouhar (2022) is much simpler then the true one generated by DEA.

- *Absence of Ties.* The model of Holý and Zouhar (2022) has a limitation in that it does not allow for rankings with ties. This means that in the second stage, we need to use a suitable DEA model that can rank all DMUs, including the efficient ones. However, this can be addressed by extending the Plackett-Luce distribution to incorporate ties, as demonstrated by Turner *et al.* (2020).

- *Sufficient Variation in Rankings.* A single realization of efficiency scores is often used in a second stage regression model. A single ranking is, however, not enough for a meaningful analysis. Repeated rankings are therefore needed, which naturally take the form of panel data. Our approach is therefore suitable only when the time dimension is present. Even with repeated rankings, however, the Plackett–Luce distribution requires that for any possible partition of DMUs into two non-empty subsets, there exists at least one DMU in the second subset that is ranked higher than at least one DMU in the first subset (see Hunter, 2004).

Our approach is fundamentally different from traditional panel regressions, but it is not intended to replace them. Particularly when it suffers from the same shortcomings highlighted by Simar and Wilson (2007), Simar and Wilson (2011), and Kneip *et al.* (2015). Instead, our approach is best used as a complement to traditional panel regressions to provide valuable insights that are not burdened by the problems specific to efficiency scores. This can be viewed as a form of robustness check, where both approaches are used to provide a more complete picture of the data. Given the controversies surrounding the second stage DEA, conducting extensive robustness checks is crucial for ensuring the reliability and validity of the results. DEA practitioners who wish to utilize the dynamic ranking model can do so easily using the `gasmodel` R package, which offers all the necessary tools for estimation, forecasting, and simulation.

As an illustration of the proposed approach, we explore the research efficiency in higher education of European Union (EU) countries through the analysis of scientific publications in 2005–2020. In the first stage, we perform DEA analysis for each year independently. We use gross domestic expenditure on R&D and the number of researchers as inputs to reflect the financial and human resources, respectively. For outputs, we use the number of publications and the number of citations to reflect the quantity and quality of scientific research, respectively. In the second stage, we investigate the influence of good governance on the research efficiency. As contextual variables, we use the six Worldwide Governance Indicators (WGI) of Kaufmann *et al.* (2011), together with the gross domestic product (GDP). We perform panel linear regression analysis of efficiency scores obtained by three DEA models proposed by Charnes *et al.* (1978), Andersen and Petersen (1993), and Hladík (2019), along with the dynamic ranking model of Holý and Zouhar (2022). All models uncover that the Voice and Accountability indicator is significantly positively correlated with research efficiency suggesting that participation in selecting the government, freedom of expression, freedom of association, and freedom of media are key factors of governance influencing research efficiency. The Government Effectiveness indicator has also positive effect, however, its significance is not confirmed by all models and this result is therefore not robust. No other significant relations are found. By utilizing the proposed approach in this study, we are able to assess the robustness of the relationship to the Voice and Accountability indicator. However, the results also indicate caution in interpreting the findings related to the Government Effectiveness indicator. Therefore, conducting extensive robustness checks such as this one is important to increase the reliability of the analysis and prevent misleading conclusions.

The rest of the paper is structured as follows. In Section 2, we present three DEA models proposed by Charnes *et al.* (1978), Andersen and Petersen (1993), and, Hladík (2019), which are utilized in the subsequent analysis. In Section 3, we present details on the dynamic ranking model of Holý and Zouhar (2022) and its estimation, along with some modifications suitable to our case. In Section 4, we conduct an empirical study to examine research efficiency in higher education and compare the proposed ranking approach with the traditional panel regression approach. We conclude the paper in Section 5.

2 First Stage: Measuring Efficiency

The first stage of DEA involves determining the relative efficiency scores of the DMUs. The number of DMUs is denoted by N. Each DMU transforms I inputs into J outputs. Let x_{ni} denote the i-th input of the n-th DMU, and y_{nj} denote the j-th output of the n-th DMU. The matrix of inputs is denoted by $X = (x_{ni})_{n=1,i=1}^{N,I}$, while the matrix of outputs is denoted by $Y = (y_{nj})_{n=1,j=1}^{N,J}$. The inputs of a single DMU n are denoted by $x_n = (x_{n1}, \dots, x_{nI})^\intercal$, and the outputs of a DMU n are denoted

by $y_n = (y_{n1}, \ldots, y_{nJ})^\intercal$. The notation X_{-n} represents the inputs of every DMU but n, while Y_{-n} represents the outputs of every DMU but n.

2.1 Basic DEA

Charnes *et al.* (1978) proposed the very first DEA model, which has since become one of the most widely used DEA models to date. This model is commonly referred to as the CCR model and is based on the assumption of constant returns to scale (CRS). The efficiency scores θ_n^{CCR} are found for each DMU n by the following linear program:

$$\begin{aligned} \theta_n^{CCR} = \max_{u,v} \ & y_n^\intercal u \\ \text{subject to} \quad & x_n^\intercal v \leq 1, \\ & Yu - Xv \leq 0, \\ & u \geq 0, \\ & v \geq 0, \end{aligned} \tag{1}$$

where u and v are vectors of weights for the outputs and inputs respectively. The efficiency scores for inefficient DMUs lie in $[0, 1)$ and are equal to 1 for inefficient DMUs.

2.2 Super-Efficiency DEA

A shortcoming of the CCR model is that it cannot differentiate between efficient DMUs, which can lead to the loss of valuable information. Andersen and Petersen (1993) proposed a super-efficiency DEA to overcome this limitation. In this model, the DMU under evaluation is excluded from the set of benchmarks, which allows efficient DMUs to achieve score greater than 1. The super-efficiency model with CRS (labeled as the AP model) is given by the following linear program:

$$\begin{aligned} \theta_n^{AP} = \max_{u,v} \ & y_n^\intercal u \\ \text{subject to} \quad & x_n^\intercal v \leq 1, \\ & Y_{-n}u - X_{-n}v \leq 0, \\ & u \geq 0, \\ & v \geq 0. \end{aligned} \tag{2}$$

The efficiency scores for inefficient DMUs are the same as those obtained from the CCR model, while the scores for efficient DMUs are greater than or equal to 1.

2.3 Universal DEA

Recently, Hladík (2019) proposed a DEA formulation that focuses on a robust optimization viewpoint. The model uses a scaled Chebyshev norm to measure efficiency as a distance to inefficiency and inefficiency as a distance to efficiency. The scores generated by this model are universal in the sense that they are naturally normalized, and therefore, can be compared across unrelated models. The universal DEA model with CRS (labeled as the H model) is given by the following linear program:

$$\begin{aligned} \theta_n^{H} = \max_{\delta,u,v} \ & 1 + \delta \\ \text{subject to} \quad & y_n^\intercal u \geq 1 + \delta, \\ & x_n^\intercal v \leq 1 - \delta, \\ & Y_{-n}u - X_{-n}v \leq 0, \\ & u \geq 0, \\ & v \geq 0. \end{aligned} \tag{3}$$

Note that Hladík (2019) also proposed a nonlinear DEA model based on the Chebyshev norm, to which (3) is a tight approximation. The efficiency scores for inefficient DMUs lie in $[0, 1)$, while the scores for efficient DMUs lie in $[1, 2]$.

The universal DEA model is closely related to the super-efficiency DEA model of Andersen and Petersen (1993). Hladík (2019) showed that the ranking of DMUs according to θ_n^{AP} is the same as the ranking according to θ_n^H. However, the models are even more connected as the efficient scores themselves can be derived by the following transformations:

$$\theta_n^H = \frac{2\theta_n^{AP}}{1 + \theta_n^{AP}}, \qquad \theta_n^{AP} = \frac{\theta_n^H}{2 - \theta_n^H}. \tag{4}$$

Applications of the universal DEA model include Holý and Šafr (2018), Frýd and Sokol (2021), and Holý (2022).

3 Second Stage: Modeling Dynamic Rankings

The second stage of DEA involves identifying the factors that affect efficiency scores and measure their impact. We assume periodic evaluation of efficiency of the same set of DMUs at times $t = 1, \dots, T$ with efficiency scores $\theta_t = (\theta_{1t}, \dots, \theta_{Nt})^\intercal$. In this paper, we propose to model rankings of DMUs, instead of their efficiency scores as is usual in the second-stage DEA. Let $R_t(n)$ denote the rank of a DMU n according to efficiency scores θ_t at time t. The complete ranking at time t is then denoted by $R_t = (R_t(1), \dots, R_t(N))^\intercal$. The inverse of this ranking is the ordering $O_t = (O_t(1), \dots, O_t(N))^\intercal$ at time t, where $O_t(r)$ represents the DMU with rank r at time t. We employ the dynamic ranking model of Holý and Zouhar (2022).

3.1 Plackett–Luce Distribution

We assume that at each time t the ranking R_t follows the Plackett–Luce distribution proposed by Luce (1959) and Plackett (1975). In the ranking literature, it is a widely used probability distribution for random variables in the form of permutations. Each DMU n at each time t has a worth parameter $w_{nt} \in \mathbb{R}$ reflecting its rank at time t. The probability of a higher rank increases with a higher worth parameter value. Specifically, the probability mass function is given by

$$f\left(R_t | w_t\right) = \prod_{r=1}^{N} \frac{\exp\left(w_{O_t(r)t}\right)}{\sum_{s=r}^{N} \exp\left(w_{O_t(s)t}\right)}. \tag{5}$$

In other words, a ranking is iteratively constructed by selecting the best DMU, followed by the second best, the third best, and so on. At each stage, the probability of selecting a particular DMU is proportional to the exponential of its worth parameter divided by the sum of the exponentials of the worth parameters of all DMUs that have not been selected yet. The log-likelihood function is given by

$$\ell\left(w_t | R_t\right) = \sum_{n=1}^{N} w_{nt} - \sum_{r=1}^{N} \ln\left(\sum_{s=r}^{N} \exp w_{O_t(s)t}\right). \tag{6}$$

The score (i.e. the gradient of the log-likelihood function) is given by

$$\nabla_n\left(w_t | R_t\right) = 1 - \sum_{r=1}^{R_t(n)} \frac{\exp\left(w_{nt}\right)}{\sum_{s=r}^{N} \exp\left(w_{O_t(s)t}\right)}, \qquad n = 1, \dots, N. \tag{7}$$

The Plackett–Luce distribution is based on the Luce's choice axiom, which states that the probability of selecting one item over another from a set of items is not influenced by the presence or absence of other items in the set (see Luce, 1977). This property of choice is known as the independence of irrelevant alternatives. Clearly, this property is not met in the case of DEA as addition or removal of DMUs from the set can influence efficiency scores and even ranking of other DMUs. As in the case of

many second-stage models, the proposed dynamic ranking model therefore does not conform to the complex data generating process of DEA efficiency scores and rankings. Nevertheless, the proposed model can be a useful tool due to its simplicity when applied with caution.

3.2 Regression and Dynamics

We let the worth parameters linearly depend on K contextual variables and also include an autoregressive and score-driven component. The worth parameters are then given by the recursion

$$w_{nt} = \omega_n + \sum_{k=1}^{K} \beta_k z_{nkt} + e_{nt}, \quad e_{nt} = \varphi e_{nt-1} + \alpha \nabla_n \left(w_{t-1} | R_{t-1} \right), \quad n = 1, \ldots, N, \quad t = 1, \ldots, T, \tag{8}$$

where ω_n are the individual effects for each DMU n, β_k are the regression parameters for the contextual variables z_{nkt}, φ is the autoregressive parameter, and α is the score parameter for the lagged score $\nabla_n \left(w_{t-1} | R_{t-1} \right)$ given by (7). The model corresponds to panel regression with fixed effects and dynamic error term. Note that the model is overparametrized as the probability mass function (5) is invariant to the addition of a constant to all worth parameters. We therefore use standardization

$$\sum_{n=1}^{N} \omega_n = 0. \tag{9}$$

Our specification differs from the model of Holý and Zouhar (2022) by introducing the separate e_{nt} component. Our specification is inspired by the regression with ARMA errors, while the specification of Holý and Zouhar (2022) resemble the ARMAX model. In our specification, the contextual variables influence only concurrent ranking, which is easier to interpret. Our model is also easier for numerical estimation as ω_n and φ are disconnected.

The e_{nt} component captures dynamic effects by the autoregressive term and the lagged score. The model therefore belongs to the class of score-driven models, also known as generalized autoregressive score (GAS) models or dynamic conditional score (DCS) models, proposed by Creal *et al.* (2013) and Harvey (2013). The score can be interpreted as a measure of the fit of the Plackett–Luce model to the observed rankings. A positive score indicates that a DMU n is ranked higher than what its worth parameter w_{nt} suggests, while a negative score suggests that it is ranked lower. A score of zero indicates that the DMU is ranked as expected according to its worth parameter. Thus, the score can be used as a correction term for the worth parameter after the ranking is observed.

3.3 Maximum Likelihood Estimation

The model is observation-driven and can be estimated by the maximum likelihood method. Let $\theta = (\omega_1, \ldots, \omega_{N-1}, \beta_1, \ldots, \beta_K, \varphi, \alpha)'$ denote the vector of the $N + K + 1$ parameters to be estimated. Note that ω_N is obtained from (9) as $\omega_N = -\sum_{n=1}^{N-1} \omega_n$. The maximum likelihood estimate $\hat{\theta}$ is then given by

$$\hat{\theta} \in \arg\max_{\theta} \sum_{t=1}^{T} \ell \left(w_t | R_t \right), \tag{10}$$

where the log-likelihood $\ell \left(w_t | R_t \right)$ is given by (6) and w_t follow (8). The problem (10) can be numerically solved by any general-purpose algorithm for nonlinear optimization. Furthermore, the standard errors of the estimated parameters are computed using the empirical Hessian of the log-likelihood evaluated at $\hat{\theta}$.

In order for the log-likelihood to have a unique maximum, it is necessary that for any possible partition of DMUs into two non-empty subsets, there exists at least one DMU in the second subset that is ranked higher than at least one DMU in the first subset (see Hunter, 2004). This condition ensures that no DMU is always ranked first, which would result in an infinite worth parameter and violate the assumptions of maximum likelihood estimation.

4 Empirical Study

Our empirical study aims to analyze research efficiency in the higher education sector by examining scientific publications on a country-level basis, with a particular focus on the EU countries between 2005 and 2020. Specifically, we seek to determine whether certain aspects of good governance have a positive impact on research efficiency.

4.1 Relevant Studies

Assessing the efficiency of research and development (R&D) is a widely studied topic in the data envelopment analysis (DEA) literature. In Table 1, we present a list of several relevant DEA papers and the key specifics of each study. We focus on the assessment of countries (and regions), although similar analyses can be performed at more detailed levels of institutions (see, e.g., Jablonsky, 2016) and projects (see, e.g., Lee *et al.*, 2009). Typically, studies on R&D efficiency use financial resources and human resources as the two main inputs. In terms of outputs, some studies focus on variables related to scientific publications (such as Hung *et al.*, 2009), some on patents (such as Cullmann *et al.*, 2012), while the majority consider both types of R&D-related outcomes.

4.2 Input, Output, and Contextual Variables

As inputs, we use the following variables:

- *R&D Expenditure* refers to the gross domestic expenditure to R&D activities performed in the higher education sector. The unit is million purchasing power standards. Holý and Šafr (2018) emphasize the importance of accounting for purchasing power parity when adjusting prices to ensure meaningful comparisons between countries with varying purchasing power. This variable reflects the financial resources.
- *Number of Researchers* refers to the total number of researchers employed in the higher education sector. The unit is full-time equivalent. This variable reflects the human resources.

As outputs, we use the following variables:

- *Number of Publications* represents the number of articles, reviews, and conference papers published. This variable reflects the quantity of scientific research.
- *Number of Citations* represents the number of citations to the published articles, reviews, and conference papers. This variable reflects the quality of scientific research.

As contextual variables, we use the six Worldwide Governance Indicators (WGI), which Kaufmann *et al.* (2011) define in the following way:

- *Voice and Accountability* captures perceptions of the extent to which a country's citizens are able to participate in selecting their government, as well as freedom of expression, freedom of association, and a free media.
- *Political Stability and Absence of Violence/Terrorism* captures perceptions of the likelihood that the government will be destabilized or overthrown by unconstitutional or violent means, including politically-motivated violence and terrorism.
- *Government Effectiveness* captures perceptions of the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies.
- *Regulatory Quality* captures perceptions of the ability of the government to formulate and implement sound policies and regulations that permit and promote private sector development.

Table 1: An overview of relevant studies.

Paper:	Aristovnik (2012)
Sample:	37 countries
Inputs:	R&D expenditure, Researchers
Outputs:	Articles, Patent applications, High-technology exports
Paper:	Chen *et al.* (2011)
Sample:	24 countries
Inputs:	R&D expenditure stocks, R&D personnel
Outputs:	Journal articles, Patent applications, Royalty and licensing fees
Paper:	Cullmann *et al.* (2012)
Sample:	28 countries
Inputs:	Detailed R&D expenditure, Researchers
Outputs:	Weighted and unweighted patents
Paper:	Ekinci and Karadayi (2017)
Sample:	28 EU countries
Inputs:	Detailed R&D expenditure, R&D personnel, Employment
Outputs:	Patents granted, Publications
Paper:	Han *et al.* (2016)
Sample:	15 Korean regions
Inputs:	R&D expenditure
Outputs:	Patent applications, Publications
Paper:	Hung *et al.* (2009)
Sample:	27 countries
Inputs:	R&D expenditure, Researchers
Outputs:	Article share, Citation share
Paper:	Holý and Šafr (2018)
Sample:	28 EU countries
Inputs:	R&D expenditure, Scientist and engineers
Outputs:	Citations, Patent applications
Paper:	Lee and Park (2005)
Sample:	27 countries
Inputs:	R&D expenditure, Researchers
Outputs:	Patents, Articles, Technology balance of receipts
Paper:	Roman (2010)
Sample:	14 regions of Bulgaria and Romania
Inputs:	R&D expenditure, R&D personnel
Outputs:	Patents
Paper:	Sharma and Thomas (2008)
Sample:	22 countries
Inputs:	R&D expenditure, Researchers
Outputs:	Patents granted, Publications
Paper:	Thomas *et al.* (2011)
Sample:	50 US states and the District of Columbia
Inputs:	R&D expenditure
Outputs:	Patents granted, Publications

- *Rule of Law* captures perceptions of the extent to which agents have confidence in and abide by the rules of society, and in particular the quality of contract enforcement, property rights, the police, and the courts, as well as the likelihood of crime and violence.
- *Control of Corruption* captures perceptions of the extent to which public power is exercised for private gain, including both petty and grand forms of corruption, as well as "capture" of the state by elites and private interests.

Finally, we also include the following variable as a contextual variable:

- *Gross Domestic Product* is used to control for the economic development of a country. To filter out the trend, we use the percentage of EU total GDP per capita based on million purchasing power standards.

We therefore have $I = 2$ input variables, $J = 2$ output variables, and $K = 7$ contextual variables. Similarly to Holý and Šafr (2018), we lag the input and contextual variables by one year, recognizing that there is typically a delay between the input variables and the corresponding output variables.

4.3 Data Sample

Our data sample contains all $N = 27$ countries of EU. The outputs are taken from 2005 to 2020, while the inputs and contextual variables are taken with a one-year lag from 2004 to 2019. We therefore have $T = 16$ time periods to analyze. The source of the R&D expenditure, the number of researchers, and the GDP is Eurostat[1]. There were 4 missing observations for the number of researchers of Greece in 2004, 2008, 2009, and 2010. We have interpolated these values using linear regression. The source of the number of documents and the number of citations is Scimago Journal & Country Rank[2]. The source of the Worldwide Governance Indicators is the World Bank[3].

4.4 Suitability of DEA Model

In order for efficiency scores to be interpretable, several criteria need to be met. We have adopted the best practices in DEA as outlined by Dyson *et al.* (2001) and Cook *et al.* (2014). We begin by establishing that the process under evaluation is well-defined. Our focus is on the research output in the form of scientific publications. The two chosen output variables encompass both the quantity and quality of scientific publications. While quantity is naturally quantifiable, measuring quality can be achieved using several metrics such as the number of citations and the h-index. However, combining indices and volume measures can pose difficulties and we have therefore decided to use the number of citations for our analysis. The two primary resources for conducting research are funding and personnel, both of which are represented by the two input variables we have selected. All input and output variables are volume measures and are isotonic (i.e. increased input reduces efficiency, while increased output increases efficiency). With a total of 4 input and output variables and 27 DMUs, our DEA model possesses sufficient discriminatory power.

Next, we examine the homogeneity assumption. Our set of DMUs encompasses all EU countries as of February 2020, although it should be noted that EU membership changed during the period under observation. Specifically, Romania and Bulgaria joined in 2007, and Croatia became a member in July 2013, whereas the United Kingdom departed in January 2020. Nevertheless, EU countries should be considered homogeneous in terms of research due to the harmonized policies and frameworks implemented by the European Commission, such as the European Research Area (ERA) and the Horizon Europe program. These initiatives aim to promote collaboration and standardization among EU member states, facilitating the dissemination of research findings and enhancing the overall quality of scientific output.

Finally, we analyze the appropriate returns to scale. Note that EU countries exhibit considerable variation in size, with Germany being the most populous country at 83.17 million people and Malta

[1] https://ec.europa.eu/eurostat/data/database
[2] https://www.scimagojr.com
[3] https://info.worldbank.org/governance/wgi

being the least populous with a population of 0.51 million as of January 2020. Our focus is on the higher education sector, which is composed of (1) universities, colleges of technology, and other institutions providing formal tertiary education programmes, (2) research institutes, centres, experimental stations and clinics that have their R&D activities under the direct control of, or administered by, tertiary education institutions (see OECD, 2015). The scientific output of a country can be seen as the sum of outputs from these individual institutions. As a result, we assume that country size does not have a significant impact on the relative scientific output and employ the constant returns to scale (CRS) assumption in our analysis.

4.5 Efficiency Scores

Table 2 reports descriptive statistics of efficiency scores and ranks. Bulgaria consistently shows high levels of efficiency across most years, which can be primarily attributed to its extremely low R&D spending, both in absolute value and relative to the number of publications, citations and even researchers. Romania is also found to be efficient in many years due to their relatively low R&D spending. Cyprus has an average of 3.10 publications per researcher, the highest among all countries, followed by Slovenia with 2.57. Moving to Western Europe, the Netherlands stands out as the country with the highest number of citations per researcher, with an average of 81.49. The final country that is ever found efficient in our sample is Luxembourg. Germany, as the largest country, dominate in absolute values of all inputs and outputs; its efficiency is, however, average. At the other end of the efficiency spectrum, we find Latvia with 0.66 publications and 9.09 citations per researcher on average, and Lithuania with 0.61 publications and 8.82 citations per researcher on average.

4.6 Independence of Irrelevant Alternatives

As discussed in Sections 1 and 3.1, DEA rankings do not adhere to the independence of irrelevant alternatives (IIA) assumption of the Plackett–Luce distribution. This means, among other things, that if a DMU is removed from the set, the ranking of the remaining DMUs can change. The question is to what extent the IIA assumption is violated in real data.

We conduct a simple experiment. We remove a single DMU from the set, compute DEA efficiency, and compare the resulting ranking with the original ranking based on the full set of DMUs. We repeat this process for all DMUs and time periods. Thus, we obtain a total of $N \cdot T = 432$ DEA rankings. In total, 87 percent of the rankings remain unchanged after removing a single DMU. The correlation coefficient between the rankings based on the reduced sets and the full set is 0.9954. Naturally, the ranking is more likely to change when DMUs with higher efficiency scores are removed. In our case, removing countries such as Bulgaria, Cyprus, Luxembourg, Netherlands, Romania, and Slovenia – all of which hold high ranks according to Table 2 – results in changes to the ranking across multiple time periods. In summary, we confirm the violation of the IIA assumption in our empirical study. Nevertheless, the extent of this violation is rather mild, as evidenced by the relatively high correlation between rankings.

4.7 Long-Term Ranking

When conducting an analysis over multiple time periods, it can be beneficial to report the long-term behavior. This could be done by simple aggregate statistics, as we did in Table 2. But it is also a perfect task for our dynamic ranking model. For this purpose, we estimate the model without any contextual variables, only in the form of a stationary time series model. We can then rank DMUs according to the unconditional values of the worth parameters, which are simply equal to ω_n. This long-term or "ultimate" ranking is visualized in Figure 1.

4.8 Panel Regression and Ranking Model

We proceed to the second stage where we find relation between the efficiency scores or their associated rankings and the contextual variables. For the efficiency scores, we employ standard panel

Table 2: Descriptive statistics of efficiency scores and ranks obtained from the universal DEA model of Hladík (2019).

	H Efficiency Score			Rank		
Country	Min	Med	Max	Min	Med	Max
Austria	0.54	0.70	0.85	11	16.0	20
Belgium	0.50	0.75	0.87	8	15.0	22
Bulgaria	0.99	1.21	1.47	1	1.5	3
Croatia	0.56	0.75	0.99	5	9.0	22
Cyprus	0.73	1.17	1.32	1	2.0	14
Czechia	0.57	0.75	0.94	7	12.5	20
Denmark	0.52	0.78	0.97	4	12.0	19
Estonia	0.42	0.62	0.72	7	22.0	25
Finland	0.48	0.63	0.71	16	20.5	22
France	0.42	0.64	0.71	15	21.0	26
Germany	0.46	0.68	0.86	7	18.0	23
Greece	0.56	0.65	0.94	7	15.0	20
Hungary	0.57	0.73	0.80	5	13.5	19
Ireland	0.54	0.78	0.95	7	10.0	21
Italy	0.61	0.77	0.96	6	10.0	14
Latvia	0.25	0.47	0.64	12	25.5	27
Lithuania	0.31	0.40	0.48	25	27.0	27
Luxembourg	0.57	0.87	1.31	1	7.0	12
Malta	0.53	0.72	0.90	5	15.0	23
Netherlands	0.67	1.00	1.17	1	4.0	8
Poland	0.36	0.58	0.66	18	23.0	27
Portugal	0.44	0.50	0.66	19	24.5	26
Romania	0.77	1.00	1.21	2	4.5	13
Slovakia	0.43	0.61	0.80	11	19.5	26
Slovenia	0.84	1.00	1.16	2	4.0	6
Spain	0.52	0.58	0.74	14	20.0	25
Sweden	0.61	0.82	0.89	5	9.0	13

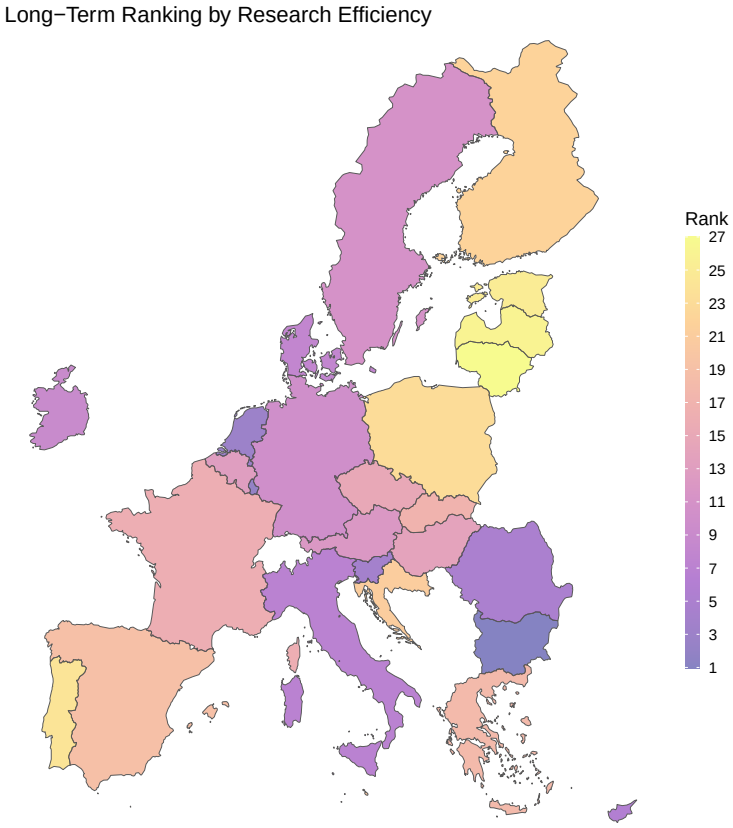

Figure 1: The long-term ranking according to the dynamic ranking model.

linear regression model with the robust estimation of the standard errors by the White method. As dependent variable, we use the efficiency scores obtained by the basic DEA model of Charnes *et al.* (1978) (denoted as CCR), the super-efficiency model of Andersen and Petersen (1993) (denoted as AP), and the universal DEA model of Hladík (2019) (denoted as H). We also use the log transform of the AP efficiency scores, which are equal to the logit transform of H efficiency scores,

$$\theta_{nt}^{Log} = \ln\left(\theta_{nt}^{AP}\right) = -\ln\left(\frac{2}{\theta_{nt}^{H}} - 1\right). \tag{11}$$

Furthermore, we use the AP efficiency scores, or equivalently the H efficiency scores, to derive rankings of the DMUs, which serve as the dependent variable in our dynamic ranking model.

The results of the estimated models are reported in Table 3. All panel linear regression models exhibit consistent signs of coefficients. They also all find the Voice and Accountability indicator to be statistically significant at the 0.05 level. Furthemore, the Government Effectiveness indicator is significant according to all panel regression models but AP. All the remaining contextual variables are found insignificant by all models. The dynamic ranking model confirms the positive and significant relation to the Voice and Accountability indicator, which is consistent with the results of all panel regression models. However, regarding the Government Effectiveness indicator, the model agrees with AP and finds it to be insignificant. The Political Stability and GDP variables have opposite signs, in contrast to the panel regression models, but remain insignificant. It is important to note that while the signs and significance of coefficients can be compared between the panel regression models and the dynamic ranking model, the estimated values cannot be directly compared due to differences in the model specifications. The coefficients φ and α, which control the dynamics, have both positive values, as expected. The estimated value of 0.86 for φ suggests that the process is stationary, but with high persistence over time.

The inference of the ranking model is derived using the empirical Hessian (denoted as Hess. in Table 3). To ensure the robustness of our findings, we additionally employ the parametric bootstrap technique to compute standard errors and p-values (denoted as Boot. in Table 3). Our bootstrap procedure is based on 1 000 simulated samples. According to Table 3, the estimated standard errors across the two methods are quite similar. An exception can be seen in the case of the Rule of Law variable; the bootstrapped standard deviation is noticeably lower here. Despite this discrepancy, the coefficient for this variable remains statistically insignificant at typical significance levels. The p-values exhibit similar behavior, and the significance of the variables remains unchanged; only the Voice and Accountability variable achieves significance at a lower level. Collectively, these findings affirm the validity of inference based on the empirical Hessian within our finite sample.

4.9 Computing Environment

The empirical study was performed in R. The CCR and AP DEA efficiency scores were obtained using the `dea()` and `sdea()` functions from the `Benchmarking` package. The H efficiency scores were obtained from the AP DEA efficiency scores using transformation (4). The panel regressions were estimated using the `plm()` function from the `plm` package with robust inference obtained using the `coeftest()` function from the `lmtest` package. The dynamic ranking model was estimated by the `gas()` and `gas_bootstrap()` functions from the `gasmodel` package. All these packages are available on CRAN.

4.10 Discussion of Results

The results of our analysis show that the Voice and Accountability indicator has a consistently positive and significant correlation with research efficiency across all models. This indicates that factors such as participation in selecting the government, freedom of expression, freedom of association, and freedom of media, which form the Voice and Accountability indicator, play a crucial role in enhancing research efficiency. In contrast, the Government Effectiveness indicator also has a positive effect on research efficiency, but its significance is not confirmed by all models. This suggests that while Government Effectiveness can enhance research efficiency, it may not be as crucial as the Voice and

Table 3: The estimated coefficients with standard errors for panel linear regressions and the dynamic ranking model.

	Efficiency Score				Rank	
	CCR	AP	H	Log	Hess.	Boot.
Voice and Accountability	0.21^{*}	0.45^{*}	0.24^{*}	0.52^{*}	3.61^{**}	3.61^{***}
	(0.09)	(0.20)	(0.10)	(0.21)	(1.21)	(1.21)
Political Stability	0.04	0.10	0.05	0.11	−0.60	−0.60
	(0.07)	(0.09)	(0.06)	(0.16)	(0.56)	(0.61)
Government Effectiveness	0.23^{**}	0.14	0.17^{*}	0.39^{*}	0.65	0.65
	(0.08)	(0.14)	(0.08)	(0.19)	(0.63)	(0.68)
Regulatory Quality	−0.09	−0.15	−0.08	−0.19	−1.21	−1.21
	(0.10)	(0.15)	(0.09)	(0.20)	(0.69)	(0.74)
Rule of Law	−0.03	−0.02	−0.01	0.02	0.02	0.02
	(0.14)	(0.19)	(0.12)	(0.27)	(0.92)	(0.25)
Control of Corruption	−0.08	−0.14	−0.08	−0.20	−0.75	−0.75
	(0.11)	(0.19)	(0.11)	(0.24)	(0.75)	(0.74)
Gross Domestic Product	−0.02	−0.23	−0.09	−0.19	1.08	1.08
	(0.25)	(0.28)	(0.19)	(0.42)	(1.59)	(1.64)
Autoregressive Parameter φ					0.86^{***}	0.86^{***}
					(0.06)	(0.10)
Score Parameter α					0.96^{***}	0.96^{***}
					(0.08)	(0.10)

Note: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Accountability indicator and lacks robustness. The findings of this study can inform policy decisions and strategic planning to enhance research performance and impact, ultimately advancing knowledge and innovation in various fields.

5 Conclusion

This paper has illustrated the usefulness of incorporating the dynamic ranking model of Holý and Zouhar (2022) in the second stage of DEA with an application to evaluating research efficiency in the higher education sector. The primary objective of the model is to serve as a complement to conventional second-stage models and provide a robustness check. While the dynamic ranking model may not be a perfect solution for all situations, it can still be a valuable addition to the DEA researcher's toolkit.

Future research efforts should be directed towards expanding the dynamic ranking model in two ways. Firstly, the model should be able to incorporate ties, which may occur due to DEA models lacking super-efficiency. Secondly, the model should be able to capture more complex interdependencies between DMUs, which can be perhaps achieved by employing Thurstone order statistics models based on the multivariate normal or multivariate extreme value distributions.

Acknowledgements

I would like Jan Zouhar for his comments. Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

Funding

The work on this paper was supported by the Czech Science Foundation under project 23-06139S and the personal and professional development support program of the Faculty of Informatics and Statistics, Prague University of Economics and Business.

References

Alvo M, Yu PLH (2014). *Statistical Methods for Ranking Data*. Springer, New York. ISBN 978-1-4939-1470-8. https://doi.org/10.1007/978-1-4939-1471-5.

Andersen P, Petersen NC (1993). "A Procedure for Ranking Efficient Units in Data Envelopment Analysis." *Management Science*, **39**(10), 1261–1264. ISSN 0025-1909. https://doi.org/10.2307/2632964.

Aristovnik A (2012). "The Relative Efficiency of Education and R&D Expenditures in the New EU Member States." *Journal of Business Economics and Management*, **13**(5), 832–848. ISSN 1611-1699. https://doi.org/10.3846/16111699.2011.620167.

Banker R, Natarajan R, Zhang D (2019). "Two-Stage Estimation of the Impact of Contextual Variables in Stochastic Frontier Production Function Models Using Data Envelopment Analysis: Second Stage OLS Versus Bootstrap Approaches." *European Journal of Operational Research*, **278**(2), 368–384. ISSN 0377-2217. https://doi.org/10.1016/j.ejor.2018.10.050.

Banker RD, Charnes A, Cooper WW (1984). "Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis." *Management Science*, **30**(9), 1078–1092. ISSN 0025-1909. https://doi.org/10.1287/mnsc.30.9.1078.

Banker RD, Natarajan R (2008). "Evaluating Contextual Variables Affecting Productivity Using Data Envelopment Analysis." *Operations Research*, **56**(1), 48–58. ISSN 0030-364X. https://doi.org/10.1287/opre.1070.0460.

Borozan D (2018). "Technical and Total Factor Energy Efficiency of European Regions: A Two-Stage Approach." *Energy*, **152**, 521–532. ISSN 0360-5442. https://doi.org/10.1016/j.energy.2018.03.159.

Charnes A, Cooper WW, Rhodes E (1978). "Measuring the Efficiency of Decision Making Units." *European Journal of Operational Research*, **2**(6), 429–444. ISSN 0377-2217. https://doi.org/10.1016/0377-2217(78)90138-8.

Chen CP, Hu JL, Yang CH (2011). "An International Comparison of R&D Efficiency of Multiple Innovative Outputs: Role of the National Innovation System." *Innovation: Management, Policy and Practice*, **13**(3), 341–360. ISSN 1447-9338. https://doi.org/10.5172/impp.2011.13.3.341.

Chen Q, Kamran SM, Fan H (2019). "Real Estate Investment and Energy Efficiency: Evidence from China's Policy Experiment." *Journal of Cleaner Production*, **217**, 440–447. ISSN 0959-6526. https://doi.org/10.1016/j.jclepro.2019.01.274.

Cook WD, Tone K, Zhu J (2014). "Data Envelopment Analysis: Prior to Choosing a Model." *Omega*, **44**, 1–4. ISSN 0305-0483. https://doi.org/10.1016/j.omega.2013.09.004.

Cox DR (1981). "Statistical Analysis of Time Series: Some Recent Developments." *Scandinavian Journal of Statistics*, **8**(2), 93–108. ISSN 0303-6898. https://doi.org/10.2307/4615819.

Creal D, Koopman SJ, Lucas A (2013). "Generalized Autoregressive Score Models with Applications." *Journal of Applied Econometrics*, **28**(5), 777–795. ISSN 0883-7252. https://doi.org/10.1002/jae.1279.

Cullmann A, Schmidt-Ehmcke J, Zloczysti P (2012). "R&D Efficiency and Barriers to Entry: A Two Stage Semi-Parametric DEA Approach." *Oxford Economic Papers*, **64**(1), 176–196. ISSN 0030-7653. https://doi.org/10.1093/oep/gpr015.

Da Silva e Souza G, Gomes EG (2015). "Management of Agricultural Research Centers in Brazil: A DEA Application Using a Dynamic GMM Approach." *European Journal of Operational Research*, **240**(3), 819–824. ISSN 0377-2217. https://doi.org/10.1016/j.ejor.2014.07.027.

Dyson RG, Allen R, Camanho AS, Podinovski VV, Sarrico CS, Shale EA (2001). "Pitfalls and Protocols in DEA." *European Journal of Operational Research*, **132**(2), 245–259. ISSN 0377-2217. https://doi.org/10.1016/S0377-2217(00)00149-1.

Ekinci Y, Karadayi MA (2017). "Analysis of the Research and Development Efficiencies of European Union Countries." *Business & Management Studies: An International Journal*, **5**(1), 1–19. ISSN 2148-2586. https://doi.org/10.15295/bmij.v5i1.97.

Fonchamnyo DC, Sama MC (2016). "Determinants of Public Spending Efficiency in Education and Health: Evidence from Selected CEMAC Sountries." *Journal of Economics and Finance*, **40**(1), 199–210. ISSN 1055-0925. https://doi.org/10.1007/s12197-014-9310-6.

Frýd L, Sokol O (2021). "Relationships Between Technical Efficiency and Subsidies for Czech Farms: A Two-Stage Robust Approach." *Socio-Economic Planning Sciences*, **78**, 101059/1–101059/9. ISSN 0038-0121. https://doi.org/10.1016/j.seps.2021.101059.

Han U, Asmild M, Kunc M (2016). "Regional R&D Efficiency in Korea from Static and Dynamic Perspectives." *Regional Studies*, **50**(7), 1170–1184. ISSN 0034-3404. https://doi.org/10.1080/00343404.2014.984670.

Harvey AC (2013). *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series*. First Edition. Cambridge University Press, New York. ISBN 978-1-107-63002-4. https://doi.org/10.1017/cbo9781139540933.

Hladík M (2019). "Universal Efficiency Scores in Data Envelopment Analysis Based on a Robust Approach." *Expert Systems with Applications*, **122**, 242–252. ISSN 0957-4174. https://doi.org/10.1016/j.eswa.2019.01.019.

Holý V (2022). "The Impact of Operating Environment on Efficiency of Public Libraries." *Central European Journal of Operations Research*, **30**(1), 395–414. ISSN 1613-9178. https://doi.org/10.1007/s10100-020-00696-4.

Holý V, Šafr K (2018). "Are Economically Advanced Countries More Efficient in Basic and Applied Research?" *Central European Journal of Operations Research*, **26**(4), 933–950. ISSN 1435-246X. https://doi.org/10.1007/s10100-018-0559-2.

Holý V, Zouhar J (2022). "Modelling Time-Varying Rankings with Autoregressive and Score-Driven Dynamics." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **71**(5), 1427–1450. ISSN 0035-9254. https://doi.org/10.1111/rssc.12584.

Hung WC, Lee LC, Tsai MH (2009). "An International Comparison of Relative Contributions to Academic Productivity." *Scientometrics*, **81**(3), 703–718. ISSN 0138-9130. https://doi.org/10.1007/s11192-008-2210-9.

Hunter DR (2004). "MM Algorithms for Generalized Bradley-Terry Models." *The Annals of Statistics*, **32**(1), 384–406. ISSN 0090-5364. https://doi.org/10.1214/aos/1079120141.

Jablonsky J (2016). "Efficiency Analysis in Multi-Period Systems: An Application to Performance Evaluation in Czech Higher Education." *Central European Journal of Operations Research*, **24**(2), 283–296. ISSN 1435-246X. https://doi.org/10.1007/s10100-015-0401-z.

Kaufmann D, Kraay A, Mastruzzi M (2011). "The Worldwide Governance Indicators: Methodology and Analytical Issues." *Hague Journal on the Rule of Law*, **3**(2), 220–246. ISSN 1876-4045. https://doi.org/10.1017/s1876404511200046.

Kneip A, Simar L, Wilson PW (2015). "When Bias Kills the Variance: Central Limit Theorems for DEA and FDH Efficiency Scores." *Econometric Theory*, **31**(2), 394–422. ISSN 0266-4666. https://doi.org/10.1017/s0266466614000413.

Lee H, Park Y (2005). "An International Comparison of R&D Efficiency: DEA Approach." *Asian Journal of Technology Innovation*, **13**(2), 207–222. ISSN 1976-1597. https://doi.org/10.1080/19761597.2005.9668614.

Lee H, Park Y, Choi H (2009). "Comparative Evaluation of Performance of National R&D Programs with Heterogeneous Objectives: A DEA Approach." *European Journal of Operational Research*, **196**(3), 847–855. ISSN 0377-2217. https://doi.org/10.1016/j.ejor.2008.06.016.

Luce RD (1959). *Individual Choice Behavior: A Theoretical Analysis*. First Edition. Wiley, New York. ISBN 978-0-486-44136-8. https://books.google.com/books/about/Individual{_}choice{_}behavior.html?id=a80DAQAAIAAJ.

Luce RD (1977). "The Choice Axiom after Twenty Years." *Journal of Mathematical Psychology*, **15**(3), 215–233. ISSN 0022-2496. https://doi.org/10.1016/0022-2496(77)90032-3.

Mamatzakis E, Kalyvas AN, Piesse J (2013). "Does Regulation in Credit, Labour and Business Matter for Bank Performance in the EU-10 Economies?" *International Journal of the Economics of Business*, **20**(3), 341–385. ISSN 1357-1516. https://doi.org/10.1080/13571516.2013.835981.

McDonald J (2009). "Using Least Squares and Tobit in Second Stage DEA Efficiency Analyses." *European Journal of Operational Research*, **197**(2), 792–798. ISSN 0377-2217. https://doi.org/10.1016/j.ejor.2008.07.039.

Moradi-Motlagh A, Emrouznejad A (2022). "The Origins and Development of Statistical Approaches in Non-Parametric Frontier Models: A Survey of the First Two Decades of Scholarly Literature (1998-2020)." *Annals of Operations Research*, **318**(1), 713–741. ISSN 1572-9338. https://doi.org/10.1007/s10479-022-04659-7.

OECD (2015). "Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development, The Measurement of Scientific, Technological and Innovation Activities." *Technical report*, Paris. https://doi.org/10.1787/9789264239012-en.

Pirani N, Zahiri M, Engali KA, Torabipour A (2018). "Hospital Efficiency Measurement Before and After Health Sector Evolution Plan in Southwest of Iran: A DEA-Panel Data Study." *Acta Informatica Medica*, **26**(2), 106–110. ISSN 0353-8109. https://doi.org/10.5455/aim.2018.26.106-110.

Plackett RL (1975). "The Analysis of Permutations." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **24**(2), 193–202. ISSN 0035-9254. https://doi.org/10.2307/2346567.

Poveda AC (2011). "Economic Development and Growth in Colombia: An Empirical Analysis with Super-Efficiency DEA and Panel Data Models." *Socio-Economic Planning Sciences*, **45**(4), 154–164. ISSN 0038-0121. https://doi.org/10.1016/j.seps.2011.07.003.

Roman M (2010). "Regional Efficiency of Knowledge Economy in the New EU Countries: The Romanian and Bulgarian Case." *Romanian Journal of Regional Science*, **4**(1), 33–53. ISSN 1843-8520. https://ideas.repec.org/a/rrs/journl/v4y2010i1p33-53.html.

Sharma S, Thomas VJ (2008). "Inter-Country R&D Efficiency Analysis: An Application of Data Envelopment Analysis." *Scientometrics*, **76**(3), 483–501. ISSN 0138-9130. https://doi.org/10.1007/s11192-007-1896-4.

Simar L, Wilson PW (2007). "Estimation and Inference in Two-Stage, Semi-Parametric Models of Production Processes." *Journal of Econometrics*, **136**(1), 31–64. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2005.07.009.

Simar L, Wilson PW (2011). "Two-Stage DEA: Caveat Emptor." *Journal of Productivity Analysis*, **36**(2), 205–218. ISSN 0895-562X. https://doi.org/10.1007/s11123-011-0230-6.

Song M, Zhang G, Zeng W, Liu J, Fang K (2016). "Railway Transportation and Environmental Efficiency in China." *Transportation Research Part D: Transport and Environment*, **48**, 488–498. ISSN 1361-9209. https://doi.org/10.1016/j.trd.2015.07.003.

Thomas VJ, Sharma S, Jain SK (2011). "Using Patents and Publications to Assess R&D Efficiency in the States of the USA." *World Patent Information*, **33**(1), 4–10. ISSN 0172-2190. https://doi.org/10.1016/j.wpi.2010.01.005.

Turner HL, van Etten J, Firth D, Kosmidis I (2020). "Modelling Rankings in R: The PlackettLuce Package." *Computational Statistics*, **35**, 1027–1057. ISSN 0943-4062. https://doi.org/10.1007/s00180-020-00959-3.

Zhang YJ, Sun YF, Huang J (2018). "Energy Efficiency, Carbon Emission Performance, and Technology Gaps: Evidence from CDM Project Investment." *Energy Policy*, **115**, 119–130. ISSN 0301-4215. https://doi.org/10.1016/j.enpol.2017.12.056.

Modeling Price Clustering in High-Frequency Prices

Vladimír Holý
Prague University of Economics and Business
Winston Churchill Square 1938/4, 130 67 Prague 3, Czechia
vladimir.holy@vse.cz

Petra Tomanová
Prague University of Economics and Business
Winston Churchill Square 1938/4, 130 67 Prague 3, Czechia
petra.tomanova@vse.cz

Abstract: The price clustering phenomenon manifesting itself as an increased occurrence of specific prices is widely observed and well-documented for various financial instruments and markets. In the literature, however, it is rarely incorporated into price models. We consider that there are several types of agents trading only in specific multiples of the tick size resulting in an increased occurrence of these multiples in prices. For example, stocks on the NYSE and NASDAQ exchanges are traded with precision to one cent but multiples of five cents and ten cents occur much more often in prices. To capture this behavior, we propose a discrete price model based on a mixture of double Poisson distributions with dynamic volatility and dynamic proportions of agent types. The model is estimated by the maximum likelihood method. In an empirical study of DJIA stocks, we find that higher instantaneous volatility leads to weaker price clustering at the ultra-high frequency. This is in sharp contrast with results at low frequencies which show that daily realized volatility has a positive impact on price clustering.

Keywords: High-Frequency Data, Price Clustering, Generalized Autoregressive Score Model, Double Poisson Distribution.

JEL Classification: C22, C46, C58.

1 Introduction

Over the last two decades, there has been a growing interest in modeling prices at the highest possible frequency which reaches fractions of a second for the most traded assets. The so-called ultra-high-frequency data possess many unique characteristics which need to be accounted for by econometricians. Notably, the prices are irregularly spaced with discrete values. Other empirical properties of high-frequency prices which can be incorporated into models include intraday seasonality, jumps in prices, price reversal, and the market microstructure noise. For related models, see, e.g. Russell and Engle (2005), Robert and Rosenbaum (2011), Barndorff-Nielsen *et al.* (2012), Shephard and Yang (2017), Koopman *et al.* (2017), Koopman *et al.* (2018), and Buccheri *et al.* (2021).

We focus on one particular empirical phenomenon observed in high-frequency prices – price clustering. In general, price clustering refers to an increased occurence of some values in prices. A notable type of price clustering is an increased occurrence of specific multiples of the tick size, i.e. the minimum price change. For example, on the NYSE and NASDAQ exchanges, stocks are traded with precision to one cent but multiples of five cents (nickels) and ten cents (dimes) tend to occur much more often in prices. In other words, while one would expect the distribution of the second digit to be uniform, the probability of 0 and 5 is actually higher than 0.1 for each. This behavior can be captured by some agents trading only in multiples of five cents and some only in multiples of ten cents. It is well documented in the literature that this type of price clustering is present in stock markets (see, e.g. Lien *et al.*, 2019), commodity markets (see, e.g. Bharati *et al.*, 2012), foreign exchange markets (see, e.g. Sopranzetti and Datar, 2002), and cryptocurrency markets (see, e.g. Urquhart, 2017). Moreover, price clustering does not appear only in spot prices but in futures (see,

e.g. Schwartz *et al.*, 2004), options (see, e.g. ap Gwilym and Verousis, 2013), and swaps (see, e.g. Liu and Witte, 2013) as well. From a methodological point of view, almost all papers on price clustering deal only with basic methods and descriptive statistics of the phenomenon. The only paper, to our knowledge, that incorporates price clustering into a price model is the recent theoretical study of Song *et al.* (2020) which introduced the sticky double exponential jump diffusion process to assess the impact of price clustering on the probability of default.

Our goal is to propose a discrete dynamic model relating price clustering to the distribution of prices and to study the high-frequency behavior of price clustering. We take a fundamentally very different approach than Song *et al.* (2020) and incorporate the mechanism of an increased occurrence of specific multiples of the tick size directly into the model. This allows us to treat the price clustering phenomenon as dynamic and driven by specified factors rather than given. We also operate within the time series framework rather than the theory of continuous-time stochastic processes. In contrast to the existing literature on price modeling, we do not model log returns or price differences but rather prices themselves. Prices are naturally discrete and positive. When represented as integers, they also exhibit underdispersion, i.e. the variance lower than the mean. To accommodate for such features, we utilize the double Poisson distribution of Efron (1986). It is a less known distribution as noted by Sellers and Morris (2017) but was utilized in the context of time series by Heinen (2003), Xu *et al.* (2012) and Bourguignon *et al.* (2019). Modeling prices directly enables us to incorporate price clustering in the model. Specifically, we consider that prices follow a mixture of several double Poisson distributions with specific supports corresponding to agents trading in different multiples of the tick size. This mixture distribution has a location parameter, a dispersion parameter and parameters determining portions of trader types. In our model, we introduce time variation to all these parameters. We consider the location parameter to be equal to the last observed price resulting in zero expected returns. For the dispersion parameter, we employ dynamics in the fashion of the generalized autoregressive conditional heteroskedasticity (GARCH) model of Bollerslev (1986). Specifically, we utilize the class of generalized autoregressive score (GAS) models of Creal *et al.* (2013) and Harvey (2013) which allows to base dynamic models on any underlying distribution. In the high-frequency literature, the GAS framework was utilized by Koopman *et al.* (2018) for discrete price changes and Buccheri *et al.* (2021) for log prices. To account for irregularly spaced observations, we include the last trade duration as an explanatory variable similarly to Engle (2000). Finally, we relate the trader portion parameters to the volatility process and other variables such as the price, the last trade duration and the volume. The resulting observation-driven model is estimated by the maximum likelihood method.

In the empirical study, we analyze 30 Dow Jones Industrial Average (DJIA) stocks in the first half of 2020. We first focus on price clustering from a daily perspective which is a common approach in the price clustering literature. Using a panel regression with fixed effects, we find a positive effect of daily volatility measured by realized kernels of Barndorff-Nielsen *et al.* (2008) on price clustering. This finding is in line with the results of ap Gwilym *et al.* (1998); Davis *et al.* (2014); Box and Griffith (2016); Hu *et al.* (2017); Lien *et al.* (2019) among others. Next, we estimate the proposed high-frequency price model and arrive at a different conclusion – the instantaneous volatility obtained by the model has a negative effect on price clustering. The main message of the empirical study is therefore that the degree of aggregation plays a pivotal role in the relation between price clustering and volatility. While high daily realized volatility correlates with high price clustering, high instantaneous volatility has the opposite effect. The other explanatory variables have the expected effect in both the daily and high-frequency cases – the volume has a positive effect on price clustering while the price and the last trade duration are insignificant.

The rest of the paper is structured as follows. In Section 2, we review the literature dealing with high-frequency price models and price clustering. In Section 3, we propose the dynamic model accommodating for price clustering based on the double Poisson distribution. In Section 4, we use this model to study determinants of price clustering in high-frequency stock prices. We conclude the paper in Section 5.

2 Literature Review

2.1 Some High-Frequency Price Models

In the literature, several models addressing specifics of ultra-high-frequency data have been proposed. One of the key issues is irregularly spaced transactions and discreteness of prices. The seminal study of Engle and Russell (1998) proposed the autoregressive conditional duration (ACD) model to capture the autocorrelation structure of trade durations, i.e. times between consecutive trades. Engle (2000) combined the ACD model with the GARCH model and jointly modeled prices with trade durations. Russell and Engle (2005) again modeled prices jointly with trade durations but addressed discreteness of prices and utilized the multinomial distribution for price changes.

Another approach is to model the price process in continuous time. Robert and Rosenbaum (2011) considered that the latent efficient price is a continuous Itô semimartingale but is observed at the discrete grid through the mechanism of uncertainty zones. Barndorff-Nielsen *et al.* (2012) considered the price process to be discrete outright and developed a continuous-time integer-valued Lévy process suitable for ultra-high-frequency data. Shephard and Yang (2017) also utilized integer-valued Lévy processes and focused on frequent and quick reversal of prices.

Transaction data at a fixed frequency can also be analyzed as equally spaced time series with missing observations. In this setting, Koopman *et al.* (2017) proposed a state space model with dynamic volatility and captured discrete price changes by the Skellam distribution. Koopman *et al.* (2018) continued with this approach and modeled dependence between discrete stock price changes using a discrete copula. Buccheri *et al.* (2021) also dealt with multivariate analysis and proposed a model for log prices accommodating for asynchronous trading and the market microstructure noise. The latter two papers utilized the GAS framework.

2.2 Price Clustering

The first academic paper on price clustering was written by Osborne (1962), where the author described the price clustering phenomenon as a pronounced tendency for prices to cluster on whole numbers, halves, quarters, and odd one-eighths in descending preference, like the markings on a ruler. Since then, there have been many studies focusing on this phenomenon – from Niederhoffer (1965) to very recent papers of Li *et al.* (2020), Song *et al.* (2020), and Das and Kadapakkam (2020) – showing that price clustering is remarkably persistent in time and across various markets.

Song *et al.* (2020) pointed out that, however, all studies are entirely focused on empirically examining price clustering in different financial markets. Except for the purely theoretical paper of Song *et al.* (2020) proposing the sticky double exponential jump-diffusion process to analyze the probability of default for financial variables, the studies related to price clustering are based on basic general methods and do not aim to incorporate the phenomenon into the dynamic price model.

The prevalent approach to price clustering examination is a linear regression model estimated by ordinary least squares (OLS) method. Ball *et al.* (1985), Kandel *et al.* (2001), and from the recent literature Urquhart (2017), Hu *et al.* (2019), and Li *et al.* (2020), used the classical regression with dummy variables to estimate frequency of each level of rounding. The vast of the literature regressed price clustering on explanatory variables such as volatility and trade size, where price clustering is defined as the excess occurrence of multiples of nickles or dimes (see, e.g. Schwartz *et al.*, 2004 and Ikenberry and Weston, 2008 followed by Chung and Chiang, 2006, Brooks *et al.*, 2013, and Hu *et al.*, 2017) or simply their frequency (see, e.g. Palao and Pardo, 2012 and Davis *et al.*, 2014). However, different definitions of the dependent variable representing price clustering can be found in the literature. Baig *et al.* (2019) defined the clustering as a sum of round clustering at prices ending by digit 0 and strategic clustering measured as a number of trades which decimals are equal to 01 or 99. ap Gwilym and Verousis (2013) defined the dependent variable as the percentage of price observations at integers, whereas ap Gwilym *et al.* (1998) estimated the percentage of trades that occur at an odd tick. Ahn *et al.* (2005) regressed abnormal even price frequencies in transaction and quote prices on the firm and trading characteristics, and similarly, Chiao and Wang (2009) performed the analysis on the limit-order data. Cooney *et al.* (2003) estimated cross-sectional regressions of

the difference in the percentage of even and odd limit orders on stock price and proxies for investor uncertainty.

Several extensions of the OLS method were employed to overcome certain issues. Verousis and ap Gwilym (2013) and Mishra and Tripathy (2018) argued that one encounters a simultaneity issue between trade size and price clustering when striving to examine a causal relationship between them. Hence, Verousis and ap Gwilym (2013) followed by Mishra and Tripathy (2018) used the two-stage least squares (2SLS) method. Moreover, to reflect the endogeneity of quote clustering in the spread model and the endogeneity of the spread in the quote clustering model, Chung *et al.* (2004, 2005) estimated a structural model using three-stage least squares (3SLS) method. Meng *et al.* (2013) used the 3SLS method to formally examine the hypothesis of a substitution effect between price clustering and size clustering in the CDS market. Finally, Mbanga (2019) estimated robust regressions that eliminate gross outliers to examine the day-of-the-week effect in Bitcoin price clustering.

Another direction arises from the need to analyze panel data. The prevailing approach is a fixed effects regression. Das and Kadapakkam (2020) included both firm and time fixed effects, whereas Box and Griffith (2016) included fixed effects only for time and report that once they also included firm fixed effects, the results remained unchanged. Blau (2019) and Blau and Griffith (2016) included month and year fixed effects respectively, and used robust standard errors that account for clustering across both the cross-sectional observations and time-series observations. On the other hand, Ohta (2006) picked random effects model over the fixed effects model based on the results from the Hausman specification test.

A substantial part of the literature models price clustering as a binary variable. For that case, the straightforward approach is to use the logit or probit model. From one of the first papers using logistic regression to analyze price clustering, Ball *et al.* (1985) modeled three dependent variables taking value 1 if the price is rounded to the whole dollar, half-dollar, or quarter, respectively. Christie and Schultz (1994) estimated logistic regressions that predict the probability of a firm being quoted using odd eighths. Aitken *et al.* (1996) employed multivariate logistic regression to model three binary dependent variables that are equal to one if the final digit is 0; 0 or 5; and 0, 2, 4, 6, or 8 (even numbers), respectively. Brown and Mitchell (2008) examined the influence of Chinese culture on price clustering by logistic regressions where a binary dependent variable is equal to 1 if the last sale price ends in 4 and 0 if it ends in 8 since many Chinese consider the number 8 as *lucky* while 4 is considered as *unlucky*. From the literature employing probit models, Kahn *et al.* (1999) analyzed the propensity to set retail deposit interest rates at integer levels and Sopranzetti and Datar (2002) analyzed the propensity for exchange rates to cluster on even digits. Moreover, Capelle-Blancard and Chaudhury (2007) modeled a binary dependent variable that is equal to one if the transaction price ends with 00, whereas Liu (2011) and Narayan and Smyth (2013) set the variable equal to one if the price ends at either 0 or 5, and 0 otherwise. Alexander and Peterson (2007) followed by Lien *et al.* (2019) used a bivariate probit model to take into account the dependence between price and trade-size clustering.

Finally, Blau (2019) estimated a vector autoregressive process and examined the impulses of price clustering in response to an exogenous shock to investor sentiment. Besides the classical regression approaches, Harris (1991) and Hameed and Terry (1998) analyzed the cross-sectional data by static discrete price model.

To the best of our knowledge, the literature still lacks a discrete dynamic model to study the high-frequency behavior of the price clustering. Thus, in the next section, we propose a novel model which models high-frequency prices directly at the highest possible frequency and allows us to study the main drivers of price clustering such as price, volatility, volume, and trading frequency in the form of trade durations.

3 Dynamic Price Clustering Model

3.1 Double Poisson Distribution

Let us start with the static version of our model for prices. In the first step, we transform the observed prices to have integer values. For example, on the NYSE and NASDAQ exchanges, the prices are recorded with precision to two decimal places and we therefore multiply them by 100 to obtain integer values. The minimum possible change in the transformed prices is 1. Empirically, the transformed prices exhibit strong *underdispersion*, i.e. the variance lower than the mean. In our application, the transformed prices are in the order of thousands and tens of thousands while the price changes are in the order of units and tens. We therefore need to base our model on a count distribution allowing for underdispersion. For a review of such distributions, we refer to Sellers and Morris (2017). Although not without its limitations, the double Poisson distribution is the best candidate for our case as the alternative distributions have too many shortcomings. For example, the condensed Poisson distribution is based on only one parameter, the generalized Poisson distribution can handle only limited underdispersion and the gamma count distribution as well as the Conway–Maxwell–Poisson distribution do not have the moments available in a closed form.

The double Poisson distribution was proposed in Efron (1986) and has a *location* parameter $\mu > 0$ and a *dispersion* parameter α. We adopt a slightly different parametrization than Efron (1986) and use the logarithmic transformation for the dispersion parameter making α unrestricted. For $\alpha = 0$, the distribution reduces to the Poisson distribution. Values $\alpha > 0$ result in underdispersion while values $\alpha < 0$ result in overdispersion. Let Y be a random variable and y an observed value. The probability mass function is given by

$$\mathrm{P}[Y = y|\mu, \alpha] = \frac{1}{C(\mu, \alpha)} \frac{y^y}{y!} \left(\frac{\mu}{y}\right)^{e^\alpha y} e^{\frac{\alpha}{2} + e^\alpha y - e^\alpha \mu - y}, \tag{1}$$

where $C(\mu, \alpha)$ is the *normalizing constant* given by

$$C(\mu, \alpha) = \sum_{y=0}^{\infty} \frac{y^y}{y!} \left(\frac{\mu}{y}\right)^{e^\alpha y} e^{\frac{\alpha}{2} + e^\alpha y - e^\alpha \mu - y}. \tag{2}$$

The log likelihood for observation y is then given by

$$\ell(y; \mu, \alpha) = -\ln\left(C(\mu, \alpha)\right) + y\ln(y) - \ln(y!) + e^\alpha y \ln\left(\frac{\mu}{y}\right) + \frac{\alpha}{2} + e^\alpha y - e^\alpha \mu - y. \tag{3}$$

Unfortunately, the normalizing constant is not available in a closed form. However, as Efron (1986) shows, it is very close to 1 (at least for some combinations of μ and α) and can be approximated by

$$C(\mu, \alpha) \simeq 1 + \frac{1 - e^\alpha}{12 e^\alpha \mu} \left(1 + \frac{1}{e^\alpha \mu}\right). \tag{4}$$

Zou *et al.* (2013) notes that approximation (4) is not very accurate for low values of the mean and suggest approximating the normalizing constant alternatively by cutting off the infinite sum, i.e.

$$C(\mu, \alpha) \simeq \sum_{y=0}^{m} \frac{y^y}{y!} \left(\frac{\mu}{y}\right)^{e^\alpha y} e^{\frac{\alpha}{2} + e^\alpha y - e^\alpha \mu - y}, \tag{5}$$

where m should be at least twice as large as the sample mean. In our case of high mean, approximation (4) is sufficient while approximation (5) would be computationally very demanding and we therefore resort to the former one. The expected value and variance can be approximated by

$$\mathrm{E}[Y] \simeq \mu, \qquad \mathrm{var}[Y] \simeq \mu e^{-\alpha}. \tag{6}$$

The score can be approximated by

$$\nabla(y; \mu, \alpha) \simeq \left(\frac{e^\alpha}{\mu}(y - \mu), e^\alpha \left(y\ln(\mu) - \mu - y\ln(y) + y\right) + \frac{1}{2}\right)'. \tag{7}$$

The Fisher information can be approximated by

$$\mathcal{I}(\mu, \alpha) \simeq \begin{pmatrix} \frac{e^{\alpha}}{\mu} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}. \tag{8}$$

3.2 Mixture Distribution for Price Clustering

Next, we propose a mixture of several double Poisson distributions corresponding to trading in different multiples of tick sizes accommodating for price clustering. We consider that there are three types of traders – one who can trade in cents, one who can trade only in multiples of 5 cents, and one who can trade only in multiples of 10 cents. In Appendix A, we treat a more general case with any number of trader types and tick size multiples. The distribution of prices corresponding to each trader type is based on the double Poisson distribution modified to have support consisting only of multiples of $k \in \{1, 5, 10\}$ while keeping the expected value $\mathrm{E}[Y] \simeq \mu$ and the variance $\mathrm{var}[Y] \simeq \mu e^{-\alpha}$ regardless of k. For a detailed derivation of the distribution, see Appendix A. The distribution of prices for trader type $k \in \{1, 5, 10\}$ is given by

$$\mathrm{P}\left[Y^{[k]} = y \middle| \mu, \alpha\right] = \mathbb{I}\{k \mid y\}\, \mathrm{P}\left[Z^{[k]} = \frac{y}{k} \middle| \mu, \alpha\right], \qquad Z^{[k]} \sim \mathrm{DP}\left(\frac{\mu}{k}, \alpha + \ln(k)\right), \tag{9}$$

where DP denotes the double Poisson distribution and $\mathbb{I}\{k \mid y\}$ is equal to 1 if y is divisible by k and 0 otherwise. Note that for $k = 1$, it is the standard double Poisson distribution. Finally, the distribution of all prices is the mixture

$$\mathrm{P}\left[Y = y | \mu, \alpha, \varphi_1, \varphi_5, \varphi_{10}\right] = \sum_{k \in \{1,5,10\}} \varphi_k \mathrm{P}\left[Y^{[k]} = y \middle| \mu, \alpha\right], \tag{10}$$

where the parameter space is restricted by $\mu > 0$, $\varphi_1 \geq 0$, $\varphi_5 \geq 0$, $\varphi_{10} \geq 0$ and $\varphi_1 + \varphi_5 + \varphi_{10} = 1$. Parameters φ_k, $k \in \{1, 5, 10\}$, are the portions of trader types and parameters μ with α have the same interpretation as in the double Poisson distribution. The log likelihood for observation y is given by

$$\begin{aligned} \ell\left(y; \mu, \alpha, \varphi_1, \varphi_5, \varphi_{10}\right) = {} & e^{\alpha} y \ln\left(\frac{\mu}{y}\right) + \frac{\alpha}{2} + e^{\alpha} y - e^{\alpha} \mu \\ & + \ln\left(\sum_{k \in \{1,5,10\}} \varphi_k \mathbb{I}\{k|y\} \frac{\sqrt{k}}{C\left(\frac{\mu}{k}, k\alpha\right)} \frac{\left(\frac{y}{k}\right)^{\frac{y}{k}}}{\left(\frac{y}{k}\right)!} e^{-\frac{y}{k}}\right). \end{aligned} \tag{11}$$

Note that the last logarithm in (11) is not dependent on parameters μ and α besides the normalizing constant making the approximation of the score quite simple. Additionaly, parameters φ_1, φ_5 and φ_{10} appear only in the last logarithm in (11) making the approximation of the score for parameters μ and α independent of parameters φ_1, φ_5 and φ_{10}. The approximations of the expected value and the variance as well as the score and the Fisher information for the parameters μ and α of the mixture distribution are therefore the same as for the regular double Poisson distribution presented in (6), (7), and (8) respectively when assuming $C(\mu/k, k\alpha) = 1$.

Figure 1 illustrates the probability mass function of the mixture distribution. As an example, we choose the portions of trader types as $\varphi_1 = 0.95$, $\varphi_5 = 0.02$, and $\varphi_{10} = 0.03$. The majority of traders therefore operate at one cent precision but some are restricted to trading in only five or ten cents, similarly to the observed behavior in the empirical study. The mixture distribution is based on the regular double Poisson distribution with all probabilities decreased by 5 percent. Prices ending with 5 or 0 are then inflated according to distribution given by (9) for $k = 5$ with all probabilities decreased by 98 percent. Finally, prices ending with 0 are further inflated according to distribution given by (9) for $k = 10$ with all probabilities decreased by 97 percent. Figure 1 shows how the probabilities can be decomposed for the three trader types.

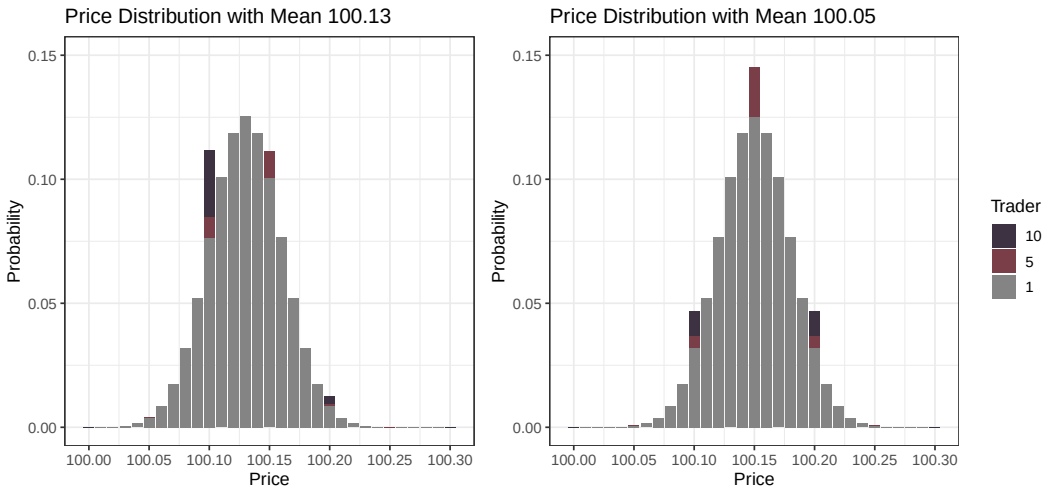

Figure 1: Illustration of the probability mass function for the mixture double Poisson distribution with parameters $\mu = 10\,013$ (left plot), $\mu = 10\,005$ (right plot), $\alpha = 7$, $\varphi_1 = 0.95$, $\varphi_5 = 0.02$, and $\varphi_{10} = 0.03$. The prices are reported in the original form with two decimal places.

3.3 Dynamics of Time-Varying Parameters

Finally, we introduce time variation into parameters $\mu, \alpha, \varphi_1, \varphi_5, \varphi_{10}$. We denote the random prices as $Y_t \in \mathbb{N}_0$, $t = 1, \ldots, n$ and the observed values as $y_t \in \mathbb{N}_0$, $t = 1, \ldots, n$. We also utilize observed trade durations $z_t \in \mathbb{R}^+$, $t = 1, \ldots, n$ and observed volumes $v_t \in \mathbb{R}^+$, $t = 1, \ldots, n$. We assume that Y_t follow the mixture double Poisson distribution proposed in Section 3.2 with time-varying parameters μ_t, α_t, $\varphi_{1,t}$, $\varphi_{5,t}$ and $\varphi_{10,t}$. The dynamics of the location parameter μ_t is given by

$$\mu_t = y_{t-1}. \tag{12}$$

This means that the expected value of the price is (approximately) equal to the last observed price, i.e. the expected value of the return is zero. This is a common assumption for high-frequency returns (see, e.g. Koopman *et al.*, 2017).

For the dynamics of the dispersion parameter α_t, we utilize the generalized autoregressive score (GAS) model of Creal *et al.* (2013), also known as the dynamic conditional score (DCS) model by Harvey (2013). The GAS model is an observation-driven model providing a general framework for modeling time-varying parameters for any underlying probability distribution. It captures dynamics of time-varying parameters by the autoregressive term and the score of the conditional density function. Blasques *et al.* (2015) investigated information-theoretic optimality properties of the score function and showed that only parameter updates based on the score will always reduce the local Kullback–Leibler divergence between the true conditional density and the model-implied conditional density. Creal *et al.* (2013) suggested to scale the score based on the Fisher information. As the Fisher information for the parameter α_t is constant in our case, the score is already normalized and we therefore omit the scaling. Using (7) and (12), we let the dispersion parameter α_t follow the recursion

$$\alpha_t = c + b\alpha_{t-1} + a\left(e^{\alpha_{t-1}}\left(y_{t-1}\ln\left(\frac{y_{t-2}}{y_{t-1}}\right) - y_{t-2} + y_{t-1}\right) + \frac{1}{2}\right) + d\ln(z_t), \tag{13}$$

where c is the constant parameter, b is the autoregressive parameter, a is the score parameter and d is the duration parameter. This volatility dynamics corresponds to the generalized autoregressive conditional heteroskedasticity (GARCH) model of Bollerslev (1986). Similarly to Engle (2000), we also include the preceding trade duration z_t as an explanatory variable to account for irregularly spaced observations. To prevent extreme values of durations, we use the logarithmic transformation.

The portions of trader types are driven by process

$$\eta_t = f\eta_{t-1} + g_1 \ln(\mu_t) + g_2 \left(\ln(\mu_t) - \alpha_t\right) + g_3 \ln(z_t) + g_4 \ln(v_t), \tag{14}$$

where f is the autoregressive parameter, g_1 is the parameter for the logarithm of the expected price, g_2 is the parameter for the logarithm of the variance of the price process $\mu_t e^{-\alpha_t}$, g_3 is the parameter for the logarithm of the preceding trade duration, and g_4 is the parameter for the logarithm of the volume v_t. The portions of trader types are then standardized as

$$\varphi_{1,t} = \frac{e^{\eta_t}}{e^{\eta_t} + h_5 + h_{10}}, \quad \varphi_{5,t} = \frac{h_5}{e^{\eta_t} + h_5 + h_{10}}, \quad \varphi_{10,t} = \frac{h_{10}}{e^{\eta_t} + h_5 + h_{10}}, \tag{15}$$

where $h_5 \geq 0$ and $h_{10} \geq 0$ are parameters capturing representation of 5 and 10 trader types. The model can be straightforwardly extended to include additional explanatory variables in (13) and (14).

3.4 Maximum Likelihood Estimation

The proposed model based on the mixture distribution for price clustering (10) with dynamics given by (12), (13) and (15) can be straightforwardly estimated by the conditional maximum likelihood method. Let $\theta = (c, b, a, d, f, g_1, g_2, g_3, g_4, h_5, h_{10})'$ denote the static vector of all parameters. The parameter vector θ is then estimated by the conditional maximum likelihood

$$\hat{\theta} \in \arg\max_{\theta} \sum_{t=1}^{n} \ell\left(y_t; \mu_t, \alpha_t, \varphi_{1,t}, \varphi_{5,t}, \varphi_{10,t}\right), \tag{16}$$

where $\ell(y_t; \mu_t, \alpha_t, \varphi_{1,t}, \varphi_{5,t}, \varphi_{10,t})$ is given by (11).

Note that process (12) implies non-stationarity of our model, even though our main focus is on processes (13) and (14), in which we assume $b < 1$ and $f < 1$, respectively. This reflects a commonly used assumption of non-stationarity of prices. Non-stationary GAS models were discussed and applied e.g. by Gorgi *et al.* (2019), Harvey *et al.* (2019), and Holý and Zouhar (2022). In our case, the use of a non-stationary model is necessary as we require to model prices directly in order to capture price clustering.

For the numerical optimization in the empirical study, we utilize the PRincipal AXIS algorithm of Brent (1972). To improve numerical performance, we standardize the explanatory variables to unit mean. We also run the estimation procedure several times with different starting values to avoid local maxima.

4 Empirical Results

4.1 Data Sample

The empirical study is conducted on transaction data extracted from the NYSE TAQ database which contains intraday data for all securities listed on the New York Stock Exchange (NYSE), American Stock Exchange (AMEX), and Nasdaq Stock Market (NASDAQ). We analyze 30 stocks that form the Dow Jones Industrial Average (DJIA) index in June 2020. The extracted data span over six months from January 2 to June 30, 2020, except for Raytheon Technologies (RTX)[1] for which the data are available from April 3, 2020.

We follow the standard cleaning procedure for the NYSE TAQ dataset described in Barndorff-Nielsen *et al.* (2009) since data cleaning is an important step of high-frequency data analysis (Hansen and Lunde, 2006). Before the standard data pre-processing is conducted, we delete entries that are identified as preferred or warrants (trades with the non-empty suffix indicator). Then we follow a common data cleaning steps and discard (i) entries outside the main opening hours (9:30 – 16:00), (ii) entries with the transaction price equal to zero, (iii) entries occurring on a different exchange than

[1]The RTX company results from the merge of the United Technologies Corporation and the Raytheon Company on April 3, 2020.

it is primarily listed, (iv) entries with corrected trades, (v) entries with abnormal sale condition, (vi) entries for which the price deviated by more than 10 mean absolute deviations from a rolling centered median of 50 observations, and (vi) duplicate entries in terms of the time stamp. In the last step, we remain the entry with mode price instead of the originally suggested median price due to avoiding distortion of the last decimal digit of prices.

The first and last step has a negligible impact on our data and steps ii, iv, and vi have no impact at all. However, the third step causes a large deletion of the data which is, however, in line with Barndorff-Nielsen *et al.* (2009). The basic descriptive statistics after data pre-processing are shown in Appendix B. Number of observations ranges from 216 618 (TRV) to 3 099 279 (MSFT). Price clustering in terms of the excess occurrence of multiples of five cents and ten cents in prices ranges from 1.45 % (KO) to 11.52 % (BA). First, we analyze the price clustering using a common approach of fixed effects model on daily data in Section 4.2 to investigate whether the results for our dataset are in line with the existing literature. Then, we estimate the proposed dynamic price model in Section 4.3.

4.2 Analysis Based on Daily Data

In this section, we investigate the main determinants of price clustering for which pervasive evidence is documented in the literature, namely price, volatility, trading frequency (which we measure in terms of trade durations), and volume. We use a panel regression with fixed effects to take into account the unobserved heterogeneity in both dimensions – stocks and days.

Let us define price clustering $p_{i,t}$ as the excess relative frequency of multiples of five cents and ten cents in prices of stock i at day t. We model $p_{i,t}$ as

$$p_{i,t} = \gamma_i + \delta_t + \beta_1 \ln(\overline{\mu}_{i,t}) + \beta_2 \ln(\overline{\alpha}_{i,t}) + \beta_3 \ln(\overline{z}_{i,t}) + \beta_4 \ln(\overline{v}_{i,t}) + \varepsilon_{i,t}, \tag{17}$$

where γ_i is a stock specific effect for the stock i, δ_t is a time effect for day t, and $\varepsilon_{i,t}$ is the error term. Parameters β_1 – β_4 corresponds to logarithmic explanatory variables, where $\overline{\mu}_{i,t}$ is an average price, $\overline{z}_{i,t}$ is an average duration and $\overline{v}_{i,t}$ is an average volume, where all averages are calculated for each stock i at each day t. Daily volatility $\overline{\alpha}_{i,t}$ is estimated by realized kernel estimator of Barndorff-Nielsen *et al.* (2008). We use Parzen kernel as suggested by Barndorff-Nielsen *et al.* (2009). See Holý and Tomanová (2023) for a comprehensive overview of quadratic covariation estimators.

Table 1 reports estimated coefficients of three variants of the fixed effects model. The first variant models price clustering on price, volatility and duration, i.e. model in (17) where volume is skipped. The second model considers only the price, duration and volume as the explanatory variables, and the third one is the full model in (17). We test the significance of the estimated coefficients using robust standard errors for which observations are clustered in both dimensions to account for serial as well as cross-sectional correlation. For illustration, Figure 2 shows fitted lines from univariate regressions with stock specific effects[2] for two stocks traded on NASDAQ – Apple Inc. (AAPL) and Microsoft (MSFT) – and two stocks traded on NYSE – Boeing (BA) and Visa Inc. (V). Figure 2 aims to depict the presence of the stock specific effects: in descending order, BA, AAPL and MSFT have the highest level of price clustering while V represents a stock with an average level of price clustering from our Dow Jones dataset.

The results show that volatility is a highly significant driver of price clustering. Moreover, the effect is positive, which is in line with the overwhelming majority of literature, in particular, with the so-called *price resolution* hypothesis. We refer to Section 4.4 for a literature review on this topic and related implications of our models. Next, Table 1 shows that volume is also highly significant and has a positive effect on price clustering according to expectations. For example, this is in line with results from the panel data analysis of Das and Kadapakkam (2020) and Box and Griffith (2016). Das and Kadapakkam (2020) suggested that the reason behind the significantly positive effect is algorithmic trading: as the trades become smaller in size due to increased algorithmic trading, lower levels of price clustering are observed.

[2]Time effects are dropped for better visibility which does not alter the main result.

Table 1: Estimated coefficients and robust standard errors of models with fixed effects for stocks and days. Label I refers to the model in (17) for which the volume parameter $\beta_4 = 0$; II to the model in (17) for which the volatility parameter $\beta_2 = 0$; III to the full model in (17).

Variable	I	II	III
Price	-1.7014^{***}	-0.8067	-0.1245
	(0.5469)	(0.6615)	(0.5323)
Volatility	0.5900^{***}		0.7008^{***}
	(0.1901)		(0.1622)
Duration	-0.0973	-0.4640^{***}	-0.0069
	(0.1672)	(0.1347)	(0.1739)
Volume		3.7544^{***}	3.9557^{***}
		(0.4869)	(0.5025)

*p<0.05; **p<0.01; ***p<0.001

The volatility is negatively correlated with duration, however, the duration does not bring additional information since it is no longer significant once the volatility is added in the model (see Model II vs. III). A similar result applies for volume and price (see Model I vs. III), where price turns to be insignificant once the volume is added to the model. We conclude that based on our model using daily data: (i) we do not find a significant effect of duration and price level on price clustering; (ii) the volatility and volume have highly significant positive effects on price clustering which are in line with expectations.

4.3 High-Frequency Analysis

Let us analyze the price clustering phenomenon at the highest possible frequency. First, we take a brief look at the relation between the individual explanatory variables and price clustering. We focus on the BA stock as its price clustering is the most pronounced. Figure 3 shows the average expected price, the average instantaneous variance obtained from the dynamic model, the average duration preceding the trade, and the average volume broken down by the second decimal of the price for the BA stock. We can clearly see that for prices ending with 0 and 5, the average variance and the average trade duration is much lower than for the other digits while the average volume is much higher. Note that succeeding durations show very similar behavior to preceding durations suggesting that price clustering tends to occur when trading is more intense.

Next, we estimate three versions of the proposed price clustering model for each of the 30 stocks. In the first version, we assume that there is no price clustering and set $f = g_1 = g_2 = g_3 = g_4 = h_5 = h_{10} = 0$. In the second version, we set only $f = g_1 = g_2 = g_3 = g_4 = 0$ and assume that there is price clustering present but is constant over time and does not depend on any variables. The third version is the dynamic model presented in Section 3.3 without any restrictions. We report the average log-likelihood and the Akaike information criterion (AIC) of the models in Table 2. We can see that adding price clustering to the model and subsequently adding dynamics to price clustering is very much worth of the extra few parameters as AIC is distinctly the lowest for the dynamic model for all stocks.

From now on we focus on the model with dynamic price clustering. Table 3 reports the estimated coefficients. We do not report standard deviations as they are very close to zero and all coefficients are significant at any reasonable level due to a huge number of observations. For all stocks, the coefficients in the volatility process c, b, a, and d have the same signs and fairly similar values

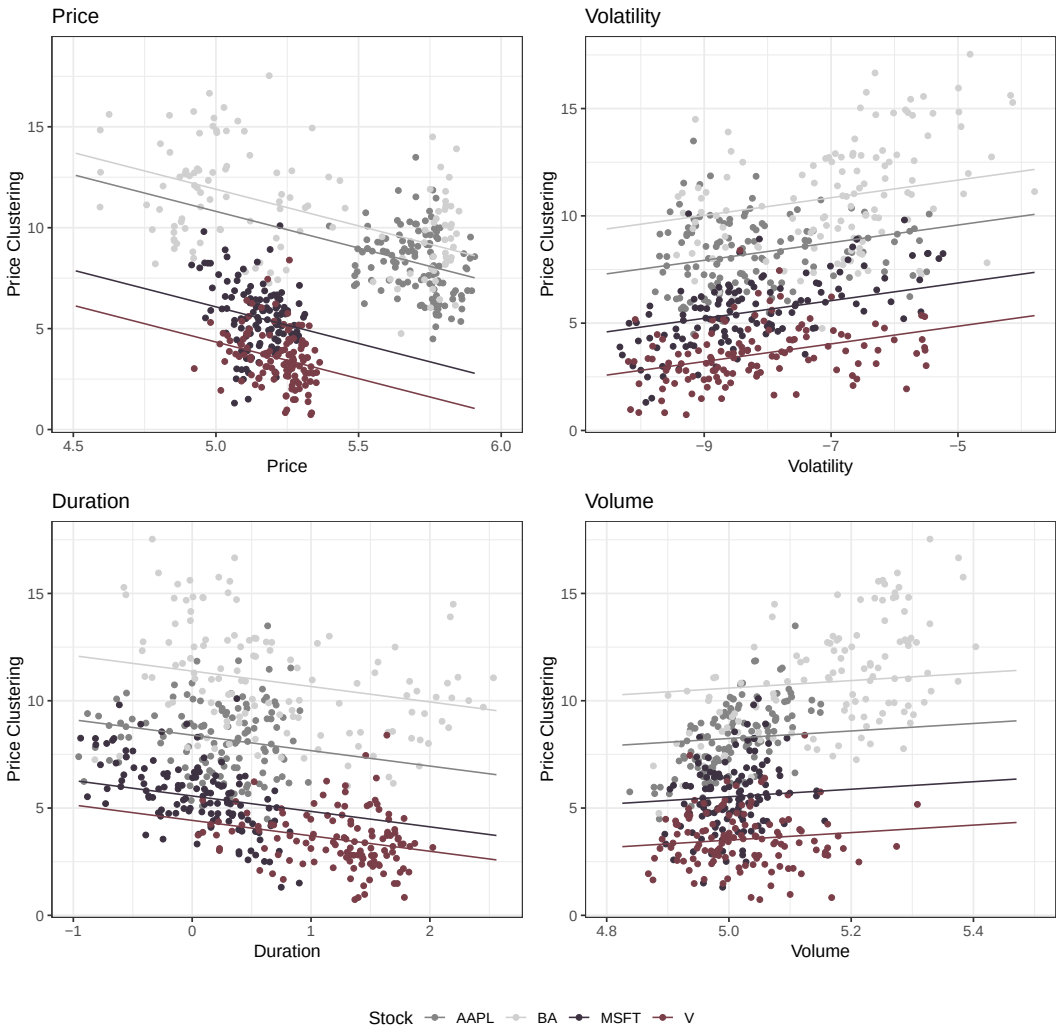

Figure 2: Daily data of four selected stocks and fitted lines from univariate panel regressions of price clustering on logarithmic price (top left), logarithmic volatility (top right), logarithmic duration (bottom left) and logarithmic volume (bottom right). All models are estimated with stock fixed effects.

demonstrating the robustness of the model. Parameter d is negative, which means that with longer durations, dispersion parameter α_t is lower and the instantaneous variance $\mu_t \exp(-\alpha_t)$ is higher. We attribute this behavior to the presence of a large amount of extremely short durations associated with small price changes. Note that for example, the BA stock has 50 percent of durations shorter than 0.1 seconds and 19 percent shorter than 0.0001 seconds. Engle (2000) observed the opposite relation between trade durations and volatility but based his results on data with a much lower frequency and without durations shorter than 1 second. This might indicate a change in the data structure over the years and a complex non-linear relation between trade durations and volatility. This topic is, however, beyond the scope of this paper.

Regarding the dynamics of price clustering, the autoregressive parameter f is stable across all stocks. The parameter for the expected price g_1 significantly varies for different stocks suggesting its low informative value. This is in line with the daily analysis in which prices were found insignificant. The parameter for the instantaneous variance g_2 is positive for all stocks. The portion of one cent traders is therefore higher with higher variance and price clustering tends to occur when prices are less volatile. This is the most interesting result as it deviates from the behavior observed in the daily analysis. The parameter for the preceding trade duration g_3 significantly varies for different stocks, similarly to g_1. When the preceding trade duration is the only explanatory variable included in the model, however, g_3 is positive for all stocks. Recall that durations have a positive effect on instantaneous variance as d is positive for all stocks. This implies that durations have an effect on instantaneous variance which in turn has an effect on price clustering. However, when controlling for instantaneous variance, durations do not bring additional information to explain price clustering. These observations are in line with the daily analysis. Finally, the parameter for the volume g_4 is negative for all stocks. As in the daily analysis, higher volume is clearly associated with higher price clustering.

We omit parameters h_5 and h_{10} controlling strength of price clustering from Table 3 as they are not very informative for readers. It is far better to look at the average values of trader portions $\bar{\varphi}_1$, $\bar{\varphi}_5$ and $\bar{\varphi}_{10}$ reported in Table 4. The average portion of ten-cent traders ranges from 0.48 percent for the TRV stock to 10.54 percent for the BA stock. The average portion of five-cent traders ranges from 0.34 percent for the TRV stock to 3.63 percent for the BA stock. An example of the progression of trader type ratios is shown in Figure 4 for the BA stock on the first trading day of 2020.

As a benchmark, we compare the proposed model with the GARCH model based on the Student's t-distribution of Bollerslev (1987). For each stock, we estimate three specifications based on prices, price differences, and logarithmic returns. The direct use of prices results in a non-stationary model, similarly to the proposed model. As our goal is to compare models based on different data transformations, we report their log-likelihoods with respect to the original discrete prices. This approach is used, e.g., by Blasques *et al.* (2022). Table 5 reports the fit of the GARCH models. Based on these results, we cannot unambiguously determine which model is the best. The GARCH models perform better for 18 stocks while the proposed price clustering model performs better for 12 stocks. Among the GARCH models, the one based on price differences has the highest log-likelihood for 17 stocks. We can conclude that the performance of our proposed model is comparable to the GARCH model based on the Student's t-distribution, regardless of the price transformation.

4.4 Implications

Several hypotheses have been established to explain the price clustering phenomenon. The *attraction hypothesis* of Goodhart and Curcio (1991) essentially states that there exists a particular preference (basic attraction) for certain numbers, especially for the rounded ones. The *negotiation hypothesis* of Harris (1991) assumes that traders use discrete price sets to lower the costs of negotiating. Once the set of prices is reduced, the traders reach agreements more easily since the amount of information that must be exchanged between negotiating traders decreases. Christie and Schultz (1994) argued in the *collusion hypothesis* that the lack of odd-eighth quotes on NASDAQ cannot be explained by the negotiation hypothesis, trading activity, or other variables thought to impact spreads, which suggests that NASDAQ dealers might implicitly collude to maintain wide spreads. However, assessing these

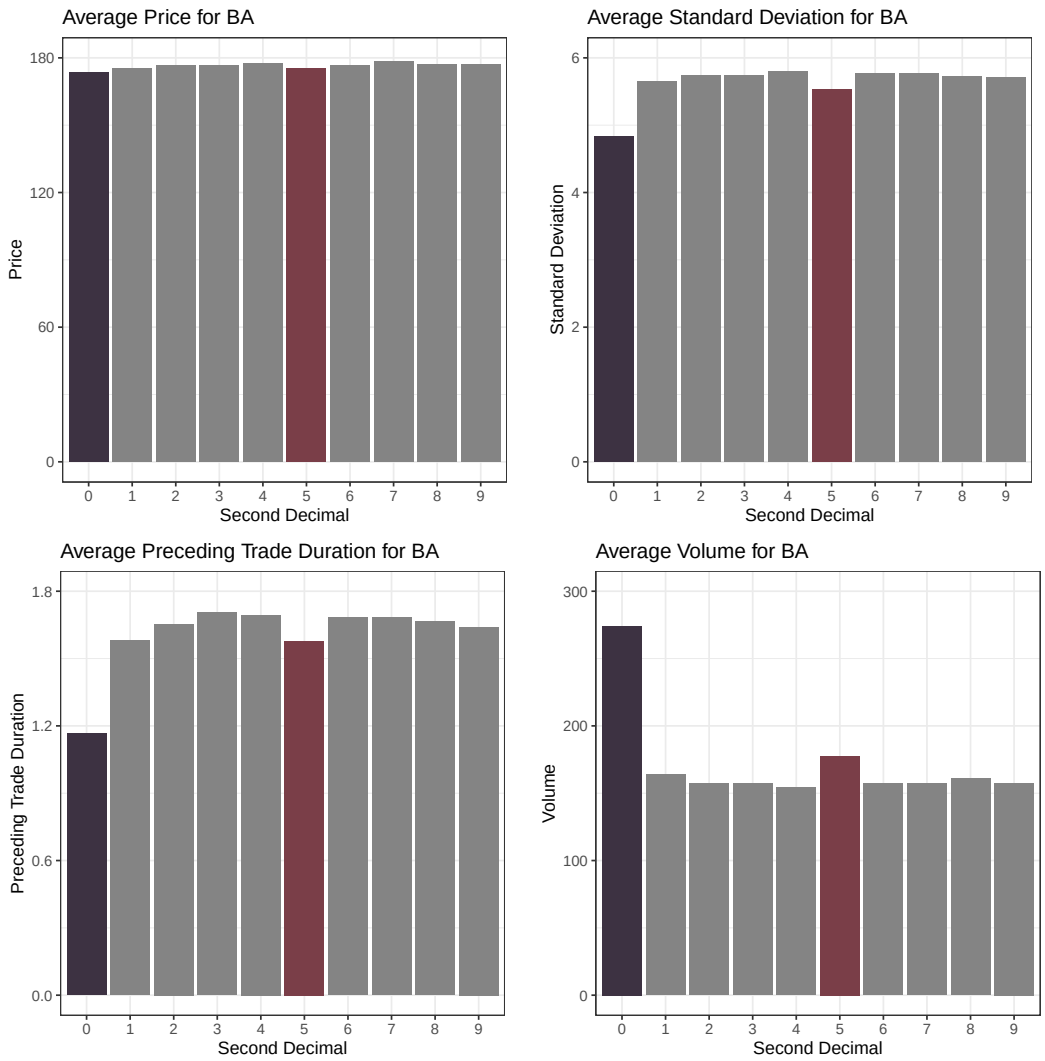

Figure 3: The average price (top left), the average standard deviation (top right), the average preceding trade duration (bottom left), and the average volume (bottom right) broken down by the second decimal digit of the BA stock prices.

Table 2: The average log-likelihood and AIC of the model without price clustering, the model with static price clustering, and the model with dynamic price clustering.

	No PC		Static PC		Dynamic PC	
Stock	Lik.	AIC	Lik.	AIC	Lik.	AIC
AAPL	-2.3548	12 080 434	-2.3292	11 949 022	-2.2953	11 775 068
AXP	-2.5117	2 284 414	-2.5054	2 278 659	-2.5008	2 274 477
BA	-2.8677	10 745 431	-2.8146	10 546 574	-2.7980	10 484 266
CAT	-2.6144	2 092 986	-2.6090	2 088 625	-2.6066	2 086 695
CSCO	-0.4650	1 525 535	-0.4636	1 520 852	-0.4606	1 510 984
CVX	-2.1513	4 094 545	-2.1476	4 087 384	-2.1444	4 081 369
DIS	-2.1382	5 620 020	-2.1295	5 597 162	-2.1231	5 580 359
DOW	-1.6701	1 981 601	-1.6685	1 979 699	-1.6668	1 977 626
GS	-3.1560	2 292 436	-3.1500	2 288 064	-3.1460	2 285 185
HD	-3.1018	3 043 787	-3.0969	3 039 026	-3.0935	3 035 655
IBM	-2.4432	2 689 846	-2.4393	2 685 501	-2.4360	2 681 876
INTC	-0.8455	3 165 039	-0.8442	3 160 026	-0.8407	3 147 091
JNJ	-2.3836	3 976 323	-2.3809	3 971 919	-2.3794	3 969 402
JPM	-2.0310	5 831 271	-2.0264	5 817 872	-2.0220	5 805 430
KO	-1.2238	2 758 822	-1.2230	2 757 081	-1.2212	2 753 113
MCD	-2.9611	2 787 186	-2.9563	2 782 675	-2.9526	2 779 159
MMM	-2.7229	2 356 746	-2.7180	2 352 455	-2.7149	2 349 844
MRK	-1.7024	3 763 397	-1.7011	3 760 622	-1.6998	3 757 820
MSFT	-1.7431	10 325 970	-1.7315	10 257 392	-1.7096	10 127 509
NKE	-2.1331	2 878 657	-2.1310	2 875 789	-2.1292	2 873 371
PFE	-0.8423	2 403 152	-0.8415	2 400 862	-0.8398	2 396 048
PG	-2.3049	3 882 847	-2.3024	3 878 712	-2.3007	3 875 756
RTX	-1.7279	1 602 691	-1.7238	1 598 954	-1.7207	1 596 096
TRV	-2.7871	1 135 714	-2.7861	1 135 292	-2.7848	1 134 797
UNH	-3.3882	3 716 005	-3.3834	3 710 792	-3.3816	3 708 786
V	-2.6542	5 058 066	-2.6485	5 047 243	-2.6434	5 037 606
VZ	-1.3165	3 025 851	-1.3156	3 023 684	-1.3141	3 020 201
WBA	-1.1807	1 507 681	-1.1787	1 505 148	-1.1748	1 500 227
WMT	-2.1291	3 718 970	-2.1262	3 714 030	-2.1233	3 708 967
XOM	-1.1969	4 893 716	-1.1951	4 886 388	-1.1895	4 863 486

Table 3: Estimated coefficients of the proposed dynamic price clustering model. Specifically, c is the constant parameter, b is the autoregressive parameter, a is the score parameter, and d is the duration parameter, which enter the dynamic equation for the dispersion parameter α_t. Further, f is the autoregressive parameter, g_1 is the parameter for the logarithm of the expected price, g_2 is the parameter for the logarithm of the variance of the price process $\mu_t e^{-\alpha_t}$, g_3 is the parameter for the logarithm of the preceding trade duration, and g_4 is the parameter for the logarithm of the volume v_t, which enter the process for the portions of trader types.

Stock	c	b	a	d	f	g_1	g_2	g_3	g_4
AAPL	6.09	0.08	0.29	-0.40	0.66	-0.33	0.61	-0.11	-0.69
AXP	5.32	0.10	0.35	-0.28	0.34	1.05	0.05	0.06	-0.74
BA	5.00	0.09	0.30	-0.29	0.39	-0.14	0.18	0.03	-0.71
CAT	5.60	0.04	0.26	-0.27	0.25	-0.28	0.29	-0.00	-0.71
CSCO	6.11	0.17	0.08	-0.35	0.81	0.28	0.40	-0.08	-0.37
CVX	5.79	0.11	0.35	-0.26	0.36	0.70	0.28	0.02	-0.64
DIS	6.05	0.12	0.33	-0.25	0.39	0.45	0.21	0.05	-0.64
DOW	5.71	0.15	0.33	-0.22	0.34	0.55	0.18	0.04	-0.70
GS	4.89	0.05	0.27	-0.31	0.29	1.06	0.12	0.02	-0.90
HD	5.01	0.08	0.29	-0.28	0.29	0.64	0.10	0.05	-0.83
IBM	5.79	0.08	0.32	-0.28	0.32	0.33	0.10	0.06	-0.80
INTC	6.29	0.14	0.13	-0.35	0.74	0.38	0.34	-0.06	-0.63
JNJ	5.76	0.13	0.40	-0.25	0.23	1.12	0.08	0.05	-0.74
JPM	5.76	0.17	0.44	-0.26	0.33	0.48	0.67	-0.07	-0.56
KO	5.97	0.25	0.33	-0.18	0.55	-0.05	0.62	0.01	-0.43
MCD	5.16	0.08	0.34	-0.27	0.31	1.69	0.02	0.06	-0.88
MMM	5.55	0.05	0.29	-0.27	0.25	0.93	0.06	0.06	-0.87
MRK	5.89	0.20	0.40	-0.23	0.45	0.20	0.71	-0.09	-0.53
MSFT	6.37	0.11	0.27	-0.38	0.72	0.17	0.39	-0.05	-0.72
NKE	6.01	0.10	0.35	-0.25	0.29	0.45	0.05	0.08	-0.70
PFE	5.19	0.38	0.28	-0.15	0.70	0.36	0.34	0.03	-0.33
PG	5.84	0.11	0.40	-0.23	0.41	-0.13	1.35	-0.25	-0.53
RTX	6.12	0.15	0.34	-0.24	0.32	-0.50	0.02	0.10	-0.65
TRV	5.07	0.05	0.30	-0.30	0.96	-0.66	0.57	-0.14	-0.02
UNH	4.58	0.07	0.28	-0.27	0.39	0.08	0.11	0.03	-0.73
V	5.91	0.06	0.32	-0.27	0.32	0.65	0.09	0.06	-0.91
VZ	5.79	0.27	0.38	-0.19	0.44	2.53	0.71	-0.05	-0.53
WBA	5.82	0.12	0.18	-0.34	0.68	0.38	0.26	-0.00	-0.62
WMT	6.19	0.12	0.38	-0.24	0.34	-2.33	0.15	0.08	-0.71
XOM	6.11	0.24	0.31	-0.18	0.57	-0.51	0.98	-0.06	-0.52

Table 4: The average values of the time-varying parameters of the proposed dynamic price clustering model, where μ_t is the location parameter for the price process and h_1, h_5 and h_{10} are parameters capturing representation of 1, 5, and 10 trader types. Values of $\bar{\mu}$ are in dollars and values of $\bar{\varphi}_1$, $\bar{\varphi}_5$, and $\bar{\varphi}_{10}$ are in percent.

Stock	$\bar{\mu}$	$\bar{\alpha}$	$\bar{\varphi}_1$	$\bar{\varphi}_5$	$\bar{\varphi}_{10}$
AAPL	291.80	8.43	92.37	0.72	6.91
AXP	98.43	7.01	95.05	1.81	3.15
BA	175.80	6.85	85.84	3.63	10.54
CAT	117.91	6.99	95.40	1.85	2.75
CSCO	42.22	10.26	98.07	0.54	1.38
CVX	88.26	7.62	96.26	1.32	2.42
DIS	112.50	7.90	94.18	2.25	3.56
DOW	36.46	7.69	97.36	1.36	1.29
GS	193.14	6.40	95.49	1.06	3.45
HD	215.44	6.63	95.72	1.44	2.84
IBM	124.05	7.38	96.01	1.64	2.34
INTC	57.72	9.82	97.83	0.65	1.52
JNJ	140.37	7.63	96.78	1.36	1.85
JPM	103.65	8.03	95.47	2.54	1.98
KO	48.33	8.88	98.29	0.67	1.04
MCD	183.86	6.74	96.12	0.72	3.16
MMM	149.92	7.02	95.87	1.17	2.96
MRK	79.28	8.43	97.69	1.23	1.08
MSFT	169.29	9.11	94.57	0.84	4.59
NKE	88.58	7.67	96.89	1.60	1.51
PFE	34.46	9.30	98.07	1.01	0.91
PG	116.41	7.61	96.82	1.87	1.31
RTX	62.28	8.13	95.83	2.06	2.11
TRV	111.57	6.59	99.18	0.34	0.48
UNH	271.73	6.29	96.01	1.08	2.91
V	180.13	7.34	95.59	1.03	3.37
VZ	55.60	8.85	97.90	1.13	0.97
WBA	45.73	8.91	97.49	0.72	1.79
WMT	118.65	7.98	96.57	1.56	1.87
XOM	45.82	8.86	97.02	1.54	1.44

Table 5: The average log-likelihood and AIC of the GARCH model based on prices, price differences, and logarithmic returns.

	GARCH Prices		GARCH Diff.		GARCH Returns	
Stock	Lik.	AIC	Lik.	AIC	Lik.	AIC
AAPL	-2.5746	13 208 099	-5.2727	27 049 463	-4.3884	22 512 972
AXP	-2.4059	2 188 119	-2.3377	2 126 170	-2.3763	2 161 238
BA	-2.7394	10 264 832	-2.6292	9 851 685	-2.6574	9 957 573
CAT	-2.5208	2 018 017	-2.4113	1 930 365	-2.4510	1 962 110
CSCO	-1.5249	5 002 357	-2.0551	6 741 593	-2.1396	7 018 554
CVX	-2.0584	3 917 643	-2.0497	3 901 118	-2.0949	3 987 076
DIS	-2.0628	5 421 767	-2.0354	5 349 755	-2.0673	5 433 680
DOW	-1.6886	2 003 496	-4.5152	5 357 203	-2.3970	2 844 000
GS	-3.0694	2 229 509	-2.9351	2 131 952	-2.9922	2 173 438
HD	-2.9571	2 901 792	-2.8380	2 784 909	-2.8896	2 835 555
IBM	-2.3741	2 613 784	-2.2938	2 525 361	-2.3335	2 569 071
INTC	-1.9481	7 292 699	-2.6590	9 953 934	-2.6006	9 735 148
JNJ	-2.2352	3 728 754	-2.1734	3 625 695	-2.2070	3 681 711
JPM	-1.9525	5 605 765	-1.9862	5 702 612	-1.9881	5 707 947
KO	-1.9314	4 354 157	-3.8296	8 633 353	-3.9745	8 959 837
MCD	-2.8049	2 640 119	-2.6921	2 534 034	-2.7435	2 582 349
MMM	-2.6323	2 278 283	-2.5214	2 182 343	-2.5606	2 216 259
MRK	-1.7283	3 820 785	-4.2919	9 488 111	-2.4060	5 318 993
MSFT	-2.5715	15 233 489	-4.0619	24 062 626	-3.6716	21 750 487
NKE	-2.0372	2 749 121	-2.0211	2 727 453	-2.0467	2 761 959
PFE	-2.1853	6 235 010	-3.4096	9 728 002	-4.0341	11 509 851
PG	-2.1326	3 592 595	-2.0727	3 491 644	-2.1039	3 544 311
RTX	-1.7441	1 617 736	-2.7315	2 533 649	-2.0640	1 914 459
TRV	-2.6650	1 085 969	-2.5548	1 041 069	-2.6048	1 061 427
UNH	-3.2196	3 531 152	-3.0794	3 377 429	-3.1228	3 425 016
V	-2.5259	4 813 622	-2.4277	4 626 415	-2.4777	4 721 700
VZ	-1.9834	4 558 623	-3.8440	8 834 862	-6.5427	15 037 233
WBA	-2.3269	2 971 343	-3.0885	3 943 860	-3.2108	4 100 059
WMT	-2.0344	3 553 640	-2.0067	3 505 265	-2.0152	3 520 046
XOM	-1.9437	7 947 244	-4.0050	16 375 114	-4.0749	16 661 060

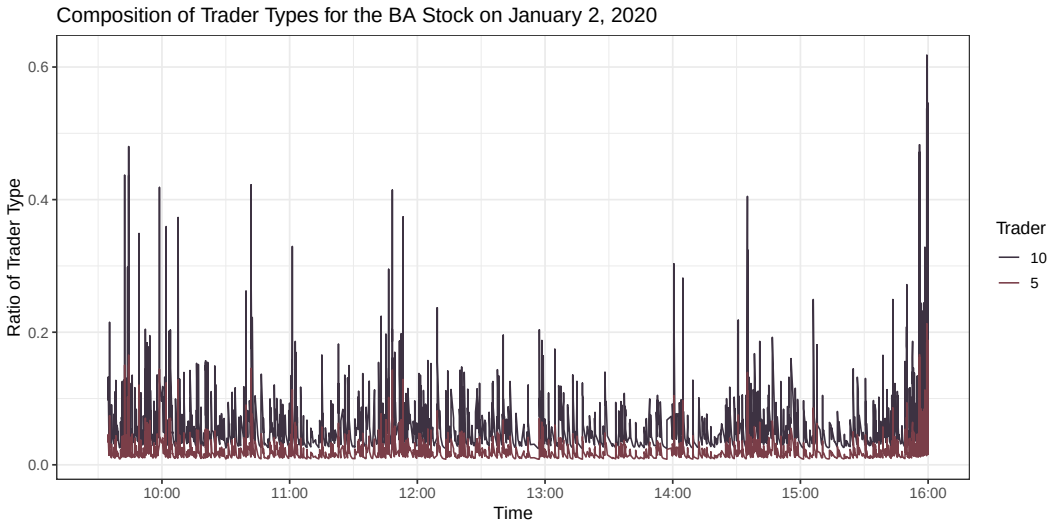

Figure 4: The time-varying portions of trader types obtained from the proposed price clustering model for the BA stock on January 2, 2020.

hypotheses is out of the scope of our paper. In this section, we focus only on the most studied hypothesis in the literature – the price resolution hypothesis.

The *price resolution hypothesis* of Ball *et al.* (1985) considers the source of price clustering to be the uncertainty. It states that when the amount of information in the market is low and the volatility becomes higher, the market participants incline to round their prices, and consequently, the price clustering increases. This hypothesis was confirmed by many studies. The studies found that price clustering increases with volatility using different data and measures. For example, Ahn *et al.* (2005) computed the volatility as the inverse of the daily return standard deviation, while Ikenberry and Weston (2008) used the standard deviation of returns over the sample period, Box and Griffith (2016) used the standard deviation of 15-minute continuously compounded midpoint returns over the trading day, Schwartz *et al.* (2004) used the difference in the high and low prices for the day, and Lien *et al.* (2019) utilized the transitory volatility defined as the coefficient of variation of intraday trade prices. Davis *et al.* (2014) found that price clustering is positively related to volatility, however, only when a non–high-frequency trading firm provides liquidity. On the contrary to the vast majority of the literature, Blau (2019) reported based on panel regressions that the volatility is negatively related to price clustering, where the volatility is measured as the standard deviation of residual returns obtained from estimating a Fama and French 3-factor model.

Our results from the daily analysis show that the realized volatility is highly significant and positively related to the price clustering. This finding is in line with the price resolution hypothesis. Interestingly, instantaneous volatility obtained from the proposed dynamic price model has a negative effect on price clustering. The results do not contradict since they explore price clustering from different perspectives. The result based on daily data holds for low-frequency traders whose price resolution is influenced by the uncertainty in a negative way, i.e. the higher daily volatility, the higher price clustering. On the other hand, the presence of high-frequency traders is typically associated with increased volatility (see, e.g., Roşu, 2019; Shkilko and Sokolov, 2020; Boehmer *et al.*, 2021). Moreover, high-frequency traders generally do not incline to price rounding (see Davis *et al.*, 2014). Consequently, the higher the instantaneous volatility is, the higher portion of high-frequency traders is, which lowers the price clustering.

5 Conclusion

We have proposed a dynamic price model to capture agents trading in different multiples of the tick size. In the literature, this empirical phenomenon known as price clustering was mostly approached only by basic descriptive statistics rather than a proper price model. By analyzing 30 DJIA stocks from both daily and high-frequency perspectives, we have revealed dissension between the two time scales. While daily realized volatility has a positive effect on price clustering, instantaneous volatility obtained by the proposed model has a negative effect. We argue that volatility on lower frequency affects low-frequency traders through the resolution hypothesis while volatility on higher frequency affects only high-frequency traders who do not tend to price clustering.

Our price model brings several advantages over existing approaches to price clustering since it operates at the highest possible frequency. Consequently, (i) it allows practitioners and academics to study price clustering from a new perspective, and (ii) it can be seen as an extension or a new component for high-frequency price modeling. For example, our price model can be used for simulations of intra-day price processes, which typically serve for optimization of trading strategies, assessment and calibration of models, asset pricing, a decision-making process of investments and hedging strategies, pricing derivatives and other financial instruments. We show that price clustering is not negligible since the excess relative frequency of multiples of five and ten cents can make up to 11.52 percent in our dataset. Consequently, neglecting the price clustering phenomenon in the simulations of the price process can have serious consequences for the subsequent usage resulting in unexpected losses caused by, e.g., biased estimates for prices of derivatives, a sub-optimal choice of trading or hedging strategies and models. Our high-frequency price model can also help academics and practitioners to investigate and understand the price clustering phenomenon as such. Finally, high-frequency traders and market makers can directly incorporate our approach into their price models to take into account this phenomenon.

We believe the model to be sufficient for its purpose – capturing price clustering and allowing to explain it. For the model to be able to compete with other high-frequency price models, however, it would have to be improved. The main limitation lies in the underlying distribution. We have to study how well the double Poisson distribution, which we have used, captures the observed prices in more detail. However, due to our specific problem, we require the distribution to be defined on positive integers and allow for underdispersion. The range of possible alternatives is therefore severely limited as it is not a typical situation in count data analysis. Furthermore, the specification of the dynamics could be enhanced. We could include a separate model for durations and we could add a seasonality component to the volatility. One possible direction for the future research is therefore to comprehensively assess the suitability of the double Poisson distribution for prices, extend the specification of the proposed dynamic model, and compare it with various models for price differences.

Concerning the empirical study, our focus has been on the price variance, whether it is daily realized volatility or instantaneous variance. Nevertheless, we have also included the expected price, the preceding trade duration, and the volume as control variables. These are the most common variables in the price clustering literature. However, other factors such as the spread and the investor sentiment could also be considered. In the context of the proposed high-frequency price model, any variable could be straightforwardly included in the price clustering dynamics. Price clustering at levels other than 5 and 10 cents can be added to the model as well. Analyzing the effects of these factors and price clustering levels is the second possible direction of the future research.

Acknowledgements

Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

Funding

This research was supported by the Czech Science Foundation under project 19-02773S, the Internal Grant Agency of the Prague University of Economics and Business under project F4/53/2019, and the Institutional Support Funds for the long-term conceptual development of the Faculty of Informatics, Prague University of Economics and Business.

References

Ahn HJ, Cai J, Cheung YL (2005). "Price Clustering on the Limit-Order Book: Evidence from the Stock Exchange of Hong Kong." *Journal of Financial Markets*, **8**(4), 421–451. ISSN 1386-4181. https://doi.org/10.1016/j.finmar.2005.07.001.

Aitken M, Brown P, Buckland C, Izan HY, Walter T (1996). "Price Clustering on the Australian Stock Exchange." *Pacific-Basin Finance Journal*, **4**(2-3), 297–314. ISSN 0927-538X. https://doi.org/10.1016/0927-538x(96)00016-9.

Alexander GJ, Peterson MA (2007). "An Analysis of Trade-Size Clustering and Its Relation to Stealth Trading." *Journal of Financial Economics*, **84**(2), 435–471. ISSN 0304-405X. https://doi.org/10.1016/j.jfineco.2006.02.005.

ap Gwilym O, Clare A, Thomas S (1998). "Extreme Price Clustering in the London Equity Index Futures and Options Markets." *Journal of Banking & Finance*, **22**(9), 1193–1206. ISSN 0378-4266. https://doi.org/10.1016/S0378-4266(98)00054-5.

ap Gwilym O, Verousis T (2013). "Price Clustering in Individual Equity Options: Moneyness, Maturity, and Price Level." *Journal of Futures Markets*, **33**(1), 55–76. ISSN 0270-7314. https://doi.org/10.1002/fut.21547.

Baig A, Blau BM, Sabah N (2019). "Price Clustering and Sentiment in Bitcoin." *Finance Research Letters*, **29**, 111–116. ISSN 1544-6123. https://doi.org/10.1016/j.frl.2019.03.013.

Ball CA, Torous WN, Tschoegl AE (1985). "The Degree of Price Resolution: The Case of the Gold Market." *Journal of Futures Markets*, **5**(1), 29–43. ISSN 0270-7314. https://doi.org/10.1002/fut.3990050105.

Barndorff-Nielsen OE, Hansen PR, Lunde A, Shephard N (2008). "Designing Realized Kernels to Measure the ex post Variation of Equity Prices in the Presence of Noise." *Econometrica*, **76**(6), 1481–1536. ISSN 0012-9682. https://doi.org/10.3982/ecta6495.

Barndorff-Nielsen OE, Hansen PR, Lunde A, Shephard N (2009). "Realized Kernels in Practice: Trades and Quotes." *Econometrics Journal*, **12**(3), 1–32. ISSN 1368-4221. https://doi.org/10.1111/j.1368-423X.2008.00275.x.

Barndorff-Nielsen OE, Pollard DG, Shephard N (2012). "Integer-Valued Lévy Processes and Low Latency Financial Econometrics." *Quantitative Finance*, **12**(4), 587–605. ISSN 1469-7688. https://doi.org/10.1080/14697688.2012.664935.

Bharati R, Crain SJ, Kaminski V (2012). "Clustering in Crude Oil Prices and the Target Pricing Zone Hypothesis." *Energy Economics*, **34**(4), 1115–1123. ISSN 0140-9883. https://doi.org/10.1016/j.eneco.2011.09.009.

Blasques F, Holý V, Tomanová P (2022). "Zero-Inflated Autoregressive Conditional Duration Model for Discrete Trade Durations with Excessive Zeros." https://arxiv.org/abs/1812.07318.

Blasques F, Koopman SJ, Lucas A (2015). "Information-Theoretic Optimality of Observation-Driven Time Series Models for Continuous Responses." *Biometrika*, **102**(2), 325–343. ISSN 0006-3444. https://doi.org/10.1093/biomet/asu076.

Blau BM (2019). "Price Clustering and Investor Sentiment." *Journal of Behavioral Finance*, **20**(1), 19–30. ISSN 1542-7560. https://doi.org/10.1080/15427560.2018.1431887.

Blau BM, Griffith TG (2016). "Price Clustering and the Stability of Stock Prices." *Journal of Business Research*, **69**(10), 3933–3942. ISSN 0148-2963. https://doi.org/10.1016/j.jbusres.2016.06.008.

Boehmer E, Fong K, Wu J (2021). "Algorithmic Trading and Market Quality: International Evidence." *Journal of Financial and Quantitative Analysis*, **56**(8), 2659–2688. ISSN 0022-1090. https://doi.org/10.1017/s0022109020000782.

Bollerslev T (1986). "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics*, **31**(3), 307–327. ISSN 0304-4076. https://doi.org/10.1016/0304-4076(86)90063-1.

Bollerslev T (1987). "A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return." *Review of Economics and Statistics*, **69**(3), 542–547. ISSN 0034-6535. https://doi.org/10.2307/1925546.

Bourguignon M, Rodrigues J, Santos-Neto M (2019). "Extended Poisson INAR(1) Processes with Equidispersion, Underdispersion and Overdispersion." *Journal of Applied Statistics*, **46**(1), 101–118. ISSN 0266-4763. https://doi.org/10.1080/02664763.2018.1458216.

Box T, Griffith T (2016). "Price Clustering Asymmetries in Limit Order Flows." *Financial Management*, **45**(4), 1041–1066. ISSN 0046-3892. https://doi.org/10.1111/fima.12136.

Brent RP (1972). *Algorithms for Minimization Without Derivatives*. Prentice-Hall, Englewood Cliffs. ISBN 978-0-13-022335-7. https://books.google.com/books?id=Ee5QAAAAMAAJ.

Brooks R, Harris E, Joymungul Y (2013). "Price Clustering in Australian Water Markets." *Applied Economics*, **45**(6), 677–685. ISSN 0003-6846. https://doi.org/10.1080/00036846.2011.610747.

Brown P, Mitchell J (2008). "Culture and Stock Price Clustering: Evidence from The Peoples' Republic of China." *Pacific-Basin Finance Journal*, **16**(1-2), 95–120. ISSN 0927-538X. https://doi.org/10.1016/j.pacfin.2007.04.005.

Buccheri G, Bormetti G, Corsi F, Lillo F (2021). "A Score-Driven Conditional Correlation Model for Noisy and Asynchronous Data: An Application to High-Frequency Covariance Dynamics." *Journal of Business & Economic Statistics*, **39**(4), 920–936. ISSN 0735-0015. https://doi.org/10.1080/07350015.2020.1739530.

Capelle-Blancard G, Chaudhury M (2007). "Price Clustering in the CAC 40 Index Options Market." *Applied Financial Economics*, **17**(15), 1201–1210. ISSN 0960-3107. https://doi.org/10.1080/09603100600949218.

Chiao C, Wang ZM (2009). "Price Clustering: Evidence Using Comprehensive Limit-Order Data." *Financial Review*, **44**(1), 1–29. ISSN 0732-8516. https://doi.org/10.1111/j.1540-6288.2008.00208.x.

Christie WG, Schultz PH (1994). "Why do NASDAQ Market Makers Avoid Odd-Eighth Quotes?" *The Journal of Finance*, **49**(5), 1813–1840. ISSN 0022-1082. https://doi.org/10.1111/j.1540-6261.1994.tb04782.x.

Chung H, Chiang S (2006). "Price Clustering in E-Mini and Floor-Traded Index Futures." *Journal of Futures Markets*, **26**(3), 269–295. ISSN 0270-7314. https://doi.org/10.1002/fut.20196.

Chung KH, Kim KA, Kitsabunnarat P (2005). "Liquidity and Quote Clustering in a Market with Multiple Tick Sizes." *Journal of Financial Research*, **28**(2), 177–195. ISSN 0270-2592. https://doi.org/10.1111/j.1475-6803.2005.00120.x.

Chung KH, Van Ness BF, Van Ness RA (2004). "Trading Costs and Quote Clustering on the NYSE and NASDAQ after Decimalization." *Journal of Financial Research*, **27**(3), 309–328. ISSN 0270-2592. https://doi.org/10.1111/j.1475-6803.2004.00096.x.

Cooney JW, Van Ness B, Van Ness R (2003). "Do Investors Prefer Even-Eighth Prices? Evidence from NYSE Limit Orders." *Journal of Banking & Finance*, **27**(4), 719–748. ISSN 0378-4266. https://doi.org/10.1016/S0378-4266(01)00262-X.

Creal D, Koopman SJ, Lucas A (2013). "Generalized Autoregressive Score Models with Applications." *Journal of Applied Econometrics*, **28**(5), 777–795. ISSN 0883-7252. https://doi.org/10.1002/jae.1279.

Das S, Kadapakkam PR (2020). "Machine over Mind? Stock Price Clustering in the Era of Algorithmic Trading." *The North American Journal of Economics and Finance*, **51**, 100831/1–100831/15. ISSN 1062-9408. https://doi.org/10.1016/j.najef.2018.08.014.

Davis RL, Van Ness BF, Van Ness RA (2014). "Clustering of Trade Prices by High-Frequency and Non-High-Frequency Trading Firms." *Financial Review*, **49**(2), 421–433. ISSN 0732-8516. https://doi.org/10.1111/fire.12042.

Efron B (1986). "Double Exponential Families and Their Use in Generalized Linear Regression." *Journal of the American Statistical Association*, **81**(395), 709–721. ISSN 0162-1459. https://doi.org/10.1080/01621459.1986.10478327.

Engle RF (2000). "The Econometrics of Ultra-High-Frequency Data." *Econometrica*, **68**(1), 1–22. ISSN 0012-9682. https://doi.org/10.1111/1468-0262.00091.

Engle RF, Russell JR (1998). "Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data." *Econometrica*, **66**(5), 1127–1162. ISSN 0012-9682. https://doi.org/10.2307/2999632.

Goodhart C, Curcio R (1991). "The Clustering of Bid/Ask Prices and the Spread in the Foreign Exchange Market." https://www.fmg.ac.uk/publications/discussion-papers/clustering-bidask-prices-and-spread-foreign-exchange-market.

Gorgi P, Koopman SJ, Lit R (2019). "The Analysis and Forecasting of Tennis Matches by Using a High Dimensional Dynamic Model." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **182**(4), 1393–1409. ISSN 0964-1998. https://doi.org/10.1111/rssa.12464.

Hameed A, Terry E (1998). "The Effect of Tick Size on Price Clustering and Trading Volume." *Journal of Business Finance & Accounting*, **25**(7-8), 849–867. ISSN 0306-686X. https://doi.org/10.1111/1468-5957.00216.

Hansen PR, Lunde A (2006). "Realized Variance and Market Microstructure Noise." *Journal of Business & Economic Statistics*, **24**(2), 127–161. ISSN 0735-0015. https://doi.org/10.1198/073500106000000071.

Harris L (1991). "Stock Price Clustering and Discreteness." *Review of Financial Studies*, **4**(3), 389–415. ISSN 0893-9454. https://doi.org/10.1093/rfs/4.3.389.

Harvey A, Hurn S, Thiele S (2019). "Modeling Directional (Circular) Time Series." https://doi.org/10.17863/cam.43915.

Harvey AC (2013). *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series*. First Edition. Cambridge University Press, New York. ISBN 978-1-107-63002-4. https://doi.org/10.1017/cbo9781139540933.

Heinen A (2003). "Modelling Time Series Count Data: An Autoregressive Conditional Poisson Model." https://ssrn.com/abstract=1117187.

Holý V, Tomanová P (2023). "Streaming Approach to Quadratic Covariation Estimation Using Financial Ultra-High-Frequency Data." *Computational Economics*, **62**(1), 463–485. ISSN 0927-7099. https://doi.org/10.1007/s10614-021-10210-w.

Holý V, Zouhar J (2022). "Modelling Time-Varying Rankings with Autoregressive and Score-Driven Dynamics." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **71**(5), 1427–1450. ISSN 0035-9254. https://doi.org/10.1111/rssc.12584.

Hu B, Jiang C, McInish T, Zhou H (2017). "Price Clustering on the Shanghai Stock Exchange." *Applied Economics*, **49**(28), 2766–2778. ISSN 0003-6846. https://doi.org/10.1080/00036846.2016.1248284.

Hu B, McInish T, Miller J, Zeng L (2019). "Intraday Price Behavior of Cryptocurrencies." *Finance Research Letters*, **28**, 337–342. ISSN 1544-6123. https://doi.org/10.1016/j.frl.2018.06.002.

Ikenberry DL, Weston JP (2008). "Clustering in US Stock Prices After Decimalisation." *European Financial Management*, **14**(1), 30–54. ISSN 1354-7798. https://doi.org/10.1111/j.1468-036X.2007.00410.x.

Kahn C, Pennacchi G, Sopranzetti B (1999). "Bank Deposit Rate Clustering: Theory and Empirical Evidence." *The Journal of Finance*, **54**(6), 2185–2214. ISSN 0022-1082. https://doi.org/10.1111/0022-1082.00185.

Kandel S, Sarig O, Wohl A (2001). "Do Investors Prefer Round Stock Prices? Evidence from Israeli IPO Auctions." *Journal of Banking & Finance*, **25**(8), 1543–1551. ISSN 0378-4266. https://doi.org/10.1016/s0378-4266(00)00131-x.

Koopman SJ, Lit R, Lucas A (2017). "Intraday Stochastic Volatility in Discrete Price Changes: The Dynamic Skellam Model." *Journal of the American Statistical Association*, **112**(520), 1490–1503. ISSN 0162-1459. https://doi.org/10.1080/01621459.2017.1302878.

Koopman SJ, Lit R, Lucas A, Opschoor A (2018). "Dynamic Discrete Copula Models for High-Frequency Stock Price Changes." *Journal of Applied Econometrics*, **33**(7), 966–985. ISSN 0883-7252. https://doi.org/10.1002/jae.2645.

Li X, Li S, Xu C (2020). "Price Clustering in Bitcoin Market - An Extension." *Finance Research Letters*, **32**, 101072/1:101072/9. ISSN 1544-6123. https://doi.org/10.1016/j.frl.2018.12.020.

Lien D, Hung PH, Hung IC (2019). "Order Price Clustering, Size Clustering, and Stock Price Movements: Evidence from the Taiwan Stock Exchange." *Journal of Empirical Finance*, **52**, 149–177. ISSN 0927-5398. https://doi.org/10.1016/j.jempfin.2019.03.005.

Liu HC (2011). "Timing of Price Clustering and Trader Behavior in the Foreign Exchange Market: Evidence from Taiwan." *Journal of Economics and Finance*, **35**(2), 198–210. ISSN 1055-0925. https://doi.org/10.1007/s12197-009-9096-0.

Liu HC, Witte MD (2013). "Price Clustering in the U.S. Dollar/Taiwan Dollar Swap Market." *Financial Review*, **48**(1), 77–96. ISSN 0732-8516. https://doi.org/10.1111/j.1540-6288.2012.00353.x.

Mbanga CL (2019). "The Day-of-the-Week Pattern of Price Clustering in Bitcoin." *Applied Economics Letters*, **26**(10), 807–811. ISSN 1350-4851. https://doi.org/10.1080/13504851.2018.1497844.

Meng L, Verousis T, ap Gwilym O (2013). "A Substitution Effect Between Price Clustering and Size Clustering in Credit Default Swaps." *Journal of International Financial Markets, Institutions and Money*, **24**(1), 139–152. ISSN 1042-4431. https://doi.org/10.1016/j.intfin.2012.11.011.

Mishra AK, Tripathy T (2018). "Price and Trade Size Clustering: Evidence from the National Stock Exchange of India." *The Quarterly Review of Economics and Finance*, **68**, 63–72. ISSN 1062-9769. https://doi.org/10.1016/j.qref.2017.11.006.

Narayan PK, Smyth R (2013). "Has Political Instability Contributed to Price Clustering on Fiji's Stock Market?" *Journal of Asian Economics*, **28**, 125–130. ISSN 1049-0078. https://doi.org/10.1016/j.asieco.2013.07.002.

Niederhoffer V (1965). "Clustering of Stock Prices." *Operations Research*, **13**(2), 258–265. ISSN 0030-364X. https://doi.org/10.1287/opre.13.2.258.

Ohta W (2006). "An Analysis of Intraday Patterns in Price Clustering on the Tokyo Stock Exchange." *Journal of Banking & Finance*, **30**(3), 1023–1039. ISSN 0378-4266. https://doi.org/10.1016/j.jbankfin.2005.07.017.

Osborne MFM (1962). "Periodic Structure in the Brownian Motion of Stock Prices." *Operations Research*, **10**(3), 345–379. ISSN 0030-364X. https://doi.org/10.1287/opre.10.3.345.

Palao F, Pardo A (2012). "Assessing Price Clustering in European Carbon Markets." *Applied Energy*, **92**, 51–56. ISSN 0306-2619. https://doi.org/10.1016/j.apenergy.2011.10.022.

Robert CY, Rosenbaum M (2011). "A New Approach for the Dynamics of Ultra-High-Frequency Data: The Model with Uncertainty Zones." *Journal of Financial Econometrics*, **9**(2), 344–366. ISSN 1479-8409. https://doi.org/10.1093/jjfinec/nbq023.

Roşu I (2019). "Fast and Slow Informed Trading." *Journal of Financial Markets*, **43**, 1–30. ISSN 1386-4181. https://doi.org/10.1016/j.finmar.2019.02.003.

Russell JR, Engle RF (2005). "A Discrete-State Continuous-Time Model of Financial Transactions Prices and Times: The Autoregressive Conditional Multinomial-Autoregressive Conditional Duration Model." *Journal of Business & Economic Statistics*, **23**(2), 166–180. ISSN 0735-0015. https://doi.org/10.1198/073500104000000541.

Schwartz AL, Van Ness BF, Van Ness RA (2004). "Clustering in the Futures Market: Evidence From S&P 500 Futures Contracts." *Journal of Futures Markets*, **24**(5), 413–428. ISSN 0270-7314. https://doi.org/10.1002/fut.10129.

Sellers KF, Morris DS (2017). "Underdispersion Models: Models That Are "Under the Radar"." *Communications in Statistics - Theory and Methods*, **46**(24), 12075–12086. ISSN 0361-0926. https://doi.org/10.1080/03610926.2017.1291976.

Shephard N, Yang JJ (2017). "Continuous Time Analysis of Fleeting Discrete Price Moves." *Journal of the American Statistical Association*, **112**(519), 1090–1106. ISSN 0162-1459. https://doi.org/10.1080/01621459.2016.1192544.

Shkilko A, Sokolov K (2020). "Every Cloud Has A Silver Lining: Fast Trading, Microwave Connectivity, and Trading Costs." *The Journal of Finance*, **75**(6), 2899–2927. ISSN 0022-1082. https://doi.org/10.1111/jofi.12969.

Song S, Wang Y, Xu G (2020). "On the Probability of Default in a Market with Price Clustering and Jump Risk." *Mathematics and Financial Economics*, **14**(2), 225–247. ISSN 1862-9679. https://doi.org/10.1007/s11579-019-00253-x.

Sopranzetti BJ, Datar V (2002). "Price Clustering in Foreign Exchange Spot Markets." *Journal of Financial Markets*, **5**(4), 411–417. ISSN 1386-4181. https://doi.org/10.1016/S1386-4181(01)00032-5.

Urquhart A (2017). "Price Clustering in Bitcoin." *Economics Letters*, **159**, 145–148. ISSN 0165-1765. https://doi.org/10.1016/j.econlet.2017.07.035.

Verousis T, ap Gwilym O (2013). "Trade Size Clustering and the Cost of Trading at the London Stock Exchange." *International Review of Financial Analysis*, **27**, 91–102. ISSN 1057-5219. https://doi.org/10.1016/j.irfa.2012.08.007.

Xu HY, Xie M, Goh TN, Fu X (2012). "A Model for Integer-Valued Time Series with Conditional Overdispersion." *Computational Statistics & Data Analysis*, **56**(12), 4229–4242. ISSN 0167-9473. https://doi.org/10.1016/j.csda.2012.04.011.

Zou Y, Geedipally SR, Lord D (2013). "Evaluating the Double Poisson Generalized Linear Model." *Accident Analysis and Prevention*, **59**, 497–505. ISSN 0001-4575. https://doi.org/10.1016/j.aap.2013.07.017.

A Derivation of Distribution for Specific Trader Types

Let there be m types of traders that can trade only in $k_1, \ldots, k_m$ multiples of the tick size respectively. For trader type $k \in \{k_1, \ldots, k_m\}$, we derive the distribution of prices $\mathrm{P}\left[Y^{[k]} = y \middle| \mu, \alpha\right]$. We require the distribution to be based on the double Poisson distribution, to have the support consisting of multiples of k, to have the expected value $\mathrm{E}\left[Y^{[k]}\right] \simeq \mu$ and to have the variance $\mathrm{var}\left[Y^{[k]}\right] \simeq \mu e^{-\alpha}$. We can modify any integer distribution $\mathrm{P}\left[Z^{[k]} = y \middle| \mu, \alpha\right]$ to have support consisting only of multiples of k as

$$\mathrm{P}\left[Y^{[k]} = y \middle| \mu, \alpha\right] = \mathbb{I}\left\{k \mid y\right\} \mathrm{P}\left[Z^{[k]} = \frac{y}{k} \middle| \mu, \alpha\right], \tag{18}$$

where $\mathbb{I}\left\{k \mid y\right\}$ is equal to 1 if y is divisible by k and 0 otherwise. We assume that $Z^{[k]}$ follows the double Poisson distribution with parameters $\mu^{[k]}$ and $\alpha^{[k]}$, i.e. $Z^{[k]} \sim \mathrm{DP}\left(\mu^{[k]}, \alpha^{[k]}\right)$. The expected value of $Y^{[k]}$ is

$$\begin{aligned}
\mathrm{E}\left[Y^{[k]}\right] &= \sum_{y=0}^{\infty} y \mathrm{P}\left[Y^{[k]} = y \middle| \mu, \alpha\right] \\
&= \sum_{y=0}^{\infty} y \mathbb{I}\left\{k \mid y\right\} \mathrm{P}\left[Z^{[k]} = \frac{y}{k} \middle| \mu, \alpha\right] \\
&= \sum_{y=0}^{\infty} k y \mathrm{P}\left[Z^{[k]} = y \middle| \mu, \alpha\right] \\
&= k \mathrm{E}\left[Z^{[k]}\right] \\
&\simeq k \mu^{[k]}.
\end{aligned} \tag{19}$$

The variance of $Y^{[k]}$ is

$$\begin{aligned}
\mathrm{var}\left[Y^{[k]}\right] &= \sum_{y=0}^{\infty} \left(y - \mathrm{E}\left[Y^{[k]}\right]\right)^2 \mathrm{P}\left[Y^{[k]} = y \middle| \mu, \alpha\right] \\
&= \sum_{y=0}^{\infty} \left(y - \mathrm{E}\left[Y^{[k]}\right]\right)^2 \mathbb{I}\left\{k \mid y\right\} \mathrm{P}\left[Z^{[k]} = \frac{y}{k} \middle| \mu, \alpha\right] \\
&= \sum_{y=0}^{\infty} \left(k y - k \mathrm{E}\left[Z^{[k]}\right]\right)^2 \mathrm{P}\left[Z^{[k]} = y \middle| \mu, \alpha\right] \\
&= k^2 \mathrm{var}\left[Z^{[k]}\right] \\
&\simeq k^2 \mu^{[k]} e^{-\alpha^{[k]}}.
\end{aligned} \tag{20}$$

Our last requirements $\mathrm{E}\left[Y^{[k]}\right] \simeq \mu$ with $\mathrm{var}\left[Y^{[k]}\right] \simeq \mu e^{-\alpha}$ lead to the system of equations

$$\begin{aligned}
\mu &= k \mu^{[k]} \\
\mu e^{-\alpha} &= k^2 \mu^{[k]} e^{-\alpha^{[k]}}
\end{aligned} \tag{21}$$

with the solution

$$\mu^{[k]} = \frac{\mu}{k}, \qquad \alpha^{[k]} = \alpha + \ln(k). \tag{22}$$

Everything together gives us the distribution

$$\mathrm{P}\left[Y^{[k]} = y \middle| \mu, \alpha\right] = \mathbb{I}\{k \mid y\}\,\mathrm{P}\left[Z^{[k]} = \frac{y}{k} \middle| \mu, \alpha\right], \qquad Z^{[k]} \sim \mathrm{DP}\left(\frac{\mu}{k}, \alpha + \ln(k)\right). \tag{23}$$

Note that the mixture distribution of all prices

$$\mathrm{P}\left[Y = y | \mu, \alpha, \varphi_{k_1}, \ldots, \varphi_{k_m}\right] = \sum_{k \in \{k_1, \ldots, k_m\}} \varphi_k \mathrm{P}\left[Y^{[k]} = y \middle| \mu, \alpha\right] \tag{24}$$

has approximately the same expected value and variance as the distribution of $Y^{[k]}$. This is based on the identity

$$\begin{aligned}
\mathrm{E}\left[g(Y)\right] &= \sum_{y=0}^{\infty} g(y) \mathrm{P}\left[Y = y | \mu, \alpha, \varphi_{k_1}, \ldots, \varphi_{k_m}\right] \\
&= \sum_{y=0}^{\infty} g(y) \sum_{k \in \{k_1, \ldots, k_m\}} \varphi_k \mathrm{P}\left[Y^{[k]} = y \middle| \mu, \alpha\right] \\
&= \sum_{k \in \{k_1, \ldots, k_m\}} \varphi_k \sum_{y=0}^{\infty} g(y) \mathrm{P}\left[Y^{[k]} = y \middle| \mu, \alpha\right] \\
&= \sum_{k \in \{k_1, \ldots, k_m\}} \varphi_k \mathrm{E}\left[g\left(Y^{[k]}\right)\right] \\
&= \mathrm{E}\left[g\left(Y^{[k]}\right)\right],
\end{aligned} \tag{25}$$

where $g(\cdot)$ is any function satisfying that $\mathrm{E}\left[g\left(Y^{[k]}\right)\right]$ are the same for all k.

B Descriptive Statistics of Cleaned Data

Descriptive statistics are given in Table 6.

Table 6: The table reports a number of observations (#Trades), sample mean (Mean P) and standard deviation (SD P) of prices, sample mean (Mean D) and standard deviation (SD D) of durations, and price clustering (PC) calculated as the excess relative frequency of multiples of five cents and ten cents in prices.

Stock	#Trades	Mean P	SD P	Mean D	SD D	PC [%]
AAPL	2 671 590	291.71	34.67	1.10	2.71	8.28
AXP	467 623	98.59	17.84	6.24	13.30	4.02
BA	1 886 402	176.06	60.40	1.55	5.01	11.52
CAT	413 148	118.07	14.32	7.08	15.25	3.79
CSCO	1 712 341	42.26	4.14	1.71	5.54	2.08
CVX	964 508	88.34	15.87	3.03	6.42	3.03
DIS	1 327 041	112.56	16.38	2.20	4.67	4.80
DOW	606 116	36.53	8.33	4.82	10.44	2.24
GS	376 058	193.37	31.28	7.78	16.89	3.53
HD	503 522	215.65	28.69	5.80	12.12	3.57
IBM	563 345	124.12	15.30	5.19	10.60	3.34
INTC	2 065 813	57.84	5.53	1.42	4.22	2.13
JNJ	846 988	140.42	9.87	3.45	7.14	2.68
JPM	1 448 425	103.69	17.38	2.02	4.30	3.71
KO	1 140 050	48.35	5.78	2.56	6.63	1.45
MCD	483 508	184.02	20.76	6.05	12.52	3.16
MMM	445 633	150.04	14.01	6.56	13.99	3.51
MRK	1 118 215	79.30	5.40	2.61	5.55	1.94
MSFT	3 099 279	169.25	16.27	0.94	2.40	5.96
NKE	687 619	88.66	11.31	4.25	8.78	2.50
PFE	1 439 433	34.48	3.06	2.03	5.73	1.56
PG	855 186	116.44	6.85	3.41	7.14	2.79
RTX	470 061	62.28	4.87	3.04	5.49	3.37
TRV	216 618	111.75	17.08	13.49	29.25	1.84
UNH	561 256	271.93	26.72	5.21	12.16	3.39
V	965 718	180.20	18.36	3.03	6.07	3.68
VZ	1 162 040	55.61	2.60	2.52	6.15	1.62
WBA	668 123	45.78	4.87	4.38	10.54	2.22
WMT	886 255	118.66	6.14	3.30	6.82	2.85
XOM	2 057 230	45.84	9.29	1.42	3.43	2.25

An Intraday GARCH Model for Discrete Price Changes and Irregularly Spaced Observations

Vladimír Holý
Prague University of Economics and Business
Winston Churchill Square 1938/4, 130 67 Prague 3, Czechia
vladimir.holy@vse.cz

Abstract: We develop a novel observation-driven model for high-frequency prices. We account for irregularly spaced observations, simultaneous transactions, discreteness of prices, and market microstructure noise. The relation between trade durations and price volatility, as well as intraday patterns of trade durations and price volatility, is captured using smoothing splines. The dynamic model is based on the zero-inflated Skellam distribution with time-varying volatility in a score-driven framework. Market microstructure noise is filtered by including a moving average component. The model is estimated by the maximum likelihood method. In an empirical study of the IBM stock, we demonstrate that the model provides a good fit to the data. Besides modeling intraday volatility, it can also be used to measure daily realized volatility.

Keywords: Ultra-High-Frequency Data, Trade Duration, Price Volatility, UHF-GARCH Model, Score-Driven Model, Skellam Distribution.

JEL Classification: C22, C41, C58, G12.

1 Introduction

Modeling intraday volatility presents several challenges in contrast to modeling volatility at the daily level as high-frequency data have distinct characteristics. A widely used tool for modeling daily volatility is the class of generalized autoregressive conditional heteroskedasticity (GARCH) models with seminal contributions by Engle (1982), Bollerslev (1986, 1987), and Nelson (1991). A variety of intraday GARCH models extending daily models therefore emerged, following the call for research in this direction by Engle (2002). In this paper, we focus on the following four characteristics of high-frequency prices in the context of intraday GARCH models:

Irregularly spaced observations. Engle (2000) coined the term ultra-high-frequency (UHF) data, which refer to records of every transaction made resulting in irregularly spaced observations. Such data require special treatment in econometric modeling. Engle and Russell (1998) proposed to model times between successive transactions, also known as trade durations, by the autoregressive conditional duration (ACD) model. Furthermore, Engle (2000) proposed to model the variance per time unit using irregularly spaced observations by the UHF-GARCH model. Ghysels and Jasiak (1998) proposed an alternative GARCH model for UHF data in which the total variance is modeled but the GARCH parameters are functions of the expected duration. Meddahi *et al.* (2006) highlighted the differences between these two models. The UHF-GARCH model of Engle (2000) was further applied e.g. by Racicot *et al.* (2008) and Huptas (2016).

Simultaneous transactions. A particular issue of UHF data is the occurence of transactions with the same timestamp resulting in zero durations. Engle and Russell (1998) considered these transactions to be split transactions which belong to a single trade and decided to aggregate them. Note that zero duration does not necessarily mean zero return as transactions can be executed at the same time at different prices. Blasques *et al.* (2022a) further studied the issue of zero durations and pointed out that, depending on the precision of timestamps in data, zero durations may account for the majority of observations and aggregation is not a suitable solution. When measuring price variance per time unit, as Engle (2000) did, returns are divided by the square root of the corresponding trade duration. Zero durations with nonzero returns of course disrupt this concept of variance per time unit.

Discretness of prices. Financial assets are traded on a discrete grid of prices. On the NYSE and NASDAQ exchanges, e.g., stocks are traded with precision to one cent. This discreteness has a large impact on the distribution of returns (see, e.g., Münnix *et al.*, 2010 for empirical evidence). Consequently, a strand of literature emerged that focuses on dynamic volatility models for discrete price changes based on the Skellam distribution and its generalizations. Koopman *et al.* (2017) modified the Skellam distribution by transferring probability mass between 0, 1, and -1 values and used it in a dynamic state space model for price changes. Koopman *et al.* (2018) took a multidimensional approach and modeled price changes by a score-driven model based on a discrete copula with Skellam margins. Alomani *et al.* (2018) used the Skellam GARCH model for drug crimes. Gonçalves and Mendes-Lopes (2020) studied more general integer GARCH processes with applications to polio cases and Olympic medals won. Cui *et al.* (2021) used a GARCH model based on the Skellam distribution with modified probabilities for daily price changes. Doukhan *et al.* (2021) studied theoretical properties of integer GARCH processes. Catania *et al.* (2022) used the zero-inflated Skellam distribution in a hidden Markov model for multivariate price changes. Note that none of these studies utilize UHF data and are limited only to a fixed frequency – e.g., 1 second in Koopman *et al.* (2017), 10 second in Koopman *et al.* (2018), and 15 second in Catania *et al.* (2022). In contrast to time series models, Skellam models in continuous time were analyzed by Barndorff-Nielsen *et al.* (2012) and Shephard and Yang (2017). An alternative approach was adopted by Holý and Tomanová (2022) who modeled prices directly, instead of price changes or logarithmic returns, by the double Poisson distribution.

Market microstructure noise. A well documented feature of high frequency data is market microstructure noise – a deviation from the fundamental efficient price (see, e.g., Hansen and Lunde, 2006 for an in-depth study). It is caused by price discreteness but also by bid-ask bounce, asymmetric information of traders, and other informational effects. It plays a key role in nonparametric estimation of quadratic variation and integrated variance as it significantly biases realized variance at higher frequencies (see, e.g., Holý and Tomanová, 2023 for an overview of noise-robust estimators). Regarding parametric processes, independent market microstructure noise induces a moving average component of order one. Specifically, Aït-Sahalia *et al.* (2005) showed that Wiener process contaminated by independent market microstructure noise sampled at discrete times corresponds to ARIMA(0,1,1) process and Holý and Tomanová (2019) showed that discretized noisy Ornstein–Uhlenbeck process corresponds to ARIMA(1,0,1) process.

Table 1 lists notable high-frequency models and summarizes their features. Note that none of these models address all four of the above high-frequency characteristics. The goal of this paper is therefore to combine the UHF-GARCH approach with the Skellam-GARCH approach while accounting for simultaneous transactions and market microstructure noise.

Our approach starts with nonparametric estimation of diurnal trends in trade durations and squared price changes using smoothing splines. When both these time series are adjusted for diurnal trends, their relation is estimated using smoothing splines. Next, we build our dynamic model. The original (unadjusted) price changes are assumed to follow the zero-inflated Skellam distribution of Skellam (1946) with time-varying mean and variance and static probability of zero-inflation. The dynamic mean follows MA(1) process to capture the effects of market microstructure noise. As high-frequency data exhibit zero expected returns, we set the intercept to zero. In the Skellam distribution, the variance is required to be higher than the absolute value of the mean, which is suitable for high-frequency data. However, to avoid inconvenient restrictions on the parameter space, we propose to parametrize the distribution in terms of the overdispersion parameter, i.e. the excessive variance. The dynamic overdispersion then follows score-driven model, developed by Creal *et al.* (2013) and Harvey (2013). The estimated diurnal pattern of squared price changes and their relation to trade durations are further plugged into this dynamics. The used relation to trade durations simultaneously captures adjustment of variance to time unit and the residual dependency on trade durations, which were modeled separately by Engle (2000). The proposed joint modeling removes the problems with zero trade durations, which can be quite frequent in high-frequency data. The proposed model belongs to the class of observation-driven models and can be estimated by the maximum likelihood method, which makes it suitable even for large datasets.

In an empirical study, we focus on the IBM stock (just as, e.g., Engle and Russell, 1998; Engle,

Table 1: An overview of selected high-frequency time series models and their features – using ultra-high-frequency data with irregularly spaced observations (Irreg), accounting for simultaneous transactions with zero trade durations (Simul), accounting for discrete prices or price changes (Discrete), accounting for market microstructure noise (Noise), joint modeling of volatility and trade durations (Duration), joint modeling of volatility and trade volume (Volume), and multivariate modeling (Multi).

Paper	Irreg	Simul	Discrete	Noise	Duration	Volume	Multi
Ghysels and Jasiak (1998)	✓	✗	✗	✗	✓	✗	✗
Engle (2000)	✓	✗	✗	✗	✓	✗	✗
Grammig and Wellner (2002)	✓	✗	✗	✗	✓	✗	✗
Manganelli (2005)	✓	✗	✗	✗	✓	✓	✗
Russell and Engle (2005)	✓	✗	✓	✗	✓	✗	✗
Liu and Maheu (2012)	✓	✗	✗	✓	✓	✗	✗
Huptas (2016)	✓	✗	✗	✗	✓	✗	✗
Koopman *et al.* (2017)	✗	✗	✓	✗	✗	✗	✗
Koopman *et al.* (2018)	✗	✗	✓	✗	✗	✗	✓
Buccheri *et al.* (2021)	✗	✗	✗	✓	✗	✗	✓
Catania *et al.* (2022)	✗	✗	✓	✗	✗	✗	✓
Holý and Tomanová (2022)	✓	✗	✓	✗	✗	✗	✗
This study	✓	✓	✓	✓	✗	✗	✗

2000) from March to July, 2022. However, we also report results for 6 other stocks traded on the NYSE and NASDAQ exchanges. We estimate intraday models with various specifications for each of the 105 trading days in our dataset. For the IBM stock, the average number of observations in a day is 63 673. We show that the proposed model is a good fit and all its components are justifiable. We also demonstrate how the results can be used as an alternative to daily realized measures of volatility such as the realized kernel of Barndorff-Nielsen *et al.* (2008). Finally, we find that the relation between price volatility and trade durations is the same as described by Engle (2000), eventhough the magnitude of high-frequency data has increased considerabely since then.

2 Methodology

2.1 Nonparametric Temporal Adjustment

Let t_i, $i = 0, \ldots, n$, denote times of transactions and p_i, $i = 0, \ldots, n$, prices (with precision to two decimal places). Furthermore, let $d_i = t_i - t_{i-1}$, $i = 1, \ldots, n$, denote trade durations and $y_i = 100(p_i - p_{i-1})$, $i = 1, \ldots, n$, (integer) price changes.

First, we estimate the intraday pattern of trade durations. On each day, we standardize trade durations as $\bar{d}_i = d_i / \frac{1}{n}\sum_{i=1}^{n} d_i$. Using the complete dataset, we then estimate the (possibly nonlinear) dependence of $\bar{d}_i$ on t_i by the cubic smoothing spline method (see, e.g., Hastie *et al.*, 2008, Section 5.4). The chosen nonparametric method, however, is not essential to our model and alternatives can be used as well. We obtain the fitted function $\hat{f}_{\text{dur}}(t_i)$ and adjust trade durations as $\tilde{d}_i = \bar{d}_i / \hat{f}_{\text{dur}}(t_i)$.

Next, we estimate the intraday pattern of squared price changes. To be precise, squared price changes with substracted price changes in absolute value, $z_i = y_i^2 - |y_i| = y_i(y_i - \text{sgn}(y_i))$. This transformation corresponds to the overdispersion parameter, which plays a central role in our dynamic model. As in the case of trade durations, we standardize modified squared price changes as $\bar{z}_i = z_i / \frac{1}{n}\sum_{i=1}^{n} z_i$ and then estimate the dependence of $\bar{z}_i$ on t_i by the cubic smoothing spline method. We obtain the fitted function $\hat{f}_{\text{disp}}(t_i)$ and adjust modified squared returns as $\tilde{z}_i = z_i / \hat{f}_{\text{disp}}(t_i)$.

Finally, we estimate the relation between modified squared price changes and trade durations, i.e. dependence of $\tilde{z}_i$ on $\tilde{d}_i$. Again, we use the cubic smoothing spline method and obtain the fitted

function $\hat{f}_{\text{rel}}(\tilde{d}_i)$.

2.2 Zero-Inflated Skellam Distribution

The probability theory and statistics literature does not offer many distributions defined on integer support (without the nonnegativity or positivity constraint). The most used representative is the Skellam distribution of Skellam (1946), which is the distribution of the difference between two independent variables following the Poisson distribution with rates λ_1 and λ_2 respectively. Regarding dynamic models, it can be used when a time series of counts is nonstationary, but its first difference is stationary – a typical feature of high-frequency prices.

The Skellam distribution is often parametrized in terms of mean $\mu = \lambda_1 - \lambda_2$ and variance $\sigma^2 = \lambda_1 + \lambda_2$ rather than rates λ_1 and λ_2 (see, e.g., Koopman *et al.*, 2017, 2018; Alomani *et al.*, 2018). However, in this parametrization, it is required that $\sigma^2 > |\mu|$. When μ is nonzero, this condition can be hard to satisfy in dynamic models. For this reason, we propose an alternative parametrization with overdispersion parameter $\delta = \sigma^2 - |\mu| = \min\{2\lambda_1, 2\lambda_2\} > 0$ representing excessive variance.

In any case, only two parameters of the distribution do not offer much flexibility needed for high frequency prices. Koopman *et al.* (2017) deflate the probability of 0 and inflate probability of 1 and -1 using an additional parameter. On the other hand, Karlis and Ntzoufras (2006, 2009) and Catania *et al.* (2022) inflate the probability of 0 and deflate the probabilities of all other values using an additional parameter, in the fashion of the zero-inflated model of Lambert (1992). As our data exhibit increased occurence of zero values (in comparison to the fitted Skellam distribution), we follow the latter approach and introduce the zero-inflation parameter π to the distribution.

The probability mass function of the zero-inflated Skellam distribution with the mean-overdispersion parametrization is given by

$$\mathrm{P}[Y = y \mid \mu, \delta, \pi] = \begin{cases} \pi + (1-\pi)\exp(-|\mu| - \delta) I_0\left(\sqrt{\delta^2 + 2|\mu|\delta}\right) & \text{for } y = 0, \\ (1-\pi)\exp(-|\mu| - \delta)\left(\frac{|\mu|+\mu+\delta}{|\mu|-\mu+\delta}\right)^{\frac{y}{2}} I_y\left(\sqrt{\delta^2 + 2|\mu|\delta}\right) & \text{for } y \neq 0, \end{cases} \tag{1}$$

where $I_\cdot(\cdot)$ is the modified Bessel function of the first kind. The first two moments are given by

$$\mathrm{E}[Y] = (1-\pi)\mu, \qquad \mathrm{var}[Y] = (1-\pi)\left(|\mu| + \delta + \pi\mu^2\right). \tag{2}$$

2.3 Time-Varying Parameters

In the dynamic model, we let the mean parameter μ and the overdispersion parameter δ be time-varying but keep the zero-inflation parameter π static.

Strong negative first order autocorrelation, insignificant autocorrelation of higher order, and decaying negative partial autocorrealtion is typical for ultra-high-frequency price changes or returns and is caused by market microstrucure noise (see, e.g., Aït-Sahalia *et al.*, 2005; Hansen and Lunde, 2006). It can be effectively captured by MA(1) process. Another typical feature of high-frequency data is zero mean of price changes or returns in long term (see, e.g., Koopman *et al.*, 2017). We therefore model dynamics of the mean parameter as MA(1) process with zero intercept,

$$\mu_i = \theta\left(y_{i-1} - \mu_{i-1}\right), \tag{3}$$

where θ is the moving average parameter.

For the dynamics of the overdispersion parameter, we adopt a GARCH-like structure and include the temporal adjustments presented in Section 2.1. To avoid any restrictions on the parameter space, we model the logarithm of the overdispersion parameter, which is in line with the multiplicative form of the temporal adjustments. Similarly to Koopman *et al.* (2018), we let the overdispersion parameter be driven by lagged conditional score, i.e. the gradient of the log-likelihood, of the Skellam distribution. Our model therefore belongs to the class of score-driven models (see Creal *et al.*, 2013;

Harvey, 2013)[1]. All put together, the dynamics of the overdispersion parameter is given by

$$\ln(\delta_i) = \omega + \ln\left(\hat{f}_{\text{disp}}(t_i)\right) + \ln\left(\hat{f}_{\text{rel}}(\tilde{d}_i)\right) + \varepsilon_i, \qquad \varepsilon_i = \varphi\varepsilon_{i-1} + \alpha\nabla_{\ln(\delta)}\left(y_{i-1}; \mu_{i-1}, \delta_{i-1}, \pi\right), \tag{4}$$

where ω is the intercept, φ is the autoregressive parameter, α is the score parameter, and $\nabla_{\ln(\delta)}(\cdot)$ is the score given by

$$\begin{aligned}\nabla_{\ln(\delta)}(y; \mu, \delta, \pi) &= \frac{\partial \ln \mathrm{P}[Y = y \mid \mu, \delta, \pi]}{\partial \ln(\delta)} \\ &= \begin{cases} \frac{\delta(\pi-1)\left(\sqrt{\delta^2+2|\mu|\delta} I_0\left(\sqrt{\delta^2+2|\mu|\delta}\right) - (|\mu|+\delta) I_1\left(\sqrt{\delta^2+2|\mu|\delta}\right)\right)}{\sqrt{\delta^2+2|\mu|\delta}\left((1-\pi) I_0\left(\sqrt{\delta^2+2|\mu|\delta}\right) + \pi \exp(|\mu|+\delta)\right)} & \text{for } y = 0, \\ \frac{\delta^2+|\mu|\delta}{2\sqrt{\delta^2+2|\mu|\delta}} \frac{I_{y-1}\left(\sqrt{\delta^2+2|\mu|\delta}\right) + I_{y+1}\left(\sqrt{\delta^2+2|\mu|\delta}\right)}{I_y\left(\sqrt{\delta^2+2|\mu|\delta}\right)} - \frac{\mu y}{\delta+2|\mu|} - \delta & \text{for } y \neq 0. \end{cases} \end{aligned} \tag{5}$$

Although the formula for the score is quite complex in the case of the Skellam distribution, its interpretation is clear – it is a correction term improving the fit of the distribution after an observation is realized. The use of the conditional score in dynamic models is optimal in the sense of the Kullback–Leibler divergence between the true and the model-implied distribution (see Blasques *et al.*, 2015, 2021).

2.4 Maximum Likelihood Estimation

There are five parameters in the model to be estimated – θ, ω, φ, α, and π. The model is observation-driven and we find the parameters by maximizing the log-likelihood,

$$\ell(\theta, \omega, \varphi, \alpha, \pi \mid y_1, \ldots, y_n) = \frac{1}{n}\sum_{i=1}^{n} \ln \mathrm{P}[Y_i = y_i \mid \mu_i, \delta_i, \pi], \tag{6}$$

where μ_i and δ_i are given by (3) and (4) respectively. As μ_i and ε_i are defined recursively, it is needed to set their initial values. We set them to their long-term average, i.e. $\mu_0 = \varepsilon_0 = 0$. The particular choice for the initialization is not, however, that important as their impact quickly fades out and is overall negligible in the tens of thousands or even hundreds of thousands of observations we have. We numerically find the optimal values of the parameters using the Nelder–Mead algorithm. It is, however, possible to use any general-purpose algorithm solving nonlinear optimization problems.

Deriving asymptotic properties of the maximum likelihood estimates is beyond the scope of the paper. We refer to Alzaid and Omair (2010) for the theoretical results on static case of the Skellam distribution and Blasques *et al.* (2018, 2022b) for the results on score-driven models in general. Tailoring these results to our specific model is, however, not straightforward.

3 Empirical Study

3.1 Analyzed Data Sample

As Engle and Russell (1998), Engle (2000), and many other papers, we focus our analysis on the IBM stock traded on the New York Stock Exchange (NYSE). The stock is included in the Dow Jones Industrial Average (DJIA), S&P 100, and S&P 500 indices. We use tick-by-tick transaction data from March to July, 2022 – a total of 105 trading days. The source of the data is Refinitiv Eikon[2]. Furthermore, we report results for the CAT, MA, and, MCD stocks traded on NYSE and the CSCO, EA, and INTC stocks traded on NASDAQ in Appendix A.

We preform standard data cleaning steps, as described e.g. in Barndorff-Nielsen *et al.* (2009). Namely, we remove observations outside the standard trading hours 9:30–16:00 EST, remove observations in the first 5 minutes after the opening (we further discuss this in Section 3.3), remove

[1] Besides Koopman *et al.* (2018), score-driven model based on the Skellam distribution was also used by Koopman and Lit (2019) in an application to football results.

[2] Formerly operated by Thomson Reuters.

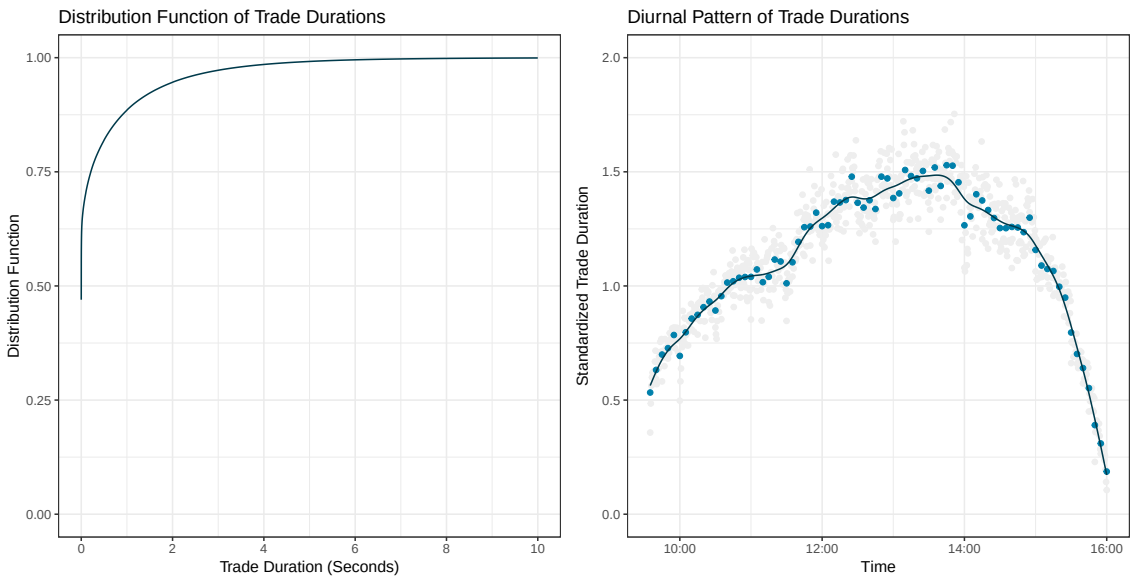

Figure 1: The empirical distribution function of trade durations (left) and average trade durations in 5 minute and 30 second intraday intervals with a smoothed curve (right) for the IBM stock.

observations without recorded price, remove outliers (when price exceeds 10 mean absolute deviations from a rolling centred median of 50 observations), and round prices to the nearest cent.

After data cleaning, we get the total of 6 685 657 transactions over 105 trading days for the IBM stock, which corresponds to 2.721 transactions per second. The two busiest days are July 19 with 258 217 transactions and April 20 with 184 250 transactions. Both these days follow announcements of quarterly results on July 18 and April 19 respectively. The quietest day is March 28 with just 35 333 transactions. The median value is 56 894 transactions per trading day.

The subsequent analysis is performed using R. The temporal adjustment is performed by the `smooth.spline()` function from the `stats` package (R Core Team, 2022). The dynamic model is estimated by the `gas()` function from the `gasmodel` package (Holý, 2022) with a one-line modification[3].

3.2 Trade Durations

We start the empirical study by a brief look at trade durations. The data are recorded with a time precision of one millisecond and we report trade durations in seconds (with precision to three decimal places). The left plot of Figure 1 shows the empirical distribution of trade durations. Most transactions occur in close succession – 47 percent of trade durations are equal to zero and 88 percent are lower than one second. Thus, aggregating simultaneous transactions would almost halve the number of observations. Using a similar dataset for the IBM stock, Blasques *et al.* (2022a) found that 95 percent of zero trade durations are caused by split transactions while 5 percent are unrelated transactions. We decide to keep simultaneous transactions in our dataset.

The right plot of Figure 1 shows diurnal pattern of trade durations – a typical hill shape. The market is most active after opening and before closing while after noon there is a quiet period. This is consistent with the duration literature.

3.3 Price Changes

Next, we move on to empirical properties of price changes. The left plot of Figure 2 shows the empirical probability mass function of price changes. The price changes at ultra-high-frequency are

[3]The score for μ in the zero-inflated Skellam distribution is replaced by $y - \mu$ to mimic the moving average process.

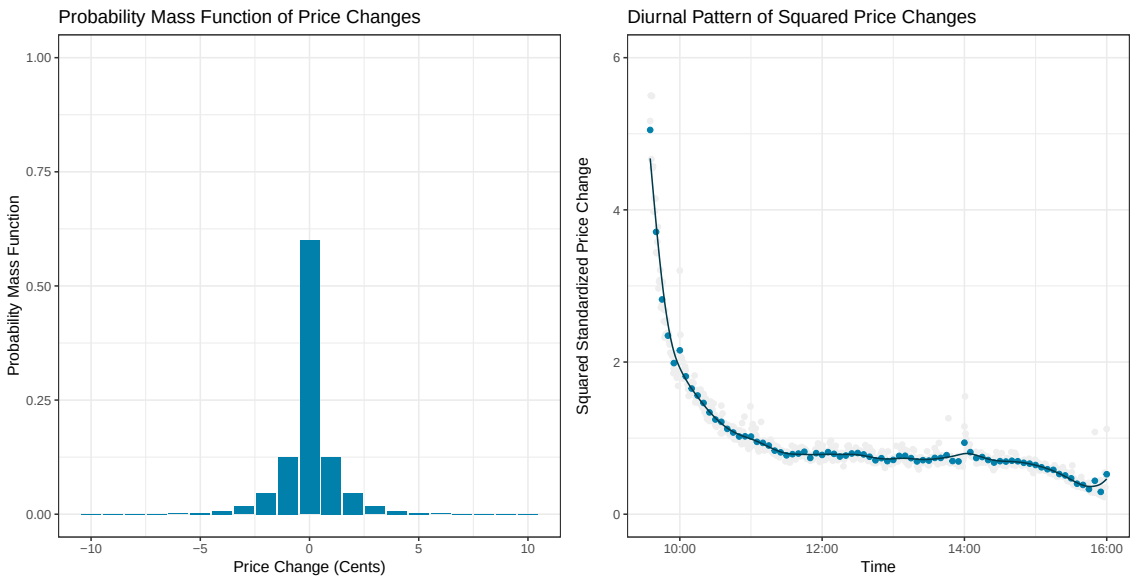

Figure 2: The empirical probability mass function of price changes (left) and average squared price changes in 5 minute and 30 second intraday intervals with a smoothed curve (right) for the IBM stock.

quite low – 60 percent of price changes are zero and 99 percent of price changes are between -3 and 3 cents. The most extreme price changes are -66 and 68 cents. Note, however, that some higher price changes were labeled as outliers and removed during data cleaning.

The right plot of Figure 2 shows diurnal pattern of squared price changes. In this plot, we do not substract the absolute value of price change – however, the plot showing the modified price change looks almost identical. There is extreme volatility after the opening, which quickly declines. As smoothing splines have trouble capturing this steep decrease, we remove the first 5 minutes from data, i.e. we focus only on 9:35–16:00 EST. Right before the closing, volatility slightly increases. There is also a slight increase around 14:00 associated with news relevant to the IBM stock[4].

There is strong serial correlation present in both price changes and squared price changes. The autocorrelation of price changes is -0.352 for the first order and very close to zero for higher orders. The partial autocorrelation, on the other hand, decreases gradually. The autocorrelation of squared price changes is 0.403 for the first order and gradually decreases for higher orders. The partial autocorrelation also decreases gradually. This suggests MA(1) dynamics for the mean process and richer dynamics for the volatility process.

Price variance (squared price changes) naturally increases with trade duration. This relation is vizualized in the left plot of Figure 3. However, this increase is slower than linear. The right plot of Figure 3 shows that price variance per second (squared price changes divided by trade duration) decreases with trade duration. This is in line with Engle (2000) who estimated a positive linear dependence of variance per time unit on the inverse of trade duration. We refrain from this approach due to problems with zero values. We would be dividing by zero twice – when calculating squared price changes per second and when inverting trade durations. Note that for the purposes of the right plot of Figure 3, we add 0.001 to the values of trade durations. Of course, this is a completely arbitrary transformation, which has a large impact on behavior near zero (which is cropped in the right plot of Figure 3). Instead, we directly estimate relation between price variance and trade durations and thus avoid problems with zero values.

[4]In the case of the IBM stock, the increase is not that major. In the case of other stocks, however, this could be much larger jump (or multiple jumps at various times), which smoothing splines could fail to capture; see Appendix A.

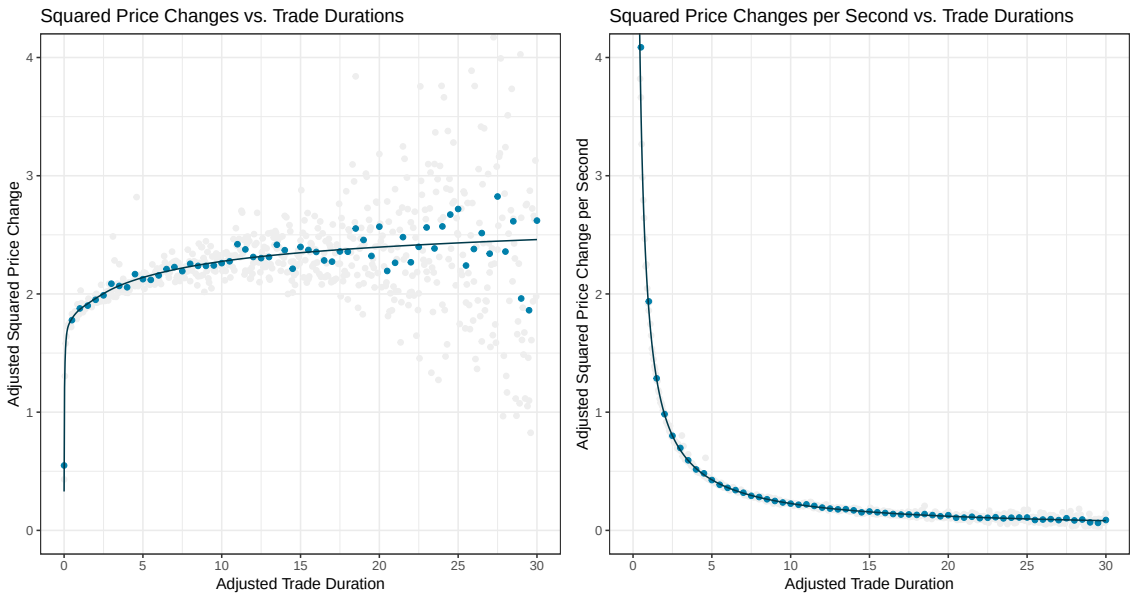

Figure 3: The average diurnally adjusted squared price changes (left) and diurnally adjusted squared price changes per second (right) in 50 millisecond and half second intervals of diurnally adjusted trade durations with a smoothed curve for the IBM stock.

3.4 Dynamics of Intraday Price Volatility

For each trading day, we estimate 10 specifications of the proposed dynamic model – with different parametrizations and different parameters set to zero. The features of the models are summarized in Table 2. In Table 3, we report the minimum, maximum, and median values of estimated parameters. Note that, we do not report p-values as all parameters are significant due to huge numbers of observations (with the exception of π for a single day, as further mentioned below). In Table 4, we assess fit of the models using the average log-likelihood and residual autocorrelation tests in the form of R^2 statistic. Note that the number of parameters (in any of our model specifications) is negligible compared to the number of observations. For this reason, we do not report AIC or BIC.

First, let us focus on the parametrization of the model. We compare the mean-variance parametrization (models I–V), used e.g. by Koopman *et al.* (2017, 2018) and Alomani *et al.* (2018), with the proposed mean-overdispersion parametrization (models VI–X). When the mean is not dynamic and is set to zero, both parametrizations are equivalent. The only difference lies in the temporal adjustment, which is based on the squared differences for the mean-variance parametrization and the squared differences with substracted mean in absolute value for the mean-overdispersion parametrization. The results show that this difference is not that distinct – model I has very similar log-likelihood to model VI and model IV to model IX. When the mean is dynamic, however, the mean-overdispersion is superior – models VII, VIII, and X clearly outperform their counterpart models II, III, and V in terms of log-likelihood. The problem, of course, lies in bounds on parameter space imposed by the mean-variance parametrization.

Next, we asses the impact of the individual parameters. As discussed in Section 3.3, the autocorrelation and partial autocorrelation functions of price changes suggest MA(1) structure for the mean process. Indeed, restricting θ to zero causes considerable decrease in log-likelihood as evident between models IV/V and IX/X. The autocorrelation in residuals also significantly increases. As expected, the estimated θ is negative for all trading days. Interestingly, its value is much closer to zero in the case of the mean-variance parametrization than the mean-overdispersion parametrization. This suppression of θ is caused by the lower bound on the variance process. In the mean-overdispersion parametrization, there is no such restriction and the mean process is able to reach its full potential.

Table 2: Summary of the features of the estimated models.

Model	Mean	Volatility	Zeros	Parametrization
I	Static	Static	Unaltered	Variance
II	Dynamic	Dynamic	Unaltered	Variance
III	Dynamic	Static	Inflated	Variance
IV	Static	Dynamic	Inflated	Variance
V	Dynamic	Dynamic	Inflated	Variance
VI	Static	Static	Unaltered	Overdispersion
VII	Dynamic	Dynamic	Unaltered	Overdispersion
VIII	Dynamic	Static	Inflated	Overdispersion
IX	Static	Dynamic	Inflated	Overdispersion
X	Dynamic	Dynamic	Inflated	Overdispersion

The comparison of log-likelihood and autocorrelation in squared residuals between models III/V and VIII/X reveals that volatility (whether paramterized in terms of variance or overdispersion) should not be treated as constant. Similarly to θ, there is a difference in estimated values of α and φ between the mean-variance and mean-overdispersion parametrizations. Model X has higher persistence in comparison to model V. Again, this can be atributed to the lower bound on the variance process in the mean-variance parametrization.

In each model allowing for zero inflation, π is positive for all days except one, July 28[5]. This suggests that there is an increased occurence of zero price changes in general and the underlying distribution should accomodate this. Among the three components studied in this section – dynamic mean, dynamic volatility, and zero inflation – setting parameter π to zero decreases the log-likelihood the least, but still distinctly.

Overall, model X performs the best in terms of the log-likelihood among our 10 candidates. The proposed specification for the mean and overdispersion processes also overwhelmingly reduces residual autocorrelation in price changes and squared price changes. Due to huge number of observations, however, it is difficult to obtain statistical significance of no autocorrelation. The associated Ljung–Box test rejects no autocorrelation in residuals of model X for all days and lags at 0.01 significance level. The associated ARCH-LM test suggests no autocorrelation in squared residuals of model X for 68 percent of days for lag 1 but only 4 percent for lag 100 at 0.01 significance level. Nevertheless, the R^2 static is very low in all cases and the model captures mean and volatility dynamics quite well.

3.5 Daily Measures of Price Volatility

The proposed approach can naturally be used to model intraday dynamics of prices but also to estimate volatility at daily level as a model-based alternative to various nonparametric volatility measures. A standard nonparametric measure of daily volatility is the realized variance – the sum of squared returns. However, this measure is biased by market microstructure noise and generally not recommended to use at ultra-high-frequency (see, e.g., Hansen and Lunde, 2006). At lower frequency such as 5 minutes, however, it can be sufficient as the impact of market microstructure noise is reduced (see, e.g., Liu *et al.*, 2015). A widely used realized measure that is robust to market microstructure noise is the realized kernel of Barndorff-Nielsen *et al.* (2008)[6].

In this section, we compare the realized variance and the realized kernel based on the modified Tukey–Hanning kernel with realized measures implied by our model. The total variance based on the proposed model is given by

$$TMV = \sum_{i=1}^{n} (1-\pi)\left(|\mu_i| + \delta_i + \pi\mu_i^2\right). \tag{7}$$

[5]However, other stocks may exhibit different behaviors, and zero inflation may not be necessary; see Appendix A

[6]For details on practical use of the realized kernel, see Barndorff-Nielsen *et al.* (2009). For the multivariate case, see Barndorff-Nielsen *et al.* (2011). Other noise-robust realized measures such as the multi-scale and pre-averaging estimators are fairly similar as they can all be expressed in a quadratic form (see, e.g., Holý and Tomanová, 2023).

Table 3: The minimum, median, and maximum values of estimated parameters of various daily models for the IBM stock.

		Variance Models					Overdispersion Models				
Coef.	Trans.	I	II	III	IV	V	VI	VII	VIII	IX	X
	Min		-0.149	-0.064		-0.181		-0.449	-0.567		-0.527
θ	Med		-0.103	-0.028		-0.116		-0.302	-0.383		-0.343
	Max		-0.050	-0.006		-0.055		-0.216	-0.305		-0.261
	Min	-0.379	-0.384	-0.379	-0.400	-0.386	-0.212	-0.513	-0.663	-0.196	-0.513
ω	Med	0.200	0.192	0.401	0.299	0.340	0.384	0.068	0.095	0.491	0.170
	Max	0.676	0.712	1.056	0.860	0.958	0.892	0.613	0.885	1.066	0.800
	Min		0.471		0.644	0.468		0.942		0.708	0.954
φ	Med		0.761		0.834	0.787		0.982		0.872	0.981
	Max		0.963		0.984	0.963		0.997		0.988	0.997
	Min		0.103		0.095	0.102		0.065		0.100	0.077
α	Med		0.495		0.498	0.517		0.165		0.488	0.192
	Max		0.667		0.681	0.701		0.259		0.697	0.287
	Min			0.000	0.000	0.000			0.000	0.000	0.000
π	Med			0.154	0.119	0.118			0.132	0.111	0.119
	Max			0.299	0.235	0.246			0.250	0.223	0.221

Table 4: The R^2 statistics of residuals and squared residuals regressed on their lagged values with the average log-likelihood of an observation for various daily models for the IBM stock.

		Variance Models					Overdispersion Models				
Statistic	Lag	I	II	III	IV	V	VI	VII	VIII	IX	X
	1	0.118	0.040	0.104	0.077	0.041	0.116	0.004	0.002	0.075	0.003
AR R^2	10	0.151	0.055	0.136	0.097	0.055	0.149	0.008	0.004	0.095	0.007
	100	0.154	0.057	0.139	0.098	0.057	0.153	0.010	0.007	0.096	0.009
	1	0.104	0.003	0.097	0.005	0.003	0.104	0.000	0.001	0.006	0.000
ARCH R^2	10	0.150	0.008	0.145	0.007	0.007	0.154	0.004	0.028	0.007	0.003
	100	0.181	0.021	0.176	0.016	0.018	0.189	0.007	0.049	0.015	0.006
Log-Likelihood		-1.264	-1.200	-1.245	-1.212	-1.193	-1.267	-1.177	-1.187	-1.209	-1.170

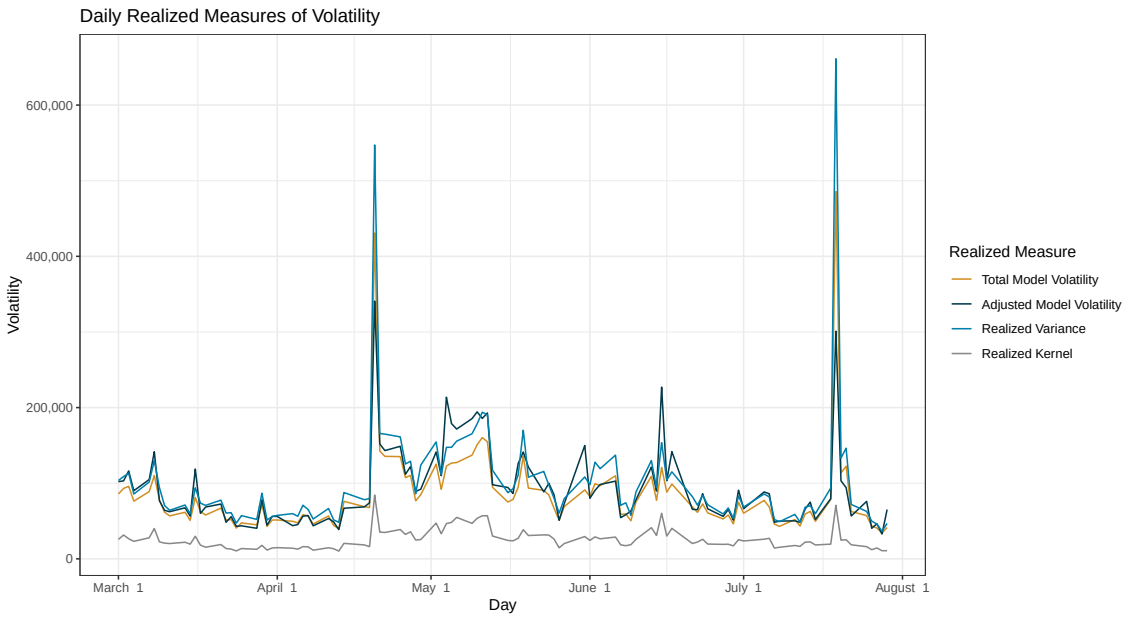

Figure 4: The daily values of various volatility realized measures for the IBM stock.

We can also measure volatility by the total overdispersion adjusted for temporal effects (both diurnal and duration) given by

$$AMV = \sum_{i=1}^{n} \frac{\delta_i}{\hat{f}_{\text{disp}}(t_i)\hat{f}_{\text{rel}}(\tilde{d}_i)} = \sum_{i=1}^{n} \exp\left(\omega + \varepsilon_i\right). \tag{8}$$

In the latter realized measure, market microstructure noise is filtered by removing the MA(1) component and the effect of trade durations.

Figure 4 shows daily volatility obtained by these measures. The largest variance for all measures is on April 20 (following the announcement of the first quarter results on April 19) and on July 19 (following the announcement of the second quarter results on July 18). We can see that all measures tend to move together but have different scale. This is also supported by a simple correlation analysis. The highest correlations are 0.998 between the total model volatility and the realized variance and 0.965 between the adjusted model volatility and the realized kernel. Other correlations lie between 0.821 and 0.882. We can conclude that the total model volatility is similar to the realized variance as they are both influenced by market microstructure noise. On the other hand, the adjusted model volatility is robust to market microstructure noise, just as the realized kernel. The main benefit of the proposed model-based approach is that we can decompose the variance into individual components according to (2) and (4).

4 Conclusion

We have proposed a dynamic model for intraday stock prices that takes into account irregularly spaced observations, simultaneous transactions, discreteness of prices, and market microstructure noise. In this model, we have combined two streams of the literature dealing with UHF-GARCH and Skellam-GARCH models respectively and further developed them. We have shown that the model finds its use not only in analysis of intraday dynamics but also in estimation of daily volatility.

Suggestions for future research follow Table 1. Our model can be extended to include dynamics of trade durations and possibly trade volumes. Another direction lies in multivariate modeling. This is, however, quite challenging due to nonsynchronicity of ultra-high-frequency data.

Acknowledgements

Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

Funding

The work on this paper was supported by the Czech Science Foundation under project 23-06139S and the personal and professional development support program of the Faculty of Informatics and Statistics, Prague University of Economics and Business.

References

Aït-Sahalia Y, Mykland PA, Zhang L (2005). "How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise." *The Review of Financial Studies*, **18**(2), 351–416. ISSN 0893-9454. https://doi.org/10.1093/rfs/hhi016.

Alomani GA, Alzaid AA, Omair MA (2018). "A Skellam GARCH model." *Brazilian Journal of Probability and Statistics*, **32**(1), 200–214. ISSN 0103-0752. https://doi.org/10.1214/16-bjps338.

Alzaid AA, Omair MA (2010). "On the Poisson Difference Distribution Inference and Applications." *Bulletin of the Malaysian Mathematical Sciences Society*, **33**(1), 17–45. ISSN 0126-6705. http://eudml.org/doc/244475.

Barndorff-Nielsen OE, Hansen PR, Lunde A, Shephard N (2008). "Designing Realized Kernels to Measure the ex post Variation of Equity Prices in the Presence of Noise." *Econometrica*, **76**(6), 1481–1536. ISSN 0012-9682. https://doi.org/10.3982/ecta6495.

Barndorff-Nielsen OE, Hansen PR, Lunde A, Shephard N (2009). "Realized Kernels in Practice: Trades and Quotes." *Econometrics Journal*, **12**(3), 1–32. ISSN 1368-4221. https://doi.org/10.1111/j.1368-423X.2008.00275.x.

Barndorff-Nielsen OE, Hansen PR, Lunde A, Shephard N (2011). "Multivariate Realised Kernels: Consistent Positive Semi-Definite Estimators of the Covariation of Equity Prices with Noise and Non-Synchronous Trading." *Journal of Econometrics*, **162**(2), 149–169. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2010.07.009.

Barndorff-Nielsen OE, Pollard DG, Shephard N (2012). "Integer-Valued Lévy Processes and Low Latency Financial Econometrics." *Quantitative Finance*, **12**(4), 587–605. ISSN 1469-7688. https://doi.org/10.1080/14697688.2012.664935.

Blasques F, Gorgi P, Koopman SJ, Wintenberger O (2018). "Feasible Invertibility Conditions and Maximum Likelihood Estimation for Observation-Driven Models." *Electronic Journal of Statistics*, **12**(1), 1019–1052. ISSN 1935-7524. https://doi.org/10.1214/18-ejs1416.

Blasques F, Holý V, Tomanová P (2022a). "Zero-Inflated Autoregressive Conditional Duration Model for Discrete Trade Durations with Excessive Zeros." https://arxiv.org/abs/1812.07318.

Blasques F, Koopman SJ, Lucas A (2015). "Information-Theoretic Optimality of Observation-Driven Time Series Models for Continuous Responses." *Biometrika*, **102**(2), 325–343. ISSN 0006-3444. https://doi.org/10.1093/biomet/asu076.

Blasques F, Lucas A, van Vlodrop AC (2021). "Finite Sample Optimality of Score-Driven Volatility Models: Some Monte Carlo Evidence." *Econometrics and Statistics*, **19**, 47–57. ISSN 2452-3062. https://doi.org/10.1016/j.ecosta.2020.03.010.

Blasques F, van Brummelen J, Koopman SJ, Lucas A (2022b). "Maximum Likelihood Estimation for Score-Driven Models." *Journal of Econometrics*, **227**(2), 325–346. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2021.06.003.

Bollerslev T (1986). "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics*, **31**(3), 307–327. ISSN 0304-4076. https://doi.org/10.1016/0304-4076(86)90063-1.

Bollerslev T (1987). "A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return." *Review of Economics and Statistics*, **69**(3), 542–547. ISSN 0034-6535. https://doi.org/10.2307/1925546.

Buccheri G, Bormetti G, Corsi F, Lillo F (2021). "A Score-Driven Conditional Correlation Model for Noisy and Asynchronous Data: An Application to High-Frequency Covariance Dynamics." *Journal of Business & Economic Statistics*, **39**(4), 920–936. ISSN 0735-0015. https://doi.org/10.1080/07350015.2020.1739530.

Catania L, Di Mari R, Santucci de Magistris P (2022). "Dynamic Discrete Mixtures for High-Frequency Prices." *Journal of Business & Economic Statistics*, **40**(2), 559–577. ISSN 0735-0015. https://doi.org/10.1080/07350015.2020.1840994.

Creal D, Koopman SJ, Lucas A (2013). "Generalized Autoregressive Score Models with Applications." *Journal of Applied Econometrics*, **28**(5), 777–795. ISSN 0883-7252. https://doi.org/10.1002/jae.1279.

Cui Y, Li Q, Zhu F (2021). "Modeling Z-Valued Time Series Based on New Versions of the Skellam INGARCH Model." *Brazilian Journal of Probability and Statistics*, **35**(2), 293–314. ISSN 0103-0752. https://doi.org/10.1214/20-bjps473.

Doukhan P, Khan NM, Neumann MH (2021). "Mixing Properties of Integer-Valued GARCH Processes." *Alea - Latin American Journal of Probability and Mathematical Statistics*, **18**(1), 401–420. ISSN 1980-0436. https://doi.org/10.30757/alea.v18-18.

Engle R (2002). "New Frontiers for ARCH Models." *Journal of Applied Econometrics*, **17**(5), 425–446. ISSN 0883-7252. https://doi.org/10.1002/jae.683.

Engle RF (1982). "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation." *Econometrica*, **50**(4), 987–1007. ISSN 0012-9682. https://doi.org/10.2307/1912773.

Engle RF (2000). "The Econometrics of Ultra-High-Frequency Data." *Econometrica*, **68**(1), 1–22. ISSN 0012-9682. https://doi.org/10.1111/1468-0262.00091.

Engle RF, Russell JR (1998). "Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data." *Econometrica*, **66**(5), 1127–1162. ISSN 0012-9682. https://doi.org/10.2307/2999632.

Ghysels E, Jasiak J (1998). "GARCH for Irregularly Spaced Financial Data: The ACD-GARCH Model." *Studies in Nonlinear Dynamics and Econometrics*, **2**(4), 133–149. ISSN 1081-1826. https://doi.org/10.2202/1558-3708.1035.

Gonçalves E, Mendes-Lopes N (2020). "Signed Compound Poisson Integer-Valued GARCH Processes." *Communications in Statistics - Theory and Methods*, **49**(22), 5468–5492. ISSN 0361-0926. https://doi.org/10.1080/03610926.2019.1619767.

Grammig J, Wellner M (2002). "Modeling the Interdependence of Volatility and Inter-Transaction Duration Processes." *Journal of Econometrics*, **106**(2), 369–400. https://doi.org/10.1016/S0304-4076(01)00105-1.

Hansen PR, Lunde A (2006). "Realized Variance and Market Microstructure Noise." *Journal of Business & Economic Statistics*, **24**(2), 127–161. ISSN 0735-0015. https://doi.org/10.1198/073500106000000071.

Harvey AC (2013). *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series*. First Edition. Cambridge University Press, New York. ISBN 978-1-107-63002-4. https://doi.org/10.1017/cbo9781139540933.

Hastie T, Tibshirani R, Friedman J (2008). *The Elements of Statistical Learning*. Second Edition. Springer, New York. ISBN 978-0-387-84857-0. https://doi.org/10.1007/978-0-387-84858-7.

Holý V (2022). "Package 'gasmodel'." https://cran.r-project.org/package=gasmodel.

Holý V, Tomanová P (2019). "Estimation of Ornstein-Uhlenbeck Process Using Ultra-High-Frequency Data with Application to Intraday Pairs Trading Strategy." https://arxiv.org/abs/1811.09312.

Holý V, Tomanová P (2022). "Modeling Price Clustering in High-Frequency Prices." *Quantitative Finance*, **22**(9), 1649–1663. ISSN 1469-7688. https://doi.org/10.1080/14697688.2022.2050285.

Holý V, Tomanová P (2023). "Streaming Approach to Quadratic Covariation Estimation Using Financial Ultra-High-Frequency Data." *Computational Economics*, **62**(1), 463–485. ISSN 0927-7099. https://doi.org/10.1007/s10614-021-10210-w.

Huptas R (2016). "The UHF-GARCH-Type Model in the Analysis of Intraday Volatility and Price Durations - the Bayesian Approach." *Central European Journal of Economic Modelling and Econometrics*, **8**(1), 1–20. ISSN 2080-0886. https://doi.org/10.24425/cejeme.2016.119184.

Karlis D, Ntzoufras I (2006). "Bayesian Analysis of the Differences of Count Data." *Statistics in Medicine*, **25**(11), 1885–1905. ISSN 0277-6715. https://doi.org/10.1002/sim.2382.

Karlis D, Ntzoufras I (2009). "Bayesian Modelling of Football Outcomes: Using the Skellam's Distribution for the Goal Difference." *IMA Journal of Management Mathematics*, **20**(2), 133–145. ISSN 1471-6798. https://doi.org/10.1093/imaman/dpn026.

Koopman SJ, Lit R (2019). "Forecasting Football Match Results in National League Competitions Using Score-Driven Time Series Models." *International Journal of Forecasting*, **35**(2), 797–809. ISSN 0169-2070. https://doi.org/10.1016/j.ijforecast.2018.10.011.

Koopman SJ, Lit R, Lucas A (2017). "Intraday Stochastic Volatility in Discrete Price Changes: The Dynamic Skellam Model." *Journal of the American Statistical Association*, **112**(520), 1490–1503. ISSN 0162-1459. https://doi.org/10.1080/01621459.2017.1302878.

Koopman SJ, Lit R, Lucas A, Opschoor A (2018). "Dynamic Discrete Copula Models for High-Frequency Stock Price Changes." *Journal of Applied Econometrics*, **33**(7), 966–985. ISSN 0883-7252. https://doi.org/10.1002/jae.2645.

Lambert D (1992). "Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing." *Technometrics*, **34**(1), 1–14. ISSN 0040-1706. https://doi.org/10.2307/1269547.

Liu C, Maheu JM (2012). "Intraday Dynamics of Volatility and Duration: Evidence from Chinese Stocks." *Pacific-Basin Finance Journal*, **20**(3), 329–348. ISSN 0927-538X. https://doi.org/10.1016/j.pacfin.2011.11.001.

Liu LY, Patton AJ, Sheppard K (2015). "Does Anything Beat 5-Minute RV? A Comparison of Realized Measures Across Multiple Asset Classes." *Journal of Econometrics*, **187**(1), 293–311. ISSN 1872-6895. https://doi.org/10.1016/j.jeconom.2015.02.008.

Manganelli S (2005). "Duration, Volume and Volatility Impact of Trades." *Journal of Financial Markets*, **8**(4), 377–399. ISSN 1386-4181. https://doi.org/10.1016/j.finmar.2005.06.002.

Meddahi N, Renault E, Werker B (2006). "GARCH and Irregularly Spaced Data." *Economics Letters*, **90**(2), 200–204. ISSN 0165-1765. https://doi.org/10.1016/j.econlet.2005.07.027.

Münnix MC, Schfer R, Guhr T (2010). "Impact of the Tick-Size on Financial Returns and Correlations." *Physica A: Statistical Mechanics and Its Applications*, **389**(21), 4828–4843. ISSN 0378-4371. https://doi.org/10.1016/j.physa.2010.06.037.

Nelson DB (1991). "Conditional Heteroskedasticity in Asset Returns: A New Approach." *Econometrica*, **59**(2), 347. ISSN 0012-9682. https://doi.org/10.2307/2938260.

R Core Team (2022). "R: A Language and Environment for Statistical Computing." https://www.r-project.org.

Racicot FÉ, Théoret R, Coën A (2008). "Forecasting Irregularly Spaced UHF Financial Data: Realized Volatility vs UHF-GARCH models." *International Advances in Economic Research*, **14**(1), 112–124. ISSN 1083-0898. https://doi.org/10.1007/s11294-008-9134-2.

Russell JR, Engle RF (2005). "A Discrete-State Continuous-Time Model of Financial Transactions Prices and Times: The Autoregressive Conditional Multinomial-Autoregressive Conditional Duration Model." *Journal of Business & Economic Statistics*, **23**(2), 166–180. ISSN 0735-0015. https://doi.org/10.1198/073500104000000541.

Shephard N, Yang JJ (2017). "Continuous Time Analysis of Fleeting Discrete Price Moves." *Journal of the American Statistical Association*, **112**(519), 1090–1106. ISSN 0162-1459. https://doi.org/10.1080/01621459.2016.1192544.

Skellam JG (1946). "The Frequency Distribution of the Difference Between Two Poisson Variates Belonging to Different Populations." *Journal of the Royal Statistical Society*, **109**(3), 296. ISSN 0952-8385. https://doi.org/10.2307/2981372.

A Evidence from Further Stocks

In this appendix, we report the results for additional stocks: Caterpillar (CA), traded on NYSE with an average of 2.320 transactions per second; Cisco (CSCO), traded on NASDAQ with an average of 5.738 transactions per second; Electronic Arts (EA), traded on NASDAQ with an average of 1.518 transactions per second; Intel (INTC), traded on NASDAQ with an average of 8.683 transactions per second; Mastercard (MA), traded on NYSE with an average of 2.732 transactions per second; and McDonald's (MCD), traded on NYSE with an average of 2.402 transactions per second.

In general, these results closely resemble those observed for the IBM stock. Nonetheless, there are two distinctions. First, while smoothing splines effectively capture the diurnal patterns of price volatility in the IBM stock, they struggle to account for the impact of news events occurring at regular times. This discrepancy is particularly pronounced when analyzing the INTC stock. Nonetheless, this isn't a significant limitation for our study. Second, zero-inflation is not necessary in most days for the CSCO and INTC stocks, which are the two most frequently traded stocks in our sample. Although zero price changes occur more frequently for these stocks compared to others, a regular Skellam distribution suffices. In other aspects, the results reinforce the implications drawn from the analysis of the IBM stock.

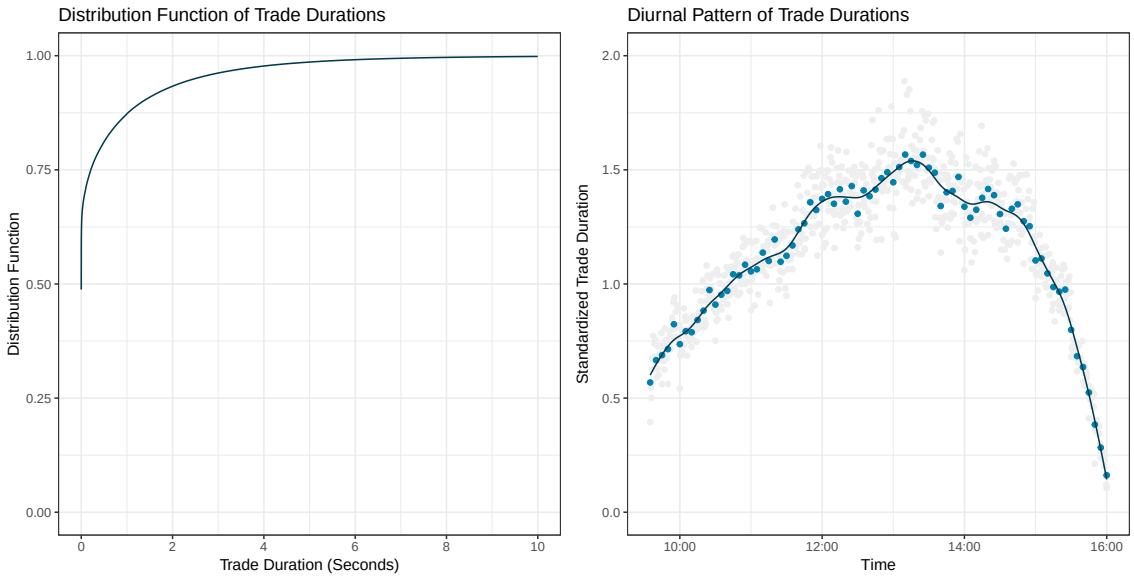

Figure 5: The empirical distribution function of trade durations (left) and average trade durations in 5 minute and 30 second intraday intervals with a smoothed curve (right) for the CAT stock.

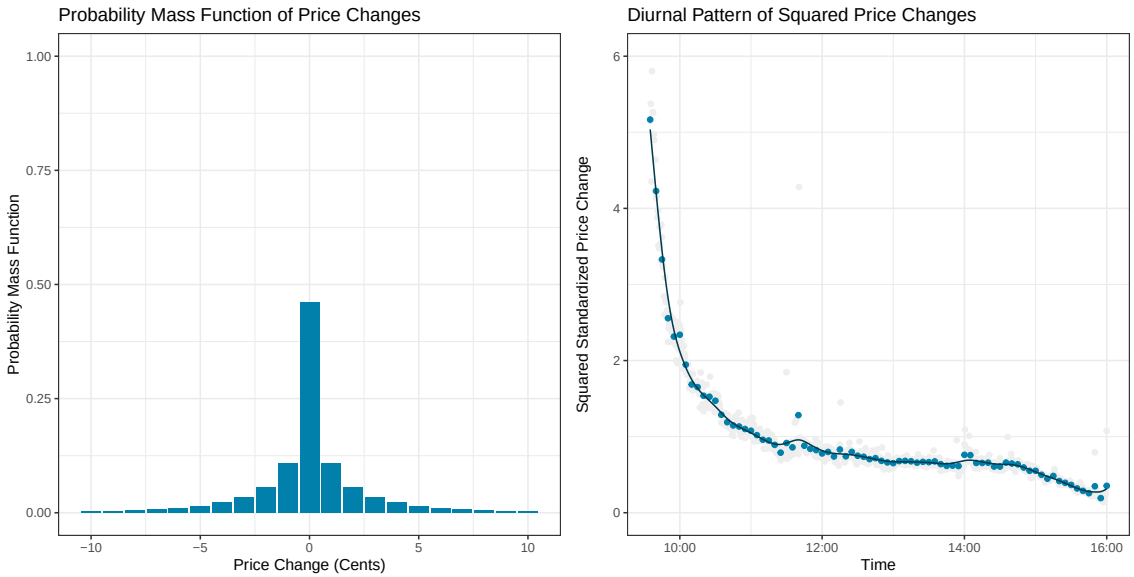

Figure 6: The empirical probability mass function of price changes (left) and average squared price changes in 5 minute and 30 second intraday intervals with a smoothed curve (right) for the CAT stock.

Figure 7: The average diurnally adjusted squared price changes (left) and diurnally adjusted squared price changes per second (right) in 50 millisecond and half second intervals of diurnally adjusted trade durations with a smoothed curve for the CAT stock.

Table 5: The minimum, median, and maximum values of estimated parameters of various daily models for the CAT stock.

		Variance Models					Overdispersion Models				
Coef.	Trans.	I	II	III	IV	V	VI	VII	VIII	IX	X
θ	Min		-0.195	-0.187		-0.268		-0.372	-0.554		-0.509
	Med		-0.120	-0.059		-0.155		-0.283	-0.432		-0.387
	Max		-0.074	-0.013		-0.096		-0.217	-0.347		-0.289
ω	Min	1.122	1.064	1.546	1.318	1.357	1.238	0.904	1.190	1.423	1.178
	Med	1.724	1.633	2.147	1.915	1.982	1.834	1.497	1.829	2.023	1.774
	Max	2.817	2.703	3.302	3.070	3.136	2.906	2.655	3.077	3.194	3.005
φ	Min		0.724		0.821	0.735		0.910		0.859	0.937
	Med		0.908		0.950	0.933		0.957		0.950	0.971
	Max		0.986		0.996	0.990		0.993		0.995	0.997
α	Min		0.026		0.021	0.029		0.025		0.020	0.025
	Med		0.200		0.199	0.221		0.210		0.190	0.204
	Max		0.434		0.421	0.463		0.287		0.416	0.297
π	Min			0.217	0.177	0.176			0.202	0.166	0.179
	Med			0.277	0.237	0.237			0.256	0.226	0.232
	Max			0.343	0.317	0.318			0.325	0.313	0.309

Table 6: The R^2 statistics of residuals and squared residuals regressed on their lagged values with the average log-likelihood of an observation for various daily models for the CAT stock.

		Variance Models					Overdispersion Models				
Statistic	Lag	I	II	III	IV	V	VI	VII	VIII	IX	X
AR R^2	1	0.117	0.040	0.092	0.092	0.043	0.115	0.005	0.003	0.091	0.004
	10	0.153	0.056	0.123	0.117	0.058	0.150	0.011	0.005	0.115	0.006
	100	0.156	0.058	0.126	0.119	0.060	0.154	0.013	0.008	0.117	0.009
ARCH R^2	1	0.099	0.007	0.085	0.016	0.008	0.099	0.001	0.003	0.017	0.001
	10	0.137	0.009	0.127	0.017	0.010	0.139	0.003	0.038	0.018	0.003
	100	0.173	0.017	0.162	0.022	0.016	0.177	0.008	0.065	0.023	0.006
Log-Likelihood		-2.057	-1.947	-1.948	-1.908	-1.881	-2.050	-1.923	-1.889	-1.907	-1.857

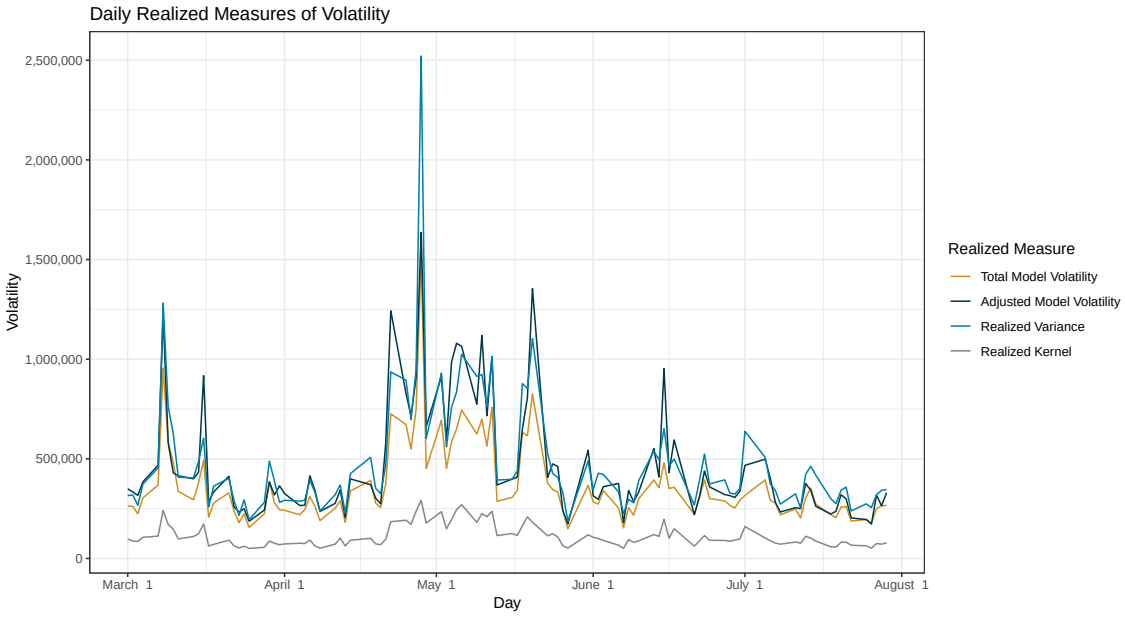

Figure 8: The daily values of various volatility realized measures for the CAT stock.

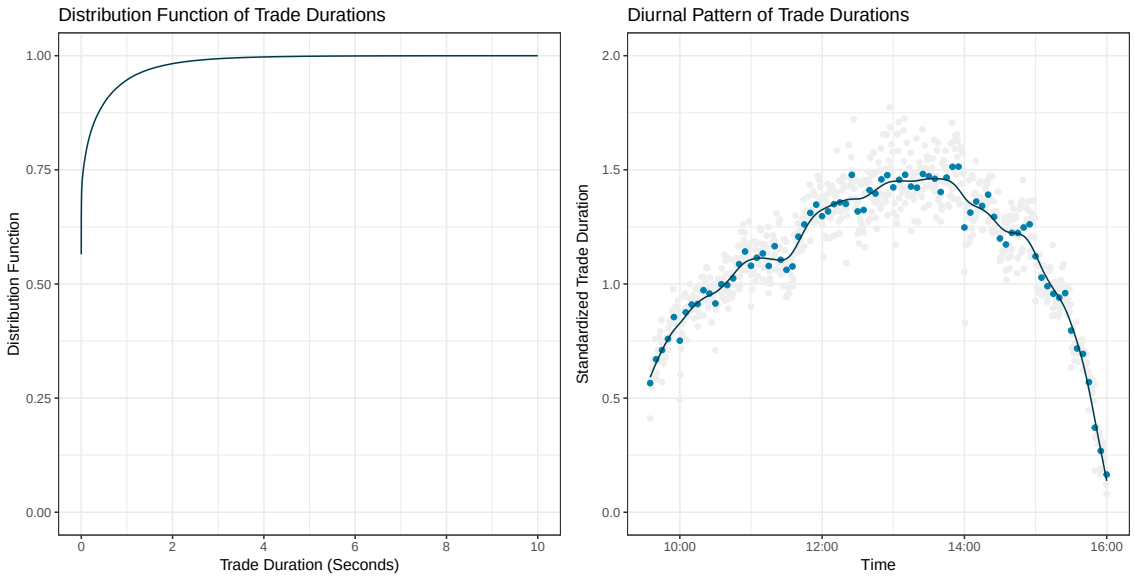

Figure 9: The empirical distribution function of trade durations (left) and average trade durations in 5 minute and 30 second intraday intervals with a smoothed curve (right) for the CSCO stock.

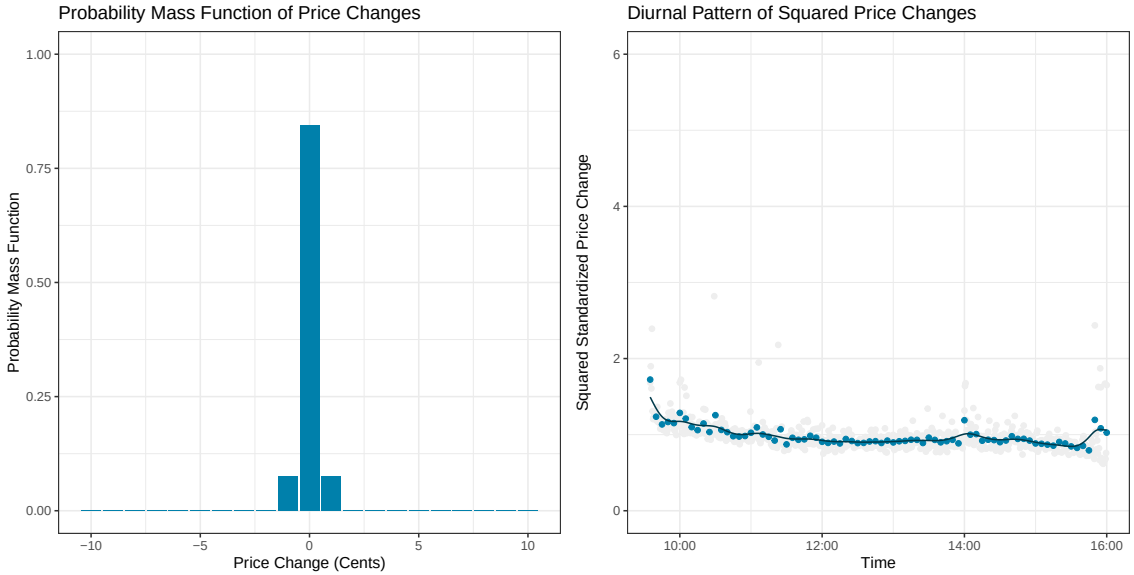

Figure 10: The empirical probability mass function of price changes (left) and average squared price changes in 5 minute and 30 second intraday intervals with a smoothed curve (right) for the CSCO stock.

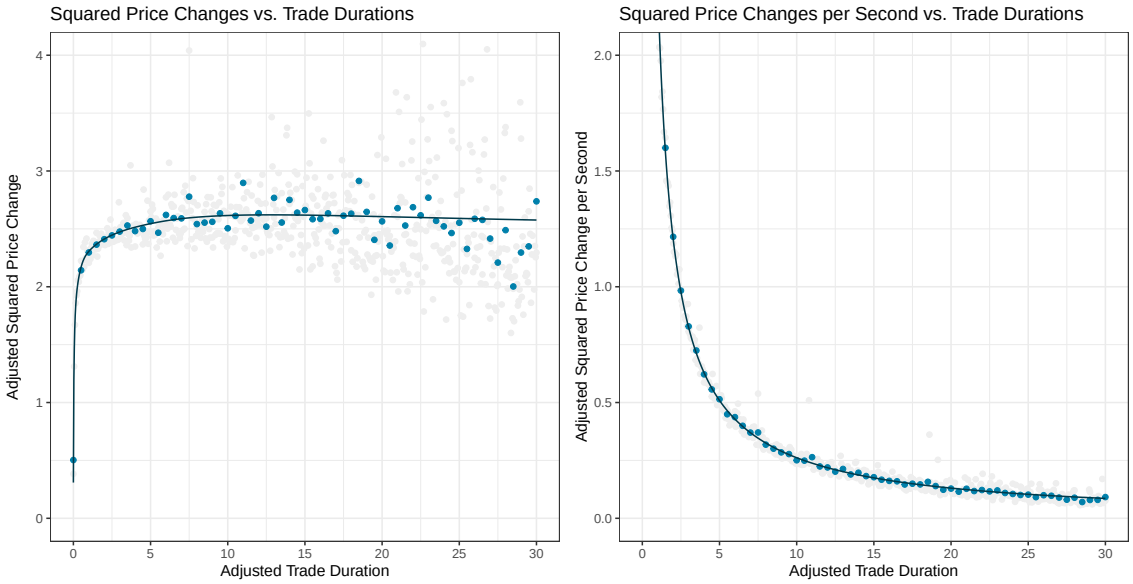

Figure 11: The average diurnally adjusted squared price changes (left) and diurnally adjusted squared price changes per second (right) in 50 millisecond and half second intervals of diurnally adjusted trade durations with a smoothed curve for the CSCO stock.

Table 7: The minimum, median, and maximum values of estimated parameters of various daily models for the CSCO stock.

		Variance Models					Overdispersion Models				
Coef.	Trans.	I	II	III	IV	V	VI	VII	VIII	IX	X
	Min		-0.091	-0.013		-0.090		-0.571	-0.620		-0.571
θ	Med		-0.056	-0.007		-0.056		-0.466	-0.483		-0.469
	Max		-0.015	-0.002		-0.018		-0.252	-0.266		-0.252
	Min	-1.784	-1.876	-1.784	-1.924	-1.876	-1.729	-2.640	-2.695	-1.796	-2.640
ω	Med	-1.659	-1.714	-1.659	-1.736	-1.716	-1.601	-2.251	-2.316	-1.622	-2.250
	Max	-1.466	-1.467	-1.014	-1.451	-1.345	-1.386	-1.750	-1.856	-1.289	-1.750
	Min		0.533		0.670	0.533		0.978		0.731	0.978
φ	Med		0.752		0.843	0.752		0.998		0.887	0.998
	Max		0.984		0.986	0.983		1.000		0.986	1.000
	Min		0.100		0.113	0.100		0.014		0.111	0.014
α	Med		0.789		0.746	0.788		0.079		0.720	0.079
	Max		1.323		1.071	1.323		0.240		1.120	0.299
	Min			0.000	0.000	0.000			0.000	0.000	0.000
π	Med			0.000	0.000	0.000			0.000	0.000	0.000
	Max			0.317	0.179	0.176			0.154	0.199	0.152

Table 8: The R^2 statistics of residuals and squared residuals regressed on their lagged values with the average log-likelihood of an observation for various daily models for the CSCO stock.

		Variance Models					Overdispersion Models				
Statistic	Lag	I	II	III	IV	V	VI	VII	VIII	IX	X
AR R^2	1	0.126	0.058	0.122	0.075	0.058	0.123	0.000	0.000	0.072	0.000
	10	0.174	0.084	0.170	0.101	0.083	0.165	0.004	0.003	0.095	0.004
	100	0.178	0.085	0.173	0.102	0.084	0.168	0.005	0.004	0.096	0.005
ARCH R^2	1	0.094	0.005	0.093	0.006	0.005	0.091	0.000	0.000	0.008	0.000
	10	0.130	0.011	0.129	0.008	0.010	0.125	0.002	0.004	0.010	0.002
	100	0.158	0.020	0.157	0.015	0.019	0.148	0.006	0.018	0.014	0.006
Log-Likelihood		-0.512	-0.480	-0.510	-0.488	-0.480	-0.516	-0.448	-0.451	-0.489	-0.448

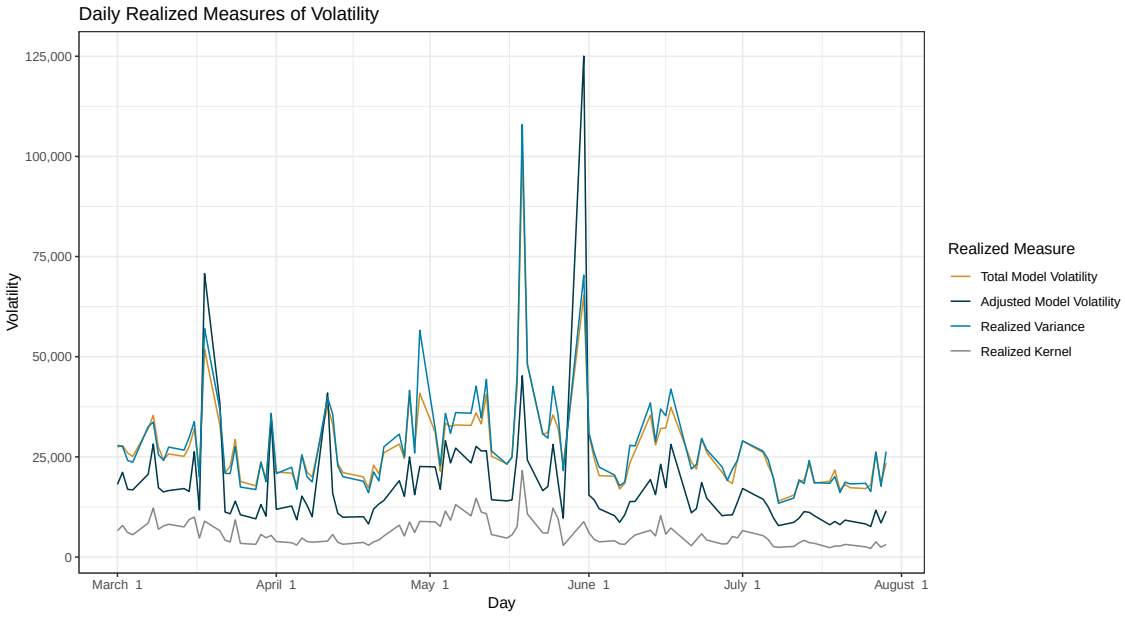

Figure 12: The daily values of various volatility realized measures for the CSCO stock.

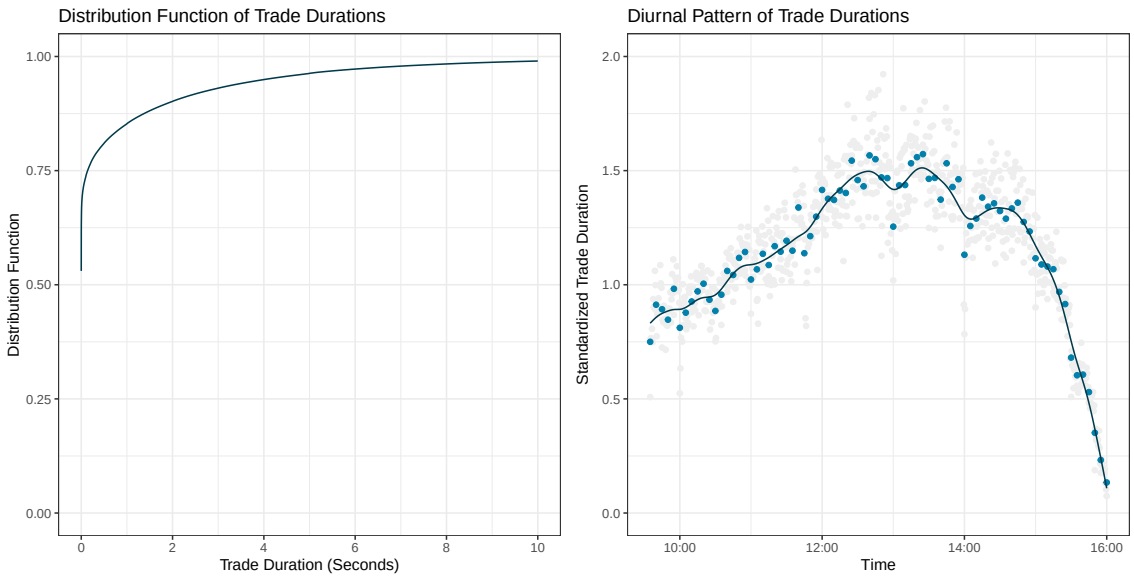

Figure 13: The empirical distribution function of trade durations (left) and average trade durations in 5 minute and 30 second intraday intervals with a smoothed curve (right) for the EA stock.

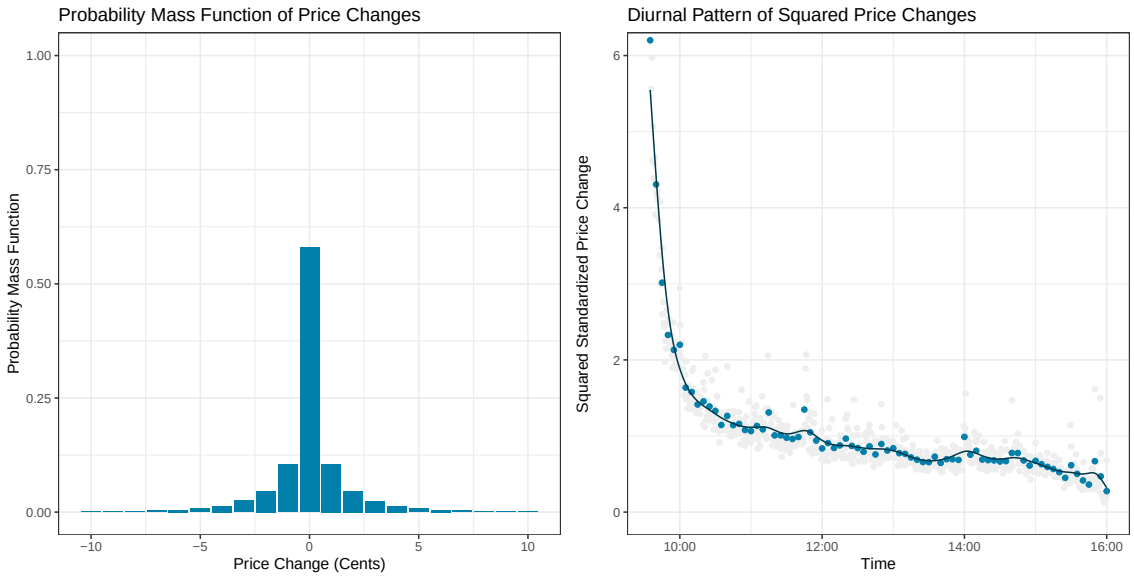

Figure 14: The empirical probability mass function of price changes (left) and average squared price changes in 5 minute and 30 second intraday intervals with a smoothed curve (right) for the EA stock.

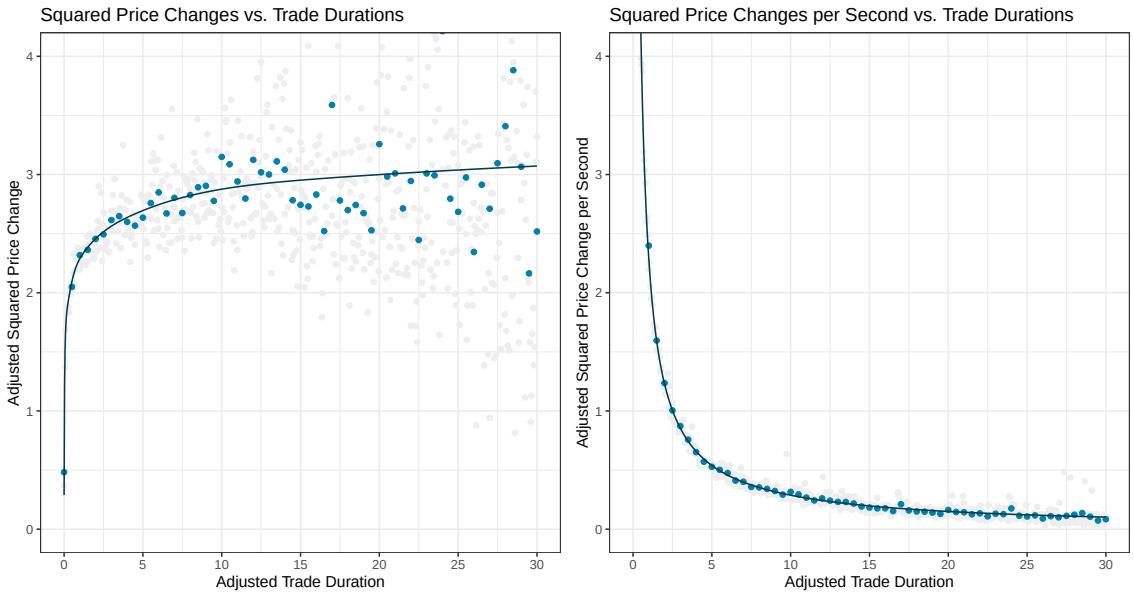

Figure 15: The average diurnally adjusted squared price changes (left) and diurnally adjusted squared price changes per second (right) in 50 millisecond and half second intervals of diurnally adjusted trade durations with a smoothed curve for the EA stock.

Table 9: The minimum, median, and maximum values of estimated parameters of various daily models for the EA stock.

		Variance Models					Overdispersion Models				
Coef.	Trans.	I	II	III	IV	V	VI	VII	VIII	IX	X
	Min		-0.124	-0.095		-0.202		-0.365	-0.631		-0.536
θ	Med		-0.073	-0.035		-0.103		-0.242	-0.416		-0.352
	Max		-0.020	-0.003		-0.032		-0.133	-0.242		-0.179
	Min	0.262	0.115	0.633	0.468	0.476	0.361	-0.183	0.232	0.506	0.176
ω	Med	0.878	0.761	1.324	1.091	1.135	0.979	0.650	1.008	1.194	0.981
	Max	2.135	1.917	3.201	3.074	2.738	2.270	1.628	2.497	3.063	2.086
	Min		0.683		0.629	0.671		0.847		0.545	0.897
φ	Med		0.916		0.950	0.945		0.954		0.949	0.971
	Max		0.994		0.999	0.998		0.997		1.000	0.999
	Min		0.043		0.017	0.020		0.031		0.008	0.032
α	Med		0.177		0.184	0.190		0.188		0.184	0.189
	Max		0.433		0.491	0.496		0.346		0.484	0.347
	Min			0.161	0.137	0.138			0.150	0.127	0.140
π	Med			0.317	0.263	0.266			0.292	0.249	0.261
	Max			0.468	0.445	0.416			0.435	0.390	0.387

Table 10: The R^2 statistics of residuals and squared residuals regressed on their lagged values with the average log-likelihood of an observation for various daily models for the EA stock.

		Variance Models					Overdispersion Models				
Statistic	Lag	I	II	III	IV	V	VI	VII	VIII	IX	X
AR R^2	1	0.101	0.043	0.087	0.074	0.045	0.099	0.006	0.002	0.072	0.004
	10	0.134	0.058	0.118	0.093	0.060	0.131	0.013	0.005	0.091	0.007
	100	0.141	0.061	0.125	0.096	0.063	0.139	0.017	0.010	0.094	0.011
ARCH R^2	1	0.094	0.008	0.086	0.013	0.009	0.092	0.001	0.002	0.014	0.001
	10	0.136	0.010	0.130	0.015	0.012	0.135	0.003	0.028	0.016	0.003
	100	0.170	0.020	0.164	0.022	0.019	0.170	0.011	0.052	0.023	0.009
Log-Likelihood		-1.590	-1.491	-1.514	-1.462	-1.446	-1.582	-1.467	-1.455	-1.460	-1.422

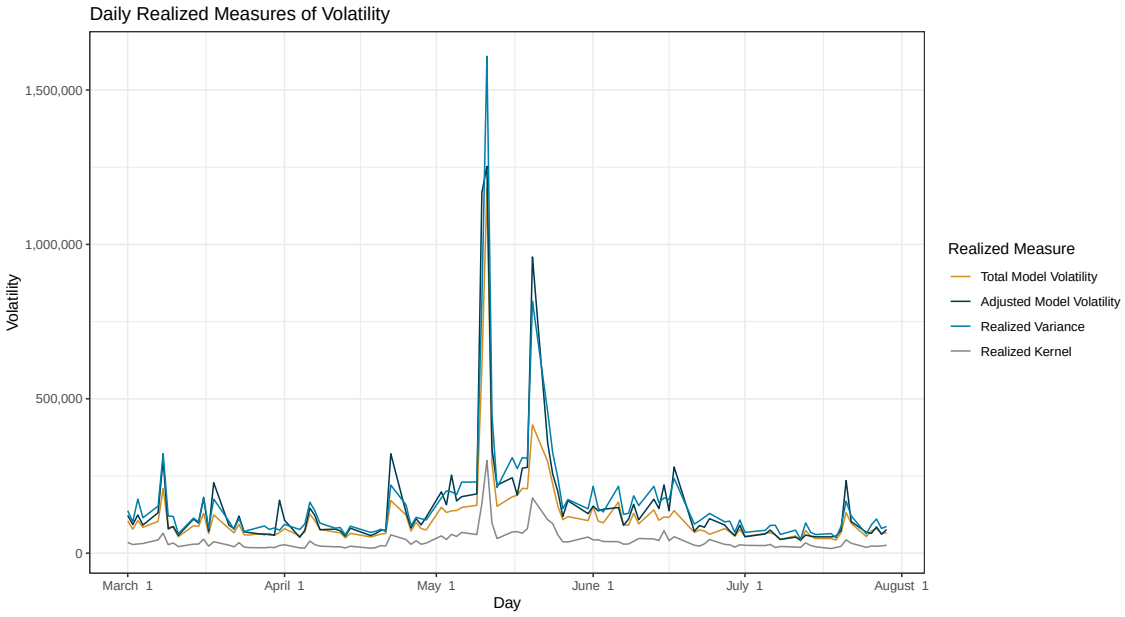

Figure 16: The daily values of various volatility realized measures for the EA stock.

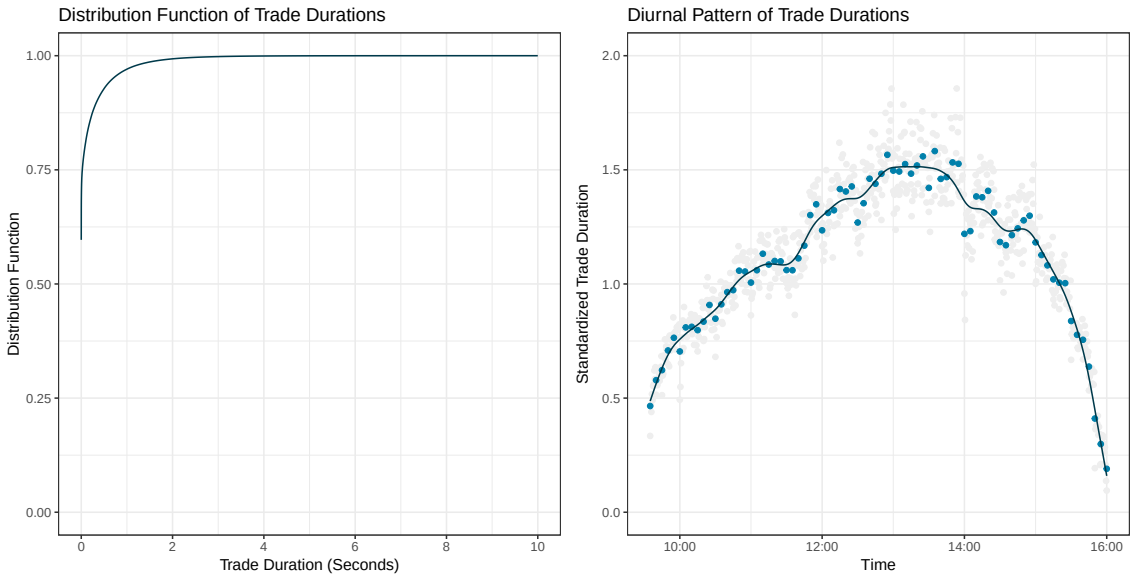

Figure 17: The empirical distribution function of trade durations (left) and average trade durations in 5 minute and 30 second intraday intervals with a smoothed curve (right) for the INTC stock.

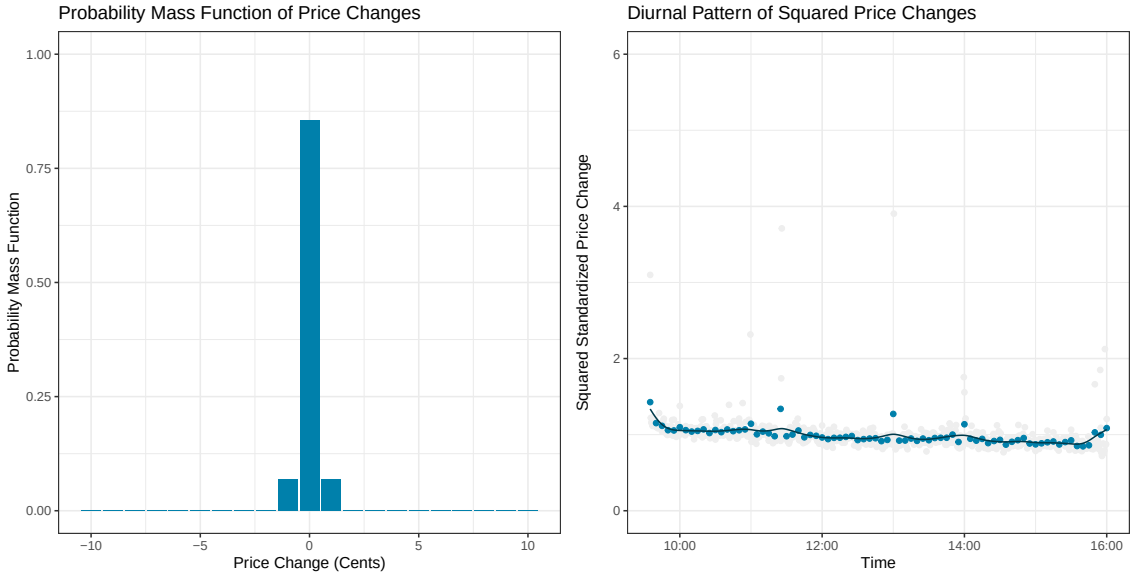

Figure 18: The empirical probability mass function of price changes (left) and average squared price changes in 5 minute and 30 second intraday intervals with a smoothed curve (right) for the INTC stock.

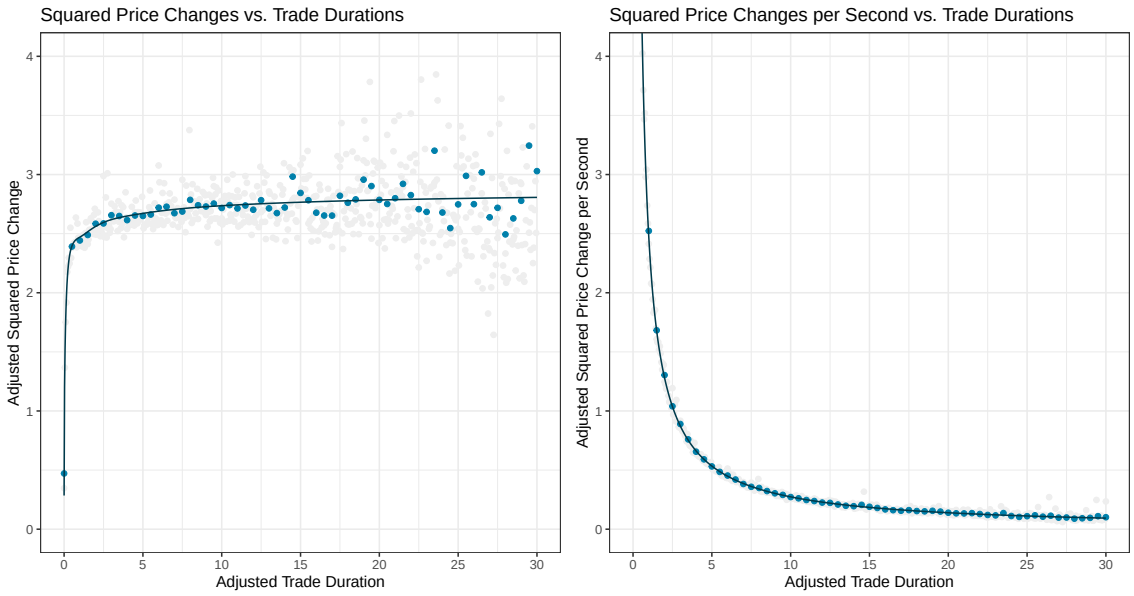

Figure 19: The average diurnally adjusted squared price changes (left) and diurnally adjusted squared price changes per second (right) in 50 millisecond and half second intervals of diurnally adjusted trade durations with a smoothed curve for the INTC stock.

Table 11: The minimum, median, and maximum values of estimated parameters of various daily models for the INTC stock.

		Variance Models					Overdispersion Models				
Coef.	Trans.	I	II	III	IV	V	VI	VII	VIII	IX	X
	Min		-0.080	-0.012		-0.081		-0.693	-0.697		-0.693
θ	Med		-0.050	-0.006		-0.050		-0.519	-0.526		-0.519
	Max		-0.012	-0.001		-0.011		-0.329	-0.337		-0.329
	Min	-1.930	-2.131	-1.930	-2.105	-2.061	-1.904	-3.160	-3.047	-2.051	-3.160
ω	Med	-1.760	-1.849	-1.760	-1.875	-1.839	-1.723	-2.516	-2.548	-1.812	-2.516
	Max	-1.570	-1.611	-1.570	-1.638	-1.616	-1.531	-1.194	-2.145	-1.509	-1.212
	Min		0.540		0.740	0.540		0.989		0.798	0.988
φ	Med		0.772		0.872	0.772		0.999		0.893	0.999
	Max		0.974		0.982	0.977		1.000		0.992	1.000
	Min		0.102		0.104	0.107		0.010		0.107	0.010
α	Med		0.895		0.842	0.894		0.047		0.839	0.047
	Max		1.325		1.018	1.324		0.208		1.027	0.208
	Min			0.000	0.000	0.000			0.000	0.000	0.000
π	Med			0.000	0.000	0.000			0.000	0.000	0.000
	Max			0.000	0.063	0.102			0.028	0.074	0.016

Table 12: The R^2 statistics of residuals and squared residuals regressed on their lagged values with the average log-likelihood of an observation for various daily models for the INTC stock.

		Variance Models					Overdispersion Models				
Statistic	Lag	I	II	III	IV	V	VI	VII	VIII	IX	X
AR R^2	1	0.132	0.061	0.129	0.079	0.061	0.131	0.000	0.000	0.077	0.000
	10	0.186	0.091	0.182	0.109	0.090	0.183	0.003	0.003	0.106	0.003
	100	0.188	0.092	0.185	0.109	0.091	0.185	0.004	0.004	0.106	0.004
ARCH R^2	1	0.096	0.005	0.095	0.008	0.005	0.095	0.000	0.000	0.009	0.000
	10	0.143	0.012	0.142	0.010	0.012	0.139	0.002	0.003	0.011	0.002
	100	0.175	0.029	0.174	0.022	0.029	0.165	0.006	0.016	0.020	0.007
Log-Likelihood		-0.464	-0.436	-0.461	-0.442	-0.435	-0.467	-0.399	-0.401	-0.443	-0.399

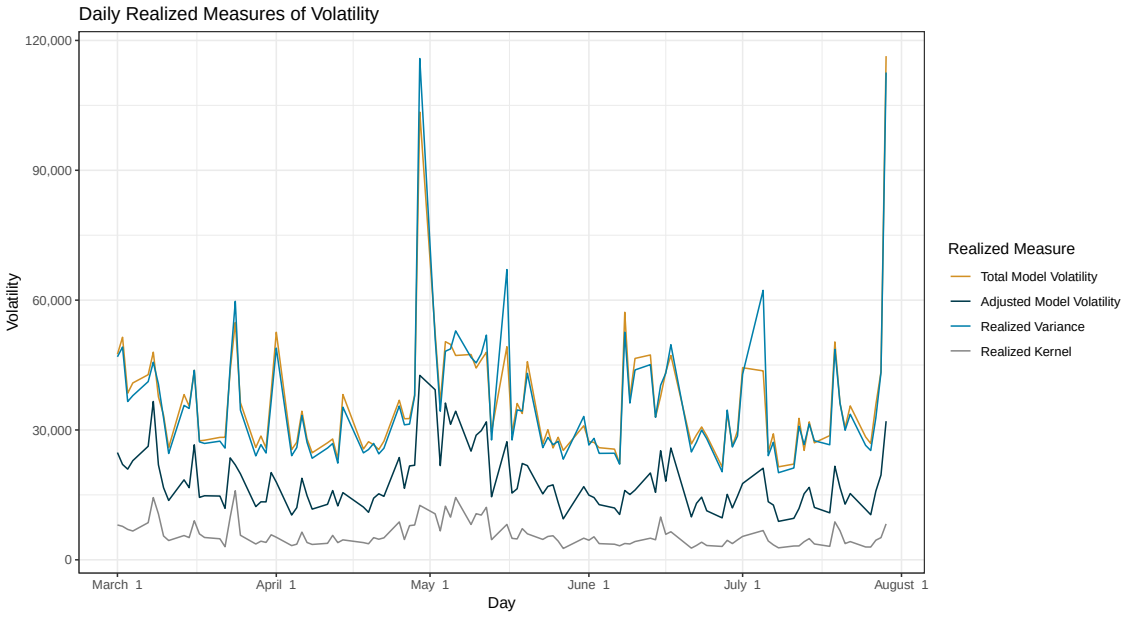

Figure 20: The daily values of various volatility realized measures for the INTC stock.

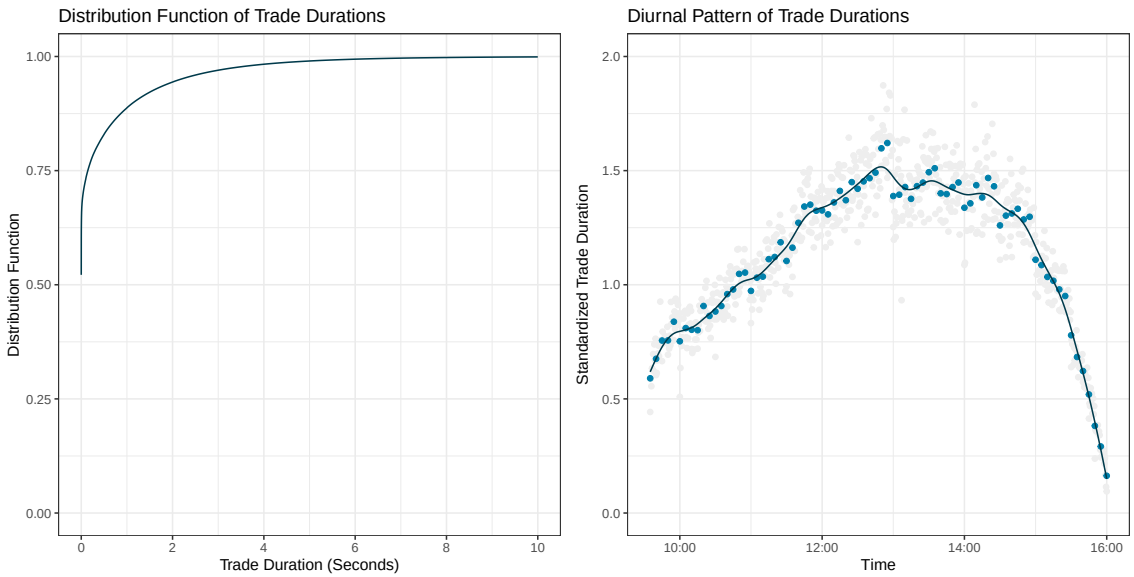

Figure 21: The empirical distribution function of trade durations (left) and average trade durations in 5 minute and 30 second intraday intervals with a smoothed curve (right) for the MA stock.

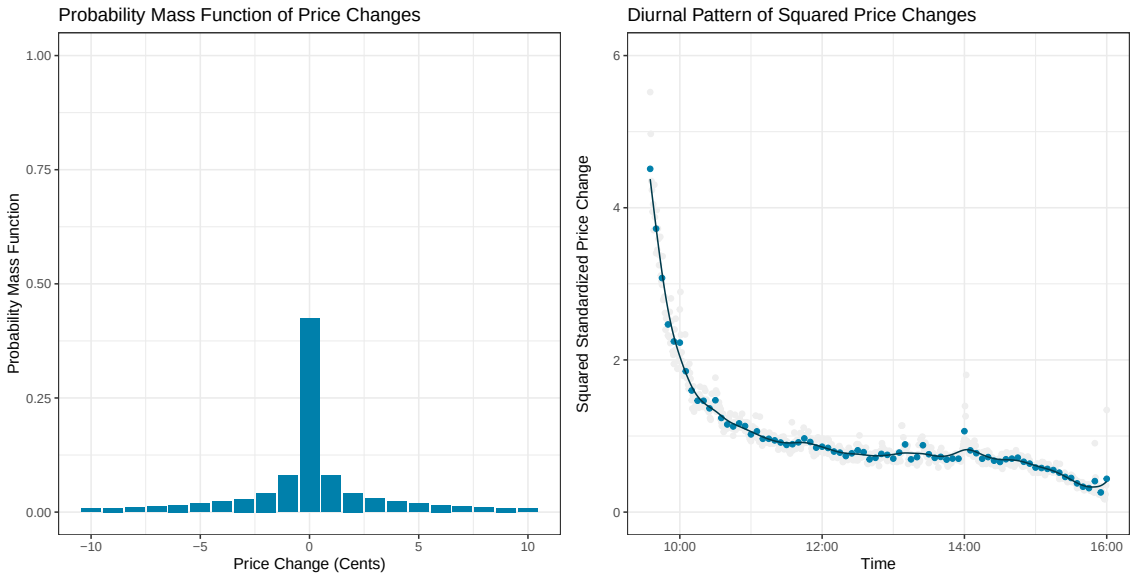

Figure 22: The empirical probability mass function of price changes (left) and average squared price changes in 5 minute and 30 second intraday intervals with a smoothed curve (right) for the MA stock.

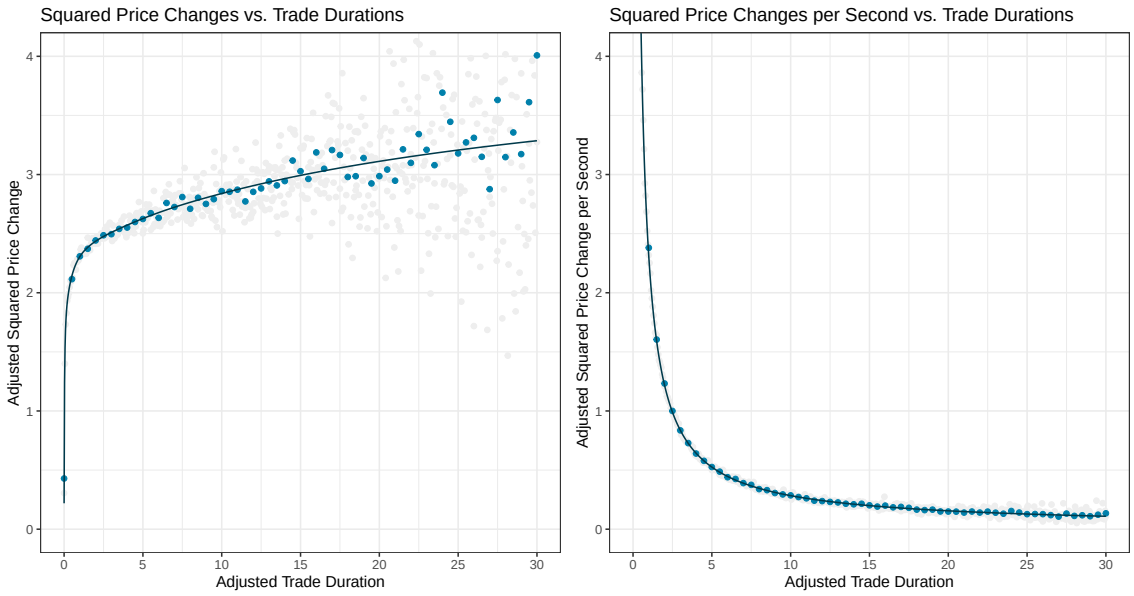

Figure 23: The average diurnally adjusted squared price changes (left) and diurnally adjusted squared price changes per second (right) in 50 millisecond and half second intervals of diurnally adjusted trade durations with a smoothed curve for the MA stock.

Table 13: The minimum, median, and maximum values of estimated parameters of various daily models for the MA stock.

		Variance Models					Overdispersion Models				
Coef.	Trans.	I	II	III	IV	V	VI	VII	VIII	IX	X
	Min		-0.197	-0.238		-0.351		-0.432	-0.552		-0.528
θ	Med		-0.118	-0.106		-0.167		-0.312	-0.468		-0.445
	Max		-0.078	-0.041		-0.098		-0.216	-0.392		-0.325
	Min	2.495	2.264	2.923	1.602	2.689	2.539	2.032	2.671	0.745	2.261
ω	Med	3.218	2.949	3.722	3.444	3.456	3.263	2.741	3.421	3.500	3.204
	Max	3.974	4.061	4.000	3.994	4.004	3.903	3.800	3.892	3.913	3.916
	Min		0.420		0.442	0.287		0.800		0.385	0.866
φ	Med		0.953		0.986	0.963		0.976		0.982	0.989
	Max		1.000		1.000	1.000		1.000		1.000	1.000
	Min		0.001		0.000	0.000		0.000		0.000	0.000
α	Med		0.047		0.021	0.031		0.050		0.020	0.033
	Max		0.169		0.149	0.155		0.167		0.147	0.154
	Min			0.250	0.000	0.231			0.244	0.234	0.228
π	Med			0.332	0.322	0.323			0.329	0.317	0.319
	Max			0.388	0.394	0.398			0.386	0.395	0.387

Table 14: The R^2 statistics of residuals and squared residuals regressed on their lagged values with the average log-likelihood of an observation for various daily models for the MA stock.

		Variance Models					Overdispersion Models				
Statistic	Lag	I	II	III	IV	V	VI	VII	VIII	IX	X
AR R^2	1	0.125	0.055	0.086	0.118	0.061	0.123	0.005	0.007	0.118	0.008
	10	0.167	0.081	0.121	0.157	0.087	0.165	0.015	0.010	0.156	0.012
	100	0.170	0.083	0.123	0.159	0.089	0.167	0.017	0.013	0.157	0.014
ARCH R^2	1	0.090	0.030	0.061	0.056	0.030	0.088	0.009	0.004	0.058	0.004
	10	0.122	0.038	0.099	0.068	0.045	0.121	0.020	0.035	0.071	0.020
	100	0.149	0.044	0.126	0.077	0.054	0.149	0.025	0.058	0.081	0.026
Log-Likelihood		-2.736	-2.626	-2.466	-2.451	-2.415	-2.734	-2.586	-2.413	-2.455	-2.382

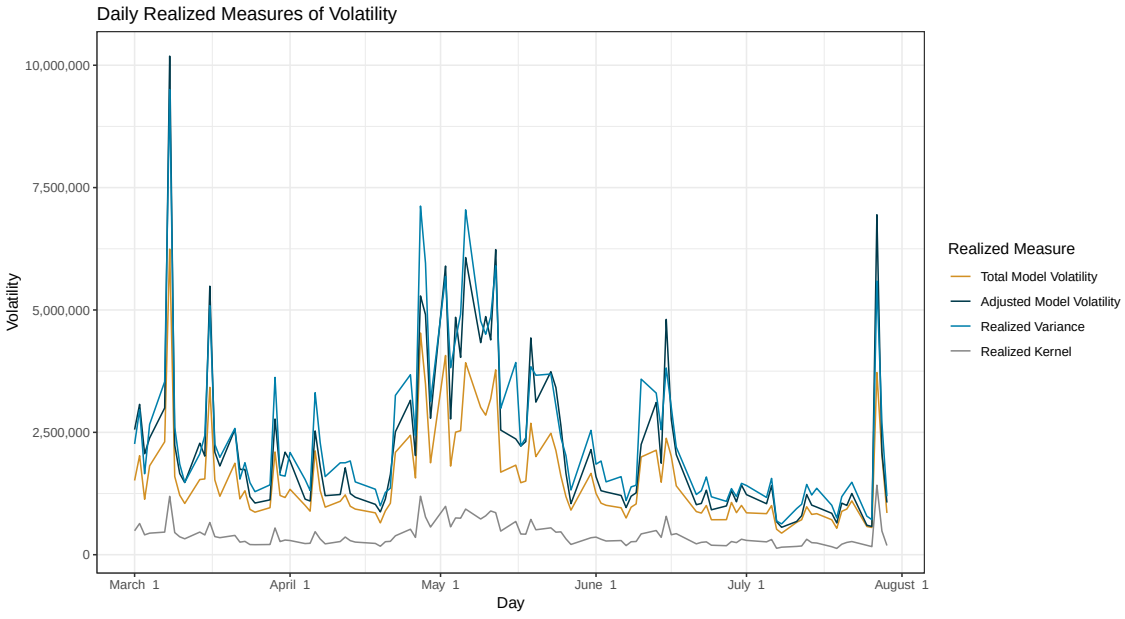

Figure 24: The daily values of various volatility realized measures for the MA stock.

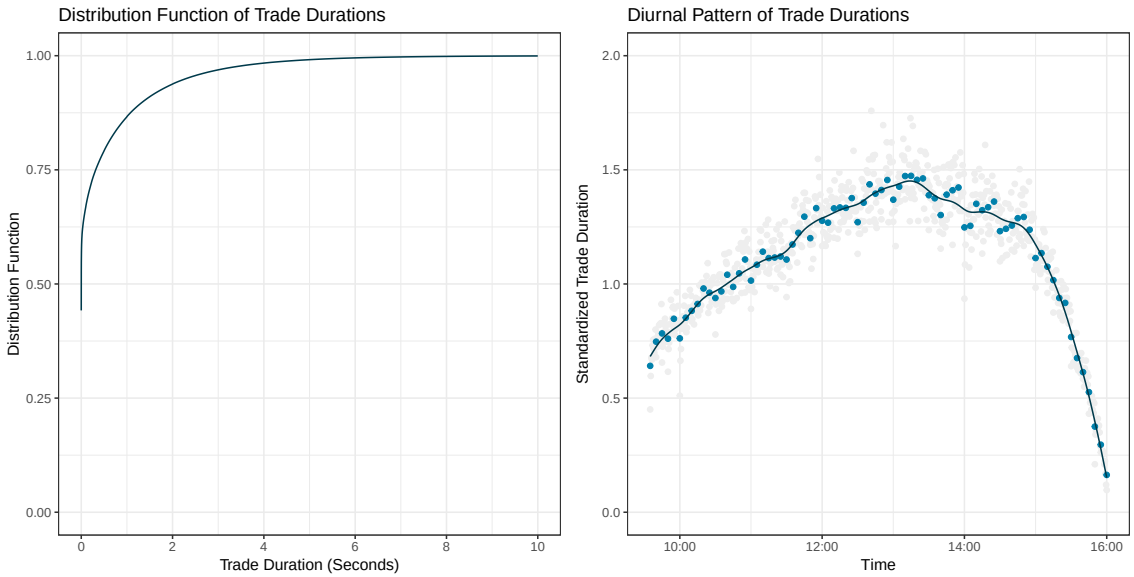

Figure 25: The empirical distribution function of trade durations (left) and average trade durations in 5 minute and 30 second intraday intervals with a smoothed curve (right) for the MCD stock.

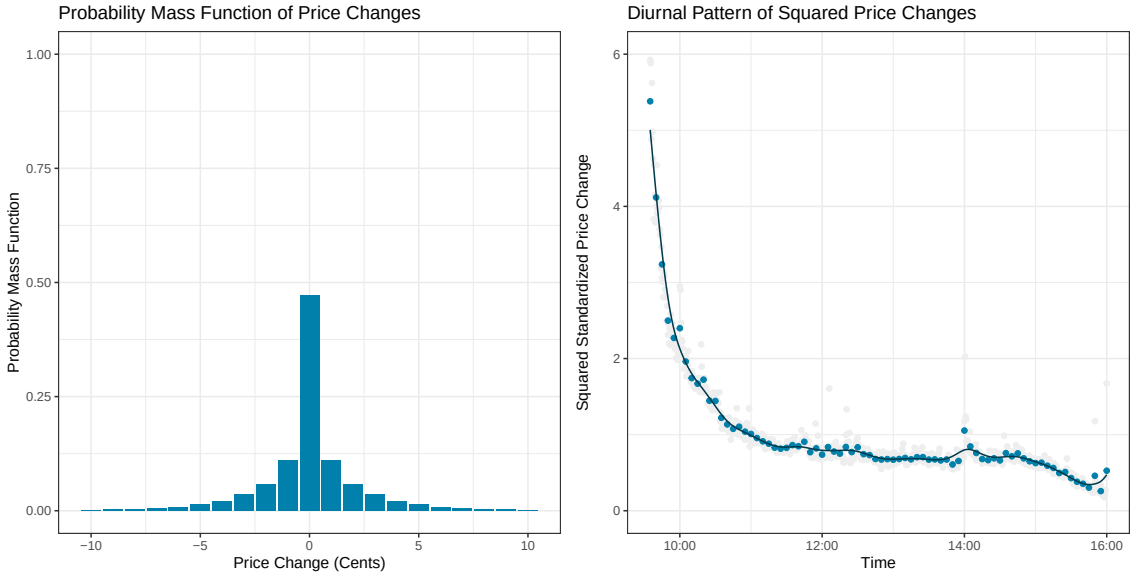

Figure 26: The empirical probability mass function of price changes (left) and average squared price changes in 5 minute and 30 second intraday intervals with a smoothed curve (right) for the MCD stock.

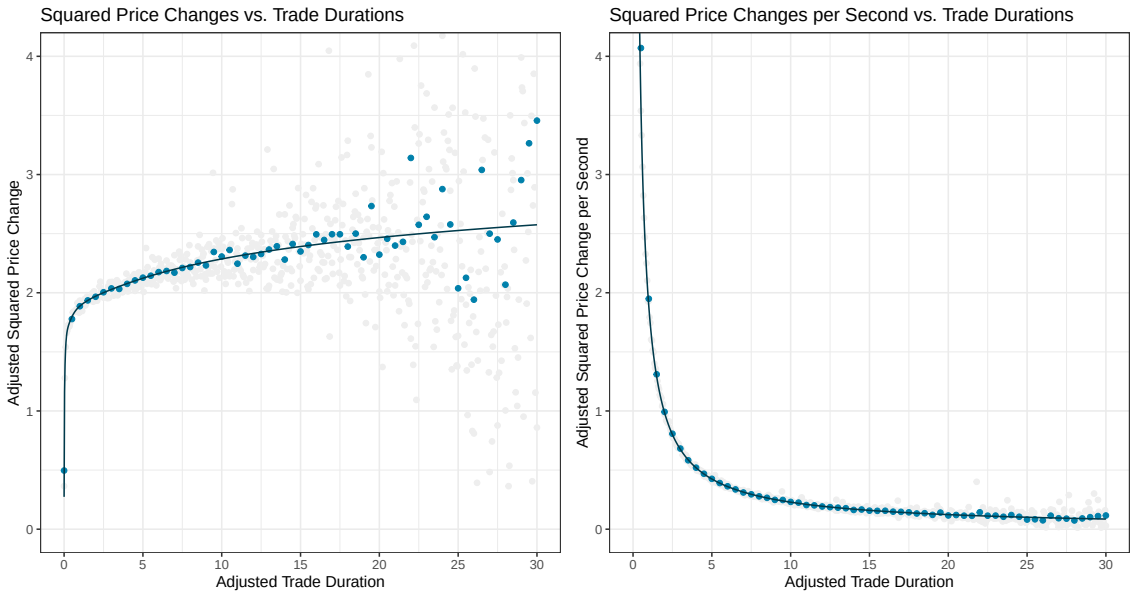

Figure 27: The average diurnally adjusted squared price changes (left) and diurnally adjusted squared price changes per second (right) in 50 millisecond and half second intervals of diurnally adjusted trade durations with a smoothed curve for the MCD stock.

Table 15: The minimum, median, and maximum values of estimated parameters of various daily models for the MCD stock.

		Variance Models					Overdispersion Models				
Coef.	Trans.	I	II	III	IV	V	VI	VII	VIII	IX	X
	Min		-0.173	-0.135		-0.231		-0.377	-0.568		-0.490
θ	Med		-0.112	-0.045		-0.138		-0.297	-0.443		-0.394
	Max		-0.054	-0.008		-0.064		-0.232	-0.358		-0.305
	Min	1.017	0.980	1.275	1.127	1.159	1.125	0.776	0.961	1.226	0.987
ω	Med	1.466	1.383	1.802	1.578	1.616	1.573	1.212	1.473	1.684	1.446
	Max	2.556	2.381	2.938	2.744	2.832	2.648	2.353	2.688	2.911	2.630
	Min		0.745		0.821	0.783		0.904		0.817	0.942
φ	Med		0.899		0.946	0.924		0.958		0.945	0.974
	Max		0.992		0.996	0.996		0.999		0.999	0.998
	Min		0.030		0.029	0.018		0.021		0.027	0.019
α	Med		0.223		0.217	0.235		0.211		0.214	0.211
	Max		0.425		0.461	0.466		0.278		0.435	0.308
	Min			0.185	0.144	0.139			0.169	0.133	0.148
π	Med			0.251	0.205	0.203			0.225	0.191	0.204
	Max			0.343	0.310	0.307			0.325	0.303	0.299

Table 16: The R^2 statistics of residuals and squared residuals regressed on their lagged values with the average log-likelihood of an observation for various daily models for the MCD stock.

		Variance Models					Overdispersion Models				
Statistic	Lag	I	II	III	IV	V	VI	VII	VIII	IX	X
AR R^2	1	0.123	0.047	0.102	0.095	0.050	0.120	0.005	0.003	0.094	0.004
	10	0.162	0.065	0.138	0.122	0.067	0.158	0.011	0.005	0.119	0.006
	100	0.166	0.067	0.141	0.124	0.069	0.162	0.013	0.008	0.121	0.008
ARCH R^2	1	0.098	0.007	0.086	0.014	0.008	0.097	0.001	0.002	0.014	0.000
	10	0.147	0.008	0.139	0.015	0.010	0.148	0.003	0.040	0.016	0.002
	100	0.190	0.020	0.181	0.021	0.018	0.192	0.007	0.073	0.022	0.006
Log-Likelihood		-1.941	-1.833	-1.861	-1.814	-1.788	-1.933	-1.806	-1.792	-1.812	-1.760

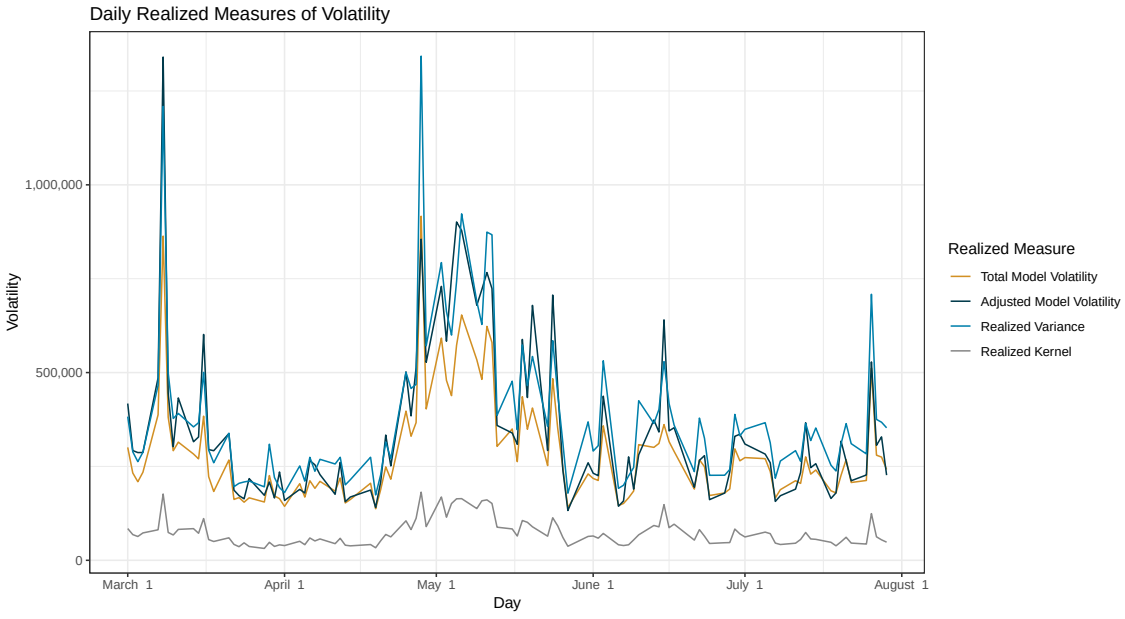

Figure 28: The daily values of various volatility realized measures for the MCD stock.

Zero-Inflated Autoregressive Conditional Duration Model for Discrete Trade Durations with Excessive Zeros

Francisco Blasques
VU University Amsterdam and Tinbergen Institute
De Boelelaan 1105, NL-1081HV Amsterdam, The Netherlands
f.blasques@vu.nl

Vladimír Holý
Prague University of Economics and Business
Winston Churchill Square 1938/4, 130 67 Prague 3, Czechia
vladimir.holy@vse.cz

Petra Tomanová
Prague University of Economics and Business
Winston Churchill Square 1938/4, 130 67 Prague 3, Czechia
petra.tomanova@vse.cz

Abstract: In finance, durations between successive transactions are usually modeled by the autoregressive conditional duration model based on a continuous distribution omitting zero values. Zero or close-to-zero durations can be caused by either split transactions or independent transactions. We propose a discrete model allowing for excessive zero values based on the zero-inflated negative binomial distribution with score dynamics. This model allows to distinguish between the processes generating split and standard transactions. We use the existing theory on score models to establish the invertibility of the score filter and verify that sufficient conditions hold for the consistency and asymptotic normality of the maximum likelihood of the model parameters. In an empirical study, we find that split transactions cause between 92 and 98 percent of zero and close-to-zero values. Furthermore, the loss of decimal places in the proposed approach is less severe than the incorrect treatment of zero values in continuous models.

Keywords: Financial High-Frequency Data, Autoregressive Conditional Duration Model, Zero-Inflated Negative Binomial Distribution, Generalized Autoregressive Score Model.

JEL Classification: C22, C41, C58.

1 Introduction

An important aspect of financial high-frequency data analysis is modeling of durations between various events. These include times of recording of transactions (trade durations), times when price changes by a given level (price durations), and times when volume reaches a given level (volume durations). Financial durations exhibit strong serial correlation, i.e. long durations are usually followed by long durations and short durations are followed by short durations. To capture this time dependence, Engle and Russell (1998) proposed the autoregressive conditional duration (ACD) model.

We focus on trade durations and one of their particular empirical characteristics – the frequent occurrence of zero durations, i.e. trades executed at the same time. Zero durations are typically assumed to be caused by so-called split transactions, i.e. large trades broken into two or more smaller trades (see e.g. Pacurar, 2008). Subsequently, observations with the same timestamp are merged and the resulting prices are calculated as the average of prices weighted by volume. From the perspective of time series of trade durations, zero values are simply discarded. There is an obvious issue with this approach – unrelated transactions that just occur at the same time but do not originate from the same source might be merged as well and their zero durations discarded. Nevertheless, this is the most common approach in the ACD literature dating back to Engle and Russell (1998). Dealing with zero values is even a necessity for ACD models based on distributions that do not contain zero

in their support. Alternatively, Bauwens (2006) suggested setting zero durations to a small given value instead of discarding them. This transformation allows to keep all observations in the dataset but is quite arbitrary and distorts the distribution of durations near zero. From an economic point of view, however, it makes sense to consider split transactions as one single trade (see e.g. Grammig and Wellner, 2002).

Datasets analyzed by Engle and Russell (1998) and others at the turn of the millennium had timestamps with precision to one second. Nowadays it is standard that transactions are recorded with precision to one millisecond, one microsecond, or even one nanosecond by some exchanges. This high detail causes an additional problem – split transactions do not have to occur at the exact same time. An anecdotal evidence is presented in Table 1. This has already been recognized e.g. by Grammig and Wellner (2002) who treated successive trades with either non-increasing or non-decreasing prices within one second as one large trade. Let us take a closer look at a recent dataset consisting of 6 stocks traded on the EURONEXT, NYSE, and NASDAQ exchanges with precision to one millisecond obtained from the Thomson Reuters database. The right plot of Figure 1 shows the density of the logarithm of durations estimated by the Parzen–Rosenblatt window method. Values equal to exactly zero are omitted from this figure. The density of log-durations is concentrated in two areas for each stock – a "hill" in the middle of the plot and a "wave" in the left part of the plot. The "wave" shape is caused by discreteness of the data and captures durations close to zero. The left-most spike corresponds to 0.001 seconds, the next to it to the right to 0.002, and so on. For better readability of these close-to-zero durations, the left plot of Figure 1 shows their occurrence in data. First of all, we can see that exactly zero durations make up between 43 and 67 percent of all durations for the individual stocks. Durations equal to 0.001 are also quite frequent and make up between 5 and 8 percent. Durations equal to 0.002 make up about 2 percent and 0.003 durations about 1 percent. Other descriptive statistics are reported in Table 3. The main message here is that Figure 1 suggests that durations are generated by two processes – one process generates dispersed values corresponding to unrelated transactions and the other process generates zero or close-to-zero values corresponding to split transactions.

The traditional approach which assumes that all split transactions have exactly zero duration and all zero durations correspond to a split transaction is therefore not very suitable. Firstly, as mentioned above, discarding all zero durations might also discard zero durations corresponding to unrelated transactions. Secondly, and more importantly, keeping all positive durations might also keep close-to-zero durations corresponding to split transactions. Discarding all zeros and no positive values then leads to distorted distribution caused by inaccurate representation of values near zero.

We propose to model durations by a mixture of two processes generating unrelated and split transactions respectively. We artificially reduce the precision of durations by rounding down the values to hundredths of a second, i.e. centiseconds, and operate within a discrete framework. With this reduced precision, we assume that all close-to-zero durations corresponding to split transactions fall into the new group of exactly zero durations, i.e. their original values are lower than 0.01 seconds. We then employ a zero-inflated distribution of Lambert (1992) for modeling of durations. This distribution assumes that one process generates integer values greater or equal to zero and another process generates only zero values. The probability of unrelated transactions with zero durations is then determined by the distribution of positive values while the probability of split transactions with zero durations is given by the inflation parameter of the zero-inflated distribution. We are therefore able to estimate the ratio between unrelated and split transactions. In the empirical study, we demonstrate that the loss of precision of durations is redeemed by the simplicity of our model and its ability to accommodate for both unrelated and split transactions.

Given the discussion above, we propose in this paper a new zero-inflated autoregressive conditional duration (ZIACD) model. We base our model on the negative binomial distribution to accommodate for overdispersion in durations (see Boswell and Patil, 1970; Cameron and Trivedi, 1986; Christou and Fokianos, 2014). The excessive zero durations caused by split transactions are captured by the zero-inflated modification of the negative binomial distribution (see Greene, 1994). We let the scale, dispersion, and inflation parameters of the distribution be time-varying and follow the dynamics of generalized autoregressive score (GAS) models, also known as dynamic conditional score models (see

Table 1: An excerpt from the limit order book of the MSFT stock on June 21, 2012.

Message Time	Order ID	Event	Direction	Size	Price
09:30:01.146	16333185	Submission	Buy	300	$30.99
⋮					
09:30:01.370	16333185	Execution	Buy	100	$30.99
09:30:01.377	16333185	Execution	Buy	200	$30.99
⋮					
09:30:03.550	16576783	Submission	Sell	3000	$30.99
⋮					
09:30:03.553	16576783	Execution	Sell	400	$30.99
09:30:03.555	16576783	Execution	Sell	400	$30.99
09:30:03.555	16576783	Execution	Sell	300	$30.99
09:30:03.627	16576783	Deletion	Sell	1900	$30.99

Creal *et al.*, 2013; Harvey, 2013). In the GAS framework, time-varying parameters are dependent on their lagged values and a scaled score of the conditional observation density.

In this paper, we establish the invertibility of the GAS filter for the ZIACD model and the consistency and asymptotic normality of the maximum likelihood estimator for the case of time-varying scale parameter and static dispersion and zero-inflation parameters. In an empirical study of the stock market, we demonstrate that the proposed ZIACD model for durations rounded to centiseconds is usable in practice and is superior to continuous models with the incorrect treatment of zero values.

The rest of the paper is structured as follows. In Section 3, we review the related literature on ACD and GAS models. In Section 3, we propose the ZIACD model based on the zero-inflated negative binomial distribution. In Section 4, we verify the asymptotic properties of the maximum likelihood estimator for the case of time-varying scale. In Section 5, we describe characteristics of financial durations data, fit the proposed ZIACD model within a discrete framework, and compare it to a continuous model. In Section 6, we discuss the use of the proposed ZIACD model for low-precision data and alternative mixture ACD models as topics for future research. We conclude the paper in Section 7.

2 Literature Review

In this section, we examine two fundamental cornerstones of our paper: the Autoregressive Conditional Duration (ACD) model and the Generalized Autoregressive Score (GAS) model. These established models serve as the foundation for our novel contribution, the zero-inflated autoregressive conditional duration (ZIACD) model.

2.1 Autoregressive Conditional Duration Models

Since the seminal paper of Engle and Russell (1998), many extensions of the original ACD model have been proposed in the literature. Bauwens and Giot (2000) introduced the *logarithmic ACD model* utilizing the logarithmic transformation and exogenous variables. Logarithmic model with a slightly different dynamic was considered by Lunde (1999). Other proposed models include the *fractionally integrated ACD model* of Jasiak (1998), *threshold ACD model* of Zhang *et al.* (2001), *Box-Cox ACD model* of Hautsch (2001, 2003), *asymmetric ACD model* of Bauwens and Giot (2003), *additive and multiplicative ACD model* of Hautsch (2012), and *directional ACD model* of Jeyasreedharan *et al.* (2014). Time-varying and non-stationary ACD models were studied by Bortoluzzo *et al.* (2010) and Mishra and Ramanathan (2017). Joint models for durations and prices were proposed by Engle (2000), Grammig and Wellner (2002), Russell and Engle (2005) and Herrera and Schipp (2013).

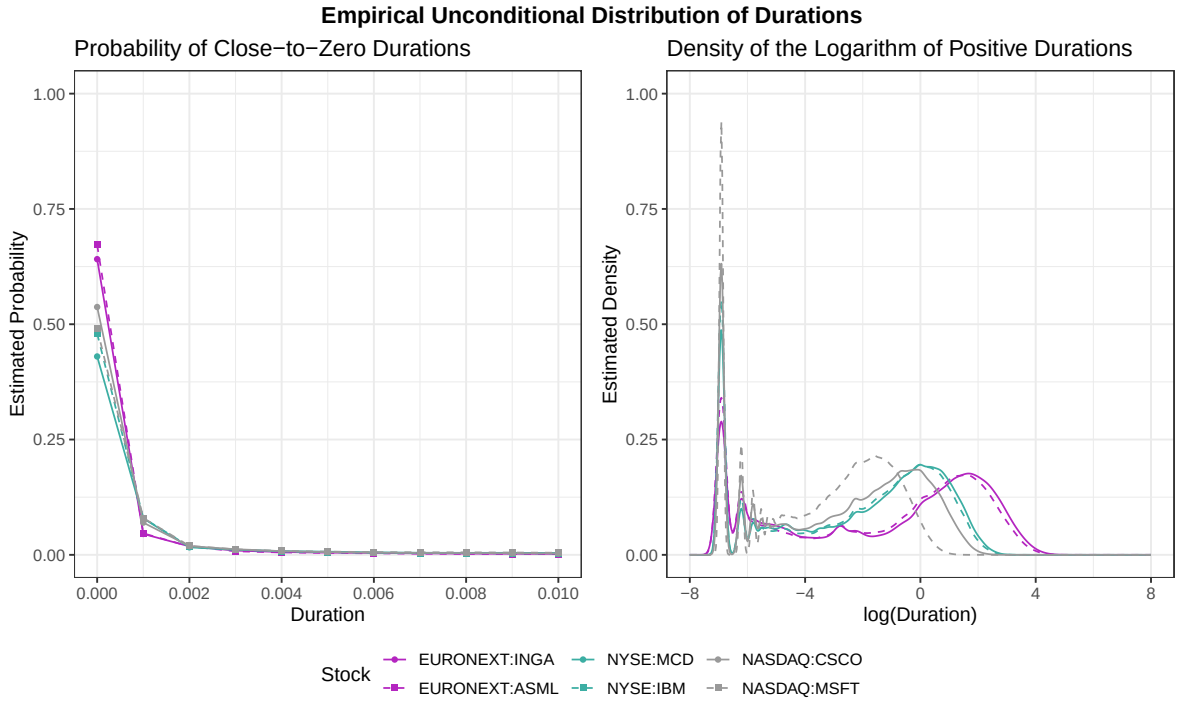

Figure 1: The probability of durations between 0 and 0.01 seconds (left plot) and the density function of logarithmic durations estimated using the Gaussian kernel (right plot) in June, 2021. Zero durations are excluded.

Ghysels *et al.* (2004) proposed the *stochastic volatility duration model*, which accounts for mean and variance dynamics in financial duration processes. Additionally, Bauwens and Veredas (2004) introduced the *stochastic conditional duration (SCD) model*, which was further extended by Feng (2004) and Xu *et al.* (2011). Feng (2004) proposed the SCD model with leverage effect and Xu *et al.* (2011) added an interaction element between the duration process and the latent autoregressive process. Hujer *et al.* (2005) proposed *Markov switching ACD model* that extends the traditional ACD model by introducing an unobservable stochastic process modeled by a Markov chain. Chen *et al.* (2013) proposed *Markov-switching multifractal duration model*, which allows for modeling long memory in the duration process. Fernandes and Grammig (2006) developed a family of ACD models that encompasses most common specifications, where the nesting relies on a Box-Cox transformation.

Numerous studies in the literature also explore the incorporation of information about zero durations. Zhang *et al.* (2001) included an indicator of multiple transactions as an explanatory variable in their ACD model. Veredas *et al.* (2002) noticed that many simultaneous transactions occur at round prices suggesting many traders post limit orders to be executed at round prices – this is an empirical phenomenon known as price clustering (see e.g. the literature review in Holý and Tomanová, 2022). More recently, Liu *et al.* (2018) examined the effect of zero durations on integrated volatility estimation.

The first ACD models analyzed by Engle and Russell (1998) utilize the exponential and Weibull distributions. However, since then, various continuous distributions have been employed in duration modeling; an overview can be found in Table 2. Additionally, several studies in the literature have proposed ACD models based on mixtures of distributions. De Luca and Zuccolotto (2003) and De Luca and Gallo (2004) suggested using a mixture of two exponential distributions to capture distinct behaviors of informed and uninformed traders. This work was further extended by De Luca and Gallo (2009), proposing the incorporation of the two exponential distributions with time-varying weights. On the other hand, to account for the unobserved market heterogeneity of traders, Gómez-Déniz and Pérez-Rodríguez (2016, 2017) proposed finite and infinite mixture of distributions based

Table 2: The use of continuous distributions in ACD models.

Article	Distribution	Parameters
Engle and Russell (1998)	Exponential	1
Engle and Russell (1998)	Weibull	2
Lunde (1999)	Generalized Gamma	3
Grammig and Maurer (2000)	Burr	3
Hautsch (2001)	Generalized F	4
Bhatti (2010)	Birnbaum–Saunders	2
Xu (2013)	Log-Normal	2
Leiva *et al.* (2014)	Power-Exponential B–S	3
Leiva *et al.* (2014)	Student's t B–S	3
Zheng *et al.* (2016)	Fréchet	2

on non-exponentials, specifically a mixture of an inverse Gaussian distribution. For a survey of duration analysis, see Pacurar (2008), Bauwens and Hautsch (2009), Hautsch (2012), and Saranjeet and Ramanathan (2018).

2.2 Generalized Autoregressive Score Models

Generalized autoregressive score (GAS) models (Creal *et al.*, 2013), also known as *dynamic conditional score* models (Harvey, 2013), capture dynamics of time-varying parameters by the autoregressive term and the scaled score of the conditional observation density (see Section 3.3 for further details). GAS models belong to the class of observation-driven models, as defined by Cox (1981), and thus have their advantages, e.g. observation-driven models can be estimated in a straightforward manner by the maximum likelihood method and their parameters are perfectly predictable given the past information. Moreover, Blasques *et al.* (2015) investigated information-theoretic optimality properties of the score function of the predictive likelihood and showed that only parameter updates based on the score will always reduce the local Kullback–Leibler divergence between the true conditional density and the model-implied conditional density. Koopman *et al.* (2016) find that observation-driven models based on the score perform comparably to parameter-driven models in terms of predictive accuracy.

The GAS specification includes many commonly used econometric models. For example, the GAS model with the normal distribution, the inverse of the Fisher information scaling and time-varying variance results in the GARCH model while the GAS model with the exponential distribution, the inverse of the Fisher information scaling and time-varying expected value results in the ACD model (Creal *et al.*, 2013). The GAS framework can be utilized for discrete models as well. Koopman *et al.* (2018) used discrete copulas based on the Skellam distribution for high-frequency stock price changes. Koopman and Lit (2019) used the bivariate Poisson distribution for a number of goals in football matches and the Skellam distribution for a score difference. Gorgi (2018) used the Poisson distribution as well as the negative binomial distribution for offensive conduct reports. Holý and Tomanová (2022) used a mixture of double Poisson distributions to model price clustering in high-frequency prices.

Andres and Harvey (2012) specified ACD-like models belonging to the GAS framework and applied them to intra-day stock market data, considering both range and duration. Tomanová and Holý (2021) utilized the GAS model based on the generalized gamma distribution in the spirit of ACD models, and demonstrated that this approach outperforms the traditional method that assumes times between arrivals follow the exponential distribution with a constant rate, making it a superior choice for modeling arrivals in queueing systems.

A comprehensive list of papers on GAS models can be found at `http://gasmodel.com`.

3 Zero-Inflated ACD Model

Let $T_0 \leq T_1 \leq \cdots \leq T_n$ be random variables denoting times of transactions. Trade durations are then defined as $X_i = T_i - T_{i-1}$ for $i = 1, \ldots, n$. As we operate in a discrete framework, we assume

$T_i \in \mathbb{N}_0$, $i = 0, \ldots, n$ and $X_i \in \mathbb{N}_0$, $i = 1, \ldots, n$.[1] We further assume trade durations X_i to follow some given discrete distribution with conditional probability mass function $P[X_i = x_i|\theta]$, where x_i are observations and $\theta = (\theta_1, \ldots, \theta_l)'$ are parameters. First, we consider trade durations to follow the negative binomial distribution. Next, we extend the negative binomial distribution to capture excessive zeros using the zero-inflated model. Finally, we let parameters be time-varying with the generalized autoregressive score dynamics.

3.1 Negative Binomial Distribution

Non-negative integer variables are commonly analyzed using count data models based on specific underlying distribution, most notably the Poisson distribution and the negative binomial distribution (see Cameron and Trivedi, 2013). A distinctive feature of the Poisson distribution is that its expected value is equal to its variance. This characteristic is too strict in many applications as count data often exhibit overdispersion, a higher variance than the expected value. A generalization of the Poisson distribution overcoming this limitation is the negative binomial distribution with one parameter determining its expected value and another parameter determining its excess dispersion.

The *negative binomial (NB) distribution* can be derived in many ways (see Boswell and Patil, 1970). We use the NB2 parameterization of Cameron and Trivedi (1986) derived from the Poisson-gamma mixture distribution. It is the most common parametrization used in the negative binomial regression according to Cameron and Trivedi (2013). The probability mass function with scale parameter $\mu > 0$ and dispersion parameter $\alpha \geq 0$ is

$$\mathrm{P}[X_i = x_i|\mu, \alpha] = \frac{\Gamma(x_i + \alpha^{-1})}{\Gamma(x_i + 1)\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu}\right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1} + \mu}\right)^{x_i} \quad \text{for } x_i = 0, 1, 2, \ldots. \tag{1}$$

The expected value and variance is

$$\begin{aligned} \mathrm{E}[X_i] &= \mu, \\ \mathrm{var}[X_i] &= \mu(1 + \alpha\mu). \end{aligned} \tag{2}$$

Special cases of the negative binomial distribution include the Poisson distribution for $\alpha = 0$ and the geometric distribution for $\alpha = 1$.

3.2 Zero-Inflation

The zero-inflated distribution is an extension of a discrete distribution allowing the probability of zero values to be higher than the probability given by the original distribution. In the zero-inflated distribution, values are generated by two components – one component generates only zero values while the other component generates integer values (including zero values) according to the original distribution. Lambert (1992) proposed the zero-inflated Poisson model and Greene (1994) used zero-inflated model for the negative binomial distribution.

The *zero-inflated negative binomial distribution* is a discrete distribution with three parameters: scale parameter $\mu > 0$, dispersion parameter $\alpha \geq 0$ and probability of excessive zero values $\pi \in [0, 1)$. The variable X_i follows the zero-inflated negative binomial distribution if

$$\begin{aligned} X_i &\sim 0 && \text{with probability } \pi, \\ X_i &\sim \mathrm{NB}(\mu, \alpha) && \text{with probability } 1 - \pi. \end{aligned} \tag{3}$$

The first process generates only zeros and corresponds to split transactions, while the second process generates values from the negative binomial distribution and corresponds to regular transactions. The

[1] Note that this assumption is not restrictive since durations are naturally discrete and non-negative. Thus when expressed in the units corresponding to precision of the timestamps (e.g. seconds, milliseconds, ...), the durations are natural numbers (with zero).

probability mass function is

$$\begin{aligned}
\mathrm{P}[X_i = 0|\mu,\alpha,\pi] &= \pi + (1-\pi)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}, \\
\mathrm{P}[X_i = x_i|\mu,\alpha,\pi] &= (1-\pi)\frac{\Gamma(x_i+\alpha^{-1})}{\Gamma(x_i+1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{x_i} \quad \text{for } x_i = 1,2,\ldots.
\end{aligned} \tag{4}$$

The expected value and variance is

$$\begin{aligned}
\mathrm{E}[X_i] &= \mu(1-\pi), \\
\mathrm{var}[X_i] &= \mu(1-\pi)(1+\pi\mu+\alpha\mu).
\end{aligned} \tag{5}$$

The score vector is given by

$$\nabla(x_i,\mu,\alpha,\pi) = \begin{pmatrix}
(\pi-1)(\alpha\mu+1)^{-1}\Big(1+\pi(\alpha\mu+1)^{\alpha^{-1}}-\pi\Big)^{-1} \\
\alpha^{-2}\Big(\ln(\alpha\mu+1)-\alpha\mu(\alpha\mu+1)^{-1}\Big)\Big(1-\pi(\pi-1)^{-1}(\alpha\mu+1)^{\alpha^{-1}}\Big)^{-1} \\
\Big((\alpha\mu+1)^{\alpha^{-1}}-1\Big)\Big(1+\pi(\alpha\mu+1)^{\alpha^{-1}}-\pi\Big)^{-1}
\end{pmatrix} \tag{6}$$

for $x_i = 0$ and

$$\nabla(x_i,\mu,\alpha,\pi) = \begin{pmatrix}
\mu^{-1}(x_i-\mu)(\alpha\mu+1)^{-1} \\
\alpha^{-2}\Big(\ln(\alpha\mu+1)+\alpha(x_i-\mu)(\alpha\mu+1)^{-1}+\psi_0(\alpha^{-1})-\psi_0(x_i+\alpha^{-1})\Big) \\
(\pi-1)^{-1}
\end{pmatrix} \tag{7}$$

for $x_i = 1,2,\ldots$.

3.3 Score-Driven Dynamics

Generalized autoregressive score (GAS) models (Creal *et al.*, 2013) capture dynamics of time-varying parameters $\tilde{f}_i = (\tilde{f}_{i,1},\ldots,\tilde{f}_{i,k})'$ by the autoregressive term and the scaled score of the conditional observation density (or the conditional observation probability mass function in the case of discrete distribution). The time-varying parameters $\tilde{f}_i$ follow the recursion

$$\tilde{f}_{i+1} = C + B\tilde{f}_i + AS(\tilde{f}_i)\nabla(x_i,\tilde{f}_i), \tag{8}$$

where $C = (c_1,\ldots,c_k)'$ are the constant parameters, $B = \mathrm{diag}(b_1,\ldots,b_k)$ are the autoregressive parameters, $A = \mathrm{diag}(a_1,\ldots,a_k)$ are the score parameters, $S(\tilde{f}_i)$ is the scaling function for the score and $\nabla(x_i,\tilde{f}_i)$ is the score. In the original paper of Creal *et al.* (2013), authors noted that via the choice of the scaling function $S(\tilde{f}_i)$, the GAS model allows for additional flexibility in how the score is used for updating $\tilde{f}_i$. The commonly used scaling functions in the GAS literature are based on the Fisher information matrix. We explored this option, however, we have not found it very suitable for the GAS model with the negative binomial distribution since the Fisher information for the parameter α does not have a closed-form. Consequently, the approximation of the Fisher information brings undue computational complexity resulting in an overly time-consuming optimization procedure. In order to keep our model simple, from now on we avoid the scaling, which is also a widely used option in the GAS literature. Moreover, Holý (2020) showed that the differences of models performance based on different scaling functions are negligible in the case of the negative binomial distribution.

The long-term mean and unconditional value of the time-varying parameters is $\tilde{f} = (I-B)^{-1}C$. The parameters $\tilde{f}_i$ in (8) are assumed to be unbounded. However, some distributions require bounded parameters (e.g. variance greater than zero). The standard solution in the GAS framework is to use an unbounded parametrization $f_i = H(\tilde{f}_i)$, which follows the GAS recursion instead of the original parametrization $\tilde{f}_i$, i.e.

$$f_{i+1} = c + bf_i + as(x_i,f_i), \tag{9}$$

where c are the constant parameters, b are the autoregressive parameters, a are the score parameters, and $s(x_i, f_i)$ is the reparametrized score. The reparametrized score equals to

$$s(x_i, f_i) = \dot{H}^{-1}(\tilde{f}_i)\nabla(x_i, \tilde{f}_i), \tag{10}$$

where $\dot{H}(\tilde{f}_i) = \partial H(\tilde{f}_i)/\partial \tilde{f}_i'$ is the derivation of $H(\tilde{f}_i)$.

3.4 Zero-Inflated Autoregressive Conditional Duration Model

We consider a model where observations follow the zero-inflated negative binomial distribution with the time-varying scale parameter μ_i, time-varying dispersion parameter α_i and time-varying inflation parameter π_i specified in (4). We use an unbounded parametrization with the exponential link for the scale and dispersion parameters and logistic transformation for the inflation parameter, i.e. $f_i = (\ln(\mu_i), \ln(\alpha_i), \ln(\pi_i/(1-\pi_i)))'$. Parameters f_i are assumed to follow the recursion in (9), where the score for the zero-inflated negative binomial distribution is given by

$$s(x_i, f_i) = \begin{pmatrix} \mu_i(\pi_i - 1)(\alpha_i\mu_i + 1)^{-1}\Big(1 + \pi_i(\alpha_i\mu_i + 1)^{\alpha_i^{-1}} - \pi_i\Big)^{-1} \\ \alpha_i^{-1}\Big(\ln(\alpha_i\mu_i + 1) - \alpha_i\mu_i(\alpha_i\mu_i + 1)^{-1}\Big)\Big(1 - \pi_i(\pi_i - 1)^{-1}(\alpha_i\mu_i + 1)^{\alpha_i^{-1}}\Big)^{-1} \\ \pi_i(1 - \pi_i)\Big((\alpha_i\mu_i + 1)^{\alpha_i^{-1}} - 1\Big)\Big(1 + \pi_i(\alpha_i\mu_i + 1)^{\alpha_i^{-1}} - \pi_i\Big)^{-1} \end{pmatrix} \tag{11}$$

for $x_i = 0$ and

$$s(x_i, f_i) = \begin{pmatrix} (x_i - \mu_i)(\alpha_i\mu_i + 1)^{-1} \\ \alpha_i^{-1}\Big(\ln(\alpha_i\mu_i + 1) + \alpha_i(x_i - \mu_i)(\alpha_i\mu_i + 1)^{-1} + \psi_0(\alpha_i^{-1}) - \psi_0(x_i + \alpha_i^{-1})\Big) \\ -\pi_i \end{pmatrix} \tag{12}$$

for $x_i = 1, 2, \ldots$.

4 Estimation and Asymptotic Properties

In this section, we focus on the model with the time-varying scale parameter μ_i and static dispersion α and inflation π parameters. As such we set $f_i = \ln(\mu_i)$ and $\theta = (\alpha, \pi, c, b, a)'$. The score in (11) and (12) simplifies to

$$\begin{aligned} s(0, f_i) &= \frac{(\pi - 1)\exp(f_i)}{(\alpha\exp(f_i) + 1)\left(1 + \pi(\alpha\exp(f_i) + 1)^{\alpha^{-1}} - \pi\right)}, \\ s(x_i, f_i) &= \frac{x_i - \exp(f_i)}{\alpha\exp(f_i) + 1} \quad \text{for } x_i = 1, 2, \ldots. \end{aligned} \tag{13}$$

For this GAS model with dynamics defined in (9) and (13), we establish the invertibility of the score filter and verify that sufficient conditions hold for the consistency and asymptotic normality of the maximum likelihood of the model parameters.

The static parameter vector θ is estimated by the method of maximum likelihood

$$\hat{\theta}_n \in \arg\max_{\theta\in\Theta} \hat{L}_n(\theta), \tag{14}$$

where $\hat{L}_n(\theta)$ denotes the log likelihood function obtained from a sequence of n observations $x_1, \ldots, x_n$, which depends on the filtered time-varying parameter $\hat{f}_1(\theta), ..., \hat{f}_n(\theta)$. Since we are dealing with observation-driven filters which require an initialization value $\hat{f}_1$, we make an important distinction here between $\hat{L}_n(\theta)$ and $L_n(\theta)$. The first log likelihood is a function of the filtered parameter $\hat{f}_1(\theta), ..., \hat{f}_n(\theta)$ initialized at a given value $\hat{f}_1$. The second likelihood is a function of the filtered parameter $f_1(\theta), ..., f_n(\theta)$ initialized at the true unobserved value f_1. Of course, since f_1 is unobserved, we typically have that $\hat{f}_1 \neq f_1$. In practice, the sample log likelihood is thus given by

$$\hat{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \hat{\ell}_i(x_i, \theta) = \frac{1}{n}\sum_{i=1}^{n} \ln \mathrm{P}[X_i = x_i | \hat{f}_i(\theta), \theta]. \tag{15}$$

In our case, the log likelihood is based on the zero-inflated negative binomial distribution

$$\ln \mathrm{P}[X_i = 0|\hat{f}_i(\theta), \theta] = \ln\left(\pi + (1-\pi)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\right),$$

$$\ln \mathrm{P}[X_i = x_i|\hat{f}_i(\theta), \theta] = \ln(1-\pi) + \ln\frac{\Gamma(x_i+\alpha^{-1})}{\Gamma(x_i+1)\Gamma(\alpha^{-1})} + \frac{1}{\alpha}\ln\left(\frac{\alpha^{-1}}{\alpha^{-1}+\exp(\hat{f}_i)}\right) \tag{16}$$
$$+ x_i \ln\left(\frac{\exp(\hat{f}_i)}{\alpha^{-1}+\exp(\hat{f}_i)}\right) \quad \text{for } x_i = 1, 2, \ldots.$$

Below, we show that the maximum likelihood estimator of the ZIACD model is consistent and asymptotically normal. The proof follows the structure laid down in Blasques *et al.* (2022), but we focus on the particular case of discrete data $\{x_i\}_{i\in\mathbb{N}}$ with a probability mass function $\mathrm{P}[X_i = x_i|f_i(\theta), \theta]$. In contrast, Blasques *et al.* (2022) treat a general case for continuous data with a smooth probability density function.

4.1 Filter Invertibility

Filter invertibility is crucial for statistical inference in the context of observation-driven time-varying parameter models; see e.g. Straumann and Mikosch (2006), Wintenberger (2013), and Blasques *et al.* (2022). The filter $\{\hat{f}_i(\theta)\}_{i\in\mathbb{N}}$ initialized at some point $\hat{f}_1 \in \mathbb{R}$ is said to be invertible if $\hat{f}_i(\theta)$ converges almost surely exponentially fast to a unique limit strictly stationary and ergodic sequence $\{f_i(\theta)\}_{i\in\mathbb{Z}}$,

$$|\hat{f}_i(\theta) - f_i(\theta)| \overset{eas}{\to} 0 \quad \text{as} \quad i \to \infty.$$

Let $L_n(\theta)$ denote the log likelihood which depends on the limit time-varying parameter $f_1(\theta), ..., f_n(\theta)$

$$L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(x_i, \theta) = \frac{1}{n}\sum_{i=1}^{n} \ln \mathrm{P}[X_i = x_i|f_i(\theta), \theta],$$

and let L_∞ denote the limit log likelihood function

$$L_\infty(\theta) = \mathrm{E}[\ell_i(\theta)] = \mathrm{E}\left[\ln \mathrm{P}[X_i = x_i|f_i(\theta), \theta]\right].$$

Proposition 1 appeals to the results in Blasques *et al.* (2022) to establish the invertibility of the score filter with zero-inflated negative binomial distribution as stated in (9) and (13). The proof presented in Technical Appendix A is an application of the results in Blasques *et al.* (2022) to our current model.

Proposition 1 (Filter invertibility)**.** *Consider the score-driven model with zero-inflated negative binomial distribution in* (9) *and* (13)*. Let the observed data* $\{x_i\}_{i\in\mathbb{N}}$ *be strictly stationary and ergodic, with a logarithmic moment* $\mathrm{E}[\ln^+|x_i|] < \infty$*, and let* Θ *be a compact parameter space defined as*

$$\Theta = [\alpha^-, \alpha^+] \cdot [\pi^-, \pi^+] \cdot [c^-, c^+] \cdot [b^-, b^+] \cdot [a^-, a^+]$$

and satisfying the following restrictions

$$\frac{a^+(\pi^- - 1)^2}{2\alpha^-} + \frac{a^+|\pi^- - 1|}{(\alpha^-)^2} + b^+ < 1,$$
$$\mathrm{E}_{x_i>0}\left[\ln\left(\frac{a^+(\alpha^+ x_i + 1)}{4\alpha^-} + b^+\right)\right] < 0.$$

Then the filter $\{\hat{f}_i(\theta)\}_{i\in\mathbb{N}}$ *defined as* $\hat{f}_{i+1} = c + b\hat{f}_i + as(x_i, \hat{f}_i)$ *is invertible, uniformly in* $\theta \in \Theta$*.*

4.2 Consistency

Proposition 1 gives us sufficient elements to characterize the asymptotic behavior of the ML estimator. This section uses existing theory on score models in Blasques *et al.* (2022) to verify the strong consistency of the ML estimator $\hat{\theta}_n$ as the sample size n diverges to infinity.

For completeness, Lemma 1 states conditions for the consistency of the ML estimator. A sketch of the proof is offered in Technical Appendix A, and appropriate references are offered. This theorem naturally uses the invertibility properties established in Proposition 1 for our zero-inflated negative binomial score model. Following Blasques *et al.* (2022), this theorem allows for potential model mispecification.

Lemma 1 (Consistency of the ML estimator)**.** *Let the conditions of Proposition 1 hold. Suppose further that the observed data has one bounded moment* $\mathrm{E}[x_i] < \infty$*, and let* θ_0 *be the unique maximizer of the limit log likelihood function* $\mathrm{E}[\ell_i(x_i, \cdot)] : \Theta \to \mathbb{R}$ *over the parameter space* Θ*. Then* $\hat{\theta}_n \xrightarrow{as} \theta_0 \in \Theta$ *as* $n \to \infty$*.*

4.3 Asymptotic Normality

Finally, we shed some light on the $\sqrt{n}$-consistency rate of $\hat{\theta}_n$ and the asymptotic normality of the standardized estimator $\sqrt{n}(\hat{\theta}_n - \theta_0)$ as $n \to \infty$, when the model is well specified. For completeness, Lemma 2 summarizes standard conditions for asymptotic normality. A sketch of the proof is presented in Technical Appendix A, and we refer to Blasques *et al.* (2022) for additional details.

Lemma 2 (Asymptotic normality of the ML estimator)**.** *Let the conditions of Lemma 1 hold. Suppose that the observed data has four bounded moments* $\mathrm{E}|x_i|^4 < \infty$*, and let the true parameter lie in the interior of the parameter space, i.e.* $\theta_0 \in \mathrm{int}(\Theta)$*. Finally, let the further regularity conditions stated in Theorem 4.16 of Blasques* et al. *(2022) hold. Then the ML estimator is asymptotically Gaussian*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \mathcal{I}(\theta_0)^{-1}) \quad as \quad n \to \infty,$$

where $\mathcal{I}(\theta_0)^{-1}$ *denotes the inverse Fisher information matrix.*

5 Empirical Study

5.1 Data Overview

In our empirical study, we analyze transaction data extracted from the Thomson Reuters Eikon. Eikon provides access to real-time market data and also contains historical intraday transactions. The data are taken from June to July of 2021. We analyze 6 stocks: ING Groep (INGA) and ASML Holding (ASML) which are listed on EURONEXT; McDonald's Corporation (MCD) and International Business Machines Corporation (IBM) which are listed on NYSE; Cisco Systems, Inc. (CSCO) and Microsoft Corporation (MSFT) which are listed on NASDAQ.

We clean data using the following procedure. First, we delete observations with the timestamp outside the main trading hours and trading days. Second, for EURONEXT stocks, we delete all observations with the timestamps equal to the first timestamp of the day that occurs between 09:00:00 and 09:00:30. The reason is that the opening uncrossing (resulting from the morning auction) randomly occurs between 09:00:00 and 09:00:30. Third, we round the timestamp to the right precision (i.e. milliseconds) to fix the incorrect representation of the float.[2]

The statistical characteristics for cleaned data are presented in Table 3. The two analyzed stocks listed on the NASDAQ belong to the most liquid stocks, while the stocks listed on the EURONEXT represent the least liquid stocks in our dataset. In June 2021, exact zero durations range from 43.01 percent (MCD) to 67.19 percent (ASML) and durations lower than 1 second form up to 98.57 percent (MSFT) of the dataset. For further descriptive statistics, see Table 3.

[2]For all analyzed stocks we observed that the sorted unique duration values are: 0, 0.00099927520751953,

Table 3: Descriptive statistics of trade durations in June and July, 2021.

		EURONEXT		NYSE		NASDAQ	
Statistic	Sample	INGA	ASML	MCD	IBM	CSCO	MSFT
% = 0	June	64.11	67.19	43.01	47.97	53.73	49.05
	July	57.78	65.66	46.01	48.75	54.14	48.93
% < 0.01	June	73.70	76.30	56.98	61.63	66.86	63.86
	July	67.52	74.53	59.78	63.01	67.11	63.93
% < 0.1	June	77.53	79.77	65.02	68.81	74.82	77.81
	July	71.86	78.50	67.18	71.13	74.48	79.20
% < 1	June	82.31	84.73	82.91	85.72	91.37	98.57
	July	78.37	85.11	84.59	88.75	90.72	99.05
Mean	June	1.56	1.19	0.58	0.47	0.26	0.10
	July	1.72	0.91	0.52	0.37	0.29	0.08
Variance	June	27.85	18.69	1.90	1.43	0.54	0.05
	July	26.01	10.31	1.72	1.02	0.63	0.04
Std. Dev.	June	5.28	4.32	1.38	1.19	0.73	0.23
	July	5.10	3.21	1.31	1.01	0.79	0.20
95%-Quantile	June	9.94	7.50	3.25	2.70	1.60	0.54
	July	10.48	5.66	2.96	2.14	1.73	0.46
Obs. per Min.	June	38.48	50.50	103.55	128.53	227.47	622.69
	July	34.88	66.00	115.11	163.22	210.39	723.17
Total Obs.	June	431 441	566 303	888 400	1 102 742	1 951 673	5 342 645
	July	391 156	740 150	942 707	1 336 712	1 641 075	5 922 788

5.2 In-Sample Performance

We use the proposed ZIACD model based on the zero-inflated negative binomial distribution with the time-varying scale, dispersion, and zero inflation parameters to fit observed durations rounded down to hundredths of a second using data from June 2021. The estimated coefficients are reported in Table 4. All coefficients are significant at any reasonable level and their standard deviations are virtually zero due to huge sample sizes ranging from 431463 (INGA) to 5342667 (MSFT). We, therefore, report only the estimated values. The numbers of observations per minute are also reported in Table 3. As expected, the coefficient controlling the impact of the score a is positive for all three parameters and all six stocks. This means that the score serves as a correction term that adjusts the time-varying parameters for the observed values. The autoregressive coefficient b is also positive and quite high for all three parameters and all six stocks. In the case of the scale parameter, it is very close to one signaling high persistence of the time series.

Table 5 reports the average values of the scale, dispersion, and zero inflation parameters over time. Note that the average scale parameter (adjusted to seconds) is much higher than the sample mean reported in Table 3 as our model is able to separate zeros attributed to split transactions which subsequently do not affect the scale parameter. On average, between 53.27 percent (MCD) and 74.88 percent (ASML) of all durations are excessive zeros generated by split transactions depending on the stock. This corresponds to the ratio of excessive zeros to all zeros ranging between 91.81 percent (MSFT) and 98.13 percent (ASML). In other words, between 1.87 percent (ASML) and 8.19 percent (MSFT) of zero durations are generated by unrelated transactions which should not be discarded from the data.

Table 5 also evaluates the fit of the ZIACD model. The mean absolute error is between 0.11 seconds (MSFT) and 2.50 seconds (INGA) while the root mean square error is between 0.21 (MSFT) and 5.22 (INGA). These values are quite high when compared to the predicted value $\mu_i(1-\pi_i)$, on which both errors are based, with its mean ranging from 0.09 seconds (MSFT) to 1.58 seconds (INGA). This is caused by the fact that the predicted value is not very representative of the whole distribution as, on average, between 53.27 percent (MCD) and 74.88 percent (ASML) of all values are exactly zero while the rest have expected value between 0.22 seconds (MSFT) and 5.68 seconds (INGA). It is therefore more suitable to assess the fit of the model based on the whole distribution.

We focus on the probability of zeros given by the model. Table 5 reports the mean probabilities of zero value given by the model when the observed value is indeed zero and when the observed value is positive. For the INGA and ASML stocks, the difference between these two probabilities is lower than one percent suggesting a limited benefit of the dynamics in the zero-inflation parameter. For the more traded stocks, the difference is between 5.78 percent (MCD) and 9.58 percent (CSCO) suggesting a certain degree of predictive ability of the zero-inflation dynamics.

The left plot of Figure 2 studies the fit of the model in more detail by comparing the average conditional probabilities given by the ZIACD model with the unconditional empirical distribution. The largest deviation is -0.68 percent at 0.01 seconds for the MCD stock. This deviation is rather small but uncovers a systematic error as the probability of 0.01 durations is underestimated for all stocks. Similar underestimation is also present at 0.06 seconds for the ASML and INGA stocks traded on the EURONEXT exchange and at 0.10 seconds for all stocks. The latter two anomalies are also visible in the right plot of Figure 1 at -2.81 and -2.30 log-durations. The proposed model is therefore incorrectly specified and the true distribution of durations is much more complex. Nevertheless, the deviations of the conditional ZIACD probabilities are quite small and the model is usable in practice.

5.3 Out-of-Sample Performance

In this section, we use the models estimated using durations from June 2021 and perform one-step-ahead forecasts during July 2021 to assess their long-term behavior. The right plot of Figure 2 shows deviations of the average out-of-sample conditional probabilities given by the ZIACD model from the

0.00100016593933105, 0.00199985504150391, 0.00200009346008301, ... The Thomson Reuters data are stamped with precision to one millisecond and this strange behavior is caused by an issue related to the representation of the float, which can be easily fixed by rounding.

Table 4: The estimated coefficients of the zero-inflated negative binomial model.

		EURONEXT		NYSE		NASDAQ	
Parameter	Coef.	INGA	ASML	MCD	IBM	CSCO	MSFT
Scale	c	0.006068	0.002591	0.000011	0.000151	0.000180	0.000064
	a	0.109420	0.089377	0.032544	0.032434	0.051552	0.032155
	b	0.998958	0.999509	0.999996	0.999954	0.999913	0.999938
Dispersion	c	0.006364	0.061190	0.148700	0.129161	0.042488	0.000869
	a	0.057713	0.216293	0.289589	0.243921	0.136988	0.021367
	b	0.992826	0.927438	0.806245	0.815294	0.948153	0.998387
Zero Inflation	c	0.030722	0.017910	0.048158	0.138758	0.116703	0.119207
	a	0.164058	0.100389	2.129143	2.177550	2.672883	2.542853
	b	0.968110	0.983785	0.680476	0.668047	0.856934	0.743213

Table 5: The mean scale parameter (in seconds), the mean dispersion parameter, the mean inflation parameter (in percent), the mean ratio of zeros caused by split transactions (in percent), the mean predicted value (in seconds), the mean absolute error (in seconds), the root mean square absolute error (in seconds), the mean probabilities of zero value given by the zero-inflated negative binomial model when the observation is either zero or positive (in percent), and the mean log-likelihood.

	EURONEXT		NYSE		NASDAQ	
Variable	INGA	ASML	MCD	IBM	CSCO	MSFT
Mean Scale	5.68	4.89	1.28	1.14	0.70	0.22
Mean Dispersion	2.45	2.35	2.17	2.02	2.26	1.70
Mean Zero Inflation	72.06	74.88	53.27	58.60	63.01	58.63
Mean Split Ratio	97.78	98.13	93.49	95.09	94.23	91.81
Mean Predicted Value	1.58	1.22	0.60	0.48	0.27	0.09
Mean Absolute Error	2.50	1.96	0.72	0.59	0.31	0.11
Root Mean Square Error	5.22	4.28	1.29	1.11	0.66	0.21
P$[X_i = 0]$ When $x_i = 0$	67.48	67.97	65.60	66.47	69.15	69.66
P$[X_i = 0]$ When $x_i > 0$	67.09	67.74	59.81	60.62	59.57	61.52
Mean Log-Likelihood	-2.42	-2.17	-3.01	-2.69	-2.16	-2.00

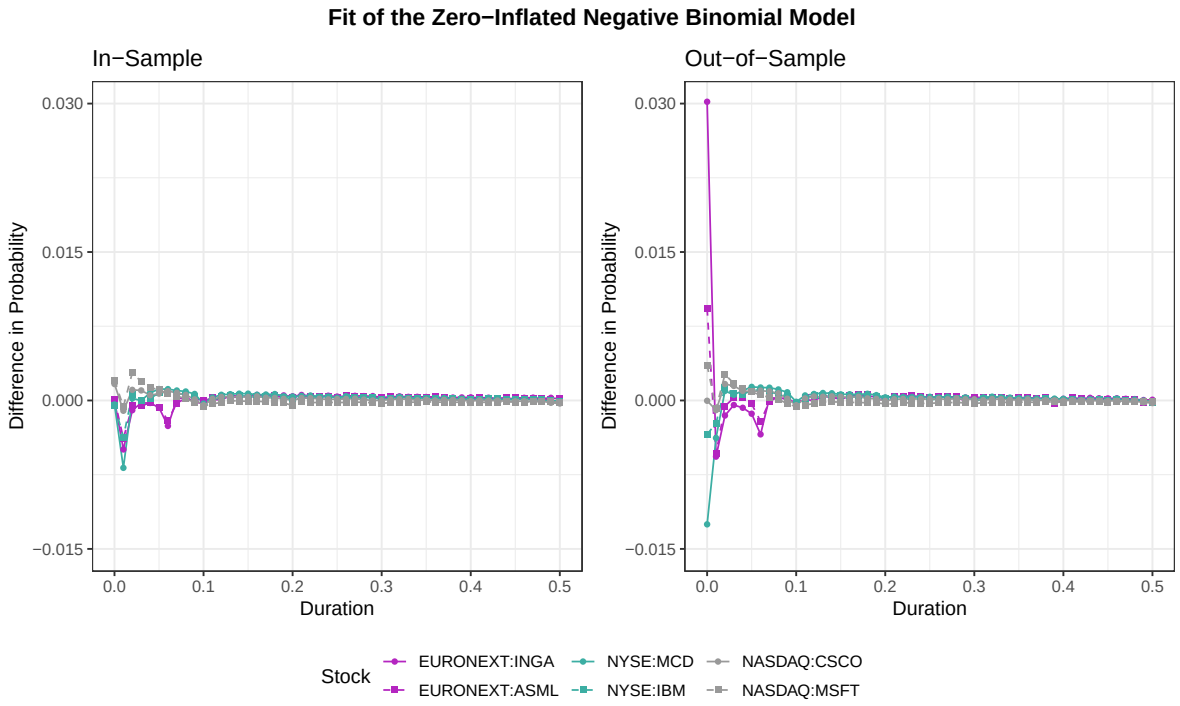

Figure 2: The in-sample and out-of-sample difference between the conditional probabilities given by the zero-inflated negative binomial model and the unconditional distribution of observations.

unconditional empirical distribution. Similarly to the left plot of Figure 2, the probabilities at 0.01, 0.06, and 0.10 seconds are systematically underestimated. However, the highest deviations are in the case of the probabilities of zero durations. The difference in probability reaches 3.02 percent (INGA) and drops down to -1.25 percent (MCD). This is related to a change in the occurrence of zero values in July. According to Table 3, the unconditional probability of zero values decreases from 64.11 to 57.78 percent for the INGA stock while it increases from 43.01 to 46.01 percent for the MCD stock. Note that the other descriptive statistics in Table 3 also change considerably.

However, this does not translate to a significant decrease in the log-likelihood. Figure 3 shows no apparent trend in the out-of-sample average daily log-likelihood, which is further supported by a simple linear regression analysis. This indicates that while the model may not be capable of accurately predicting long-term changes in the process, its forecasting performance does not significantly deteriorate over the long run. Furthermore, it should be noted that despite the overall stable performance, there is a noticeable volatility in day-to-day changes in the log-likelihood. This suggests that the accuracy of forecasts can vary significantly from one day to another.

To summarize, the proposed model is best suited for short-term predictions. For capturing changing characteristics of durations, it would be more appropriate to use a non-stationary model. As for the long-term dynamics of excessive zero probability, we leave this analysis as a topic for future research.

5.4 Model Specification

We compare the proposed ZIACD model, which is based on the zero-inflated negative binomial distribution and has all three parameters time-varying, with models imposing some restrictions. Specifically, Table 7 compares models based on the Poisson, geometric, and negative binomial distributions together with their zero-inflated versions. All parameters in these models are time-varying. On the other hand, Table 6 compares models based on the zero-inflated negative binomial distribution with some parameters static and some time-varying. We use two criteria to compare the models – the

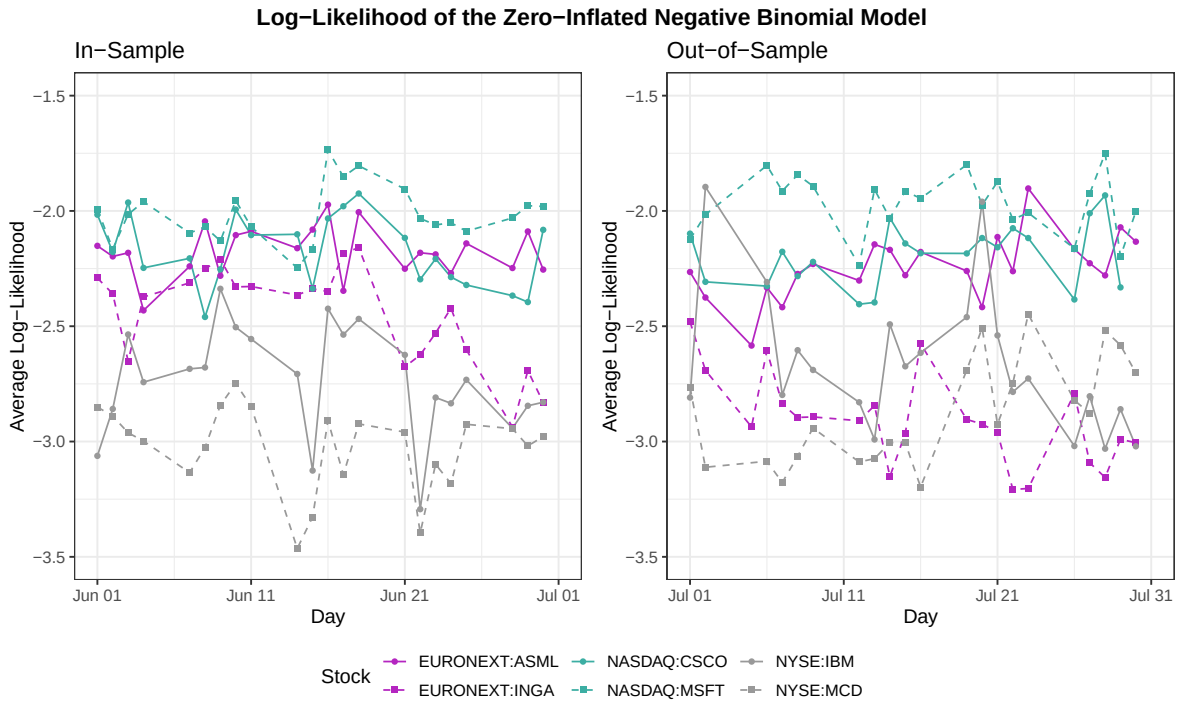

Figure 3: The in-sample and out-of-sample average daily log-likelihood of the zero-inflated negative binomial model.

difference in the Akaike information criterion (AIC) for the in-sample fit and the Diebold-Mariano (DM) statistic for the out-of-sample fit. When comparing two models, a positive difference in the AIC favors the second model over the first model while a positive value of the DM statistic favors the first model over the second model. The DM statistic has asymptotically the standard normal distribution under the null hypothesis of equivalent out-of-sample log-likelihoods. More details on these criteria are given in Technical Appendix B. Not surprisingly in such large datasets, the most general specification of the model has the best fit. We do not report p-values for the DM statistic as it is significant at any reasonable level in all cases due to huge sample sizes.

There is clear evidence of overdispersion, i.e. the variance higher than the expected value. According to Table 5, the average value of the dispersion parameter α in the zero-inflated negative binomial model ranges between 1.70 (MSFT) and 2.45 (INGA). This favors the negative binomial distribution over the Poisson distribution with fixed $\alpha = 0$ and the geometric distribution with fixed $\alpha = 1$. Overdispersion is also supported by the difference in the AIC and the DM statistic reported in Table 6. The Poisson distribution has the highest AIC for all stocks while the geometric distribution has the worst DM statistic compared to the zero-inflated negative binomial distribution. One possible reason for overdispersion could just be the presence of excessive zeros. Indeed, the zero-inflated Poisson and geometric distributions perform better than their original versions. However, they are still inferior to the zero-inflated negative binomial distribution suggesting that there is overdispersion present in non-zero values as well. Table 7 further shows that the specification with the time-varying dispersion parameter performs significantly better than the static one. This improvement of the in-sample and out-of-sample fit is, however, the smallest among all specifications in Tables 6 and 7. For some smaller data samples of less traded assets or with shorter periods of time (such as a day), the model with static dispersion parameter might be more suitable due to possible overfitting.

Our analysis also reveals the presence of excessive zeros suggesting the existence of the process generating only zero values (i.e. split transactions) alongside the process generating regular durations. According to Table 5, the average probability of excessive zeros π in the zero-inflated negative binomial model ranges between 53.27 percent (MCD) and 74.88 percent (ASML). The presence of

Table 6: The difference in the Akaike information criterion (AIC) and the Diebold–Mariano (DM) statistic for the models based on the Poisson distribution (P), the geometric distribution (G), the negative binomial distribution (NB), the zero-inflated Poisson distribution (ZIP), the zero-inflated geometric distribution (ZIG), and the zero-inflated negative binomial distribution (ZINB).

		EURONEXT		NYSE		NASDAQ	
Distribution	Crit.	INGA	ASML	MCD	IBM	CSCO	MSFT
P / ZINB	AIC	267461915.62	280418352.77	138404692.48	148394239.86	142150126.37	124884854.32
	DM	-235.37	-269.95	-384.02	-436.68	-361.32	-945.72
G / ZINB	AIC	2995721.03	3894926.03	3283604.41	4270769.74	6100642.18	10641774.31
	DM	-689.62	-981.64	-793.38	-953.95	-668.61	-1272.63
NB / ZINB	AIC	45891.96	58214.58	118521.51	153740.09	279306.75	617624.25
	DM	-119.12	-138.71	-221.39	-261.55	-287.93	-469.99
ZIP / ZINB	AIC	104329981.44	100918549.37	60299913.57	58956197.68	58305592.35	43332973.41
	DM	-170.40	-210.26	-291.16	-311.60	-336.45	-617.43
ZIG / ZINB	AIC	49991.79	50910.51	84210.77	77231.38	122581.75	112468.26
	DM	-80.79	-97.10	-108.19	-113.01	-123.58	-144.40

excessive zeros is further supported by the better in-sample and out-of-sample fit for the zero-inflated distributions as reported in Table 6. Table 7 shows that it is also suitable to let the zero-inflation parameter be time-varying as this increases the in-sample and out-of-sample fit, particularly for the more traded stocks MCD, IBM, CSCO, and MSFT. This is in line with the mean probabilities of zero value when the observation is either zero or positive reported in Table 5.

5.5 Degree of Rounding

The choice of rounding to hundredths of a second, i.e. centiseconds, is motivated by Figure 1 which shows that the majority of excessive close-to-zero durations is concentrated in values 0 and 0.001 and the occurrence of larger values quickly decreases. In this section, we study the impact of different degrees of rounding and whether this choice is appropriate. Again, we use the difference in the AIC to assess the in-sample fit and the DM statistic to assess the out-of-sample fit. When comparing two models with different degrees of rounding, we compute the log-likelihood (which is the base for both AIC and DM) with respect to the rounding to fewer decimal places. A probability under the rounding to fewer decimal places is then the sum of the corresponding probabilities under the rounding to more decimal places. We consider rounding to zero decimal places (seconds), one decimal place (deciseconds), two decimal places (centiseconds), and three decimal places (milliseconds), i.e. the original data.

Table 8 shows the impact of increasing rounding. The rounding to centiseconds is clearly preferred over no rounding, i.e. precision to milliseconds. This is caused by the inability of the ZIACD model on milliseconds to account for an excessive probability of durations between 0.001 and 0.009 seconds; mostly, however, 0.001 seconds. The choice between the rounding to centiseconds and deciseconds differs for the individual stocks. For the INGA and AMSL stocks traded on the EURONEXT exchange, the model on deciseconds performs better. The difference in the AIC and the value of the DM statistic suggesting deciseconds are significant but smaller compared to the other values in Table 8. To keep the reported results simple, we stick with the model on centiseconds. For the more trade stocks MCD, IBM, CSCO, and MSFT, the model on centiseconds clearly outperforms the model on deciseconds. Finally, deciseconds are preferred over seconds for all stocks.

Table 7: The difference in the Akaike information criterion (AIC) and the Diebold–Mariano (DM) statistic for the zero-inflated negative binomial model with all parameters static (SSS), dynamic μ (DSS), dynamic μ, α (DDS), dynamic μ, π (DSD), and dynamic μ, α, π (DDD).

		EURONEXT		NYSE		NASDAQ	
Dynamics	Crit.	INGA	ASML	MCD	IBM	CSCO	MSFT
SSS / DDD	AIC	21492.36	27667.68	264677.50	325382.37	839689.51	1780799.16
	DM	-86.74	-133.22	-302.72	-393.75	-425.24	-762.36
DSS / DDD	AIC	8135.80	7612.61	219595.15	271386.43	742989.86	1447522.61
	DM	-57.16	-71.11	-274.29	-321.08	-396.20	-619.05
DDS / DDD	AIC	5652.62	4831.46	86375.82	112311.74	212495.87	521863.12
	DM	-50.55	-56.63	-177.80	-208.29	-231.05	-420.17
DSD / DDD	AIC	1617.11	1953.18	7286.14	5069.21	6614.27	9203.83
	DM	-13.99	-24.88	-32.19	-35.19	-24.04	-32.57

Table 8: The difference in the Akaike information criterion (AIC) and the Diebold–Mariano (DM) statistic for the zero-inflated negative binomial model based on data rounded to milliseconds (ms), centiseconds (cs), deciseconds (ds), and seconds (s).

		EURONEXT		NYSE		NASDAQ	
Precision	Crit.	INGA	ASML	MCD	IBM	CSCO	MSFT
ms / cs	AIC	28004.18	37387.09	49082.83	58914.43	42215.78	178862.15
	DM	-74.72	-102.18	-101.74	-97.79	-32.09	-114.94
cs / ds	AIC	3085.80	4326.46	-18588.85	-28803.35	-66676.58	-233072.75
	DM	-11.59	-29.46	47.73	76.70	87.26	218.46
ds / s	AIC	-1071.15	-1021.23	-34331.20	-42690.00	-70647.10	-34868.11
	DM	10.75	12.33	87.23	97.62	105.13	66.52

Table 9: The difference in the Akaike information criterion (AIC) and the Diebold–Mariano (DM) statistic for the generalized gamma model (GG) with zeros discarded (Discard) or truncated (Trunc) and the zero-inflated negative binomial model (ZINB).

		EURONEXT		NYSE		NASDAQ	
Model	Crit.	INGA	ASML	MCD	IBM	CSCO	MSFT
GG Discard / ZINB Discard	AIC	14808.54	18250.42	24298.98	32947.27	69568.44	143388.69
	DM	-41.41	-48.97	-72.66	-81.80	-103.98	-143.54
GG Trunc to 0.001 / ZINB	AIC	268465.60	356891.51	360261.19	469852.61	730365.26	1358026.51
	DM	-293.92	-385.96	-342.79	-387.75	-258.53	-548.24
GG Trunc to 0.0005 / ZINB	AIC	218050.18	302587.56	325396.64	406946.28	572429.68	1111215.05
	DM	-245.36	-337.90	-203.65	-342.76	-365.74	-479.62
GG Trunc to 0.0001 / ZINB	AIC	174616.63	227173.46	283440.84	329196.87	462917.51	1073765.49
	DM	-224.66	-271.87	-154.87	-299.34	-191.52	-461.16

5.6 Comparison to Continuous Models

We compare the proposed discrete ZIACD model with continuous models based on the generalized gamma distribution (see Technical Appendix C) with GAS dynamics. The generalized gamma distribution contains the exponential, Weibull, and gamma distributions as special cases and belongs to the family of the generalized F distribution. The use of the generalized gamma distribution in ACD models was proposed by Lunde (1999). Both Bauwens *et al.* (2004) and Fernandes and Grammig (2005) found that the generalized gamma distribution is more adequate than the exponential, Weibull, and Burr distributions. The study Xu (2013) shows that the log-normal distribution does not outperform the generalized gamma distribution either. For these reasons, the generalized gamma distribution is our main candidate for the competing continuous distribution. In our comparison, we do not consider the generalized F distribution as it has four parameters and in most cases of financial durations reduces to the generalized gamma distribution as discussed by Hautsch (2003) and Hautsch (2012). We also do not consider the Birnbaum–Saunders distribution as it models the median instead of the scale parameter and therefore does not strictly belong to the traditional ACD class. Models based on continuous distributions must address the issue of zero durations. We consider two ways of dealing with zero values in continuous models – discarding them and truncating them to a given value. Furthermore, we consider three values for truncating – 0.001, 0.0005, and 0.0001 seconds. Bauwens (2006) used truncation to the half of the smallest increment, which is 0.0005 seconds in our case. Similarly to the previous section, we compute log-likelihood on a discrete grid of centiseconds. In the case of discarding zeros, we compare the generalized gamma model with the zero-inflated negative binomial model that is also estimated without zero values.

Figure 4 demonstrates the unsuitability of the approach discarding zeros. Similarly to Figure 2, the generalized gamma model is not able to capture unusually increased occurrence of 0.06 seconds (for the INGA and ASML stocks) and 0.10 seconds (for all stocks). A crucial problem, however, is significantly underestimated probabilities in the wider vicinity of zero. In the case of the zero value itself, the difference in probability reaches -10.17 percent for the AMSL stock. Note that Figure 4 has much larger scale than Figure 2. Table 9 then confirms the superiority of the ZIACD model over the continuous alternatives in terms of the difference in the AIC and the DM statistic. Concerning the treatment of zero values, we can see that it is better to truncate zeros to smaller values but it is even better to just discard them. Either way, the results imply that the loss of decimal places in the proposed ZIACD model is of much less importance than the incorrect treatment of zero values in the continuous models.

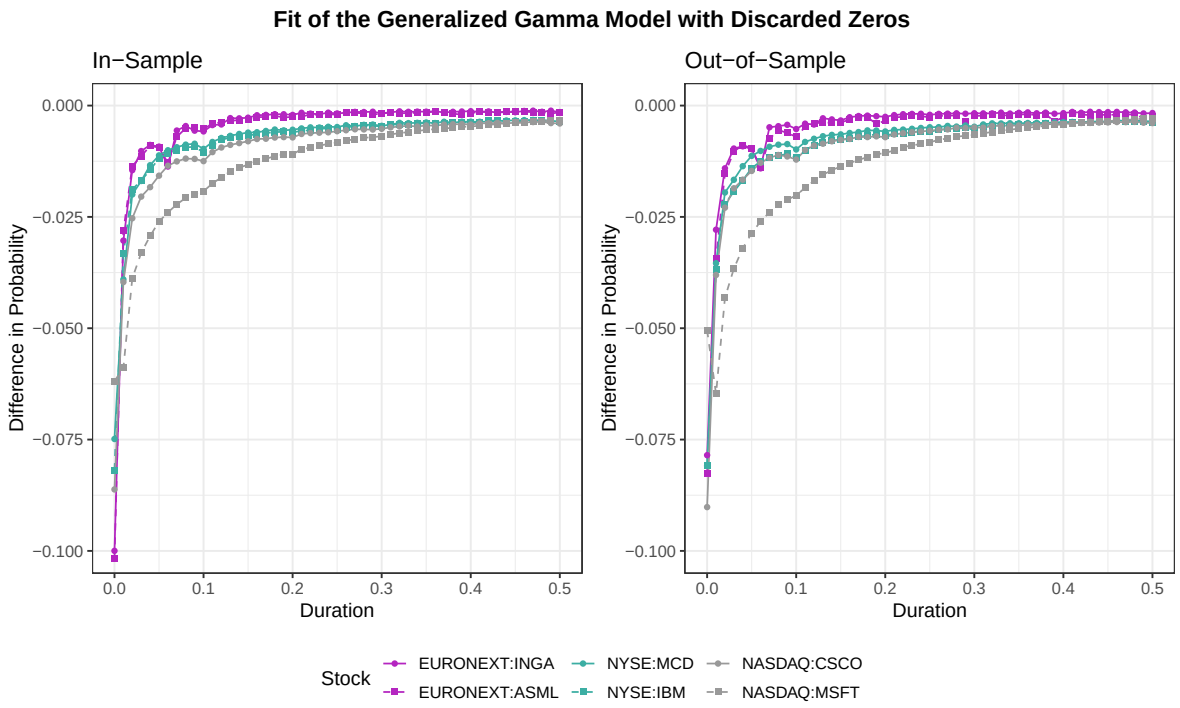

Figure 4: The in-sample and out-of-sample difference between the conditional probabilities given by the generalized gamma model with discarded zeros and the unconditional distribution of observations.

6 Discussion

6.1 Discreteness of Data

As mentioned above, our paper studies data with high-precision timestamps. Although it is nowadays quite common that exchanges record transactions with precision to one millisecond or higher, one can encounter preprocessed datasets with precision to one second due to their easier readability. In some cases, this can even be the only dataset provided by the exchange to the public[3]. For these low-precision data, it is more natural to use a discrete model such as ours rather than a continuous model.

To our knowledge, Grimshaw *et al.* (2005) is the only paper addressing the issue of rounding in financial durations analysis. They found that ignoring the discreteness of data leads to a distortion of time-dependence tests in financial durations. More loosely related, Schneeweiss *et al.* (2010) reviewed the bias-inducing effects of rounding. Tricker (1984) and Taraldsen (2011) explored the effects of rounding on the exponential distribution while Tricker (1992) dealt with the gamma distribution. Zhang *et al.* (2010) and Li and Bai (2011) found that the rounding errors in autoregressive processes can further accumulate making continuous models unreliable.

Let us conduct the following experiment to explore the influence of rounding on the estimation of GAS models based on discrete and continuous distributions. We simulate 10000 observations using a dynamic model based on the generalized gamma distribution with the time-varying scale parameter following the GAS dynamics given by $c = 0.10$, $a = 0.10$, $b = 0.90$ and the two static shape parameters $\theta = 0.50$ and $\varphi = 0.50$. The unconditional mean is then approximately equal to 2.05. Then, we round down the observations to a given number of decimal places. Finally, using rounded observations, we estimate GAS models based on the generalized gamma distribution with zero values (created by the

[3]For example, the Prague Stock Exchange currently records times of transactions with precision to one millisecond and distributes millisecond data to its members and external agencies. However, data provided to individuals have a precision of one second only.

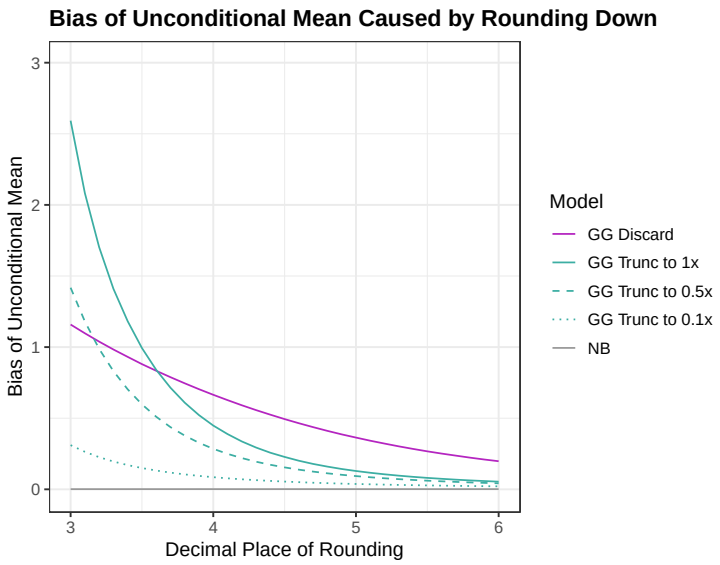

Figure 5: The bias of the unconditional mean given by the generalized gamma model (GG) with zeros discarded (Discard) or truncated (Trunc) and the negative binomial model (NB).

rounding) either discarded or truncated as well as the GAS model based on the negative binomial distribution. Note that we do not consider zero inflation in the negative binomial distribution as there are no excessive zeros generated by a different process. The simulation is repeated 1000 times. Figure 5 shows the bias of the unconditional mean of the estimated models with data rounded down to decimal places ranging from 3 up to 6. The negative binomial model, although with incorrectly specified distribution, has the smallest bias. On the other hand, the generalized gamma model with either treatment of zero values has a much higher bias which increases with rounding to fewer decimal places. This is caused by an increased occurrence of discarded or truncated zero values which significantly distorts the continuous distribution. This experiment demonstrates that it is more appropriate to use a distribution that is able to handle zero values, even though it is not the true distribution of the data generating process.

6.2 Other Mixture Models

On a final note, we discuss some potential alternatives to our proposed model that also utilize a mixture of two processes to capture unrelated and split transactions.

One possibility is to consider a hurdle model based on a continuous distribution with a point mass at zero. For example, the dynamic zero-augmented model of Hautsch *et al.* (2014)[4] or the dynamic censoring model of Harvey and Ito (2020) could be used. Hautsch *et al.* (2014) proposed a multiplicative error model based on a zero-augmented distribution and applied it to high-frequency time series of cumulated trading volumes. Harvey and Ito (2020) proposed a dynamic model with a left-shifted distribution for non-zero observations and censored negative values and applied it to daily rainfalls in northern Australia. Note that similarly to us, Harvey and Ito (2020) utilized the GAS framework. There are, however, two issues with this approach. Without any transformation of data, both these models would require split transactions to result in exactly zero durations, which is not realistic as shown in Section 1. Of course, one could follow our approach and round down durations below a given threshold, e.g. one hundredth of a second, to zero. Unlike in our approach, only durations below the threshold would be rounded and durations above would be kept continuous. The second issue is that hurdle models assume that one process generates zero values while the other

[4]The use of zero-augmented models for duration modeling was suggested by Prof. T. V. Ramanathan during the 3rd Conference and Workshop on Statistical Methods in Finance (Chennai, December 16–19, 2017).

process generates positive values only. In other words, it would not be possible to determine the ratio between zeros caused by unrelated and split transactions as all zeros would be attributed solely to split transactions. For this reason, our proposed model is superior.

A more complex approach is to assume a non-trivial process for split transactions. Both processes would then generate positive values and at least one of them would also generate zero values. This could be accomplished within either a continuous or discrete framework depending on the underlying data. The choice of a continuous distribution for the process governing split transactions would, however, be limited as zero is required to lie in its support. An exponential distribution would be an obvious starting point here. Note that the appropriately chosen process governing split transactions would not require any transformation of data, which would be a major benefit. On the other hand, the potential complexity of such a model could be a drawback. The ACD model based on a mixture of two non-trivial processes is the direction of our future research.

7 Conclusion

We analyze trade durations with split transactions manifesting themselves as zero and close-to-zero values. We round down durations to hundredths of a second and approach this problem within a discrete framework. To capture excessive zero values and autocorrelation structure in durations, we propose a model based on the zero-inflated negative binomial distribution with score dynamics for the time-varying parameters. We label this model the zero-inflated autoregressive conditional duration model or ZIACD model for short. The paper has three main contributions.

1. We extend the theory of GAS models for the zero-inflated negative binomial distribution with time-varying scale parameter. Specifically, we establish the invertibility of the score filter. We also derive sufficient conditions for the consistency and asymptotic normality of the maximum likelihood of the model parameters.

2. We argue that zero durations should not be removed from the data as they can correspond not only to split transactions but to unrelated transactions as well. Even more, split transactions can generate not only zero values but positive values as well. In the empirical study, the proposed model identifies that split transactions form between 92 and 98 percent of durations smaller than 0.01 seconds. Furthermore, between 53 and 75 percent of all durations correspond to split transactions.

3. We compare the proposed discrete approach with the commonly used continuous approach. We find that even when durations are recorded with high precision suitable for continuous modeling, the proposed discrete model estimated from rounded durations outperforms traditional continuous models based on unrounded data due to its correct treatment of zero and close-to-zero values.

Our proposed model can be utilized in joint modeling of prices and durations. It also allows studying the trading process from the market microstructure perspective. Future research should focus on more complex mixture models, whether in discrete or continuous frameworks, that do not require any transformation of data. However, it should be noted, that these complex models might lose the benefits of our ZIACD model such as simple implementability in practice and verifiability of sufficient conditions for asymptotic properties of the estimator.

Acknowledgements

Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures. We would like to thank Michal Černý and Tomáš Cipra for their comments. We would also like to thank participants of the 61st Meeting of EURO Working Group for Commodities and Financial Modelling (Kaunas, May 16–18, 2018) and the 2nd International Conference on Econometrics and Statistics (Hong Kong, June 19–21, 2018) for fruitful discussions.

Funding

The work of Francisco Blasques was supported by the Dutch Science Foundation (NWO) under project VI.Vidi.195.099. The work of Vladimír Holý was supported by the Internal Grant Agency of the University of Economics, Prague under project F4/21/2018. The work of Petra Tomanová was supported by the Czech Science Foundation under project 23-06139S.

References

Akaike H (1973). "Information Theory and an Extension of the Maximum Likelihood Principle." In *Proceedings of the 2nd International Symposium on Information Theory*, 267–281. Budapest. https://link.springer.com/chapter/10.1007/978-1-4612-1694-0{_}15.

Akaike H (1974). "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control*, **19**(6), 716–723. ISSN 0018-9286. https://doi.org/10.1109/tac.1974.1100705.

Amisano G, Giacomini R (2007). "Comparing Density Forecasts via Weighted Likelihood Ratio Tests." *Journal of Business & Economic Statistics*, **25**(2), 177–190. ISSN 0735-0015. https://doi.org/10.1198/073500106000000332.

Andrée BPJ, Blasques F, Koomen E (2017). "Smooth Transition Spatial Autoregressive Models." https://ssrn.com/abstract=2977830.

Andres P, Harvey A (2012). "The Dynamic Location/Scale Model." https://doi.org/10.17863/cam.4972.

Bao Y, Lee TH, Saltoğlu B (2007). "Comparing Density Forecast Models." *Journal of Forecasting*, **26**(3), 203–225. ISSN 0277-6693. https://doi.org/10.1002/for.1023.

Bauwens L (2006). "Econometric Analysis of Intra-Daily Trading Activity on the Tokyo Stock Exchange." *Monetary and Economic Studies*, **24**(1), 1–24. ISSN 0288-8432. http://www.imes.boj.or.jp/research/abstracts/english/me24-1-1.html.

Bauwens L, Giot P (2000). "The Logarithmic ACD Model: An Application to the Bid-Ask Quote Process of Three NYSE Stocks." *Annales d'Économie et de Statistique*, **60**, 117–149. ISSN 0769-489X. https://doi.org/10.2307/20076257.

Bauwens L, Giot P (2003). "Asymmetric ACD Models: Introducing Price Information in ACD Models." *Empirical Economics*, **28**(4), 709–731. ISSN 0377-7332. https://doi.org/10.1007/s00181-003-0155-7.

Bauwens L, Giot P, Grammig J, Veredas D (2004). "A Comparison of Financial Duration Models via Density Forecasts." *International Journal of Forecasting*, **20**(4), 589–609. ISSN 0169-2070. https://doi.org/10.1016/j.ijforecast.2003.09.014.

Bauwens L, Hautsch N (2009). "Modelling Financial High Frequency Data Using Point Processes." In *Handbook of Financial Time Series*, first Edition, Chapter 41, 953–979. Springer, Berlin, Heidelberg. ISBN 978-3-540-71296-1. https://doi.org/10.1007/978-3-540-71297-8.

Bauwens L, Veredas D (2004). "The Stochastic Conditional Duration Model: A Latent Variable Model for the Analysis of Financial Durations." *Journal of Econometrics*, **119**(2), 381–412. ISSN 0304-4076. https://doi.org/10.1016/s0304-4076(03)00201-x.

Bhatti CR (2010). "The Birnbaum-Saunders Autoregressive Conditional Duration Model." *Mathematics and Computers in Simulation*, **80**(10), 2062–2078. ISSN 0378-4754. https://doi.org/10.1016/j.matcom.2010.01.011.

Blasques F, Koopman SJ, Lucas A (2015). "Information-Theoretic Optimality of Observation-Driven Time Series Models for Continuous Responses." *Biometrika*, **102**(2), 325–343. ISSN 0006-3444. https://doi.org/10.1093/biomet/asu076.

Blasques F, van Brummelen J, Koopman SJ, Lucas A (2022). "Maximum Likelihood Estimation for Score-Driven Models." *Journal of Econometrics*, **227**(2), 325–346. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2021.06.003.

Bortoluzzo AB, Morettin PA, Toloi CMC (2010). "Time-Varying Autoregressive Conditional Duration Model." *Journal of Applied Statistics*, **37**(5), 847–864. ISSN 0266-4763. https://doi.org/10.1080/02664760902914458.

Boswell M, Patil GP (1970). "Chance Mechanisms Generating the Negative Binomial Distribution." In GP Patil (Ed.), *Random Counts in Models and Structures*, Volume 1, 3–22. Penn State University Press. http://www.psupress.org/books/titles/0-271-00114-3.html.

Bougerol P (1993). "Kalman Filtering with Random Coefficients and Contractions." *SIAM Journal on Control and Optimization*, **31**(4), 942–959. ISSN 0363-0129. https://doi.org/10.1137/0331041.

Cameron AC, Trivedi PK (1986). "Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests." *Journal of Applied Econometrics*, **1**(1), 29–53. ISSN 0883-7252. https://doi.org/10.1002/jae.3950010104.

Cameron AC, Trivedi PK (2013). *Regression Analysis of Count Data*. Second Edition. Cambridge University Press, New York. ISBN 978-1-107-01416-9. https://doi.org/10.1017/cbo9781139013567.

Chen F, Diebold FX, Schorfheide F (2013). "A Markov-Switching Multifractal Inter-Trade Duration Model, with Application to US Equities." *Journal of Econometrics*, **177**(2), 320–342. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2013.04.016.

Christou V, Fokianos K (2014). "Quasi-Likelihood Inference for Negative Binomial Time Series Models." *Journal of Time Series Analysis*, **35**(1), 55–78. ISSN 0143-9782. https://doi.org/10.1111/jtsa.12050.

Cox DR (1981). "Statistical Analysis of Time Series: Some Recent Developments." *Scandinavian Journal of Statistics*, **8**(2), 93–108. ISSN 0303-6898. https://doi.org/10.2307/4615819.

Creal D, Koopman SJ, Lucas A (2013). "Generalized Autoregressive Score Models with Applications." *Journal of Applied Econometrics*, **28**(5), 777–795. ISSN 0883-7252. https://doi.org/10.1002/jae.1279.

De Luca G, Gallo GM (2004). "Mixture Processes for Financial Intradaily Durations." *Studies in Nonlinear Dynamics and Econometrics*, **8**(2), 1–18. ISSN 1081-1826. https://doi.org/10.2202/1558-3708.1223.

De Luca G, Gallo GM (2009). "Time-Varying Mixing Weights in Mixture Autoregressive Conditional Duration Models." *Econometric Reviews*, **28**(1-3), 102–120. ISSN 0747-4938. https://doi.org/10.1080/07474930802387944.

De Luca G, Zuccolotto P (2003). "Finite and Infinite Mixtures for Financial Durations." *Metron - International Journal of Statistics*, **61**(3), 431–455. ISSN 00261424. https://ideas.repec.org/a/mtn/ancoec/030307.html.

Diebold FX, Mariano RS (1995). "Comparing Predictive Accuracy." *Journal of Business & Economic Statistics*, **13**(3), 253–263. ISSN 0735-0015. https://doi.org/10.1080/07350015.1995.10524599.

Diks C, Panchenko V, van Dijk D (2011). "Likelihood-Based Scoring Rules for Comparing Density Forecasts in Tails." *Journal of Econometrics*, **163**(2), 215–230. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2011.04.001.

Engle RF (2000). "The Econometrics of Ultra-High-Frequency Data." *Econometrica*, **68**(1), 1–22. ISSN 0012-9682. https://doi.org/10.1111/1468-0262.00091.

Engle RF, Russell JR (1998). "Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data." *Econometrica*, **66**(5), 1127–1162. ISSN 0012-9682. https://doi.org/10.2307/2999632.

Feng D (2004). "Stochastic Conditional Duration Models with "Leverage Effect" for Financial Transaction Data." *Journal of Financial Econometrics*, **2**(3), 390–421. ISSN 1479-8409. https://doi.org/10.1093/jjfinec/nbh016.

Fernandes M, Grammig J (2005). "Nonparametric Specification Tests for Conditional Duration Models." *Journal of Econometrics*, **127**(1), 35–68. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2004.06.003.

Fernandes M, Grammig J (2006). "A Family of Autoregressive Conditional Duration Models." *Journal of Econometrics*, **130**(1), 1–23. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2004.08.016.

Gallant AR, White H (1988). *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. First Edition. Basil Blackwell, Oxford. ISBN 978-0-631-15765-6. https://books.google.com/books?id=VVOqQgAACAAJ.

Ghysels E, Gouriéroux C, Jasiak J (2004). "Stochastic Volatility Duration Models." *Journal of Econometrics*, **119**(2), 413–433. ISSN 0304-4076. https://doi.org/10.1016/S0304-4076(03)00202-1.

Gómez-Déniz E, Pérez-Rodríguez JV (2016). "Conditional Duration Model and the Unobserved Market Heterogeneity of Traders: An Infinite Mixture of Non-Exponentials." *Revista Colombiana de Estadistica*, **39**(2), 307–323. ISSN 01201751. https://doi.org/10.15446/rce.v39n2.51584.

Gómez-Déniz E, Pérez-Rodríguez JV (2017). "Mixture Inverse Gaussian for Unobserved Heterogeneity in the Autoregressive Conditional Duration Model." *Communications in Statistics - Theory and Methods*, **46**(18), 9007–9025. ISSN 0361-0926. https://doi.org/10.1080/03610926.2016.1200094.

Gorgi P (2018). "Integer-Valued Autoregressive Models with Survival Probability Driven By a Stochastic Recurrence Equation." *Journal of Time Series Analysis*, **39**(2), 150–171. ISSN 0143-9782. https://doi.org/10.1111/jtsa.12272.

Grammig J, Maurer KO (2000). "Non-Monotonic Hazard Functions and the Autoregressive Conditional Duration Model." *The Econometrics Journal*, **3**(1), 16–38. ISSN 1368-4221. https://doi.org/10.1111/1368-423x.00037.

Grammig J, Wellner M (2002). "Modeling the Interdependence of Volatility and Inter-Transaction Duration Processes." *Journal of Econometrics*, **106**(2), 369–400. https://doi.org/10.1016/S0304-4076(01)00105-1.

Greene WH (1994). "Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models." http://ssrn.com/abstract=1293115.

Grimshaw SD, McDonald J, McQueen GR, Thorley S (2005). "Estimating Hazard Functions for Discrete Lifetimes." *Communications in Statistics - Simulation and Computation*, **34**(2), 451–463. ISSN 0361-0918. https://doi.org/10.1081/SAC-200055732.

Harvey AC (2013). *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series*. First Edition. Cambridge University Press, New York. ISBN 978-1-107-63002-4. https://doi.org/10.1017/cbo9781139540933.

Harvey AC, Ito R (2020). "Modeling Time Series When Some Observations Are Zero." *Journal of Econometrics*, **214**(1), 33–45. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2019.05.003.

Hautsch N (2001). "Modelling Intraday Trading Activity Using Box-Cox ACD Models." https://doi.org/10.2139/ssrn.289643. https://ssrn.com/abstract=289643.

Hautsch N (2003). "Assessing the Risk of Liquidity Suppliers on the Basis of Excess Demand Intensities." *Journal of Financial Econometrics*, **1**(2), 189–215. ISSN 1479-8409. https://doi.org/10.1093/jjfinec/nbg010.

Hautsch N (2012). *Econometrics of Financial High-Frequency Data*. First Edition. Springer, Berlin, Heidelberg. ISBN 978-3-642-21924-5. https://doi.org/10.1007/978-3-642-21925-2.

Hautsch N, Malec P, Schienle M (2014). "Capturing the Zero: A New Class of Zero-Augmented Distributions and Multiplicative Error Processes." *Journal of Financial Econometrics*, **12**(1), 89–121. ISSN 1479-8409. https://doi.org/10.1093/jjfinec/nbt002.

Herrera R, Schipp B (2013). "Value at Risk Forecasts by Extreme Value Models in a Conditional Duration Framework." *Journal of Empirical Finance*, **23**, 33–47. ISSN 0927-5398. https://doi.org/10.1016/j.jempfin.2013.05.002.

Holý V (2020). "Impact of the Parametrization and the Scaling Function in Dynamic Score-Driven Models: The Case of the Negative Binomial Distribution." In *Proceedings of the 38th International Conference Mathematical Methods in Economics*, 173–179. Mendel University in Brno, Brno. ISBN 978-80-7509-734-7. https://mme2020.mendelu.cz/wcd/w-rek-mme/mme2020{_}conference{_}proceedings{_}final.pdf.

Holý V, Tomanová P (2022). "Modeling Price Clustering in High-Frequency Prices." *Quantitative Finance*, **22**(9), 1649–1663. ISSN 1469-7688. https://doi.org/10.1080/14697688.2022.2050285.

Hujer R, Vuletic S, Kokot S (2005). "The Markov Switching ACD Model." https://doi.org/10.2139/ssrn.332381.

Jasiak J (1998). "Persistence in Intertrade Durations." *Finance*, **19**, 166–195. ISSN 1556-5068. https://doi.org/10.2139/ssrn.162008.

Jeyasreedharan N, Allen DE, Yang JW (2014). "Yet Another ACD Model: The Autoregressive Conditional Directional Duration (ACDD) Model." *Annals of Financial Economics*, **9**(1), 1450004/1–1450004/20. ISSN 2010-4952. https://doi.org/10.1142/S2010495214500043.

Konishi S, Kitagawa G (2008). *Information Criteria and Statistical Modeling*. Springer Series in Statistics. Springer, New York. ISBN 978-0-387-71886-6. https://doi.org/10.1007/978-0-387-71887-3.

Koopman SJ, Lit R (2019). "Forecasting Football Match Results in National League Competitions Using Score-Driven Time Series Models." *International Journal of Forecasting*, **35**(2), 797–809. ISSN 0169-2070. https://doi.org/10.1016/j.ijforecast.2018.10.011.

Koopman SJ, Lit R, Lucas A, Opschoor A (2018). "Dynamic Discrete Copula Models for High-Frequency Stock Price Changes." *Journal of Applied Econometrics*, **33**(7), 966–985. ISSN 0883-7252. https://doi.org/10.1002/jae.2645.

Koopman SJ, Lucas A, Scharth M (2016). "Predicting Time-Varying Parameters with Parameter-Driven and Observation-Driven Models." *Review of Economics and Statistics*, **98**(1), 97–110. ISSN 0034-6535. https://doi.org/10.1162/rest_a_00533.

Lambert D (1992). "Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing." *Technometrics*, **34**(1), 1–14. ISSN 0040-1706. https://doi.org/10.2307/1269547.

Leiva V, Saulo H, Leão J, Marchant C (2014). "A Family of Autoregressive Conditional Duration Models Applied to Financial Data." *Computational Statistics & Data Analysis*, **79**, 175–191. ISSN 0167-9473. https://doi.org/10.1016/j.csda.2014.05.016.

Li W, Bai ZD (2011). "Analysis of Accumulated Rounding Errors in Autoregressive Processes." *Journal of Time Series Analysis*, **32**(5), 518–530. ISSN 0143-9782. https://doi.org/10.1111/j.1467-9892.2010.00710.x.

Liu Z, Kong XB, Jing BY (2018). "Estimating the Integrated Volatility Using High-Frequency Data with Zero Durations." *Journal of Econometrics*, **204**(1), 18–32. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2017.12.008.

Lunde A (1999). "A Generalized Gamma Autoregressive Conditional Duration Model." https://www.researchgate.net/publication/228464216.

Mishra A, Ramanathan TV (2017). "Nonstationary Autoregressive Conditional Duration Models." *Studies in Nonlinear Dynamics and Econometrics*, **21**(4), 1–22. ISSN 1081-1826. https://doi.org/10.1515/snde-2015-0057.

Pacurar M (2008). "Autoregressive Conditional Duration Models in Finance: A Survey of the Theoretical and Empirical Literature." *Journal of Economic Surveys*, **22**(4), 711–751. ISSN 0950-0804. https://doi.org/10.1111/j.1467-6419.2007.00547.x.

Rao RR (1962). "Relations between Weak and Uniform Convergence of Measures with Applications." *The Annals of Mathematical Statistics*, **33**(2), 659–680. ISSN 0003-4851. https://doi.org/10.2307/2237541.

Russell JR, Engle RF (2005). "A Discrete-State Continuous-Time Model of Financial Transactions Prices and Times: The Autoregressive Conditional Multinomial-Autoregressive Conditional Duration Model." *Journal of Business & Economic Statistics*, **23**(2), 166–180. ISSN 0735-0015. https://doi.org/10.1198/073500104000000541.

Saranjeet KB, Ramanathan TV (2018). "Conditional Duration Models for High-Frequency Data: A Review on Recent Developments." *Journal of Economic Surveys*, **33**(1), 252–273. ISSN 0950-0804. https://doi.org/10.1111/joes.12261.

Schneeweiss H, Komlos J, Ahmad AS (2010). "Symmetric and Asymmetric Rounding: A Review and Some New Results." *AStA Advances in Statistical Analysis*, **94**(3), 247–271. ISSN 1863-8171. https://doi.org/10.1007/s10182-010-0125-2.

Sin CY, White H (1996). "Information Criteria for Selecting Possibly Misspecified Parametric Models." *Journal of Econometrics*, **71**(1-2), 207–225. ISSN 0304-4076. https://doi.org/10.1016/0304-4076(94)01701-8.

Stacy EW (1962). "A Generalization of the Gamma Distribution." *The Annals of Mathematical Statistics*, **33**(3), 1187–1192. ISSN 0003-4851. https://doi.org/10.2307/2237889.

Straumann D, Mikosch T (2006). "Quasi-Maximum-Likelihood Estimation in Conditionally Heteroscedastic Time Series: A Stochastic Recurrence Equations Approach." *The Annals of Statistics*, **34**(5), 2449–2495. ISSN 0090-5364. https://doi.org/10.1214/009053606000000803.

Taraldsen G (2011). "Analysis of Rounded Exponential Data." *Journal of Applied Statistics*, **38**(5), 977–986. ISSN 0266-4763. https://doi.org/10.1080/02664761003692431.

Tomanová P, Holý V (2021). "Clustering of Arrivals in Queueing Systems: Autoregressive Conditional Duration Approach." *Central European Journal of Operations Research*, **29**(3), 859–874. ISSN 1435-246X. https://doi.org/10.1007/s10100-021-00744-7.

Tricker AR (1992). "Estimation of Parameters for Rounded Data from Non-Normal Distributions." *Journal of Applied Statistics*, **19**(4), 465–471. ISSN 0266-4763. https://doi.org/10.1080/02664769200000041.

Tricker T (1984). "Effects of Rounding Data Sampled from the Exponential Distribution." *Journal of Applied Statistics*, **11**(1), 54–87. ISSN 0266-4763. https://doi.org/10.1080/02664768400000007.

Veredas D, Rodríguez-Poo JM, Espasa A (2002). "On the (Intradaily) Seasonality and Dynamics of a Financial Point Process: A Semiparametric Approach." https://ideas.repec.org/p/cor/louvco/2002023.html.

White H (1994). *Estimation, Inference and Specification Analysis*. First Edition. Cambridge University Press, Cambridge. ISBN 978-0-521-57446-4. https://doi.org/10.1017/CCOL0521252806.

Wintenberger O (2013). "Continuous Invertibility and Stable QML Estimation of the EGARCH(1,1) Model." *Scandinavian Journal of Statistics*, **40**(4), 846–867. ISSN 0303-6898. https://doi.org/10.1111/sjos.12038.

Xu D, Knight J, Wirjanto TS (2011). "Asymmetric Stochastic Conditional Duration Model - A Mixture-of-Normal Approach." *Journal of Financial Econometrics*, **9**(3), 469–488. ISSN 1479-8409. https://doi.org/10.1093/jjfinec/nbq026.

Xu Y (2013). "The Lognormal Autoregressive Conditional Duration (LNACD) Model and a Comparison with an Alternative ACD Models." https://ssrn.com/abstract=2382159.

Zhang B, Liu T, Bai ZD (2010). "Analysis of Rounded Data from Dependent Sequences." *Annals of the Institute of Statistical Mathematics*, **62**(6), 1143–1173. ISSN 0020-3157. https://doi.org/10.1007/s10463-009-0224-6.

Zhang MY, Russell JR, Tsay RS (2001). "A Nonlinear Autoregressive Conditional Duration Model with Applications to Financial Transaction Data." *Journal of Econometrics*, **104**(1), 179–207. ISSN 0304-4076. https://doi.org/10.1016/s0304-4076(01)00063-x.

Zheng Y, Li Y, Li G (2016). "On Fréchet Autoregressive Conditional Duration Models." *Journal of Statistical Planning and Inference*, **175**, 51–66. ISSN 0378-3758. https://doi.org/10.1016/j.jspi.2016.02.009.

A Proofs of Asymptotic Properties

Proof of Proposition 1:

Following Straumann and Mikosch (2006) and Blasques *et al.* (2022), we obtain invertibility by verifying that the conditions of Theorem 3.1 of Bougerol (1993) hold uniformly on a non-empty set

Θ, for any initialization $\hat{f}_1(\theta)$ In particular, we note that a $\ln^+$ bounded moment holds at $i = 1$ since

$$\begin{aligned}
\mathrm{E}\left[\log^+ \sup_{\theta\in\Theta}\left|c + b\hat{f}_1(\theta) + as(x_1, \hat{f}_1(\theta))\right|\right] &\leq 4\ln 2 + \mathrm{E}\left[\ln^+ \sup_{\theta\in\Theta}|c|\right] + \mathrm{E}\left[\ln^+ \sup_{\theta\in\Theta}\left|b\hat{f}_1(\theta)\right|\right] \\
&\quad + \mathrm{E}\left[\ln^+ \sup_{\theta\in\Theta}\left|as(x_1, \hat{f}_1(\theta))\right|\right] \\
&\leq 4\ln 2 + \mathrm{E}\left[\ln^+ \sup_{\theta\in\Theta}|c|\right] + \mathrm{E}\left[\ln^+ \sup_{\theta\in\Theta}|b|\right] \\
&\quad + \mathrm{E}\left[\ln^+ \sup_{\theta\in\Theta}\left|\hat{f}_1(\theta)\right|\right] + \mathrm{E}\left[\ln^+ \sup_{\theta\in\Theta}|a|\right] \\
&\quad + \mathrm{E}\left[\sup_{\theta\in\Theta}\left|s(x_1, \hat{f}_1(\theta))\right|\right] \\
&\leq 4\ln 2 + \max\{|c^-|, |c^+|\} + \max\{|b^-|, |b^+|\} \\
&\quad + \sup_{\theta\in\Theta}\left|\hat{f}_1(\theta)\right| + \max\{|a^-|, |a^+|\} + \mathrm{E}\left[\ln^+ \sup_{\theta\in\Theta}\left|s(x_1, \hat{f}_1(\theta))\right|\right] \\
&< \infty,
\end{aligned}$$

where the three inequalities follow by norm sub-additivity, as well as the $\ln^+$ sub-additive and sub-multiplicative inequalities in Lemma 2.2 of Straumann and Mikosch (2006), and the last bound follows since c, b, a are strictly positive and lie on the compact Θ and $\hat{f}_1(\theta)$ is a given real number. We also have that $\mathrm{E}\left[\ln^+ \sup_{\theta\in\Theta}|s(x_1, \hat{f}_1(\theta))|\right] < \infty$ as

$$\begin{aligned}
\mathrm{E}\left[\ln^+ \sup_{\theta\in\Theta}\left|s(x_i, \hat{f}_1(\theta), \theta)\right|\right] &= \mathrm{P}[x_i = 0] \cdot \ln^+ \sup_{\theta\in\Theta}\left|s(0, \hat{f}_1(\theta), \theta)\right| \\
&\quad + \mathrm{P}[x_i > 0] \cdot \mathrm{E}_{x_i>0}\left[\ln^+ \sup_{\theta\in\Theta}\left|s(x_i, \hat{f}_1(\theta), \theta)\right|\right] \\
&\leq \ln^+ \sup_{\theta\in\Theta}\left|s(0, \hat{f}_1(\theta), \theta)\right| + \mathrm{E}_{x_i>0}\left[\ln^+ \sup_{\theta\in\Theta}\left|s(x_i, \hat{f}_1(\theta), \theta)\right|\right] \\
&< \infty,
\end{aligned}$$

where $\mathrm{E}_{x_i>0}$ denotes the conditional expectation $\mathrm{E}_{x_i>0}[\cdot] = \mathrm{E}[\cdot|x_{i>0}]$ and

$$\begin{aligned}
\mathrm{E}\left[\ln^+ \sup_{\theta\in\Theta}|s(0, \hat{f}_1, \theta)|\right] &= \ln^+ \sup_{\theta\in\Theta}|s(0, \hat{f}_1, \theta)| \\
&= \ln^+ \sup_{\theta\in\Theta}\Big|(\pi - 1)\exp(\hat{f}_1)(\alpha\exp(\hat{f}_1) + 1)^{-1} \\
&\quad \cdot \left(1 + \pi(\alpha\exp(\hat{f}_1) + 1)^{\alpha^{-1}} - \pi\right)^{-1}\Big| \\
&\leq \ln^+ \sup_{\theta\in\Theta}|\pi - 1| + \ln^+ \sup_{\theta\in\Theta}|\exp(\hat{f}_1)| + \ln^+ \sup_{\theta\in\Theta}|(\alpha\exp(\hat{f}_1) + 1)^{-1}| \\
&\quad + \ln^+ \sup_{\theta\in\Theta}\left|\left(1 + \pi(\alpha\exp(\hat{f}_1) + 1)^{\alpha^{-1}} - \pi\right)^{-1}\right| \\
&< \infty,
\end{aligned}$$

which holds as the parameter vector θ lies on the compact set Θ, and $\hat{f}_1$ is a given point in $\mathbb{R}$, and

$$\begin{aligned}
\mathrm{E}_{x_i>0}\left[\ln^+ \sup_{\theta\in\Theta}\left|s(x_i, \hat{f}_1, \theta)\right|\right] &= \mathrm{E}_{x_1>0}\left[\ln^+ \sup_{\theta\in\Theta}\left|x_1 - \exp(\hat{f}_1)(\alpha\exp(\hat{f}_1) + 1)^{-1}\right|\right] \\
&\leq \mathrm{E}_{x_1>0}\left[\ln^+ \sup_{\theta\in\Theta}\left|x_1 - \exp(\hat{f}_1)\right|\right] \\
&\leq 2\ln(2) + \mathrm{E}_{x_1>0}\left[\ln^+|x_1|\right] + \ln^+|\exp(\hat{f}_1)| \\
&< \infty,
\end{aligned}$$

since x_1 has a logarithmic moment, Θ is compact and $\hat{f}_1 \in \mathbb{R}$.

Finally, the contraction condition of Bougerol (1993) is satisfied uniformly in $\theta \in \Theta$ since

$$\begin{aligned}
&\mathrm{E}\left[\ln \sup_f \sup_{\theta\in\Theta} \left| a\frac{\partial s(x_i, f, \theta)}{\partial f} + b \right|\right] < 0 \\
&\quad\Leftrightarrow\ \mathrm{P}[x_i = 0] \cdot \ln \sup_f \sup_{\theta\in\Theta} \left| a\frac{\partial s(0, f, \theta)}{\partial f} + b \right| \\
&\qquad + \mathrm{P}[x_i > 0] \cdot \mathrm{E}_{x_i>0}\left[\ln \sup_f \sup_{\theta\in\Theta} \left| a\frac{\partial s(x_i, f, \theta)}{\partial f} + b \right|\right] < 0
\end{aligned}$$

where

$$\begin{aligned}
&\mathrm{E}\left[\ln \sup_{\hat{f}} \sup_{\theta\in\Theta} \left| a\frac{\partial s(x_i, \hat{f}, \theta)}{\partial \hat{f}} + b \right|\right] < 0 \\
&\quad\Leftrightarrow\ \mathrm{P}[x_i = 0] \cdot \ln \sup_{\hat{f}} \sup_{\theta\in\Theta} \left| a\frac{\partial s(0, \hat{f}, \theta)}{\partial \hat{f}} + b \right| \\
&\qquad + \mathrm{P}[x_i > 0] \cdot \mathrm{E}_{x_i>0}\left[\ln \sup_{\hat{f}} \sup_{\theta\in\Theta} \left| a\frac{\partial s(x_i, \hat{f}, \theta)}{\partial \hat{f}} + b \right|\right] < 0 \\
&\quad\Leftrightarrow\ \left(\pi + (1-\pi)\left(\frac{\alpha^{-1}}{\alpha^{-1} + \hat{f}_i}\right)^{\alpha^{-1}}\right) \\
&\qquad \cdot \ln \sup_{\hat{f}} \sup_{\theta\in\Theta} \left| - a\frac{(\pi-1)^2 \exp(2\hat{f})}{(\alpha\exp(\hat{f})+1)^2\left(\pi(\alpha\exp(\hat{f})+1)^{1/\alpha} - \pi + 1\right)^2} \right. \\
&\qquad \left. - a\frac{(\pi-1)\exp(\hat{f})(\exp(\hat{f})-1)}{(\alpha\exp(\hat{f})+1)^2\left(\pi(\alpha\exp(\hat{f})+1)^{1/\alpha} - \pi + 1\right)} + b \right| \\
&\qquad + \left(1 - \pi - (1-\pi)\left(\frac{\alpha^{-1}}{\alpha^{-1} + \hat{f}_i}\right)^{\alpha^{-1}}\right) \cdot \mathrm{E}_{x_i>0}\left[\ln \sup_{\hat{f}} \sup_{\theta\in\Theta} \left| -a\frac{(\alpha x_i + 1)\exp(\hat{f})}{(\alpha\exp(\hat{f})+1)^2} + b \right|\right] < 0 \\
&\quad\Leftarrow\ \ln\left[\sup_{\theta\in\Theta}\left|a\frac{(\pi-1)^2}{2\alpha}\right| + \sup_{\theta\in\Theta}\left|a\frac{(\pi-1)}{\alpha^2}\right| + \sup_{\theta\in\Theta}|b|\right] + \mathrm{E}_{x_i>0}\left[\ln\left(\sup_{\theta\in\Theta}\left|a\frac{\alpha x_i + 1}{4\alpha}\right| + \sup_{\theta\in\Theta}|b|\right)\right] < 0.
\end{aligned}$$

This can be simplified by noting that

$$\begin{aligned}
\frac{\exp(2\hat{f})}{(\alpha\exp(\hat{f})+1)^2} &\leq \frac{1}{2\alpha}, \\
\left(\pi(\alpha\exp(\hat{f})+1)^{1/\alpha} - \pi + 1\right)^2 &\geq 1, \\
\frac{\exp(\hat{f})(\exp(\hat{f})-1)}{(\alpha\exp(\hat{f})+1)^2} &\leq \frac{1}{\alpha^2}.
\end{aligned}$$

This, in turn, implies that

$$\begin{aligned}
&\mathrm{E}\left[\ln \sup_{\hat{f}} \sup_{\theta\in\Theta} \left| a\frac{\partial s(x_i, \hat{f}, \theta)}{\partial \hat{f}} + b \right|\right] < 0 \\
&\quad\Leftarrow\ \left\{ \sup_{\theta\in\Theta} a\frac{(\pi-1)^2}{2\alpha} + \sup_{\theta\in\Theta} a\frac{|\pi-1|}{\alpha^2} + \sup_{\theta\in\Theta} b^+ < 1 \ \wedge\ \mathrm{E}_{x_i>0}\left[\ln\left(\sup_{\theta\in\Theta} a\frac{\alpha x_i + 1}{4\alpha} + b^+\right)\right] < 0 \right\} \\
&\quad\Leftarrow\ \left\{ \frac{a^+(\pi^- - 1)^2}{2\alpha^-} + \frac{a^+|\pi^- - 1|}{(\alpha^-)^2} + b^+ < 1 \ \wedge\ \mathrm{E}_{x_i>0}\left[\ln\left(\frac{a^+(\alpha^+ x_i + 1)}{4\alpha^-} + b^+\right)\right] < 0 \right\}.
\end{aligned}$$

Proof of Lemma 1:

This proof follows that of Blasques *et al.* (2022, Theorem 4.6). The existence and measurability of $\hat{\theta}_n$ is obtained through an application of White (1994, Theorem 2.11) or Gallant and White (1988, Lemma 2.1, Theorem 2.2), as Θ is compact and the log likelihood is continuous in θ and measurable in x_i. The consistency of the ML estimator, $\hat{\theta}_n(\hat{f}_1) \xrightarrow{as} \theta_0$, is obtained by White (1994, Theorem 3.4) or Gallant and White (1988, Theorem 3.3). Below, we note that we satisfy the sufficient conditions of uniform convergence of the log likelihood function

$$\sup_{\theta\in\Theta} |\hat{L}_n(\theta) - L_\infty(\theta)| \xrightarrow{as} 0 \;\forall\; \hat{f}_1 \in \mathcal{F} \quad \text{as} \quad n \to \infty,$$

and the identifiable uniqueness of the maximizer $\theta_0 \in \Theta$ introduced in White (1994),

$$\sup_{\theta:\|\theta-\theta_0\|>\epsilon} L_\infty(\theta) < L_\infty(\theta_0) \;\forall\; \epsilon > 0.$$

The uniform convergence of the criterion is obtained since, by norm sub-additivity, we can split the log likelihood as follows

$$\sup_{\theta\in\Theta} |\hat{L}_n(\theta) - L_\infty(\theta)| \leq \sup_{\theta\in\Theta} |\hat{L}_n(\theta) - L_n(\theta)| + \sup_{\theta\in\Theta} |L_n(\theta) - L_\infty(\theta)|. \tag{17}$$

The first term on the right-hand-side of (17) vanishes if $|\hat{l}_i(\theta) - l_i(\theta)| \xrightarrow{as} 0$ since

$$|\hat{L}_n(\theta) - L_n(\theta)| \leq \frac{1}{n}\sum^{n} |\hat{l}_i(\theta) - l_i(\theta)| \xrightarrow{as} 0,$$

and we have that

$$\sup_{\theta\in\Theta} |\hat{l}_i(\theta) - l_i(\theta)| \leq \sup_{\theta\in\Theta} \sup_{f} |\nabla(x_i, f, \theta)| \cdot \sup_{\theta\in\Theta} |\hat{f}_i(\theta) - f_i(\theta)| \xrightarrow{as} 0 \quad \forall\; \hat{f}_1 \in \mathcal{F} \quad \text{as} \quad n \to \infty,$$

where $\sup_{\theta\in\Theta} |\hat{f}_i(\theta) - f_i(\theta)| \xrightarrow{as} 0$ follows from the invertibility of the filter (proved in Proposition 1) and the product vanishes by the bounded logarithmic moment of the score $\mathrm{E}[\ln^+ \sup_f |\nabla(x_i, f)|] < \infty$ (see Lemma 2.1 in Straumann and Mikosch 2006). The logarithmic moment $\mathrm{E}[\ln^+ \sup_f |\nabla(x_i, f)|] < \infty$ follows as

$$\mathrm{E}\left[\ln^+ |s(0, \hat{f}_i)|\right] = \mathrm{E}\left[\ln^+ \left| \frac{\exp(\hat{f}_i)(\pi - 1)}{(\alpha \exp(\hat{f}_i) + 1)\left(1 + \pi(\alpha \exp(\hat{f}_i) + 1)^{\alpha^{-1}} - \pi\right)} \right|\right] < \infty,$$

$$\mathrm{E}_{x_i>0}\left[\ln^+ |s(x_i, \hat{f}_i)|\right] = \left| \frac{x_i - \exp(\hat{f}_i)}{\alpha \exp(\hat{f}_i) + 1} \right| < \infty \quad \text{for } x_i > 0.$$

Note that since we use unit scaling in Lemma 1, we have that $\nabla(x_i, f) = s\nabla(x_i, f)$. The uniform convergence of the second term on the right-hand-side of (17)

$$\sup_{\theta\in\Theta} |L_n(\theta) - L_\infty(\theta)| \xrightarrow{as} 0 \quad \forall\; \hat{f}_1 \in \mathcal{F} \quad \text{as} \quad n \to \infty,$$

follows by application of the ergodic theorem for separable Banach spaces in Rao (1962). We note that the $\{L_n(\cdot)\}_{t\in\mathbb{N}}$ has strictly stationary and ergodic elements as it depends on the limit strictly stationary and ergodic filter taking values in the Banach space of continuous functions $\mathbb{C}(\Theta, \mathbb{R})$ equipped with sup norm. We also note that $L_n(\cdot)$ has one bounded moment since $\mathrm{E}[L_n(\theta)] \leq \frac{1}{n}\sum^{n} \mathrm{E}[l_i(\theta)] < \infty$. In particular, the bounded moment for the log likelihood holds trivially if the data has a bounded

moment $\mathrm{E}[x_i] < \infty$ since $\ln \ell_i(x_i, \theta)$ is bounded in μ_i and bounded by a linear function in x_i,

$$\begin{aligned}
\ell_i(0,\theta) &= \ln P[X_i = 0 | \hat{f}_i(\theta), \theta] \\
&= \ln\left(\pi + (1-\pi)\left(\frac{\alpha^{-1}}{\alpha^{-1} + \exp(\hat{f}_i(\theta))}\right)^{\alpha^{-1}}\right), \\
\ell_i(x_i,\theta) &= \ln P[X_i = x_i | \hat{f}_i(\theta), \theta] \\
&= \ln(1-\pi) + \ln\frac{\Gamma(x_i + \alpha^{-1})}{\Gamma(x_i+1)\Gamma(\alpha^{-1})} \\
&\quad + \frac{1}{\alpha}\ln\left(\frac{\alpha^{-1}}{\alpha^{-1} + \exp(\hat{f}_i(\theta))}\right) + x_i \ln\left(\frac{\exp(\hat{f}_i(\theta))}{\alpha^{-1} + \exp(\hat{f}_i(\theta))}\right) \quad \text{for } x_i > 0.
\end{aligned}$$

The identifiable uniqueness (see e.g. White, 1994) follows from the compactness of Θ, the assumed uniqueness of θ_0, and the continuity of the limit likelihood function $\mathrm{E}[\ell_i(\theta)]$ in $\theta \in \Theta$.

Proof of Lemma 2:

This proof follows Blasques *et al.* (2022, Theorem 4.16). In particular, we obtain the asymptotic normality using the usual expansion argument found e.g. in White (1994, Theorem 6.2) by establishing:

(i) The consistency of $\hat{\theta}_n \xrightarrow{as} \theta_0 \in \mathrm{int}(\Theta)$, which follows immediately by Lemma 1.

(ii) The as twice continuous differentiability of $L_n(\theta, \hat{f}_1)$ in $\theta \in \Theta$, which holds trivially for our zero-inflated score model.

(iii) The asymptotic normality of the score, which can be shown to hold by verifying that,

$$\sqrt{n}\frac{\partial L_n(\theta_0)}{\partial \theta} \xrightarrow{d} N(0, \mathcal{I}(\theta_0)) \quad \text{as} \quad n \to \infty, \tag{18}$$

and

$$\sqrt{n}\left|\frac{\partial \hat{L}(\theta_0)}{\partial \theta} - \frac{\partial L(\theta_0)}{\partial \theta}\right| \xrightarrow{as} 0 \quad \text{as} \quad n \to \infty. \tag{19}$$

The asymptotic normality in (18) is obtained by application of a central limit theorem for martingale difference sequences to the score, after noting that the score

$$\frac{\partial L_n(\theta_0)}{\partial \theta} = \frac{1}{n}\sum^{n}\left(\frac{\partial \ell_i(x_i,\theta_0)}{\partial \theta} + \frac{\partial \ell_i(x_i,\theta_0)}{\partial f_i}\frac{\partial f_i(\theta_0)}{\partial \theta}\right).$$

has two bounded moments. In particular,

$$\mathrm{E}\left[\left\|\frac{\partial L_n(\theta_0)}{\partial \theta}\right\|^2\right] \leq \mathrm{E}\left[\left\|\frac{\partial \ell_i(x_i,\theta_0)}{\partial \theta}\right\|^2\right] + \mathrm{E}\left[\left\|\frac{\partial \ell_i(x_i,\theta_0)}{\partial f_i}\frac{\partial f_i(\theta_0)}{\partial \theta}\right\|^2\right] < \infty,$$

where the bounds

$$\mathrm{E}\left[\left\|\frac{\partial \ell_i(x_i,\theta_0)}{\partial \theta}\right\|^2\right] < \infty \quad \text{and} \quad \mathrm{E}\left[\left\|\frac{\partial \ell_i(x_i,\theta_0)}{\partial f_i}\frac{\partial f_i(\theta_0)}{\partial \theta}\right\|^2\right] < \infty,$$

hold, for example, under the assumption that

$$\mathrm{E}\left[\left\|\frac{\partial \ell_i(x_i,\theta_0)}{\partial f_i}\right\|^4\right] < \infty \quad \text{and} \quad \mathrm{E}\left[\left\|\frac{\partial \ell_i(x_i,\theta_0)}{\partial \theta}\right\|^4\right] < \infty;$$

by a generalized Holder's inequality as used e.g. in Blasques *et al.* (2022). For the negative binomial model it is easy to see for example that the four bounded moments for score term

$\partial \ell_i(x_i, \theta_0)/\partial f_i$ can be obtained if the data has four bounded moments, $\mathrm{E}|x_i|^4 < \infty$, by noting that

$$\begin{aligned}
\mathrm{E}\left[\sup_{\theta\in\Theta} \|s(0, \hat{f}_i, \theta)\|^4\right] &\leq \sup_{\mu} \sup_{\theta\in\Theta} \|s(0, \hat{f}_i, \theta)\|^4 \\
&= \sup_{\mu} \sup_{\theta\in\Theta} \left| (\pi - 1) \frac{\exp(\hat{f}_i)}{\alpha \exp(\hat{f}_i) + 1} \left(1 + \pi(\alpha \exp(\hat{f}_i) + 1)^{\alpha^{-1}} - \pi\right)^{-1} \right|^4 \\
&< \infty,
\end{aligned}$$

since $s(0, \hat{f}_i, \theta)$ is uniformly bounded in $\hat{f}_i$. Furthermore, by application of the so-called c_n-inequality, there exists a finite constant k such that,

$$\begin{aligned}
\mathrm{E}_{x_i>0}\left[\sup_{\theta\in\Theta} |s(x_i, \hat{f}_i, \theta)|^4\right] &= \mathrm{E}_{x_i>0}\left[\sup_{\theta\in\Theta} \left| x_i - \exp(\hat{f}_i)(\alpha \exp(\hat{f}_i) + 1)^{-1} \right|^4\right] \\
&\leq k \sup_{\theta\in\Theta} \frac{1}{\alpha} \mathrm{E}_{x_i>0}[x_i^4] + k|\alpha^{-1}|^4 \\
&< \infty.
\end{aligned}$$

Additionally, following the argument of Blasques *et al.* (2022, Theorem 4.14) and Straumann and Mikosch (2006, Lemma 2.1), the as convergence in (19) follows by the invertibility of the filter and its derivatives. The invertibility of the first derivative process can be verified by applying Theorem 2.10 in Straumann and Mikosch (2006). This theorem is analogue to Theorem 3.1 of Bougerol (1993), also used in the proof of Proposition 1 above, but it applies to perturbed stochastic sequences. For example, the updating equation for derivative process $\partial f_i/\partial c = \partial \hat{f}_i/\partial c$ takes the form

$$\frac{\partial \hat{f}_{i+1}}{\partial c} = 1 + b\frac{\partial \hat{f}_i}{\partial c} + \frac{\partial s(x_i, \hat{f}_i)}{\partial \hat{f}_i}\frac{\partial \hat{f}_i}{\partial c} = 1 + \left(b + \frac{\partial s(x_i, \hat{f}_i)}{\partial \hat{f}_i}\right)\frac{\partial \hat{f}_i}{\partial c}.$$

Hence, by application of Theorem 2.10 in Straumann and Mikosch (2006), the invertibility of this filter is ensured by (a) the invertibility of the filter $\{\hat{f}_i\}_{i\in\mathbb{N}}$ (shown in Proposition 1); (b) the contraction condition $\mathrm{E}[\ln|b + \partial s(x_i, \hat{f}_i)/\partial \hat{f}_i|] < 0$; and a logarithmic moment for $\partial^2 s(x_i, \hat{f}_i)/\partial \hat{f}_i^2$.

(iv) The uniform convergence of the Hessian, is obtained through the invertibility of the filter and its derivative processes. In particular, a sufficient condition is for the first and second derivatives of the filtering process to converge almost surely, exponentially fast, to a limit stationary and ergodic sequence,

$$\left\| \frac{\partial \hat{f}_i(\theta_0)}{\partial \theta} - \frac{\partial f_i(\theta_0)}{\partial \theta} \right\| \overset{eas}{\to} 0 \quad \text{and} \quad \sup_{\theta\in\Theta} \left\| \frac{\partial^2 \hat{f}_i(\theta)}{\partial \theta \partial \theta'} - \frac{\partial^2 f_i(\theta)}{\partial \theta \partial \theta'} \right\| \overset{eas}{\to} 0 \quad \text{as} \quad i \to \infty,$$

with four bounded moments

$$\mathrm{E}\left[\left\| \frac{\partial f_i(\theta_0)}{\partial \theta} \right\|^4\right] < \infty \quad \text{and} \quad \mathrm{E}\left[\sup_{\theta\in\Theta} \left\| \frac{\partial^2 f_i(\theta)}{\partial \theta \partial \theta'} \right\|^4\right] < \infty.$$

and to have logarithmic moments for cross derivatives,

$$\mathrm{E}\left[\sup_{\theta\in\Theta} \left\| \frac{\partial^2 \ell_i(x_i, \theta)}{\partial f_i \partial \theta'} \right\|\right] < \infty, \quad \mathrm{E}\left[\sup_{\theta\in\Theta} \left\| \frac{\partial^2 \ell_i(x_i, \theta)}{\partial f_i^2} \right\|\right] < \infty \quad \text{and} \quad \mathrm{E}\left[\sup_{\theta\in\Theta} \left\| \frac{\partial^2 \ell_i(x_i, \theta)}{\partial \theta \partial \theta'} \right\|\right] < \infty;$$

and also for the third-order derivatives of the log likelihood to have a uniform logarithmic bounded moment,

$$\mathrm{E}\left[\ln^+ \sup_{\theta\in\Theta} \left\| \frac{\partial^3 \ell_i(x_i, \theta_0)}{\partial f_i^2 \partial \theta'} \right\|\right] < \infty, \quad \mathrm{E}\left[\ln^+ \sup_{\theta\in\Theta} \left\| \frac{\partial^3 \ell_i(x_i, \theta_0)}{\partial f_i^3} \right\|\right] < \infty.$$

$$\text{and} \quad \mathrm{E}\left[\ln^+ \sup_{\theta\in\Theta}\left\|\frac{\partial^3 \ell_i(x_i,\theta_0)}{\partial\theta\partial\theta'\partial f}\right\|\right] < \infty;$$

Then by application of the ergodic theorem for separable Banach spaces in Rao (1962) to the limit Hessian (see also Blasques *et al.* 2022 and Straumann and Mikosch 2006, Theorem 2.7 for additional details), we have,

$$\sup_{\theta\in\Theta}\left\|\frac{\partial^2 L_n(\theta)}{\partial\theta\partial\theta'} - \mathrm{E}\left[\frac{\partial^2 \ell_i(\theta)}{\partial\theta\partial\theta'}\right]\right\| \xrightarrow{as} 0 \quad \text{as } n\to\infty. \tag{20}$$

(v) The non-singularity of the limit $L''_\infty(\theta) = \mathrm{E}[\ell''_i(\theta)] = \mathcal{I}(\theta)$ follows by the uniqueness of θ_0 and the independence of derivative processes (Straumann and Mikosch 2006, Theorem 2.7).

B Model Evaluation

It is well know that ranking models based on their expected log-likelihood $\mathrm{E}[\ell_i(\theta_0)]$ evaluated at the best (pseudo-true) parameter θ_0 is equivalent to model selection based on minimizing the expected Kullback-Leibler divergence between the true distribution of the data and the model-implied distribution. The sample log-likelihood is however an asymptotically biased estimator of the expected log likelihood. Under restrictive conditions, Akaike (1973, 1974) showed that the bias is approximately given by the number of parameters of the model $\dim(\theta)$. Since then, the AIC has been shown to consistently rank models according to the Kullback-Leibler divergence under considerably weaker conditions (Sin and White 1996; Konishi and Kitagawa 2008). Unfortunately, model specification and identification issues still exert a strong influence over the performance of in-sample information criteria.

For this reason, it could be interesting to consider criteria based on a validation sample. Lemma 3 highlights that log-likelihood based on an independent validation sample of m observations, $n\hat{L}_m(\hat{\theta}_n)$, is asymptotically unbiased for $n\mathrm{E}[\ell_i(\theta_0)]$. A proof can be found in Andrée *et al.* (2017)[5].

Lemma 3. *Let the conditions of Lemma 1 hold. Then* $\lim_{n,m\to\infty} \mathrm{E}\left[n\hat{L}_m(\hat{\theta}_n) - n\mathrm{E}[\ell_i(\theta_0)]\right] = 0.$

Lemma 4 uses a Diebold-Mariano test statistic (Diebold and Mariano, 1995) to test for differences in log-likelihoods across different models obtained from the validation sample (see Andrée *et al.*, 2017, for a proof). This test is also known as a logarithmic scoring rule, see e.g. Diks *et al.* (2011); Amisano and Giacomini (2007); Bao *et al.* (2007). Given two models, A and B, let $\tilde{\ell}_i^{\mathrm{A}}(\theta_0^{\mathrm{A}})$ and $\tilde{\ell}_i^{\mathrm{B}}(\theta_0^{\mathrm{B}})$ denote their respective log-likelihood contributions at a certain time i (in the validation sample) evaluated at each model's pseudo-true parameter. Define the log-likelihood difference

$$D_i^{A,B} := \tilde{\ell}_i^{\mathrm{A}}(\theta_0^A) - \tilde{\ell}_i^{\mathrm{B}}(\theta_0^B).$$

Finally, define the Diebold-Mariano test statistic

$$DM_{m,n} = \sqrt{m}\frac{\mu_D^{A,B}}{\sigma_D^{A,B}}, \qquad \mu_D^{A,B} = \frac{1}{m}\sum_{i=n+1}^{n+m} D_i^{A,B}, \qquad \sigma_D^{A,B} = \sqrt{\frac{1}{m-1}\sum_{i=n+1}^{n+m}\left(D_i^{A,B} - \mu_D^{A,B}\right)^2}.$$

Lemma 4 (Validation-Sample Test)**.** *Let Lemma 1 hold for both models A and B, such that* $\hat{\theta}_n^A \xrightarrow{as} \theta_0^A$ *and* $\hat{\theta}_n^B \xrightarrow{as} \theta_0^B$ *as* $n\to\infty$*. Then we have that*

$$\mathrm{DM}_{m,n} \xrightarrow{d} \mathcal{N}(0,1) \quad as \;\; n,m\to\infty,$$

under the null hypothesis $\mathrm{H}_0 : \mathrm{E}[D_m^{A,B}] = 0$*, where* $\sigma_D^{A,B}$ *is a consistent estimator of the standard deviation of* $D_m^{A,B}$*. If* $\mathrm{E}[D_m^{A,B}] > 0$ *then* $\mathrm{DM}_{m,n}\to\infty$ *as* $n,m\to\infty$*. Finally, if* $\mathrm{E}[D_m^{A,B}] < 0$*, then* $\mathrm{DM}_{m,n}\to-\infty$*.*

[5]For time-series data with fading memory, a burn-in period between the estimation and the validation samples can be the approximate independence between the two samples. Proofs then rely on expanding estimation, burn-in and validation samples.

C Generalized Gamma Distribution

The generalized gamma distribution is a continuous probability distribution and a three-parameter generalization of the two-parameter gamma distribution (Stacy, 1962). It also contains the exponential distribution and the Weibull distribution as special cases. It uses the scale parameter β and two shape parameters θ and φ. The probability density function is

$$p(x|\beta,\theta,\varphi) = \frac{1}{\Gamma(\theta)}\frac{\varphi}{\beta}\left(\frac{x}{\beta}\right)^{\theta\varphi-1} e^{-\left(\frac{x}{\beta}\right)^{\varphi}} \quad \text{for } x \in (0,\infty).$$

The expected value and variance is

$$\begin{aligned}
\mathrm{E}[X] &= \beta\frac{\Gamma\left(\theta+\varphi^{-1}\right)}{\Gamma(\theta)}, \\
\mathrm{var}[X] &= \beta^2\frac{\Gamma\left(\theta+2\varphi^{-1}\right)}{\Gamma(\theta)} - \left(\beta\frac{\Gamma\left(\theta+\varphi^{-1}\right)}{\Gamma(\theta)}\right)^2.
\end{aligned}$$

The score vector is

$$\nabla(x;\beta,\theta,\varphi) = \begin{pmatrix} \varphi\beta^{-1}\left(x^{\varphi}\beta^{-\varphi}-\theta\right) \\ \varphi\ln\left(x\beta^{-1}\right)-\psi_0(\theta) \\ \theta\ln\left(x\beta^{-1}\right)-x^{\varphi}\beta^{-\varphi}\ln\left(x\beta^{-1}\right)+\varphi^{-1} \end{pmatrix} \quad \text{for } x \in (0,\infty).$$

Special cases of the generalized gamma distribution include the gamma distribution for $\varphi = 1$, the Weibull distribution for $\theta = 1$ and the exponential distribution for $\theta = 1$ and $\varphi = 1$.

Clustering of Arrivals in Queueing Systems: Autoregressive Conditional Duration Approach

Petra Tomanová
Prague University of Economics and Business
Winston Churchill Square 1938/4, 130 67 Prague 3, Czechia
petra.tomanova@vse.cz

Vladimír Holý
Prague University of Economics and Business
Winston Churchill Square 1938/4, 130 67 Prague 3, Czechia
vladimir.holy@vse.cz

Abstract: Arrivals in a queueing system are typically assumed to be independent and exponentially distributed. Our analysis of an online bookshop, however, shows that there is an autocorrelation structure. First, we adjust the inter-arrival times for diurnal and seasonal patterns. Second, we model adjusted inter-arrival times by the generalized autoregressive score (GAS) model based on the generalized gamma distribution in the spirit of the autoregressive conditional duration (ACD) models. Third, in a simulation study, we investigate the effects of the dynamic arrival model on the number of customers, the busy period, and the response time in queueing systems with single and multiple servers. We find that ignoring the autocorrelation structure leads to significantly underestimated performance measures and consequently suboptimal decisions. The proposed approach serves as a general methodology for the treatment of arrivals clustering in practice.

Keywords: Inter-Arrival Times, Queueing Theory, Autoregressive Conditional Duration Model, Generalized Autoregressive Score Model, Retail Business.

JEL Classification: C41, C44, M11.

1 Introduction

In various applications of operations research, it is undeniable that the characteristics of a model evolve over time. The parameters of interest can depend on the time of day and season as well as on their past values and other past indicators. In the present paper, we focus on the latter dependency in arrivals to queueing systems from the perspective of the autoregressive conditional duration models with the generalized autoregressive score dynamics.

Many standard queueing systems consider inter-arrival times to be independent, for the sake of analytical tractability. Some studies, however, explicitly consider autocorrelation and model arrivals using the *Markovian arrival process* (MAP) (see, e.g., Adan and Kulkarni, 2003; Buchholz and Kriege, 2017; Manafzadeh Dizbin and Tan, 2019), the *Markov renewal process* (see, e.g., Tin, 1985; Patuwo *et al.*, 1993; Szekli *et al.*, 1994), the *moving average process* (see, e.g., Finch, 1963; Finch and Pearce, 1965; Pearce, 1967) or the *discrete autoregressive process* (see, e.g., Hwang and Sohraby, 2003; Kamoun, 2006; Miao and Lee, 2013). Hwang and Sohraby (2003) argue that time series models with few parameters are more suitable in practice than the MAP models, which might be overparametrized. Simulation studies investigating the autocorrelation in arrivals include Livny *et al.* (1993), Resnick and Samorodnitsky (1997), Altiok and Melamed (2001), Nielsen (2007) and Civelek *et al.* (2009). Overall, these studies show that ignoring the autocorrelation structure in a queueing system, if one is present, leads to biased performance measures.

Arrival processes are also extensively studied in the financial high-frequency literature. In this field, the duration analysis deals with the modeling of the times between successive transactions (trade durations), times until the price reaches a certain level (price durations), and times until a certain volume is traded (volume durations). Typically, the *autoregressive conditional duration*

(ACD) model of Engle and Russell (1998) is used. Its dynamics are analogous to the *generalized autoregressive conditional heteroskedasticity* (GARCH) model of Bollerslev (1986). In its basic version, the ACD model is based on the exponential distribution, but many other distributions are considered in the literature as well. Notably, Lunde (1999) introduces the generalized gamma distribution to the ACD model. Bauwens *et al.* (2004) and Fernandes and Grammig (2005) find that in financial applications, the generalized gamma distribution is more adequate than the exponential, Weibull, and Burr distributions. Hautsch (2003) further finds that the four-parameter generalized F-distribution reduces to the three-parameter generalized gamma distribution in most cases of financial durations. For a survey of financial duration analysis, see Pacurar (2008) and Saranjeet and Ramanathan (2018).

A modern approach to time series modeling is the *general autoregressive score* (GAS) model of Creal *et al.* (2013), also known as the *dynamic conditional score* (DCS) model by Harvey (2013). The GAS model is an observation-driven model providing a general framework for modeling time-varying parameters of any underlying probability distribution. It captures the dynamics of time-varying parameters by the autoregressive term and the score of the conditional density function using the shape of the density function. The theoretical properties of the GAS models together with their estimation by the maximum likelihood method are investigated, e.g., by Blasques *et al.* (2014) and Blasques *et al.* (2018). The empirical performance of the GAS models is studied, e.g., by Koopman *et al.* (2016) and Blazsek and Licht (2020). So far, there have been over 200 papers devoted to numerous models belonging to the GAS family, with various applications, see `www.gasmodel.com` for a comprehensive list.

The class of ACD models and the class of GAS models overlap. In the case of the exponential distribution, the ACD model is equivalent to the GAS model (see Creal *et al.*, 2013). For more complex distributions, however, they tend to differ, as the ACD models are driven by the lagged observation (or, when rewritten, the difference between the observation and the expected value) while the GAS models are driven by the lagged score. In general, the GAS models are very often superior than the alternatives (see, e.g., Blazsek and Villatoro, 2015; Koopman *et al.*, 2016; Chen and Xu, 2019; Gorgi *et al.*, 2019; Harvey *et al.*, 2019; Blazsek and Licht, 2020). Concerning the GAS models for positive or non-negative values that are suitable for the duration analysis, Fonseca and Cribari-Neto (2018) use the Birnbaum–Saunders distribution, Blasques *et al.* (2022) use the zero-inflated negative binomial distribution as well as the generalized gamma distribution, and Harvey and Ito (2020) use the generalized beta distribution as well as the generalized gamma distribution.

In the present paper, we put together three cornerstones – queueing theory, duration analysis, and the GAS models – and demonstrate that they fit together perfectly. The literature has already successfully incorporated GAS models with the duration analysis as discussed above, however, the perspective from queueing theory is our novel contribution. We analyze the inter-arrival times between orders from an online Czech bookshop. First, we adjust the arrivals for diurnal and seasonal patterns, using the cubic spline. Second, we find that the adjusted inter-arrival times exhibit strong clustering behavior: short inter-arrival times are usually followed by short times. To capture this autocorrelation, we use the dynamic model based on the generalized gamma distribution with the GAS dynamics in the spirit of the ACD models. We confirm that the proposed specification is quite suitable for the observed data. Third, we investigate the effects of the proposed arrivals model on queueing systems with single and multiple servers and exponential services. In a simulation study, we show that various performance measures – the number of customers in the system, the busy period of servers, and the response time – have higher mean and variance as well as heavier tails for the proposed dynamic arrivals model than for the standard static model. Lastly, we illustrate how the misspecification of the arrivals model can lead to suboptimal decisions.

The rest of this paper is structured as follows. In Section 2, we present the model based on the generalized gamma distribution with the GAS dynamics for diurnally adjusted inter-arrival times. In Section 3, we show that real data of a retail store exhibit an autocorrelation structure that is well captured by our model. In Section 4, we investigate the impact of the proposed arrivals model on the performance measures using simulations. We conclude the paper in Section 5.

2 Dynamic Model for Arrivals

2.1 Diurnal and Seasonal Adjustment

Before we use the generalized autoregressive score (GAS) model to capture the autoregressive structure of inter-arrival times, we need to deal with diurnal, weekly, and monthly seasonality patterns. To model the non-linear behavior of the diurnal and seasonal patterns and to properly adjust the inter-arrival times, the cubic spline method is used. A *cubic spline* is a piecewise cubic polynomial with continuous derivatives up to order two at each of the K fixed points, called knots, $k = 1, \ldots, K$. Bruce and Bruce (2017) point out that the cubic spline method is often a superior approach to polynomial regression since the polynomial regression often leads to undesirable *wiggliness* in the regression equation.

To take into account the specifics of raw inter-arrival times $\{\tilde{y}_i\}_{i=1}^n$, we define the cubic spline with knots at $\{\xi_k\}_{k=1}^K$ as

$$\log \tilde{y}_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+3} b_{K+3}(x_i) + \gamma t_i + \varepsilon_i, \tag{1}$$

where $\{\beta_j\}_{j=1}^{K+3}$ and γ are parameters to be estimated, ε_i is a disturbance term, t_i is the trend variable, $\{b_j\}_{j=1}^{K+3}$ are the basis functions, and x_i is the time difference in minutes between the time-stamp of the ith observation and the beginning of the week (Monday 00:00) to which the ith observation belongs. Thus, $\{x_i\}_{i=1}^n$ is able to capture both diurnal and intra-week patterns. The basis functions are equal to (i) the variable x_i, $b_1(x_i) = x_i$; (ii) its square, $b_2(x_i) = x_i^2$; (iii) its cube, $b_3(x_i) = x_i^3$; and (iv) truncated power functions, $b_{k+3}(x_i) = \max\left\{0, (x_i - \xi_k)^3\right\}, k = 1, \ldots, K$. The trend variable t_i is linear in time (not linear in observations), $t_1 = 0$ and $t_i = \sum_{j=1}^{i-1} \tilde{y}_j$ for $i = 2, \ldots, n$, to take into account any irregularity in the spacing of the observations. Moreover, the logarithmic transformation of $\tilde{y}$ ensures the non-negativity of the adjusted inter-arrival times. Equidistant intervals are used for identifying the knots, since intervals based on quantiles might lead to too few knots being allocated to off-peak hours.

The regression parameters in (1) are estimated by the *weighted least squares* (WLS) method with weights being the inter-arrival times. The WLS naturally compensates for the possibility that during a particular time interval either a small number of long inter-arrival times or a higher number of shorter inter-arrival times is observed, i.e., the number of observed inter-arrival times within a time interval depends on the values of the inter-arrival times themselves. Unlike ordinary least squares, this approach properly weights the inter-arrival times during hours that exhibit a small median but a huge dispersion. Once the parameters are estimated, the diurnally and seasonally adjusted and detrended inter-arrival times y_i are set to exponentiated residuals from regression (1).

2.2 Generalized Gamma Distribution

Next, we assume that the adjusted inter-arrival times y_i follow the generalized gamma distribution. The *generalized gamma distribution* is a continuous probability distribution for non-negative variables proposed by Stacy (1962). It is a three-parameter generalization of the two-parameter gamma distribution and contains the exponential distribution and the Weibull distribution as special cases. The distribution has the scale parameter α and the shape parameters $\psi > 0$ and $\varphi > 0$. We use the parametrization allowing for arbitrary values of α which is quite suitable for modeling its dynamics. The probability density function is

$$f(y|\alpha, \psi, \varphi) = \frac{1}{\Gamma(\psi)} \frac{\varphi}{e^{\alpha}} \left(\frac{y}{e^{\alpha}}\right)^{\psi\varphi - 1} e^{-\left(\frac{y}{e^{\alpha}}\right)^{\varphi}} \qquad \text{for } y \in (0, \infty), \tag{2}$$

where $\Gamma(\cdot)$ is the gamma function. The expected value and variance is

$$\begin{aligned} \mathrm{E}[Y] &= e^{\alpha} \frac{\Gamma\left(\psi + \varphi^{-1}\right)}{\Gamma(\psi)}, \\ \mathrm{var}[Y] &= e^{2\alpha} \frac{\Gamma\left(\psi + 2\varphi^{-1}\right)}{\Gamma(\psi)} - \left(e^{\alpha} \frac{\Gamma\left(\psi + \varphi^{-1}\right)}{\Gamma(\psi)}\right)^2. \end{aligned} \tag{3}$$

The score for the parameter α is

$$\nabla_\alpha(y,\alpha,\psi,\varphi) = \frac{\partial \log f(y|\alpha,\psi,\varphi)}{\partial\alpha} = \varphi\left(y^\varphi e^{-\varphi\alpha} - \psi\right) \qquad \text{for } y \in (0,\infty). \tag{4}$$

The Fisher information for the parameter α is

$$\mathcal{I}_\alpha(\alpha,\psi,\varphi) = \mathrm{E}\left[\nabla_\alpha(y,\alpha,\psi,\varphi)^2 \middle| \alpha,\psi,\varphi\right] = \psi\varphi^2. \tag{5}$$

Note that the Fisher information for α is not dependent on α itself. Special cases of the generalized gamma distribution include the gamma distribution for $\varphi = 1$, the Weibull distribution for $\psi = 1$, and the exponential distribution for $\psi = 1$ and $\varphi = 1$. The generalized gamma distribution itself is contained in a larger family – the generalized F-distribution with four parameters.

2.3 Generalized Autoregressive Score Dynamics

We now consider the scale parameter to be time-varying. In the *generalized autoregressive score* (GAS) framework of Creal *et al.* (2013), the time-varying parameters are linearly dependent on their lagged values and the lagged values of the score of the conditional density. Typically, only the first lag is used. In our case, the parameter α_i follows the recursion

$$\begin{aligned} \alpha_{i+1} &= c + b\alpha_i + a\nabla_\alpha(y_i,\alpha_i,\psi,\varphi) \\ &= c + b\alpha_i + a\varphi\left(y_i^\varphi e^{-\varphi\alpha_i} - \psi\right), \end{aligned} \tag{6}$$

where c is the constant parameter, b is the autoregressive parameter, a is the score parameter, and $\nabla_\alpha(y_i,\alpha_i,\psi,\varphi)$ is the score defined in (4). In the GAS framework, the score can be scaled by the inverse of the Fisher information or the square of the inverse of the Fisher information. In our case, however, both scaling functions, that based on the Fisher information and the unit scaling, lead to the same model, since the Fisher information does not depend on α_i. The score for a time-varying parameter α_i is the gradient of the log-likelihood with respect to α_i and indicates how sensitive the log-likelihood is to α_i. In the GAS model, the score drives the time variation in α_i based on the shape of the generalized gamma density function.

Let $\theta = (c,b,a,\psi,\varphi)$ denote the vector of parameters in the model. We can estimate θ straightforwardly by the maximum likelihood method. The log-likelihood function is given by

$$\ell(\theta) = \ln f(y_0|\alpha_0,\psi,\varphi) + \sum_{i=1}^{n} \ln f(y_i|\alpha_i,\psi,\varphi), \tag{7}$$

where $f(\cdot)$ is the generalized gamma density function given by (2). We deliberately set aside the first term as the time-varying parameter α_i needs to be initialized at $i = 0$. We set the value of α_0 to the long-term mean value $c/(1-b)$. Subsequent values of α_i, $i = 1,\ldots,n$ then follow recursion (6). The parameter estimates $\hat{\theta}$ are obtained as the answer to the non-linear optimization problem

$$\hat{\theta} \in \max_\theta \ell(\theta). \tag{8}$$

3 Empirical Evidence

3.1 Data Overview and Preparation

The data sample was obtained from the database of an online bookshop with one brick-and-mortar location in Prague, Czechia. The data cover the period from June 8 to December 20, 2018, resulting in 28 full weeks and 5 753 observations. The precision of the timestamp is one minute. Thus, zero inter-arrival times might occur in the data due to two or more orders that arrive within one minute. Since the generalized gamma distribution has strictly positive support, the zero inter-arrival times are set to a small positive number. Bauwens (2006) replaces the zero inter-arrival times with a value equal to one-half of the minimum positive inter-arrival time and argued that this is a more correct approach than discarding them. Hence, all 81 zero inter-arrival times are set to 0.5 minutes.

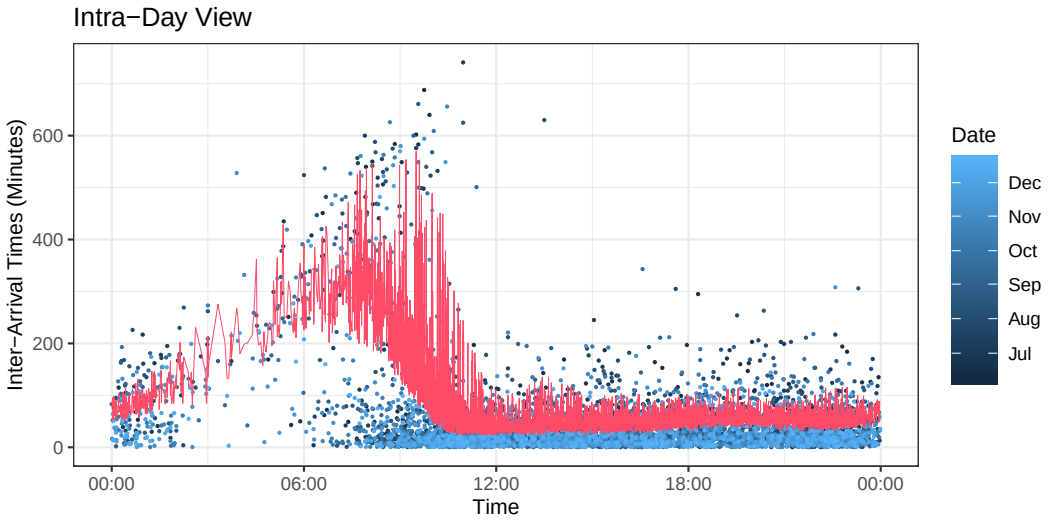

Figure 1: Intra-day view of raw inter-arrival times and their fitted diurnal/seasonal pattern.

3.2 Diurnal and Seasonal Patterns

The median of the raw inter-arrival times is 24 minutes and the mean is 49 minutes – more than double, due to the long inter-arrival times at night (specifically, the hours between midnight and 9 AM, see Figure 1). The hours between 9 and 11 AM exhibit many short inter-arrival times and several very long inter-arrival times, resulting in high dispersion (SD = 111.39). The rest of the rush hours (until 5 PM) shows a similar inter-arrival time median but much lower dispersion (SD = 35.98). Moreover, strong weekly and monthly seasonal patterns are observed. The highest order counts – and consequently lower inter-arrival time values – occur at the beginning of the week and decrease until Saturday, see Figure 2. On Sundays, order counts increase again and exhibit the highest dispersion. During the summer months, the order counts are rather low – resulting in higher inter-arrival times – and linearly increase until December.

To obtain the diurnally and seasonally adjusted and detrended inter-arrival times, the regression equation (1) with a selected number of knots is estimated. In practice, the selection of a suitable number of knots is an empirically-driven task. One must bear in mind that too many knots can result in overfitting (e.g., one knot for every hour results in too unnatural *bumpy* behavior), and, on the other hand, that too few knots can result in an inadequate fit (e.g., one knot for every two hours does not satisfactorily capture the nonlinear behavior of the data). After a little experimenting, we selected one knot for every 90 minutes, which captures all the important nonlinearities and does not produce overfitting. Note that weekly aggregation is used in (1), which results in the same daily seasonal component for Mondays, Tuesdays, etc. To ensure continuity between Sundays and Mondays, the sample is stacked three times consecutively and the adjusted inter-arrival times are computed based on the second sub-sample. Parameters are estimated by the WLS.

The fitted values are shown in Figure 1 and 2. Note that they do not coincide with the smooth cubic spline function due to a linear trend which makes the corresponding fitted line *saw-toothed*. The diurnally and seasonally adjusted and detrended inter-arrival times are computed as the exponentiated residuals of estimated equation (1) and for convenience, they are standardized to have unit mean. Their values range from 0.002 to 11.23 minutes.

3.3 Fit of the Dynamic Model

Even after the seasonal and diurnal adjustment, the inter-arrival times still tend to cluster over time – long (short) inter-arrival times are likely to be followed by long (short) inter-arrival times. This dependence is not particularly strong, but nevertheless it is statistically significant, as illustrated in

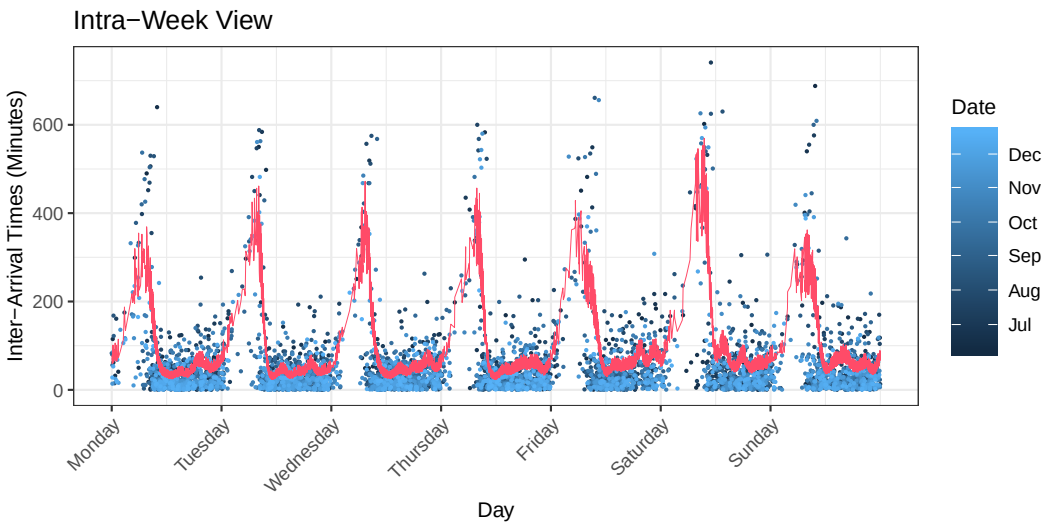

Figure 2: Intra-week view of raw inter-arrival times and their fitted diurnal/seasonal pattern.

Model		Estimate					Model Fit	
Spec.	Dist.	c	b	a	ψ	φ	Lik.	AIC
Static	Exp.	0.00	0.00	0.00	1.00	1.00	−5 753.00	11 508.00
Static	Weibull	-0.01	0.00	0.00	1.00	0.97	−5 748.93	11 501.86
Static	Gamma	0.04	0.00	0.00	0.96	1.00	−5 749.77	11 503.54
Static	G. G.	-0.12	0.00	0.00	1.08	0.93	−5 748.37	11 502.75
Dyn.	Exp.	0.00	0.76	0.06	1.00	1.00	−5 728.28	11 462.56
Dyn.	Weibull	0.00	0.75	0.06	1.00	0.97	−5 724.89	11 457.79
Dyn.	Gamma	0.01	0.76	0.06	0.97	1.00	−5 725.97	11 459.95
Dyn.	G. G.	-0.06	0.72	0.07	1.15	0.90	−5 723.31	11 456.62

Table 1: Parameter estimates of the inter-arrival time models with the log-likelihood value (Lik.) and the Akaike information criterion (AIC).

Figure 3. To capture the autocorrelation, we use the dynamic model based on the generalized gamma distribution with the GAS dynamics in (6). The parameters are estimated by the maximum likelihood method determined by the non-linear optimization problem in (8) and the log-likelihood function in (7). For comparison, we also present the results for static and dynamic models based on special cases of the generalized gamma distribution (G.G.), namely, the exponential (Exp.), Weibull, and gamma distributions.

Parameter estimates and the performance evaluation in terms of the *Akaike information criterion* (AIC) of both static and dynamic inter-arrival time models are shown in Table 1. The AIC values are at least 43.59 lower for the dynamic models than for their static counterparts. However, the differences between the dynamic models are not so striking: the highest difference is between the exponential and generalized gamma distributions (by 5.94). The best performing model is the most general one, the dynamic GAS model using the generalized gamma distribution. The dynamic models based on either the exponential or generalized gamma distributions in comparison with their static counterparts are further analyzed in the simulation study of queueing systems.

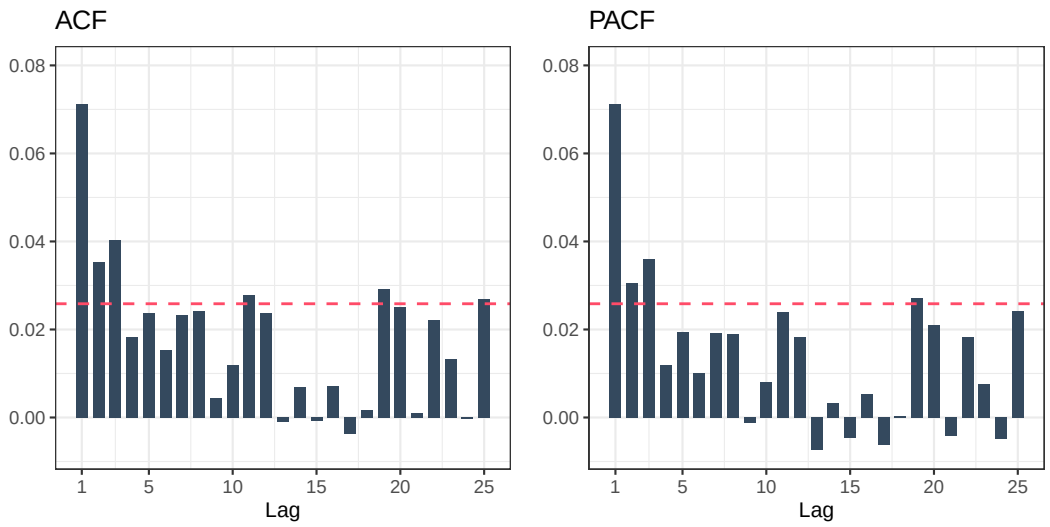

Figure 3: The autocorrelation function (ACF) and the partial autocorrelation function (PACF) of adjusted inter-arrival times. Red dashed lines indicate 5% confidence bounds.

4 Impact on Queueing Systems

4.1 System with Single Server

Using simulations, we now investigate the effects of various arrival models on performance measures in queueing systems. We consider models based on the exponential and generalized gamma distributions with the static and dynamic specifications. The coefficients of the models are taken from Table 1. In all models, the rate of arrivals is $\lambda = 1$ job per minute. First, we focus on the queueing system with a single server only. We consider the service times to be independent and exponentially distributed with the rate μ ranging from 1.1 to 1.5 jobs per minute. We simulate the arrival and service processes and measure the number of customers in the system, the busy period of the server, and the response time. The number of simulation runs is equal to 10^9, which seems to be sufficient for the reported precision of one decimal place as the results are in line with the theoretical performance measures for the static exponential scenario as well as Little's law for all scenarios.

The results are presented in Table 2. For all values of μ, the systems based on the generalized gamma distribution have higher values of performance measures than the systems based on the exponential distribution in terms of the mean, standard deviation, and 95 percent quantile. Similarly, systems with the dynamic specification have higher values of performance measures than the systems with the static specification. The left plot of Figure 4 shows how the probability mass function of the number of customers differs for the static and dynamic models. The dynamic model has a higher probability of an empty system as there is a tendency to have longer periods of low activity. It has also higher probabilities of large numbers of customers in the system, as arrivals tend to cluster. The right plot of Figure 4 shows how the density functions of the response times for the static and dynamic models differ. In the dynamic model, customers simply have to wait longer. The differences between the static and dynamic models are naturally weaker for larger μ.

These results carry a warning for practice. When the standard M/M/1 system is assumed but the arrivals actually follow the GAS model based on the generalized gamma distribution, the performance measures are significantly underestimated. For example, the mean number of customers and the mean response time are 22 percent lower than the actual value for $\mu = 1.1$ jobs per minute. It is therefore crucial to correctly specify the model for arrivals.

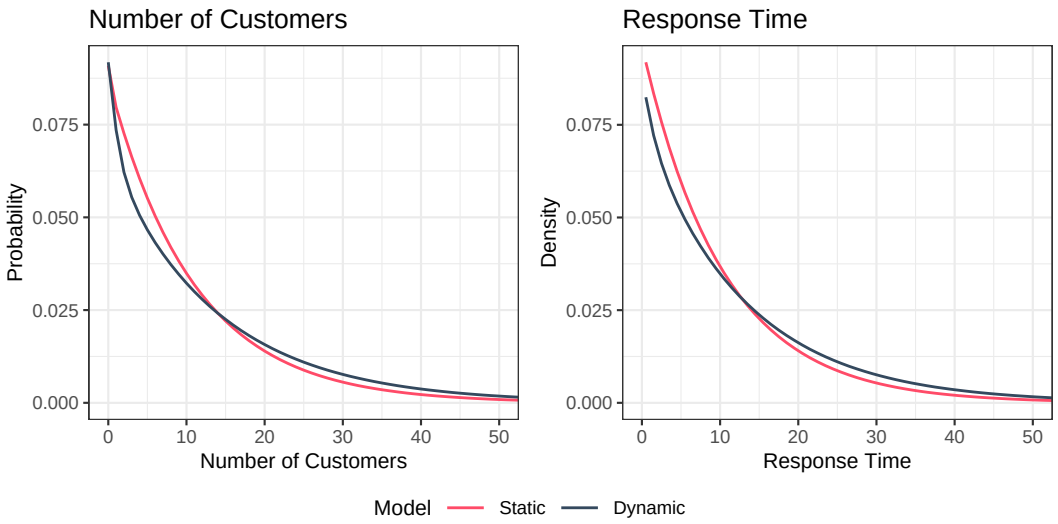

Figure 4: The probability mass functions of the number of customers in the system and density functions of the response time for the static and dynamic arrival models based on the generalized gamma distribution in a queueing system with single server and $\mu = 1.1$ jobs per minute.

Queueing System			No. of Customers			Busy Period			Response Time		
μ	Spec.	Dist.	M	SD	95%	M	SD	95%	M	SD	95%
1.1	Static	Exp.	10.0	10.5	31.0	10.0	45.8	39.8	10.0	10.0	30.0
1.1	Static	G. G.	10.4	10.9	32.0	10.4	47.6	41.4	10.4	10.4	31.1
1.1	Dyn.	Exp.	12.4	13.4	39.0	10.8	54.4	41.2	12.4	12.6	37.6
1.1	Dyn.	G. G.	12.8	13.8	41.0	11.2	56.1	43.0	12.8	13.1	39.0
1.2	Static	Exp.	5.0	5.5	16.0	5.0	16.6	22.1	5.0	5.0	15.0
1.2	Static	G. G.	5.2	5.7	17.0	5.2	17.2	22.9	5.2	5.2	15.5
1.2	Dyn.	Exp.	6.0	6.8	20.0	5.4	19.5	23.3	6.0	6.1	18.3
1.2	Dyn.	G. G.	6.2	7.1	20.0	5.6	20.2	24.4	6.2	6.4	19.0
1.3	Static	Exp.	3.3	3.8	11.0	3.3	9.2	14.8	3.3	3.3	10.0
1.3	Static	G. G.	3.4	3.9	11.0	3.4	9.6	15.3	3.4	3.4	10.3
1.3	Dyn.	Exp.	3.9	4.6	13.0	3.5	10.7	15.7	3.9	4.0	11.9
1.3	Dyn.	G. G.	4.0	4.8	14.0	3.7	11.2	16.4	4.0	4.2	12.4
1.4	Static	Exp.	2.5	3.0	8.0	2.5	6.1	11.0	2.5	2.5	7.5
1.4	Static	G. G.	2.6	3.1	9.0	2.6	6.3	11.3	2.6	2.6	7.7
1.4	Dyn.	Exp.	2.8	3.5	10.0	2.6	7.0	11.5	2.8	2.9	8.7
1.4	Dyn.	G. G.	3.0	3.7	10.0	2.7	7.3	12.1	3.0	3.1	9.1
1.5	Static	Exp.	2.0	2.4	7.0	2.0	4.5	8.6	2.0	2.0	6.0
1.5	Static	G. G.	2.1	2.5	7.0	2.1	4.6	8.9	2.1	2.1	6.2
1.5	Dyn.	Exp.	2.2	2.9	8.0	2.1	5.1	9.0	2.2	2.3	6.8
1.5	Dyn.	G. G.	2.3	3.0	8.0	2.2	5.3	9.4	2.3	2.4	7.1

Table 2: Mean values (M), standard deviations (SD) and 95%-quantiles (95%) of the number of customers in the system, the busy period of the server and the response time in various queueing systems with a single server.

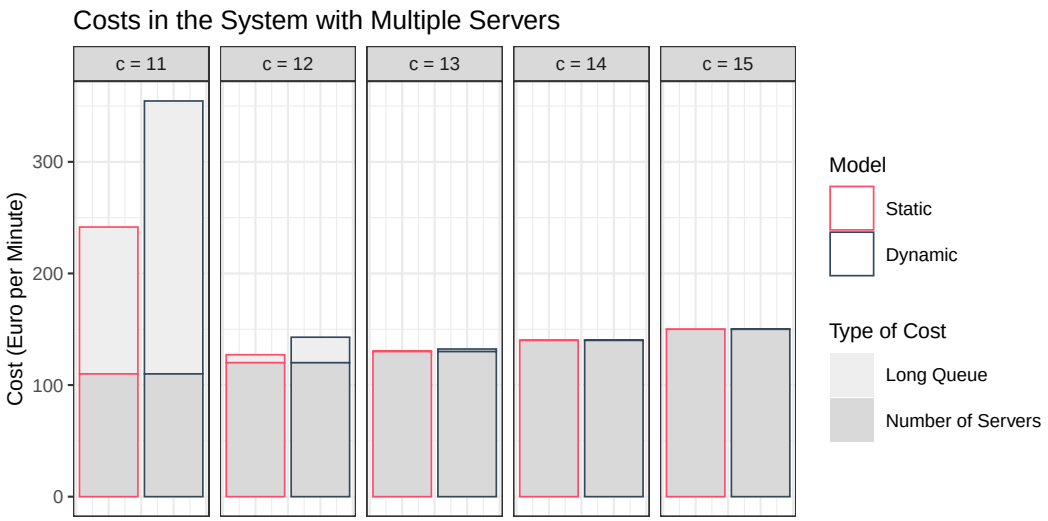

Figure 5: Costs related to the number of servers and long queues for the static and dynamic arrival models based on the generalized gamma distribution in queueing systems with multiple servers and $\mu = 0.10$ jobs per minute.

4.2 System with Multiple Servers

Next, we consider queueing systems with multiple servers. We base the simulations on the same setting as in the previous section. The only difference lies in the service structure. We let the number of servers c range from 11 to 15 and take the individual service rate to be $\mu = 0.1$ jobs per minute. Such values result in the same server utilizations $\rho = \lambda/(c\mu)$ as in the previous section. Again, we measure the number of customers in the system, the busy period of the servers, and the response time. By the busy period, we mean the full busy period, i.e., the duration of the state in which all servers are busy.

The results are presented in Table 3. They are very similar to those for a system with a single server: the generalized gamma distribution and the dynamic specification increase all performance measures. When incorrectly assuming an M/M/c system, the specification error is distinct but not as high as in the case of a single server. For example, when assuming an M/M/11 system, the mean number of customers and the mean response time are 14 percent lower than the actual value for arrivals based on the generalized gamma distribution with the dynamic specification.

In the following toy example, we illustrate how the misspecification of the arrival model can affect decision making. Let us assume that there are two types of costs associated with the operation of the system: the cost of running one server per unit of time $C_1 = 10$ euro per minute, and the cost of having a queue longer than 30 customers per unit of time $C_2 = 3\,000$ euro per minute. The analytic department is faced with the question of how many servers to operate. The composition of costs for different numbers of servers is shown in Figure 5. The optimal number of servers according to the static model is 12 while it is 13 for the dynamic model. An analyst employing the static model believes that the total optimal costs are 127.13 euro per minute while they actually are 142.87 euro per minute for the suboptimal choice of 12 servers. An analyst correctly specifying the dynamic model finds that the lowest possible costs are 132.32 euro per minute for the optimal choice of 13 servers. The decision based on the misspecified arrival model therefore results in a total cost increase of 8 percent.

4.3 Discussion of More Complex Systems

We have focused on rather simple queueing systems in order to get transparent results. The M/M/1 system is as straightforward as can be, and therefore the best choice for an illustration of the impact

Queueing System			No. of Customers			Busy Period			Response Time		
c	Spec.	Dist.	M	SD	95%	M	SD	95%	M	SD	95%
11	Static	Exp.	16.8	10.7	38.0	10.0	45.8	39.9	16.8	13.8	43.8
11	Static	G. G.	17.2	11.1	39.0	10.4	47.6	41.4	17.2	14.0	44.6
11	Dyn.	Exp.	19.1	13.5	46.0	12.4	58.6	49.7	19.1	15.7	49.9
11	Dyn.	G. G.	19.5	14.0	47.0	12.8	60.3	51.8	19.5	16.1	50.9
12	Static	Exp.	12.2	5.8	24.0	5.0	16.6	22.1	12.2	10.8	33.6
12	Static	G. G.	12.4	6.1	24.0	5.2	17.2	22.9	12.4	10.9	33.8
12	Dyn.	Exp.	13.1	7.2	27.0	6.1	21.1	27.5	13.1	11.3	35.4
12	Dyn.	G. G.	13.3	7.5	28.0	6.3	21.9	28.7	13.3	11.5	35.8
13	Static	Exp.	11.0	4.4	19.0	3.3	9.2	14.8	11.0	10.3	31.3
13	Static	G. G.	11.0	4.5	19.0	3.4	9.6	15.3	11.0	10.3	31.4
13	Dyn.	Exp.	11.4	5.2	21.0	4.0	11.7	18.4	11.4	10.5	32.0
13	Dyn.	G. G.	11.5	5.4	22.0	4.2	12.1	19.1	11.5	10.5	32.2
14	Static	Exp.	10.4	3.8	17.0	2.5	6.1	11.0	10.4	10.1	30.5
14	Static	G. G.	10.5	3.9	17.0	2.6	6.3	11.3	10.5	10.1	30.6
14	Dyn.	Exp.	10.7	4.4	19.0	3.0	7.7	13.5	10.7	10.2	30.9
14	Dyn.	G. G.	10.7	4.5	19.0	3.1	8.0	14.0	10.7	10.2	30.9
15	Static	Exp.	10.2	3.5	16.0	2.0	4.5	8.6	10.2	10.0	30.2
15	Static	G. G.	10.2	3.6	16.0	2.1	4.6	8.9	10.2	10.0	30.2
15	Dyn.	Exp.	10.3	3.9	17.0	2.3	5.6	10.5	10.3	10.1	30.4
15	Dyn.	G. G.	10.3	4.1	18.0	2.4	5.8	10.9	10.3	10.1	30.4

Table 3: Mean values (M), standard deviations (SD) and 95%-quantiles (95%) of the number of customers in the system, the full busy period of servers and the response time in various queueing systems with multiple servers and $\mu = 0.1$ jobs per minute.

of autocorrelated arrivals. The M/M/c system is used as a robustness check to show that the behavior observed for the M/M/1 system is present even for different specifications. As for the toy example of decision making in the M/M/c system, it is meant just as a simplistic illustration revealing a potential source of suboptimal decisions.

On the other hand, Tomanová (2018), Tomanová (2019b), and Tomanová (2019a) explore a much more realistic and complex queueing system specific to this case of an online bookshop. As this queueing system is tailored just for this specific application and cannot be easily transferred to others, we only summarize the main findings. Tomanová (2018) performs a process quality assessment based on process simulation and reports that the key quality target is not satisfied in almost twice as many cases when the dynamic model is considered (the target is not satisfied in 6.16 percent of them) than when the static model is considered (for which the target is not satisfied 3.23 percent of the time). The common approach – a static model which assumes that times between arrivals follow the exponential distribution with a constant rate – underestimates the probability of extreme values and thus significantly skews the basis for process quality assessment and leads to suboptimal decisions. Tomanová (2019b) also demonstrates that the clustering of arrivals increases the probability of weeks with an extreme number of arrivals, something which has a negative effect on the fulfillment of targets. Tomanová (2019a) further extends that work to making final recommendations for the management of the online bookshop. The main finding is that 21 percent of the orders are not satisfied within a working day due to insufficiently allocated resources for the first stage (pre-processing of arrivals).

5 Conclusion

We have analyzed the dependence of inter-arrival times in queueing systems and demonstrated the negative effect of misspecifying the arrival model on decision making. To capture the autocorrelation structure of the inter-arrival times, we have proposed using a dynamic model based on the generalized gamma distribution with the GAS dynamics. We have found that this approach is superior to the standard model that uses the exponential distribution with a constant rate, since it leads to a more faithful representation of the mean and extreme values of the arrival process. Our study has carried out three steps.

1. We have constructed a suitable model for capturing the diurnal and seasonal dependencies which takes into account a specific time-structure of inter-arrival times. It uses a cubic spline approach and estimates the parameters by the weighted ordinary least square method to properly adjust inter-arrival times during hours that exhibit a small median but a huge dispersion.

2. We have found that the GAS models based on the generalized gamma distribution and its special cases fit the data better than do their static counterparts. This is due to the fact that the static models ignore the autocorrelation structure, which is still present even after the proper diurnal and seasonal adjustments.

3. We have compared both static and dynamic models in a simulation study of queueing systems with single and multiple servers and exponential services. We have shown that ignoring the autocorrelation structure leads to biased performance measures. The number of customers in the system, the busy periods of the servers, and the response times, have higher means and variances as well as heavier tails for the proposed dynamic arrivals model than for the standard static model. We have also shown how a trust in the standard static model for inter-arrival times leads to suboptimal decisions and consequently to a loss of profits.

A proper treatment of arrival dependence is of great importance since its ignorance generates extra costs. Our approach is useful for process simulations and consequently for process optimization and process quality assessment.

The main limitation of this paper and a topic for future research is the theoretical treatment of queueing systems with inter-arrival times following the GAS model. In the paper, we have resorted to simulations to determine the moments, quantiles, and density functions of the performance

measures. Theoretical derivations of these quantities and functions is undoubtedly challenging but perhaps possible in some cases. Another topic for future research, which would be easier to achieve, is the use of the proposed approach in other applications. Besides retail order processing, these may include customer service, project management, manufacturing engineering, emergency services, logistics, transportation, telecommunications, computing, and others.

Acknowledgements

We would like to thank the organizers and participants of the 7th International Conference on Management (Nový Smokovec, September 26–29, 2018), the 30th European Conference on Operational Research (Dublin, June 23–26, 2019), the 15th International Symposium on Operations Research in Slovenia (Bled, September 25–27, 2019) and the 3rd International Conference on Advances in Business and Law (Dubai, November 23–24, 2019) for fruitful discussions.

Funding

The work on this paper was supported by the Internal Grant Agency of the Prague University of Economics and Business under project F4/27/2020, the Czech Science Foundation under project 19-08985S, and the Institutional Support Funds for the long-term conceptual development of the Faculty of Informatics, Prague University of Economics and Business.

References

Adan IJBF, Kulkarni VG (2003). "Single-Server Queue with Markov-Dependent Inter-Arrival and Service Times." *Queueing Systems*, **45**(2), 113–134. ISSN 0257-0130. https://doi.org/10.1023/a:1026093622185.

Altiok T, Melamed B (2001). "The Case for Modeling Correlation in Manufacturing Systems." *IIE Transactions*, **33**(9), 779–791. ISSN 0740-817X. https://doi.org/10.1080/07408170108936872.

Bauwens L (2006). "Econometric Analysis of Intra-Daily Trading Activity on the Tokyo Stock Exchange." *Monetary and Economic Studies*, **24**(1), 1–24. ISSN 0288-8432. http://www.imes.boj.or.jp/research/abstracts/english/me24-1-1.html.

Bauwens L, Giot P, Grammig J, Veredas D (2004). "A Comparison of Financial Duration Models via Density Forecasts." *International Journal of Forecasting*, **20**(4), 589–609. ISSN 0169-2070. https://doi.org/10.1016/j.ijforecast.2003.09.014.

Blasques F, Gorgi P, Koopman SJ, Wintenberger O (2018). "Feasible Invertibility Conditions and Maximum Likelihood Estimation for Observation-Driven Models." *Electronic Journal of Statistics*, **12**(1), 1019–1052. ISSN 1935-7524. https://doi.org/10.1214/18-ejs1416.

Blasques F, Holý V, Tomanová P (2022). "Zero-Inflated Autoregressive Conditional Duration Model for Discrete Trade Durations with Excessive Zeros." https://arxiv.org/abs/1812.07318.

Blasques F, Koopman SJ, Lucas A (2014). "Stationarity and Ergodicity of Univariate Generalized Autoregressive Score Processes." *Electronic Journal of Statistics*, **8**(1), 1088–1112. ISSN 1935-7524. https://doi.org/10.1214/14-ejs924.

Blazsek S, Licht A (2020). "Dynamic Conditional Score Models: A Review of Their Applications." *Applied Economics*, **52**(11), 1181–1199. ISSN 0003-6846. https://doi.org/10.1080/00036846.2019.1659498.

Blazsek S, Villatoro M (2015). "Is Beta-t-EGARCH(1,1) Superior to GARCH(1,1)?" *Applied Economics*, **47**(17), 1764–1774. ISSN 0003-6846. https://doi.org/10.1080/00036846.2014.1000536.

Bollerslev T (1986). "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics*, **31**(3), 307–327. ISSN 0304-4076. https://doi.org/10.1016/0304-4076(86)90063-1.

Bruce P, Bruce A (2017). *Practical Statistics for Data Scientists: 50 Essential Concepts*. O'Reilly Media, Sebastopol. ISBN 978-1-4919-5295-5. https://www.oreilly.com/library/view/practical-statistics-for/9781491952955/.

Buchholz P, Kriege J (2017). "Fitting Correlated Arrival and Service Times and Related Queueing Performance." *Queueing Systems*, **85**(3-4), 337–359. ISSN 0257-0130. https://doi.org/10.1007/s11134-017-9514-5.

Chen R, Xu J (2019). "Forecasting Volatility and Correlation Between Oil and Gold Prices Using a Novel Multivariate GAS Model." *Energy Economics*, **78**, 379–391. ISSN 0140-9883. https://doi.org/10.1016/j.eneco.2018.11.011.

Civelek I, Biller B, Scheller-Wolf A (2009). "The Impact of Dependence on Queueing Systems." https://www.researchgate.net/publication/228814043.

Creal D, Koopman SJ, Lucas A (2013). "Generalized Autoregressive Score Models with Applications." *Journal of Applied Econometrics*, **28**(5), 777–795. ISSN 0883-7252. https://doi.org/10.1002/jae.1279.

Engle RF, Russell JR (1998). "Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data." *Econometrica*, **66**(5), 1127–1162. ISSN 0012-9682. https://doi.org/10.2307/2999632.

Fernandes M, Grammig J (2005). "Nonparametric Specification Tests for Conditional Duration Models." *Journal of Econometrics*, **127**(1), 35–68. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2004.06.003.

Finch PD (1963). "The Single Server Queueing System with Non-Recurrent Input-Process and Erlang Service Time." *Journal of the Australian Mathematical Society*, **3**(2), 220–236. ISSN 1446-8107. https://doi.org/10.1017/s1446788700027968.

Finch PD, Pearce C (1965). "A Second Look at a Queueing System with Moving Average Input Process." *Journal of the Australian Mathematical Society*, **5**(1), 100–106. ISSN 1446-8107. https://doi.org/10.1017/s144678870002591x.

Fonseca RV, Cribari-Neto F (2018). "Bimodal Birnbaum-Saunders Generalized Autoregressive Score Model." *Journal of Applied Statistics*, **45**(14), 2585–2606. ISSN 0266-4763. https://doi.org/10.1080/02664763.2018.1428734.

Gorgi P, Koopman SJ, Lit R (2019). "The Analysis and Forecasting of Tennis Matches by Using a High Dimensional Dynamic Model." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **182**(4), 1393–1409. ISSN 0964-1998. https://doi.org/10.1111/rssa.12464.

Harvey A, Hurn S, Thiele S (2019). "Modeling Directional (Circular) Time Series." https://doi.org/10.17863/cam.43915.

Harvey AC (2013). *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series*. First Edition. Cambridge University Press, New York. ISBN 978-1-107-63002-4. https://doi.org/10.1017/cbo9781139540933.

Harvey AC, Ito R (2020). "Modeling Time Series When Some Observations Are Zero." *Journal of Econometrics*, **214**(1), 33–45. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2019.05.003.

Hautsch N (2003). "Assessing the Risk of Liquidity Suppliers on the Basis of Excess Demand Intensities." *Journal of Financial Econometrics*, **1**(2), 189–215. ISSN 1479-8409. https://doi.org/10.1093/jjfinec/nbg010.

Hwang GU, Sohraby K (2003). "On the Exact Analysis of a Discrete-Time Queueing System with Autoregressive Inputs." *Queueing Systems*, **43**(1-2), 29–41. ISSN 0257-0130. https://doi.org/10.1023/a:1021848330183.

Kamoun F (2006). "The Discrete-Time Queue with Autoregressive Inputs Revisited." *Queueing Systems*, **54**(3), 185–192. ISSN 0257-0130. https://doi.org/10.1007/s11134-006-9591-3.

Koopman SJ, Lucas A, Scharth M (2016). "Predicting Time-Varying Parameters with Parameter-Driven and Observation-Driven Models." *Review of Economics and Statistics*, **98**(1), 97–110. ISSN 0034-6535. https://doi.org/10.1162/rest_a_00533.

Livny M, Melamed B, Tsiolis AK (1993). "The Impact of Autocorrelation on Queuing Systems." *Management Science*, **39**(3), 322–339. ISSN 0025-1909. https://doi.org/10.2307/2632647.

Lunde A (1999). "A Generalized Gamma Autoregressive Conditional Duration Model." https://www.researchgate.net/publication/228464216.

Manafzadeh Dizbin N, Tan B (2019). "Modelling and Analysis of the Impact of Correlated Inter-Event Data on Production Control Using Markovian Arrival Processes." *Flexible Services and Manufacturing Journal*, **31**(4), 1042–1076. ISSN 1936-6582. https://doi.org/10.1007/s10696-018-9329-7.

Miao DWC, Lee HC (2013). "Second-Order Performance Analysis of Discrete-Time Queues Fed by DAR(2) Sources with a Focus on the Marginal Effect of the Additional Traffic Parameter." *Applied Stochastic Models in Business and Industry*, **29**(1), 45–60. ISSN 1524-1904. https://doi.org/10.1002/asmb.939.

Nielsen EH (2007). "Autocorrelation in Queuing Network-Type Production Systems - Revisited." *International Journal of Production Economics*, **110**(1-2), 138–146. ISSN 0925-5273. https://doi.org/10.1016/j.ijpe.2007.02.014.

Pacurar M (2008). "Autoregressive Conditional Duration Models in Finance: A Survey of the Theoretical and Empirical Literature." *Journal of Economic Surveys*, **22**(4), 711–751. ISSN 0950-0804. https://doi.org/10.1111/j.1467-6419.2007.00547.x.

Patuwo BE, Disney RL, McNickle DC (1993). "The Effect of Correlated Arrivals on Queues." *IIE Transactions*, **25**(3), 105–110. ISSN 0740-817X. https://doi.org/10.1080/07408179308964296.

Pearce C (1967). "An Imbedded Chain Approach to a Queue with Moving Average Input." *Operations Research*, **15**(6), 1117–1130. ISSN 0030-364X. https://doi.org/10.1287/opre.15.6.1117.

Resnick S, Samorodnitsky G (1997). "Performance Decay in a Single Server Exponential Queueing Model with Long Range Dependence." *Operations Research*, **45**(2), 235–243. ISSN 0030-364X. https://doi.org/10.1287/opre.45.2.235.

Saranjeet KB, Ramanathan TV (2018). "Conditional Duration Models for High-Frequency Data: A Review on Recent Developments." *Journal of Economic Surveys*, **33**(1), 252–273. ISSN 0950-0804. https://doi.org/10.1111/joes.12261.

Stacy EW (1962). "A Generalization of the Gamma Distribution." *The Annals of Mathematical Statistics*, **33**(3), 1187–1192. ISSN 0003-4851. https://doi.org/10.2307/2237889.

Szekli R, Disney RL, Hur S (1994). "MR/GI/1 Queues by Positively Correlated Arrival Stream." *Journal of Applied Probability*, **31**(2), 497–514. ISSN 0021-9002. https://doi.org/10.1017/s0021900200045009.

Tin P (1985). "A Queueing System with Markov-Dependent Arrivals." *Journal of Applied Probability*, **22**(3), 668–677. ISSN 0021-9002. https://doi.org/10.1017/s0021900200029417.

Tomanová P (2018). "Measuring Intensity of Order Arrivals and Process Quality Assessment of an Online Bookshop: A Case Study from the Czech Republic." In M Frankovský, J Dobrovič, R Fedorko (Eds.), *Proceedings of the 7th International Conference on Management*, 768–773. Bookman s.r.o., Prešov. ISBN 978-80-8165-301-8. http://www.managerconf.com/.

Tomanová P (2019a). "Business Process Simulation and Process Quality Assessment of Czech Online Bookshop." In *Proceedings of the 3rd International Conference on Advances in Business and Law*, 1–4. Dubai Business School, Dubai. http://publications.ud.ac.ae/index.php/ICABML-CP/article/view/403/127.

Tomanová P (2019b). "Clustering of Arrivals and Its Impact on Process Simulation." In *Proceedings of the 15th International Symposium on Operations Research in Slovenia*, 314–319. Slovenian Society Informatika, Bled. ISBN 978-961-6165-55-6. http://fgg-web.fgg.uni-lj.si/{~}/sdrobne/sor/SOR'19-Proceedings.pdf.