Modeling Inter-Event Durations in High-Frequency Financial Transactions Data via Estimating Functions

Nalini Ravishanker
Dept. of Statistics, Univ. of Connecticut, Storrs
nalini.ravishanker@uconn.edu
www.stat.uconn.edu/~nalini

Joint work with
Yaohua Zhang (UConn), Jian Zou (WPI),
A. Thavaneswaran (U. Manitoba)

Outline

- Introduction
- Estimating Function (EF) Approach for Time Series
- Practical Considerations in using the EF Approach
- Applications to High-Frequency Financial Data
- Work in Progress

Introduction

The EF framework enables modeling linear or nonlinear time series allows efficient estimation under minimal distributional assumptions.
Godambe Biometrika 1985; Thavaneswaran and Abraham JTSA 1988

Basic idea: Construct suitable unbiased martingale Estimating Functions (EFs) and solve the resulting Estimating Equations (EEs) to get optimal parameter estimates.

Recursive formulas (over time) can enable online estimation of parameters.

This talk describes modeling inter-event durations in the EF framework, and about making the EF implementation user-friendly.

Durations Between Events

Let τ_t be the time until the t^{th} event, τ_0 being the starting time.

The t^{th} duration is the time interval between two consecutive occurrences of an event:

$$x_t = \tau_t - \tau_{t-1}, \quad t = 1, 2, \cdots$$

For each event (positive integer) t, x_t is a positive-valued random variable.

Inter-event Durations in Financial Transactions Level Data

High-frequency transaction level stock prices data for several years from the Trade and Quotes (TAQ) database at Wharton Research Data Services (WRDS).

For trading days in June 2013, the data set consists of around four million observations.

We selected 3 stocks based on liquidity behavior: BAC (high), IBM (medium), and 3M (low).

We considered transactions between 9:30 AM to 4:00 PM.

IBM Raw Transaction Data- a few rows

```
41438159 20130603 9:24:04.4 100 208.41
41438160 20130603 9:24:22.3 100 208.4
41438161 20130603 9:24:23.5 100 208.4
41438162 20130603 9:29:45.3 100 208.4
41438163 20130603 9:29:45.3 100 208.41
41438164 20130603 9:30:00.0 100 208.4
41438165 20130603 9:30:00.1 100 208.4
41438166 20130603 9:30:00.2 200 208.4
41438167 20130603 9:30:00.2 900 208.4
41438168 20130603 9:30:04.0 100 208.25
```

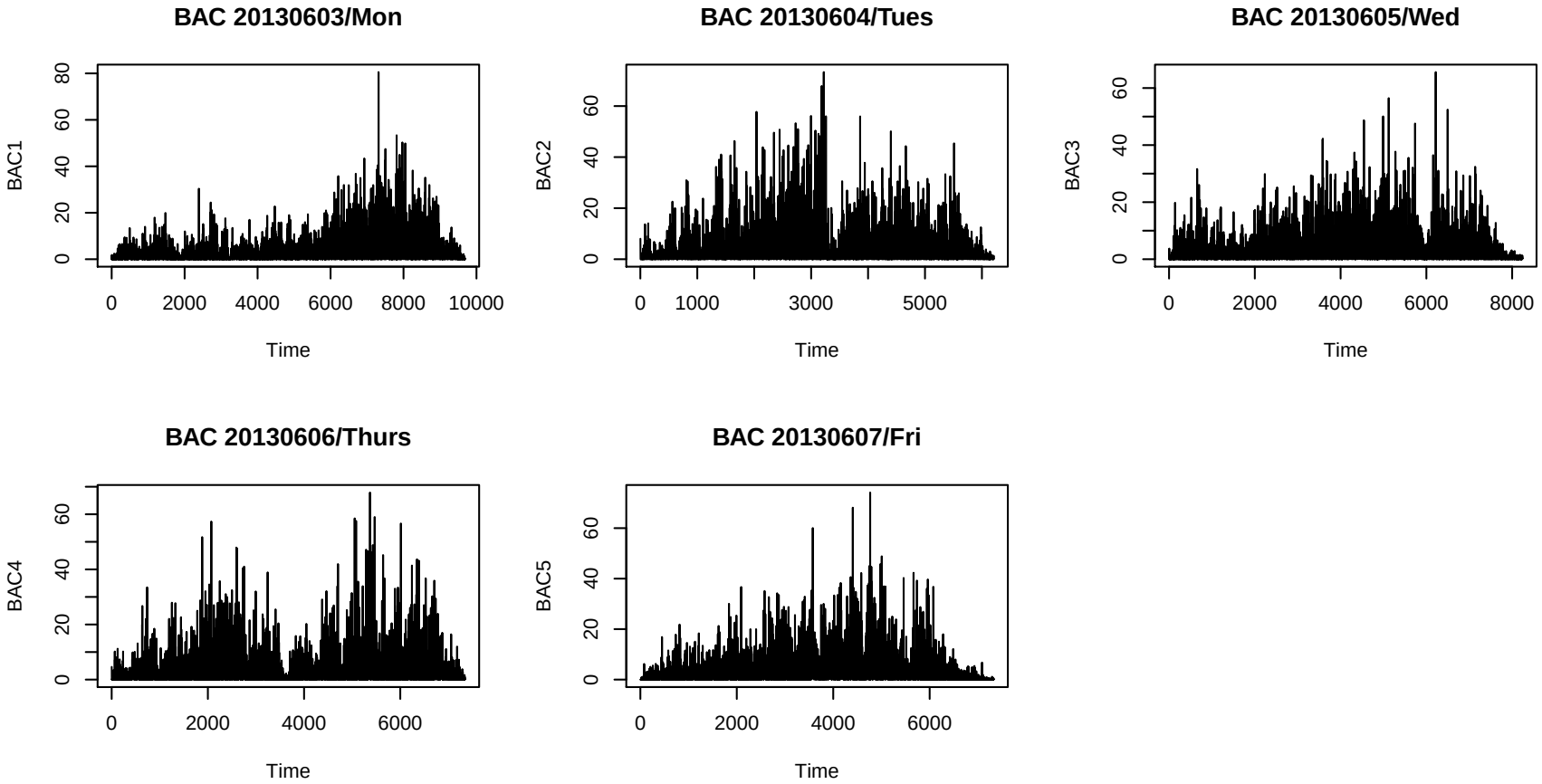
Event Definition

A practitioner may define an event, based on a certain price change, or a certain volume jump, etc., that directs his/her decision making.

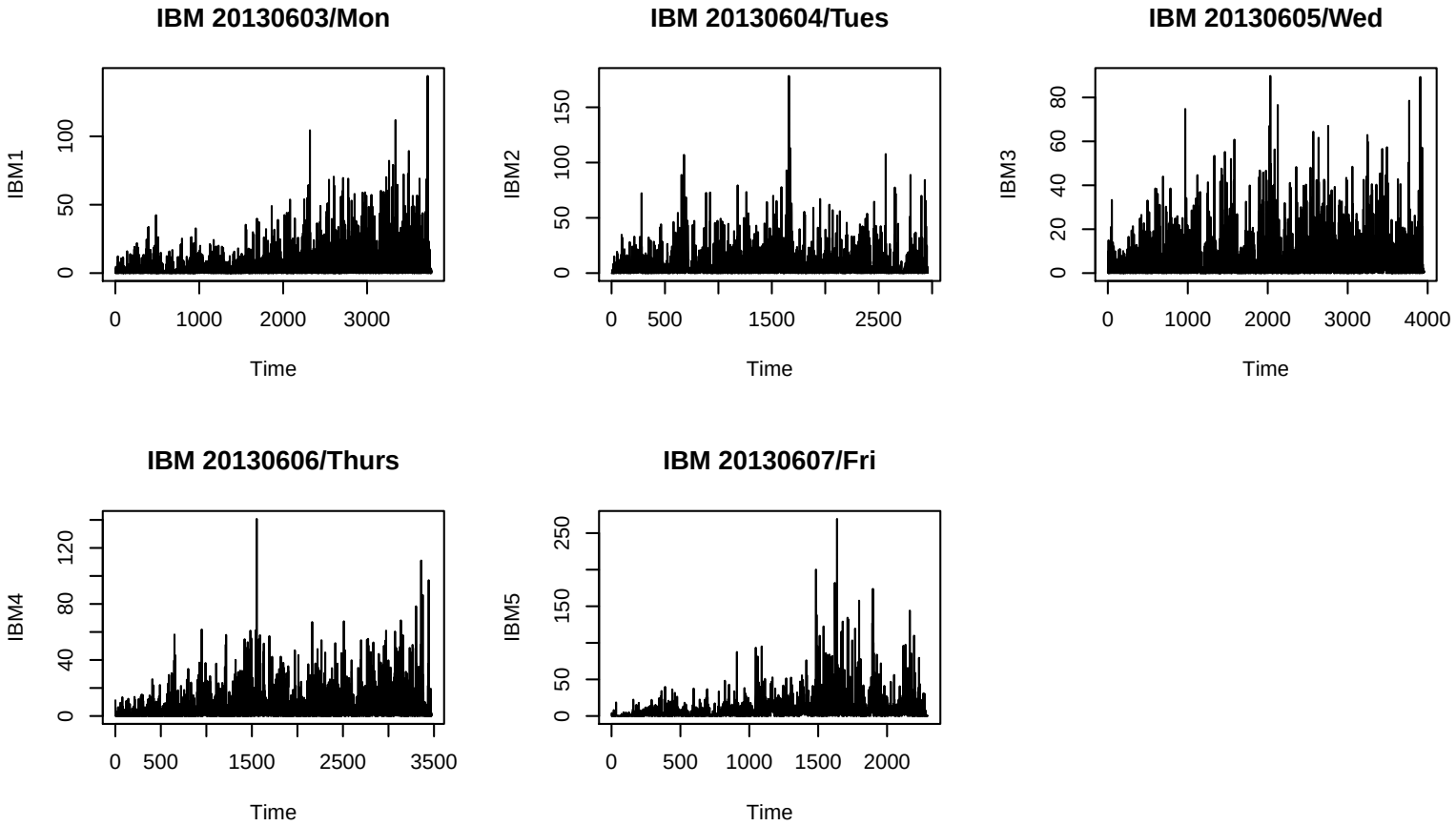
Each event will lead to a different set of durations obtained from the raw transaction-level data.

For our analysis, one event is based on a certain percent δ change over the open price of an asset.

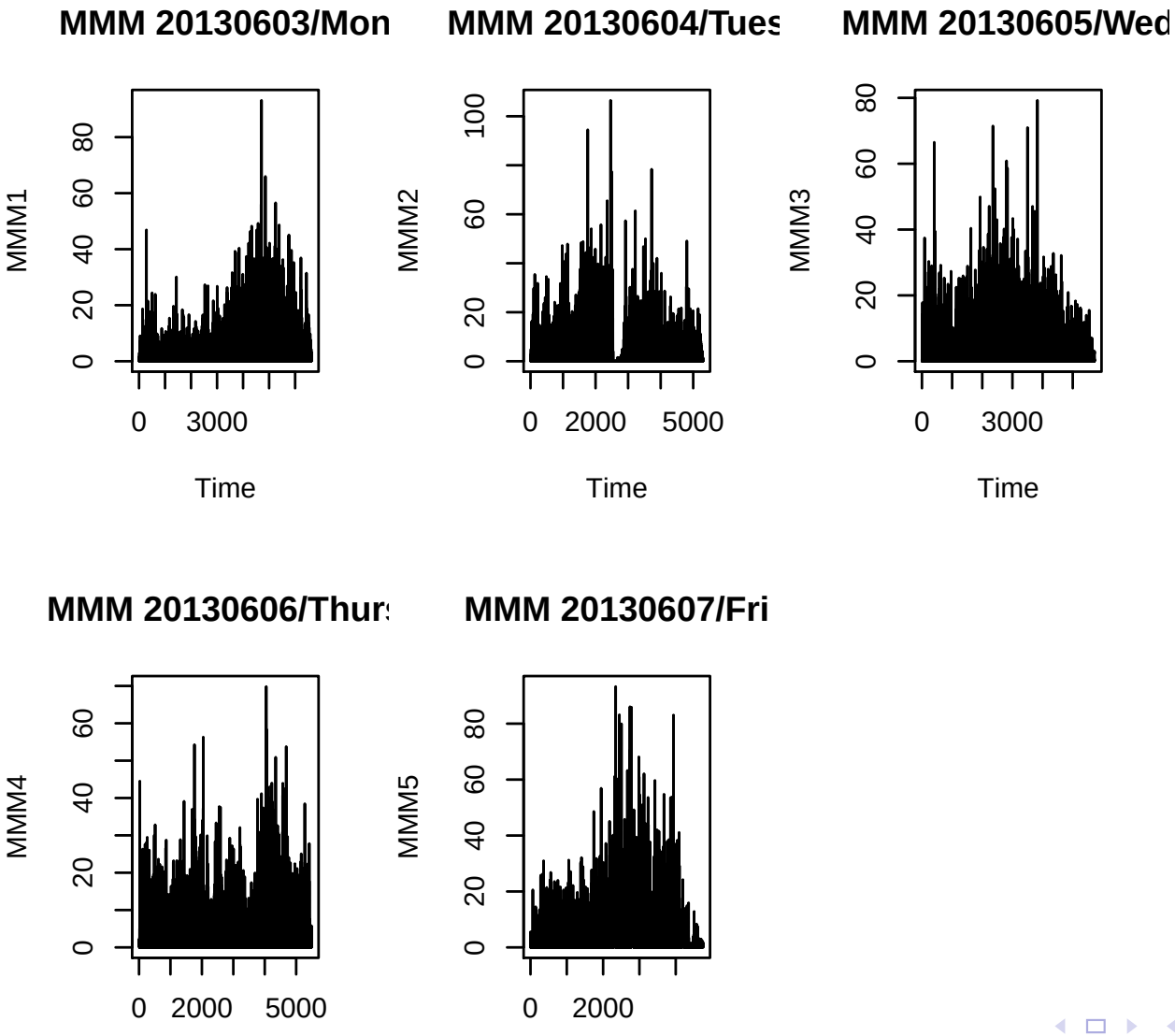
BAC-Durations Between Price Change, $\delta = 0.05/100$



IBM-Durations Between Price Change, $\delta = 0.01/100$



MMM-Durations Between Price Change, $\delta = 0.005/100$



We would like to fit suitable time series models to such durations.

Use the EF approach to estimate model parameters and do model fitting.

Results could be one tool in the financial decision-making.

Examples of Duration Models

Example 1. Log ACD(1, 1) model

$$\begin{aligned} x_t &= \exp(\psi_t)\varepsilon_t, \\ \psi_t &= E[x_t|\mathcal{F}_{t-1}] = \omega + \alpha \log(x_{t-1}) + \beta\psi_{t-1} \end{aligned} \tag{1}$$

where $\alpha + \beta < 1$.

We assume ε_t are i.i.d. non-negative random variables with $E(\varepsilon_t) = 1$ and moments up to order 4.

$\mathcal{F}_{t-1}$ is the σ-field generated by $x_1, x_2, \cdots, x_{t-1}$, assumed to be independent of ε_t.

Example 2. Log ACD(p, q) model

$$\begin{aligned} x_t &= \exp(\psi_t)\varepsilon_t \\ \psi_t &= \omega + \sum_{j=1}^{p} \alpha_j \log(x_{t-j}) + \sum_{j=1}^{q} \beta_j \psi_{t-j} \end{aligned} \tag{2}$$

where $\sum_{j=1}^{\max(p,\ q)}(\alpha_j + \beta_j) < 1$.

Bauwens and Giot 2000 Annales d' Économie et de Statistique.

Let $\boldsymbol{\theta} = (\omega, \boldsymbol{\alpha}, \boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)$

Estimating Function (EF) Approach for Time Series

Suppose $\{x_t, t = 1, \cdots, n\}$ is a realization of a discrete-time, real-valued stochastic process, whose distribution depends on a vector $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathcal{R}^k$.

Let $\mathbf{x}_n = (x_1, \cdots, x_n)'$.

Let $(\Omega, \mathcal{F}, P_{\boldsymbol{\theta}})$: underlying probability space.

Let $\mathcal{F}_t$: σ-field generated by $\{x_1, \cdots, x_t, t \geq 1\}$.

Let $\mathbf{h}_t(\mathbf{x}_t, \boldsymbol{\theta}), 1 \leq t \leq n$ be specified q-dim. martingale differences (MDs)

Let $\{x_t, t = 1, 2, \ldots\}$ have these four conditional moments:

$$\mu_t(\boldsymbol{\theta}) = \mathsf{E}\left[x_t | \mathcal{F}_{t-1}\right],$$

$$\sigma_t^2(\boldsymbol{\theta}) = \mathsf{Var}\left(x_t | \mathcal{F}_{t-1}\right),$$

$$\gamma_t(\boldsymbol{\theta}) = \frac{1}{\sigma_t^3(\boldsymbol{\theta})} \mathsf{E}\left[\left(x_t - \mu_t(\boldsymbol{\theta})\right)^3 | \mathcal{F}_{t-1}\right],$$

$$\kappa_t(\boldsymbol{\theta}) = \frac{1}{\sigma_t^4(\boldsymbol{\theta})} \mathsf{E}\left[\left(x_t - \mu_t(\boldsymbol{\theta})\right)^4 | \mathcal{F}_{t-1}\right]$$

Goal: estimate the parameter $\boldsymbol{\theta}$ based on the dependent observations $x_1, \ldots, x_n$.

Useful General References on EFs

- Godambe Ann. Math. Stat. 1960
- Durbin JRSSB 1960
- Godambe Biometrika 1985
- Lindsay Ann. Stat. 1985
- Thavaneswaran and Thompson J. Appl. Prob. 1986
- Tjøstheim Stoch. Processes and Appls. 1986
- Bera *et al*: excellent review and historical perspective Handbook Econometrics 2006

Selected Useful References on EFs for Time Series

- Thavaneswaran and Abraham JTSA 1988; Merkouris Ann. Stat. 2007; Ghahramani and Thavaneswaran JSPI 2009, 2012; Thavaneswaran *et al.* SPL 2012: estimation for linear and nonlinear time series models using linear EFs.
- Thavaneswaran and Ravishanker 2015, Handbook of Discrete-valued Time Series, Chapman & Hall/ CRC, eds. R. A. Davis, S. H. Holan, R. B. Lund, N. Ravishanker: integer-valued time series, esp. counts.
- Thavaneswaran, Ravishanker, Liang, AISM 2015: Generalized Durations Models

No distributional assumptions are required. We only need to specify the first few conditional moments of $\{x_t\}$

Steps:

For each model/data framework,
• Construct a suitable class of unbiased martingale EFs (they depend on both the observations and parameters) - easy to define for given problems;

• Find the optimal EF in this class which maximizes the *Godambe* information - our AISM paper based on Godambe/Durbin theory;

• Solve the resulting set of nonlinear EEs to obtain parameter estimates - problem specific, and straightforward.

Two Classes of Martingale Differences

$$\{m_t(\boldsymbol{\theta}) = x_t - \mu_t(\boldsymbol{\theta}), t = 1, \ldots, n\}$$

$$\{M_t(\boldsymbol{\theta}) = m_t^2(\boldsymbol{\theta}) - \sigma_t^2(\boldsymbol{\theta}), t = 1, \ldots, n\}.$$

Obtain Quadratic variations and quadratic covariation

$$\langle m \rangle_t = \mathsf{E}\left[m_t^2 | \mathcal{F}_{t-1}\right] = \sigma_t^2,$$

$$\langle M \rangle_t = \mathsf{E}\left[M_t^2 | \mathcal{F}_{t-1}\right] = \sigma_t^4(\kappa_t + 2),$$

$$\langle m, M \rangle_t = \mathsf{E}\left[m_t M_t | \mathcal{F}_{t-1}\right] = \sigma_t^3 \gamma_t.$$

We describe the form of optimal EFs which maximize Godambe information.

Class of zero mean, square integrable k-dim. martingale EFs:

$$\mathcal{M} = \left\{ \mathbf{g}(\mathbf{x}_n, \boldsymbol{\theta}) : \mathbf{g}(\mathbf{x}_n, \boldsymbol{\theta}) = \sum_{t=1}^{n} \mathbf{a}_{t-1}(\boldsymbol{\theta}) \mathbf{h}_t(\mathbf{x}_t, \boldsymbol{\theta}) \right\}, \qquad (3)$$

where $\mathbf{a}_{t-1}$ is $k \times q$ $\mathcal{F}_{t-1}$-measurable matrix, and $\mathbf{h}_t(\mathbf{x}_t, \boldsymbol{\theta})$ is a MD (such as m_t or M_t shown earlier).

Optimality Criterion

The optimal EF $\mathbf{g}^*(\boldsymbol{\theta})$ maximizes the Godambe information matrix

$$\begin{aligned} \mathbf{I}_{\mathbf{g}} &= \left(\sum_{t=1}^{n} \mathbf{a}_{t-1} E\left[\frac{\partial \mathbf{h}_t}{\partial \boldsymbol{\theta}} \middle| \mathcal{F}_{t-1} \right] \right)' \left(\sum_{t=1}^{n} E[(\mathbf{a}_{t-1}\mathbf{h}_t)(\mathbf{a}_{t-1}\mathbf{h}_t)' | \mathcal{F}_{t-1}] \right)^{-1} \\ &\times \left(\sum_{t=1}^{n} \mathbf{a}_{t-1} E\left[\frac{\partial \mathbf{h}_t}{\partial \boldsymbol{\theta}} \middle| \mathcal{F}_{t-1} \right] \right) \end{aligned} \tag{4}$$

Assume that EF $\mathbf{g}(\boldsymbol{\theta})$ is almost surely differentiable with respect to the components of $\boldsymbol{\theta}$

Optimal EF and Corresponding Information:

$$\mathbf{g}^*(\theta) = \sum_{t=1}^{n} \mathbf{a}_{t-1}^* \mathbf{h}_t = \sum_{t=1}^{n} \left(E\left[\frac{\partial \mathbf{h}_t}{\partial \theta} \middle| \mathcal{F}_{t-1} \right] \right)' (E[\mathbf{h}_t \mathbf{h}_t' | \mathcal{F}_{t-1}])^{-1} \mathbf{h}_t, \tag{5}$$

$$\mathbf{I}_{\mathbf{g}^*} = E(\mathbf{g}_n^*(\theta) \mathbf{g}_n^*(\theta)' \tag{6}$$

Solve the set of nonlinear equations $\mathbf{g}^*(\theta) = \mathbf{0}$ to get an estimate of θ. We do this using R and Matlab.

Let $\{x_t\}$ denote the time series of interest. Suppose we fit (a linear or nonlinear) model involving unknown parameters $\boldsymbol{\theta}$.

Linear EF

When the MD is $\{m_t(\boldsymbol{\theta}) = x_t - \mu_t(\boldsymbol{\theta}), t = 1, \ldots, n\}$:

$$g_m^*(\boldsymbol{\theta}) = -\sum_{t=1}^{n} \frac{\partial \mu_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{m_t}{\langle m \rangle_t} \tag{7}$$

with optimal information

$$\mathbf{I}_{g_m^*}(\boldsymbol{\theta}) = \sum_{t=1}^{n} \frac{\partial \mu_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \mu_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \frac{1}{\langle m \rangle_t} \tag{8}$$

Solve $g_m^*(\boldsymbol{\theta}) = \mathbf{0}$ to get $\widehat{\boldsymbol{\theta}}_m$.

Quadratic EF

When the MD is $\{M_t(\boldsymbol{\theta}) = m_t^2(\boldsymbol{\theta}) - \sigma_t^2(\boldsymbol{\theta}), t = 1, \ldots, n\}$:

$$g_M^*(\boldsymbol{\theta}) = -\sum_{t=1}^{n} \frac{\partial \sigma_t^2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{M_t}{\langle M \rangle_t} \tag{9}$$

with optimal information

$$\mathbf{I}_{g_M^*}(\boldsymbol{\theta}) = \sum_{t=1}^{n} \frac{\partial \sigma_t^2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \sigma_t^2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \frac{1}{\langle M \rangle_t} \tag{10}$$

Solve $g_M^*(\boldsymbol{\theta}) = \mathbf{0}$ to get $\widehat{\boldsymbol{\theta}}_M$.

We obtain and use a **combined optimal EF** which is more informative.

Practical Considerations in using EF Approaches

We use three approaches:

(i) solve the system of nonlinear EEs $g^*(\boldsymbol{\theta}) = \mathbf{0}$ using R and Matlab;
(ii) use recursive formulas for $\boldsymbol{\theta}$ using R; and
(iii) iterate recursive formulas for scalar components of $\boldsymbol{\theta}$ using R.

As in most numerical optimization problems, it is important to have good starting values.

Recursive Formulas for Fast, On-line Estimation of $\boldsymbol{\theta}$

$$\widehat{\boldsymbol{\theta}}_t \simeq \widehat{\boldsymbol{\theta}}_{t-1} - [\tfrac{\partial g_t^*(\widehat{\boldsymbol{\theta}}_{t-1})}{\partial \boldsymbol{\theta}}]^{-1} \mathbf{a}_{t-1}^*(\widehat{\boldsymbol{\theta}}_{t-1}) \mathbf{h}_t(\widehat{\boldsymbol{\theta}}_{t-1})$$

where

$$\mathbf{K}_t^{-1} = \mathbf{K}_{t-1}^{-1} - \mathbf{a}_{t-1}^*(\widehat{\boldsymbol{\theta}}_{t-1}) \mathbf{h}_t(\widehat{\boldsymbol{\theta}}_{t-1})$$

These can be easily coded in R or Matlab.

Applications to High-Frequency Financial Data

Steps for Coding the EF Approach for Different Duration Models

- Much of the code is the same code for nearly all models, and may be hard-coded.
- For different models, we only need to change $\boldsymbol{\theta}$, ψ, the conditional central moments of x_t, viz., μ_t, σ_t^2, γ_t, and κ_t and their derivatives.
- Get suitable starting values for the recursions using simple approximating time series models that we can fit easily.
- Run the recursions, or solve the nonlinear equations.
- Need high numerical accuracy routines/functions.

Simulation Studies: Log ACD(1, 1) model

We simulate $L = 100$ sets of durations data, each of length $n = 2500$, from the Log ACD(p, q) model, when ϵ_t has exponential, gamma, or Weibull distributions.

An error distribution is only assumed for the simulation study.

All three EF methods - solving the nonlinear equations, or the recursions on the $\boldsymbol{\theta}$ vector, or recursions on scalar components - all converged to the true values.

Table for Log ACD(1,1)

Para	True	Initial	Recursive Matrix			Recursive Scalar			NLEQN		
			5th	50th	95th	5th	50th	95th	5th	50th	95th
ω	0.5	0.438	0.436	0.444	0.453	0.425	0.437	0.446	0.437	0.438	0.438
α	0.15	0.140	0.139	0.143	0.146	0.137	0.140	0.143	0.140	0.140	0.144
β	0.75	0.738	0.712	0.721	0.730	0.716	0.724	0.730	0.738	0.738	0.739
ω	1.5	1.415	1.405	1.441	1.601	1.581	1.581	1.584	1.415	1.415	1.415
α	0.1	0.089	0.088	0.090	0.101	0.100	0.101	0.103	0.089	0.089	0.089
β	0.8	0.808	0.703	0.922	0.956	0.926	0.926	0.928	0.808	0.808	0.808
ω	2.5	2.467	2.407	2.473	2.489	2.459	2.474	2.481	2.467	2.467	2.467
α	0.2	0.243	0.236	0.244	0.245	0.242	0.244	0.244	0.243	0.243	0.243
β	0.6	0.539	0.532	0.539	0.541	0.537	0.539	0.540	0.538	0.539	0.539
ω	3.2	2.967	2.884	2.967	3.099	2.819	2.964	2.987	2.960	2.967	2.974
α	0.3	0.272	0.259	0.272	0.295	0.265	0.272	0.304	0.269	0.272	0.273
β	0.55	0.576	0.567	0.576	0.591	0.563	0.576	0.576	0.551	0.576	0.576

Table for Log ACD(2,1)

Table: Percentiles of parameter estimates for the Log ACD(2,1) model; $n = 2500$, $L = 100$.

Para	True	Initial	Recursive Matrix			Recursive Scalar			NLEQN		
			5th	50th	95th	5th	50th	95th	5th	50th	95th
ω	10	9.679	10.66	10.69	10.72	10.29	10.34	10.37	9.67	9.69	13.06
α_1	0.10	0.081	0.082	0.082	0.082	0.094	0.095	0.098	0.078	0.081	0.392
α_2	-0.50	-0.501	-0.483	-0.477	-0.472	-0.459	-0.455	-0.454	-0.501	-0.501	0.130
β	0.06	0.051	0.051	0.051	0.051	0.054	0.054	0.055	-0.372	0.051	0.051
ω	5.0	5.102	5.102	5.102	5.102	5.102	5.102	5.102	3.371	5.061	6.135
α_1	0.11	0.100	0.100	0.100	0.100	0.100	0.100	0.100	-0.198	0.141	0.422
α_2	0.50	0.496	0.496	0.496	0.496	0.496	0.496	0.496	0.294	0.535	1.063
β	0.20	0.191	0.191	0.191	0.191	0.191	0.191	0.191	-0.254	0.242	0.488

Parameter estimates under Log ACD(1,1); IBM, June 2013

Date	Recursive Scalar			NLEQN		
	ω	α	β	ω	α	β
20130603	-0.017	0.225	0.521	-0.053	0.188	0.524
20130604	0.098	0.279	0.355	0.093	0.117	0.255
20130605	-0.017	0.292	0.354	-0.021	0.292	0.355
20130606	-0.007	0.282	0.405	-0.025	0.267	0.410
20130607	0.083	0.233	0.601	0.082	0.175	0.512
20130610	0.137	0.270	0.494	-0.022	0.183	0.497
20130611	0.050	0.184	0.658	0.087	0.084	0.685
20130612	0.106	0.214	0.477	0.106	0.216	0.476
20130613	0.006	0.327	0.373	-0.006	0.320	0.368
20130614	0.200	0.279	0.319	0.007	0.182	0.312
20130617	0.009	0.221	0.670	0.007	0.218	0.666
20130618	0.217	0.255	0.420	-0.081	0.106	0.418
20130619	-0.018	0.306	0.402	-0.071	0.255	0.297
20130620	-0.059	0.225	0.592	-0.078	0.201	0.589
20130621	-0.184	0.270	0.538	-0.230	0.221	0.510
20130624	-0.276	0.259	0.408	-0.299	0.118	0.396
20130625	-0.163	0.289	0.460	-0.260	0.263	0.349
20130626	-0.082	0.250	0.484	-0.077	0.250	0.481
20130627	-0.095	0.285	0.460	-0.174	0.181	0.477
20130628	-0.201	0.267	0.567	-0.317	0.081	0.353

Work in progress

Construct portfolio decisions based on such estimates and fits...

Suppose $X_t(d)$ denotes the t^{th} duration for the d^{th} day;
$t = 1, \ldots, n(d)$, and $d = 1, \ldots, D$.

Fit a Log ACD(p, q) model to daily durations.
Let $\hat{X}_t(d) = \exp \hat{\psi}_t(d)\mu_\epsilon$.

For each day, get average estimated duration
$\overline{\hat{X}}(d) = \frac{1}{n(d)} \sum_{t=1}^{n(d)} \hat{X}_t(d)$.

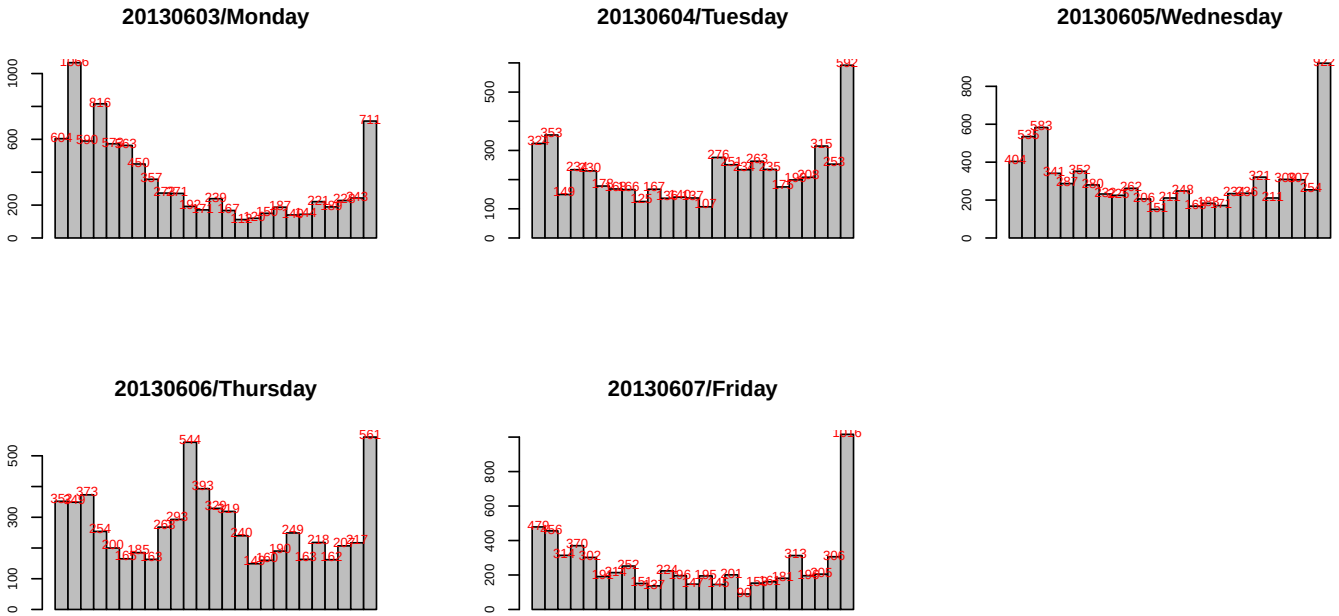
Find the empirical percentiles of $\overline{\hat{X}}(1), \ldots, \overline{\hat{X}}(D)$.

We can check whether the average observed duration for a new/hold-out day d^* lies within the 95% empirical limits, say.

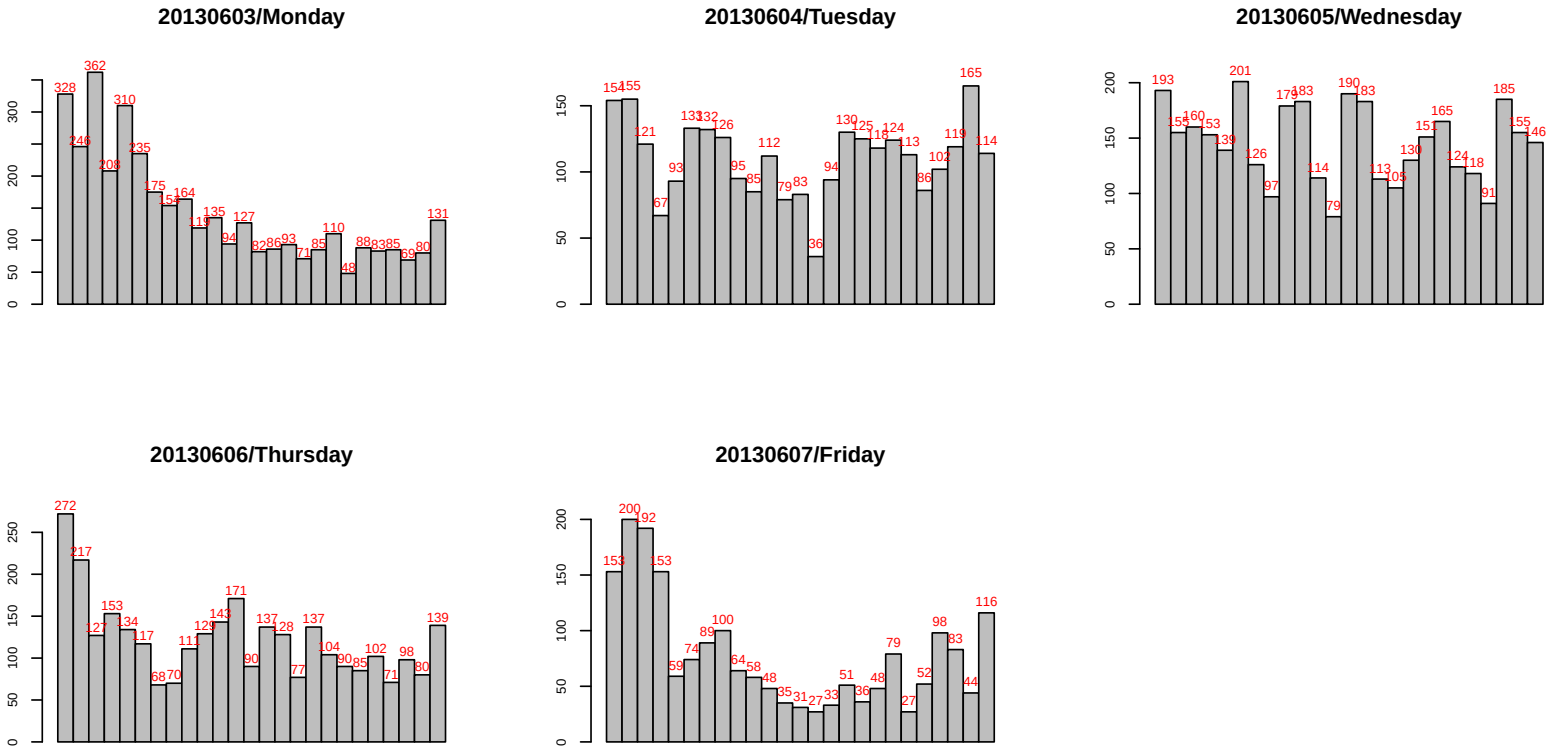
We can also study average durations in short diurnal time intervals of length ℓ minutes rather than a whole day, and count instances of whether an average observed duration is in or not the corresponding limits.

We can construct integer-valued time series based on questions of interest.

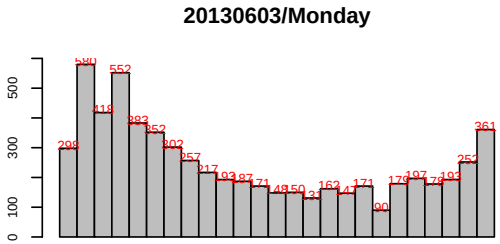
BAC-Hiistograms of Events in 15 minute intervals

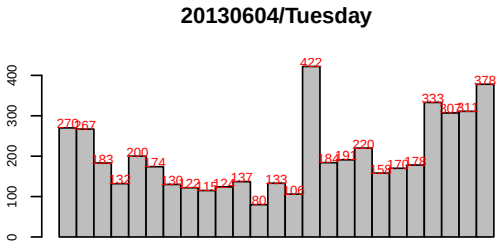


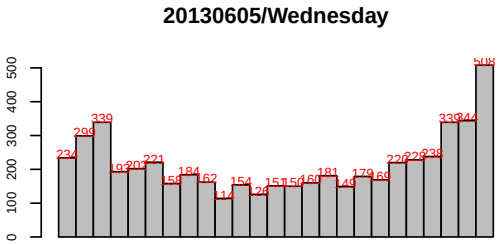
IBM-Histograms of Events in 15 minute intervals

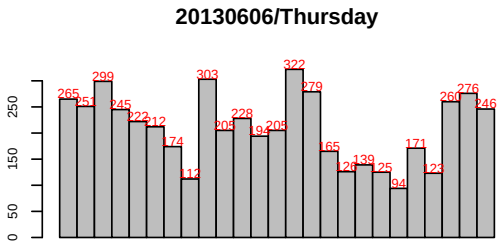


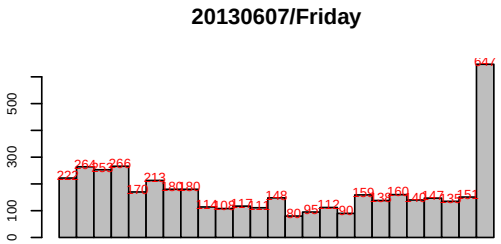
MMM-Histograms of Events in 15 minute intervals











We are now investigating ways in which these results can be incorporated into financial portfolio analysis?

Thank you!